SCIENTIFIC PROGRAMME

WEBTEC

E-BASED LEARNING, TRAINING AND BUSINESS

e-Universities: - Business and Technology (An Australian Viewpoint)

Penelope Goward, Matthew Warren and Shona Leitch

School of Computing and Mathematics
Deakin University, Australia
Email: pgoward@deakin.edu.au

ABSTRACT

This paper will briefly discuss the current environmental changes and the need to provide computer training for students and staff by offering online computer based training materials. Further, it discusses the reason why Australian Universities are using On-line teaching systems and also describes some of the technologies being used.

Keywords: *electronic commerce, higher education, training, On-line Teaching Tools.*

INTRODUCTION

Western societies are moving from an industrial to a knowledge-based economy. One of the major impacts of this change is that universities are becoming business oriented as they respond to two major sources of environmental change – reduced government funding and the increasing 'push' to use computer technology in teaching and learning (and the supporting administration systems). This paper will discuss the increasing need to provide computer on-line teaching technologies for students within an Australian context.

THE CHANGING HIGHER EDUCATION ENVIRONMENT

Higher education in Australia has changed rapidly in the last few years, in response to environmental change. A combination of factors might be suggested as contributing to the change. These factors include: changes in the Australian federal government education policy; an economic rationalist approach to managing the economy; the effects of 'globalisation' and 'internationalisation'; the introduction of computer and technology based instructional systems (particularly internet technology) into education; and what has been termed the third wave of global change: 'knowledge production and management' (Toffler 1998)(Zohar, 1997).

The changing higher education environment offers economic progress and also dilemmas. In a globalised economy the labour force has twomajor opposing factors, low wages or a well-educated and highly qualified manpower. In countries like Australia, where wages are high the development and growth of higher education as a "product" is critical to its economy. Australian universities have recognized this and so have started to move towards becoming international knowledge producing units and undertaking a larger role in society by participating in and contributing to the stimulation of the economy and its growth (Pellert, 1998).

On the dilemma side Clark (1998) argues that global forces create another level of demand on universities who have to deal with factors such as fluctuating funding sources, changing policies of governments, and varying perceptions of universities as workplace trainers and places for the socially upwardly mobile. In addition, universities not only compete with other universities internationally but with knowledge producers in other segments of society.

Further, universities have to achieve contradictory goals in order to control their own destinies. For example, universities have to do more with less money, they have to maintain a cultural heritage while meeting a high demand to develop new fields of study and thought, and develop an entrepreneurial /strategic approach rather than a government funded approach. (Burton-Jones, 1999).

To conclude, Australian universities are still changing as they respond to their new environment, and the one key area of change is in their core area of business – education. The next section will discuss the changing learning environment and its impacts.

THE CHANGING LEARNING ENVIRONMENT

Learning needs have changed because of a changed learning environment. Tsichritzis (1999) posits that there is an increasing need for an attitude of 'life-long learning' in relation to further education and professional development training to accommodate the changes in the work environment. Rather than as Tsichritzis (1999) wryly observes, the temptation for universities to see the current economic climate and 'globalisation' as temporary, and so defend the existing structures and roles, and wait for better days. Tsichritzis proposes that universities – once designed for a captive market of somewhat elite cohorts of post-secondary school students – have to address the changing spectrum of student profiles (older-age students, international students and professionals). Furthermore, Twigg (1994) argues that physical or manual work and the notion of one career will gradually disappear. He states that forecasters envisage that in an average work life people will have several different careers, each requiring new skills, new attitudes and new values. Retraining then will be the constant, as the 'technology' of each profession also changes.

Learning needs will change because it is predicted that those participating in learning and training will not always be able to attend classroom sessions, but are best taught in the workplace – on the factory floor or in the office, out at sea, and in the home. Learning will not always take place with books, overhead projector or a whiteboard. Learners will use tools such computers and all the accompanying software applications and Internet communication technologies, cable television, video-conferencing and CDs to name a few (Twigg, 1994; Bastiaens, T.J., & Martens, R. L., 2000). Examples of these technologies will be looked at later in the paper.

The ensuing outcome means that academics will require skills in using computers in order to teach in new learning environments. Furthermore, because the student cohort has changed, academics will be teaching to students who have a range of learning needs because of their varied backgrounds, ages and stages. Thus, academics will need to go beyond the traditional classroom lecture and explore the range of computer based educational materials, to accommodate the different learning environments and learning arrangements (Twigg, 1994; Bastiaens, T.J., & Martens, R. L., 2000).

THE CHALLENGE

The challenge then is: how do universities in the context of a dynamically changing global environment, who are increasingly needing to be entrepreneurial and business-like, and increasingly using computers in education and the accompanying automated administration systems, support their students and staff, to learn, work and live with computers?

It is suggested in response to this challenge, a systematic and coordinated training approach to support staff and students is taken (Mulholland & White, 2000). One of these approaches is the use of online computer based training materials, as it is accessible and easy to use (via an Intranet for example), it can support current training courses and can complement any computer training embedded in tertiary subjects. It also resolves many of the issues inherent in more traditional education paradigms. The most apparent is its ability to:

- distribute information both locally and globally
- provide flexibility, that is students and staff can study at their own pace and in their own time
- provide additional features such as electronic mail, bulletin boards, different types of media delivery and automated testing
- incorporate an administrative system for monitoring and tracking student progress.

On line learning and training also resolves a particular issue for Australia – overcoming the tyranny of distance. Australia is a vast continent, thousands of kilometers wide and separated from other countries at an equally vast distance. Online learning then offers an ability to accept students from neighbouring Asian and pacific countries, and support students in the outlying country towns and regions.

The changing higher education and learning environments have caused new challenges for universities. Each university is responding the best they can. The following is a description of Deakin University's endeavor to provide educational services that meet their student's learning needs.

DEAKIN UNIVERSITY CASE STUDY

Deakin University has become the primary provider of off campus courses to undergraduate and post graduates students within Australia and focuses on the use of new on-line teaching technologies for staff and students.

In 2000, Deakin University had the following numbers of students enrolled (Deakin University 2000):

	Total enrolments	**% of total**
Full time (on campus)	9,959	35.3%
Part time (on campus)	2,702	9.6%
Full time (off campus)	1,337	4.7%
Part time (off campus)	**10,773**	**38.2%**
Full time (multi modal)	2,502	8.9%
Part time (multi modal)	919	3.3%
Total Student Numbers	28,192	100.0%

Table 1: Deakin University student enrollments (2000)

Of particular interest is the 10,773 (38% of all enrollment) students who are enrolled in off-campus mode. These students do not physically attend lectures and historically have received study guides to work from. The development of Technology has dramatically altered the way in which materials and courses are offered to off-campus students. The off-campus students face a number of problems, these include (Leitch et al, 2000):

- a feeling of isolation;
- difficulty in contacting lecturer;
- difficulty in gaining access to the same teaching materials as on-campus students.

The use of new technology has overcome many of the problems described. The most commonly used on-line technologies used with the School of Computing and Mathematics, Deakin University are as follows.

Email

This represents the simplest Internet technology that is used by off campus students. This is used to allow students contact lecturers, exchange information. Email also allows students to form self-help groups via the use of mailing lists.
The advantage of this medium is that all students have access to email since Internet access is a pre-condition of acceptance to Deakin. The advantage of email is that it allows student to directly contact lecturers and help to reduce the feel of isolation – that is a common problem for many off campus students. However as mentioned earlier, the lack of social context cues can also create problems.

Information Repositories

Deakin University, also makes use of Web-based Information Repositories such as the Web-CT learning and teaching system. This system is used as a central location so that students can find, for example lecture notes, course news and subject resources. A screen shot of Web-CT is shown below.

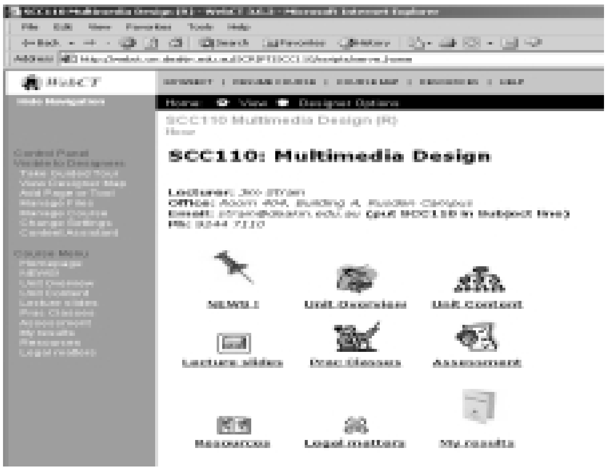


Figure 1: Example screenshot of the Web-CT system

Systems such as Web-CT are being widely adopted by Deakin University for a number of reasons (Leitch et al, 2000):

- ensures off campus and on-campus students have access to the same materials as on-campus students;
- help to ensure that there is no difference between on-campus and off campus students;
- the Web-CT system is web-based which allows easy access for all students.

Further, academics report that they find it useful as a teaching tool because it:

- does not require academics to learn “html” or web development skills;
- provides a standard user interface, that students and academics alike become familiar with.

Group Discussion Tools

There are certain academic subjects that require an element of discussion as an important part of the academic unit. This requirement is outside the ability of standards Information Repository system such as Web-CT. Therefore what is required is the use of a more dedicated system that allow students to post messages and allow those messages to be structured in a orderly manner.
The group discussion tool that is commonly used by Deakin University is called “FirstClass” (see the screen shot of the interface in Figure 2). The tool is not web-based, therefore students are required to download a dedicated browser that enables them to connect to the Deakin FirstClass server.

Figure 2: Example screenshot of the FirstClass system

The advantages of this system, is that it allows (Leitch et al, 2000):

- off-campus students to take part in discussions;
- reduces any distinction between on and off campus students;
- allows students to directly interact with each other, whether they are on or off campus students.

Web-Cam

Deakin is also involved in trying to develop teaching new technologies that can be used to assist students and staff. One of these new developments has been the use of web-based cameras. At the present a pilot study is being undertaken to connect student and staff in remote campuses (as shown by Figure 3). It is intended to expand the use of Web-cams in the future.

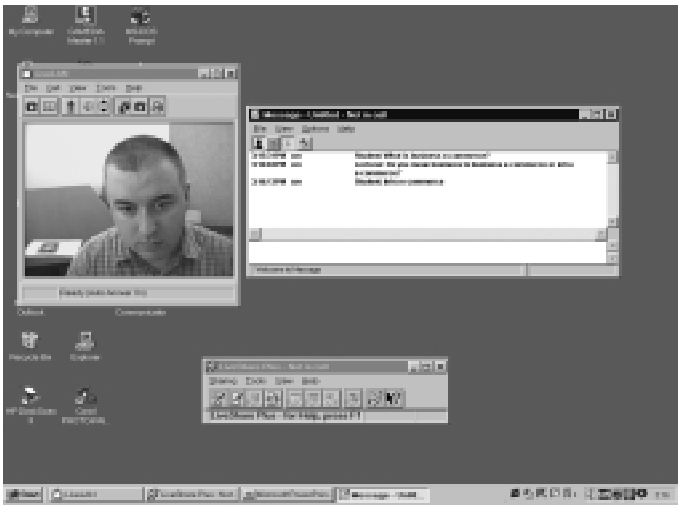


Figure 3: Web-Cam Example

The advantage of this approach is (O'Conaill et al, 1993):

- allows direct contact between staff and students;
- makes it cost effective to directly contact overseas students, an important issue in the global education markets;
- equipment required for web-cams is relatively inexpensive.

However, disadvantages of video technology can be (O'Conaill et al, 1993):

- overlapping speech between student and lecturer;
- interruptions in the sessions due to the technology;
- issue of standards, due to variety of systems available.

CONCLUSION

Higher education and learning environments are changing worldwide. The aim of this paper has been a discussion about some of the keys issues and one university's response using on-line teaching technologies to give Australian Universities competitive advantages over their rivals. It is predicted that this trend will continue to enable Australian universities to fight for their share of the global university market.

REFERENCES

Bastiaens, T.J., & Martens, R. L. (2000). Conditions for Web Based Learning with Real Events. In Abbey, B. (ed) *Instructional and Cognitive Impacts of Web-based Education*, London: Idea Group.

Burton-Jones, A. (1999). *Knowledge capitalism*. Oxford, Oxford University Press.

Clark, B. (1998). 'The Entrepreneurial University: Demand and Response.' *Tertiary Education and management* 4 (1): 5 16.

Deakin University (2000). Deakin University – Pocket Statistics, Deakin University Planning Unit, Australia.

Leitch S, Warren M.J and Coldwell J (2000).*Global Education in the Cyber Age: An Australian Example, in Proceeding of* INC (International Network Conference) 2000, Plymouth, UK, July, 2000.

Mulholland, C., & White, B. (2000*) A model to aid planning, conducting and reporting of courseware evaluation*: in proceedings from Australian Computers in Education Conference, (ACEC) Melbourne, Victoria, July, 2000.

Pellert, A. (1998). 'The Challenge of Internationalisation.' *Tertiary education and management 4* (4): 263-272.

Toffler, A. (1998). Alvin Toffler: Life Matters with Norman Swan, Radio National, The Australian Broadcasting Commission. 5/3/98.

Tsichritzis, D. (1999). 'Reengineering the University.' *Communications of the ACM* 42 (6): 93-100.

Twigg, C. A. (1994). The need for a national learning infrastructure. Originally published in Educom Review . [Online] available: http://www.educom.edu/program/nlii/keydocs/monoraph.html [1998, November 5th]

O'Conaill, B., Whittaker, S., &Wilbur, S., (1993), Conversations over video-conferences: An Evaluation of the spoken aspects of video-mediated communication, Human-Computer Interaction, 10, 401-444.

Zohar, D. (1997). *Rewiring the Corporate Brain*. San Francisco: Berrett-Koehler Publishers Inc.

Development of a web-based classroom training system, VRCATS, for the operation and maintenance of nuclear power plants in Korea

Myeong-Soo Lee
Shin-Yeoll Park
Jin-Hyuk Hong
Yong-Kwan Lee
Seong-Il Song
Korea Electric Power Research Institute
103-16, Munji-Dong Yusung-Gu
Dajeon, 305-380, KOREA
E-mail: fiatlux@kepri.re.kr

Sang-Chul Lee
Se-Jin Park
Kwang-Hyun Kim

Korea Hydro & Nuclear Power Co.
Bugu-Ri Buk-Myon
Uljin-Kun GyungBuk, 767-890
KOREA
E-mail: leedace@khnp.co.kr

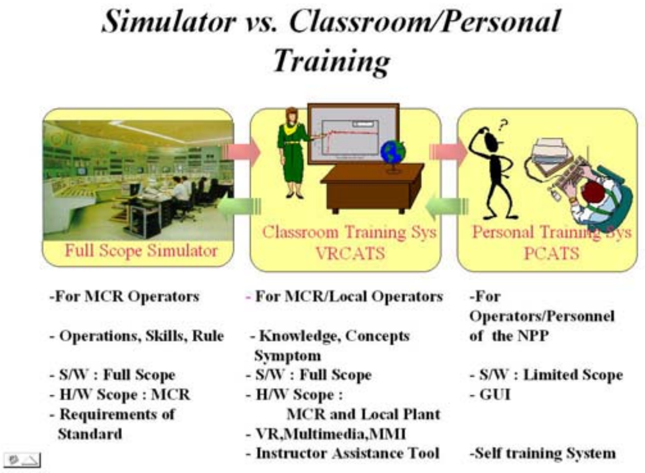
KEYWORDS
VRCATS, Real-Time Simulation, Virtual Reality, 3D, simulator

ABSTRACT

It used to be thought that the full scope or compact simulators are enough to train operators of nuclear power plants. Though most of operating skill and knowledge can be acquired by that kinds of simulator trainings for the operators, it is not easy to get the concepts of physical phenomena and principle.
KEPRI (Korea Electric Power Research Institute) has developed new classroom-training system, VRCATS (Virtual Reality Computer Assistance Training System), which is a web-based multimedia training system in the classroom to complement the traditional full scope simulator training for the operators using new multimedia technology, virtual reality, network collaboration, and etc.
This state-of the-art training system is applied to KSNP, Korea Standard Nuclear Power Plant, simulator of KHNP (Korea Hydro Nuclear Power Co.) at Uljin in Korea.

INTRODUCTION

In the past most operator training for nuclear power plant was conducted on a full-scope simulator including full sized panels and room. The purpose of the full scope simulator training is for MCR operators to increase the operation skills, knowledge of the procedure of the each reference plants.
Though the simulator training is requirements of standard (ANSI/ANS 3.5 1998), most trainings for operators are given in the classroom. The trainee can learn and understand reactor dynamics, physical phenomena, and accident scenarios from the classroom training. Traditionally, training was given with the use of OHP's and documents, which included the accidents analyses for the typical plants for the classroom training. By changing the simulation environment, simulation computer resources and operating system, classroom training can also use the full scope simulator computer, enabling instructors to use the full scope simulator with HMI (Human Machine Interface). (Ryan et al. 2000)
Figure 1 shows the comparison of the full scope simulator, classroom training, and personal self-training system.



Simulator vs. Classroom/Personal Training

Full Scope Simulator	Classroom Training Sys VRCATS	Personal Training Sys PCATS
-For MCR Operators	- For MCR/Local Operators	-For Operators/Personnel of the NPP
- Operations, Skills, Rule	- Knowledge, Concepts Symptom	
- S/W : Full Scope	- S/W : Full Scope	- S/W : Limited Scope
- H/W Scope : MCR	- H/W Scope : MCR and Local Plant	- GUI
- Requirements of Standard	- VR,Multimedia,MMI	
	- Instructor Assistance Tool	-Self training System

Figure 1 Comparison of training methods

Since KEPRI has started its power plant simulator localization project in 1994 (Y.K. Lee et al. 1995), some of state-of-the-art simulation technology have been developed and applied to full scope simulators in Korea including RETRAN based real-time NSSS T/H model, MAAP4 based real-time severe accidents model, and etc. (Lee et al. 2001a,b, c) KEPRI's new classroom training system, VRCATS, is one of them. The main features and recent achievements of the system are presented in this paper.

VRCATS (Virtual Reality Computer Assisted Training System)

During the classroom training, the instructor can access the stand-by host computer of the simulator through a network, and operate the simulator with only software panel. He can activate any malfunction that he wants to instruct, show the trends of major parameters to the trainee and discuss with them. This desktop simulator function helps trainees to understand about the basic phenomena of the accidents, so the operators can understand why some parameters are increasing or decreasing and why they should perform an operator's action. VRCATS consists of four major functions, VRPLANT, VRPANEL, CBT, and WBT. (Lee et al. 2001a)

Virtual Plant

The Virtual Plant is a cyber nuclear power plant in the cyber space. Most of major system equipments and buildings of the reference plant are modeled to navigate and walkthrough in the cyber world and linked with Engineering Data Base (P&ID, Specification, ISO Drawings, and etc.). Figure 2 shows inside of the containment with opaque function. Most of concrete walls are dimly visible. Some of major equipments can be disassembled with major piece by piece and vice versa and some of major work process like refueling operation can be simulated in the Virtual Plant.
Not only the operators can be trained to find out and operate some of the local valves and other equipments in the local area of the plant, but also new employee and existing staffs can also be trained for refreshment training using with many kinds of multimedia educational tools like VRML, MPG, AVI, and etc. Figure 3 shows the inside of the steam generator of KSNP, Korean Standard Nuclear Plant.

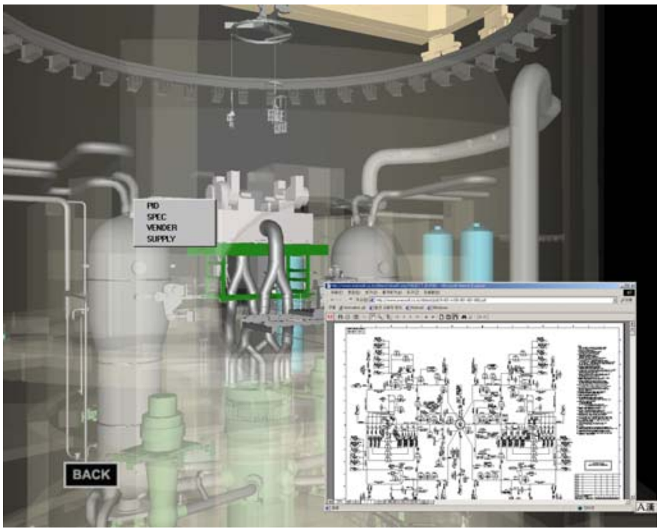


Figure 2 Inside of Containment with EDB

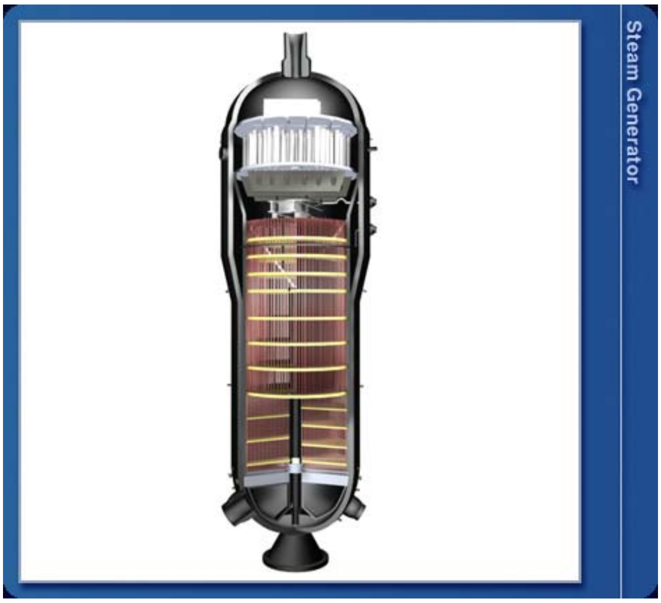


Figure 3 Steam generator VRML file of KSNP

Instructor can select and disassemble any major part of the steam generator. The models are made by EON Studio that is widely used in the E-business and shopping mall.

Virtual Panels

The instructor can give the pre-exercise briefing or reviewing to the operator in the classroom before or after conducting simulator training. When he wants to explain some switch operation of the MCR panels, he can use both 2-D soft panels switches of the instructor station or 3D switches of virtual MCR.
For the efficiency of system performance all components of the panels in MCR are not interfaced with the simulator interactively, some of pre-selected annunciators, indicators and switches, which are used for emergency operation for LOCA (EOP-E1), are dynamically modeled. Since the virtual MCR is linked with a simulator, trainees can control the simulator with the VRCATS. The operators can meet each other in the virtual MCR by connecting to the VRCATS server with their own PCs, communicate with one another by chatting with texts, voices and pictures, and collaborate in the reactor cool down response.

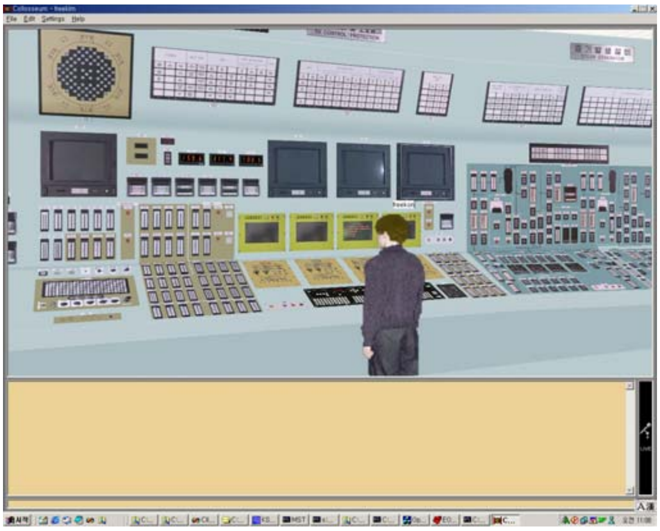


Figure 4 an avatar in the VRMCR

Even though their PCs are located in other rooms, they can see their colleagues' Avata in the same cyber-space, virtual MCR. When the instructor puts a malfunction, LOCA, they can see and hear the blinking and horning of some annunciators, and see the change in some indicating parameters. They can operate some of panel switches with collaboration by talking, chatting to each other.
In addition, virtual MCR provides much immerse environment with such virtual reality equipment as HMD and Data Glove.

Computer Based Training

Even though the text books and OHPs are some of the most effective teaching tools, with which the instructors have done their lectures very well, it is very difficult for operators without a background in nuclear engineering to understand the reactor system dynamic characteristics which is dictated by the tightly coupled non-linear system with many feedback parameters such as temperatures of RCS and fuel clad, concentrations of boric acid, poisoning and depletion, and moving of control rods.

During the classroom training, the instructor can access the stand-by host computer or specially designated replica VRCATS host computer through a network. And he can operate the simulator with only soft panel, and activate any malfunction that he wants to instruct, show the trends of major parameters to the trainees and discuss with them.

KEPRI has developed best-estimated code (RETRAN-3D) based real-time NSSS thermal hydraulic model, ARTS (Advanced Real-time Simulation Model) and applied two full scope nuclear power plant simulators, KNPEC#2 and KSNP simulators. (Lee et al. 2001a,b, c)

A desktop simulator function with this very precise NSSS thermal hydraulic model helps trainees to understand about the basic phenomena of the accidents, so the operators can understand why some parameters are increasing or decreasing and what to do such an event

One of other functions of CBT in VRCATS is a specially designed 3D reactor dynamics simulator. It shows 3D visualization of the axial and radial thermal neutron flux distributions of the reactor core and the trend of input and output parameters-including total reactivity, axial offset, fuel and moderator temperatures, and concentration of poison in real-time by connecting with the simulation stand-by computer. Therefore, understanding of the nuclear reactor dynamics with reactivity feedback effects will be one of the keys to prevent and mitigate reactor transients and accidents. Figure 5 shows 3D distribution of the thermal neutron flux when a CEA (Control Element Assembly) has slip down to the bottom. By the distortion of the flux distribution the position of the faulted CEA and the shadow effect can be very easily observed. Other major parameters, nuclear power, axial offset, total reactivity, axial thermal neutron distribution and Xenon & Samarium poison effect can be trended also. (Hong et al. 2001)

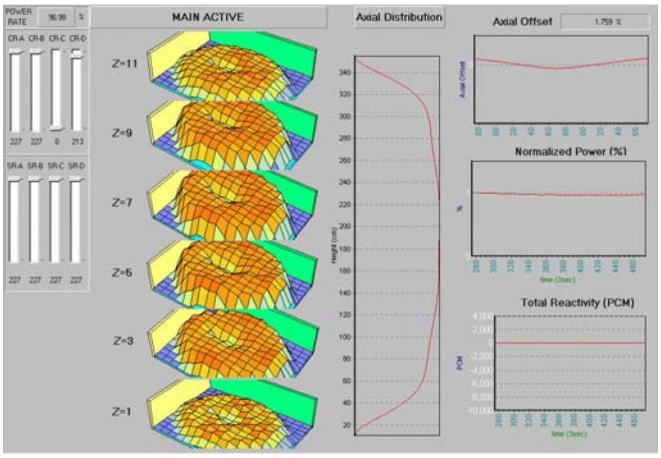


Figure 5 3D Reactor Dynamic Displays

Severe Accident Training

Ever since the core meltdown at TMI, severe accident beyond DBA (Design Based Accident) has become one of key issues in analyzing the safety of nuclear power plants. After the Chernobyl accident, the NRC (Nuclear Regulatory Commission) mandated that each power plant should develop a severe accident management plan. During severe accidents, it is essential for operators to minimize radiation leakage and avoid core meltdown. Since the scope of previous nuclear power simulators was limited to DBA, severe accident training was limited to only several pre-analyzed accident scenarios in the classroom. To overcome these limits and to provide a realistic training environment for severe accident for simulator, KEPRI developed RSAM (Real-time Severe Accident Model) in cooperation with FAI (Fauske and Associates, Inc.) (Lee et al. 2001c)

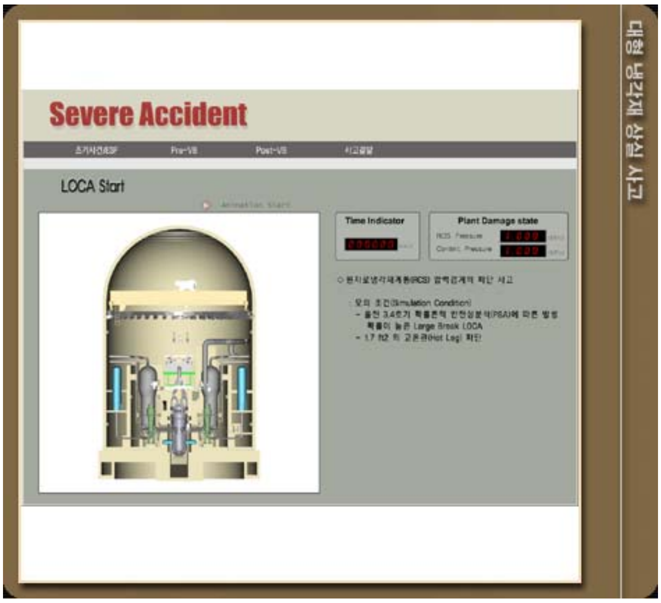


Figure 6 Severe Accident Training

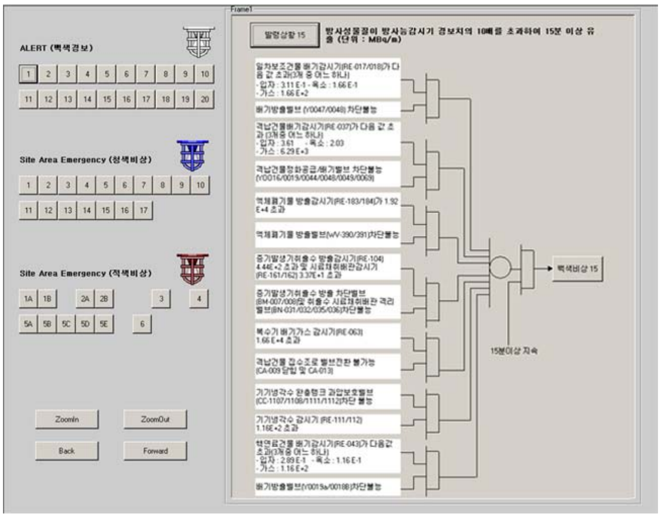


Figure 7 Status Check Tree of the Emergency Condition

It is based on MAAP 4 (Modular Accident Analysis Program) and the improved alternate source term version of MAAP4-Dose (a radiological dose analysis code), both of which are the industry standard accident analysis tools developed by FAI under the sponsorship from EPRI.

As it can simulate the various phenomena that occur during real-time severe accidents, including hydrogen generation, core meltdown, reactor vessel failure, DCH (Direct Containment Heating), hydrogen burn in containment, and radiation leakage during accidents, the trainee may not only be trained in severe accident management procedures but they also can drill to make decision and announcing the emergency condition of the plant. (Figures 6 & 7)

Personal Self-Training

Though the classroom training with the help of instructor is one of the most effective training methods, it cannot be better than self-learning on his/her own initiative. PCATS, Personal Computer Assisted Training System, can simulate the most of design base accidents (DBA) and/or sever accidents of the nuclear power plants. The operator actions can be done during the simulation of the accidents based on personal computer with single CPU. Each user can make his own mimic displays with object-oriented icon editing with true color GUI. Figure 8 shows the main display of the PCATS.

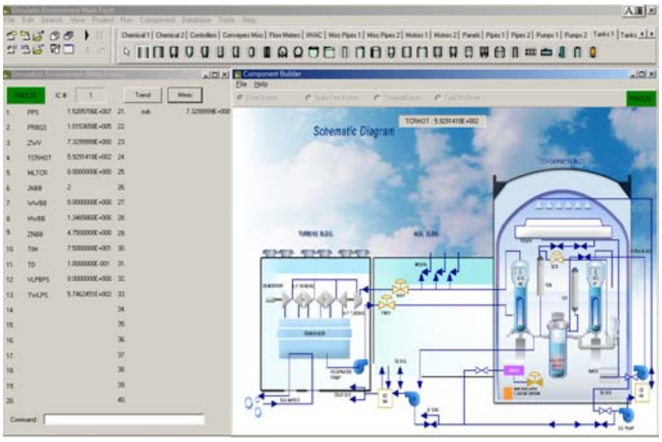


Figure 8 Personal CATS Display

SUMMARY

KEPRI developed new classroom-training system, VRCATS, which is a web-based multimedia training system in the classroom to complement the traditional full scope simulator training for the operators using virtual reality and multimedia technology. Most of major buildings and equipment, and all of control panels in MCR, main control room, of the nuclear power plant are modeled in a cyber space. 3D reactor dynamics integrated with full scope simulator and plant status emergency condition check tree can be shown in the classroom by the system.

VRCATS is integrated with KSNP, Korea Standard Nuclear Plant, simulator, and is of use for classroom training.

REFERENCES

ANSI/ANS 3.5 1998. *Nuclear Power Plant Simulations for Use in Operator Training,*ANS, La Grange Park, IL.

Jin-Hyuk Hong, and etc. 2001, “Development of the Neutronics Model for the KNPEC-2 Simulator” Summer Computer Simulation Conference 2001(Orlando, FL, July 15-19), SCS, CA, 680-684

Jody Ryan, and etc, 2000“A training simulator with soft panels” Western Multi-Conference 2000 (Sandiego, CA, Jan 23-27), SCS, CA, 94-99.

Myeong-Soo Lee, and etc. 2001, “The new research activities of KEPRI for KNPEC-2 Simulator upgrade project” Summer Computer Simulation Conference 2001(Orlando, FL, July 15-19), SCS, CA, 685-689

Myeong-Soo Lee, and etc. 2001, “The validation & verification of applying RETRAN-3D based real-time NSSS T/H model to simulator”, 10th International RETRAN Meeting (Jasckson Hole, WY Oct. 14-17), EPRI, CA, 28-40.

Myeong-Soo Lee, and etc. 2001, “The status and prospect of power plant simulation technologies in KEPRI”, 23rd KAIF-JAIF Seminar on Nuclear Industry (Seoul, Korea Sep. 24-25), KAIF, Seoul in Korea, 171-179.

Yong-Kwan Lee, and etc. 1995, “KEPCO's 3-Pack Simulator Development Plan”, Simulation Multi-Conference Proceedings 12(Pheonix, AZ, April 9-13), SCS, CA, 53-58

BIOGRAPHY

MYEONGSOO LEE was born in Incheon, Korea and had studied at Inha University majoring Mechanical engineering and obtained his M.S. degree in 1982. He worked for more than fifteen years for Korea Electric Power Corporation where he has been working in modeling & simulation, and VR training.
fiatlux@kepri.re.kr

SUPPLY CHAIN NETWORK OPTIMIZATION IN BUSINESS-TO-BUSINESS E-COMMERCE

Gang Wu
Chee Kheong Siew
Information Communication Institute of Singapore
Nanyang Technological University, Singapore 639798
E-mail: p149647525@ntu.edu.sg, ecksiew@ntu.edu.sg

ABSTRACT

To achieve high efficiency in supply chain network, an orderly and flexible flow of materials is essential. In this paper we develop a practical mixed integer programming model for a supply chain network design problem where the objective is to minimize the total production, transportation and distribution costs and the fixed costs for opening warehouses. We employ Lagrangian relaxation to the model and also present several solution procedures to find the lower bound of the objective function. These procedures are coded as integral parts of a subgradient optimization algorithm. Computational results are reported and show that the algorithm behaves well in large-scale supply chain network design problems.

1. INTRODUCTION

Electronic Commerce is a technology-enabled application environment to facilitate the exchange of business information and automate commercial transactions. It encompasses the exchange of many kinds of information, including online commercial transactions. Thus, online marketing to consumers via the Internet, known as business-to-consumer (B2C) e-commerce, is only one of several fronts in the e-business landscape. E-Commerce is also driving deep and profound changes in the structure and business practices of most organizations and the interactions between businesses. Business-to-business (B2B) e-commerce includes a variety of applications and networking technologies designed to automate and optimise interactions between business partners.

Fierce competition in today's global markets, the introduction of products with short life cycles, and the heightened expectations of customers have forced business enterprises to invest in and focus attention on their supply chains. Supply Chain Management (SCM) is a set of approaches utilized to efficiently integrate suppliers, manufacturers, warehouses and customers, so that each merchandise is produced and distributed in the right quantities, to the right locations and at the right time, in order to minimize system-wide costs while satisfying service level requirements (Simchi-Levi 2000). From this definition of SCM, we can see optimization plays a key role in the management of supply chain. One of the most important aspects of supply chain network optimization is deciding where to locate new facilities such as warehouses or factories. In this paper, we concentrate on warehouse location problem. Our problem can be described as follows: A set of plants and customers are geographically dispersed in a region. Each customer exerts a demand for a variety of products that are manufactured at various plants. A list of possible sites for setting up of warehouses is predefined. The objective is to determine where to locate the warehouses, how to ship the products from the plants to the warehouses and how to ship the products from the warehouses to the customers, so as to minimize total costs and satisfy a variety of customer demands.

There has already been a lot of work in the area of warehouse location problems. One of the simplest location problems, the p-median problem, is to locate a number of warehouses to serve a set of customers at a minimum of cost in a two-echelon (warehouses and customers) supply chain network (Bramel and Julien 1997). (Pirkul and Jayaraman 1996) develop a network flow model for the production, transportation and distribution planning in a tri-echelon system and employ Lagrangian relaxation to solve this Mixed Integer Programming (MIP) model. (Bramel and Julien 1997) introduce a multi-product warehouse location model with capacitated constraints. In this model, customers typically demand multiple units of different products. These products are shipped from potential warehouses which in turn receive these products from several manufacturing plants. A customer's demand for a product is satisfied by one of the potential open warehouses. Sanjay (Melkote and Daskin 2001) propose a capacitated facility location/network design problem (CFLNDP). This problem is derived from the classical capacitated facility location problem (CFLP) that has been discussed by many researchers. (Bramel and Julien 1997) give a general model of CFLP. (Magnanti and Wong 1990) provide an overview of solution techniques for the CFLP. Sanjay (Melkote and Mark Daskin 2001) give a mixed integer programming formulation of CFLNDP and solve it using branch-and-bound method. But all of the above models have the following limitations: they assume that a customer gets delivery for a product from only one warehouse and do not incorporate the production cost in the model.

In section 2, a new and more practical model is proposed. In this model, we incorporate the production cost and the flow variables that indicate the total number of products shipped from different plants to different warehouses and from different warehouses to different customers. In section 3, a Lagrangian relaxation of the model is provided and we provide several heuristic procedures to find the lower bound of the problem. Computational results are reported and discussed in section 4. Conclusions are provided in section 5.

2. MATHEMATICAL FORMULATION OF THE MODEL

Indices:

l =Number of plants, $l \in SL = \{1,2,\cdots,L\}$

j =Number of potential sites, $j \in SJ = \{1,2,\cdots,J\}$

i =Number of customers, $i \in SI = \{1,2,\cdots,I\}$

k =Number of products, $k \in SK = \{1,2,\cdots,K\}$

Parameters:

p = Number of warehouses to locate

c_{ljk} = Cost of shipping one unit of product k from plant l to warehouse site j

d_{jik} = Cost of shipping one unit of product k from warehouse j to customer i

f_j = Fixed cost of locating a warehouse at site j

v_{lk} = Production capacity of product k at plant l

e_{lk} = Production cost (per unit) of product k at plant l

w_{ik} = Demand for product k at customer i

s_k = Volume of one unit of product k

q_j = Capacity (in volume) of a warehouse at site j

Variables:

$$Y_j = \begin{cases} 1, & \text{if a warehouse is located at site } j, \\ 0, & \text{otherwise}, \end{cases} \quad \text{for } j \in SJ$$

U_{ljk} =Amount of product k shipped from plant l to warehouse j

N_{jik} =Amount of product k shipped from warehouse j to customer i

Then the problem can be formulated as:

$$Min \quad \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{K} c_{ljk}U_{ljk} + \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{k=1}^{K} d_{jik}N_{jik} + \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{K} e_{lk}U_{ljk} + \sum_{j=1}^{J} f_j Y_j$$

$$s.t. \quad \sum_{j=1}^{J} U_{ljk} \le v_{lk} \quad \forall l \in SL, k \in SK \qquad (1)$$

$$\sum_{l=1}^{L}\sum_{k=1}^{K} s_k U_{ljk} \le q_j \quad \forall j \in SJ \qquad (2)$$

$$\sum_{i=1}^{I}\sum_{k=1}^{K} s_k N_{jik} \le q_j \quad \forall j \in SJ \qquad (3)$$

$$\sum_{l=1}^{L} U_{ljk} = \sum_{i=1}^{I} N_{jik} \quad \forall j \in SJ, k \in SK \qquad (4)$$

$$\sum_{j=1}^{J} N_{jik} Y_j = w_{ik} \quad \forall i \in SI, k \in SK \qquad (5)$$

$$\sum_{j=1}^{J} Y_j = p \qquad (6)$$

The first term in the objective function represents the transportation costs between plants and warehouses. The second term represents the distribution costs between warehouses and customers. The third term represents the total production cost of the products and the fourth term measures the fixed cost of locating the warehouses. Constraints (1) are production capacity constraints. Constraints (2) and (3) are warehouse capacity constraints. Constraints (4) ensure that there is a conservation of the flow of products at each warehouse; that is, the amount of each product arriving at a warehouse from the plants equals the amount being shipped from the warehouse to the customers. Constraints (5) are similar to constraints (4). They ensure the demand for each product of every customer equals the amount being shipped from the warehouses to the customer. Constraint (6) ensures that we locate exactly p warehouses.

3. SOLUTION PROCEDURES

In this section, we present a solution procedure based on Lagrangian relaxation method. Lagrangian relaxation is an approach used for solving mixed integer and pure integer programming problems. The reader is referred to (Fisher 1981) for a detailed and excellent survey on the Lagrangian relaxation technique.

3.1. Lagrangian Relaxation of the Model

We relax constraints (4) (with multipliers θ_{jk}) and constraints (5) (with multipliers λ_{ik}). The resulting problem is:

$$Min \sum_{i=1}^{I}\sum_{k=1}^{K}\lambda_{ik}w_{ik} + \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[c_{ljk}+e_{lk}-\theta_{jk}\right]U_{ljk} + \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{k=1}^{K}\left[\theta_{jk}-\lambda_{ik}Y_j+d_{jik}\right]N_{jik}+\sum_{j=1}^{J}f_jY_j \quad s.t.\ (1)\text{-}(3),(6)$$

This problem can be decomposed into two separate problems.

$$P_1:\ Z_1 = Min \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[c_{ljk}+e_{lk}-\theta_{jk}\right]U_{ljk} \quad s.t.\ (1),(2)$$

$$P_2:\ Z_2 = Min \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{k=1}^{K}\left[\theta_{jk}-\lambda_{ik}Y_j+d_{jik}\right]N_{jik}+\sum_{j=1}^{J}f_jY_j \quad s.t.\ (3),(6)$$

Let $Z_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ be the optimized solution to this problem. For fixed vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, $Z_{\boldsymbol{\lambda},\boldsymbol{\theta}} = Z_1 + Z_2 + \sum_{i=1}^{I}\sum_{k=1}^{K}\lambda_{ik}w_{ik}$. It is the lower bound of the original problem.

3.2 Procedure for Solving P_1

Problem P_1 can be solved separately for each plant/product pair. For each plant/product combination, say plant l and product k, sort the J values $\overline{c_j} = c_{ljk} + e_{lk} - \theta_{jk}$. Starting with the smallest value of $\overline{c_j}$, say $\overline{c_{j'}}$, if $\overline{c_{j'}} \geq 0$, then the solution is to ship none of this product from this plant. If $\overline{c_{j'}} < 0$, then ship as much of this product k as possible along arc (l, j') subject to satisfying the capacity constraints (1) and (2). Then if the v_{lk} has not been completely shipped, do the same for the next cheapest arc, as long as it has negative cost value $\overline{c_j}$. Continue in this manner until all of the product k produced in plant l have been shipped or the cost value $\overline{c_j}$ is no longer negative. Then proceed to the next plant/product combination and repeat this procedure. Continue until all the plant/product combinations have been scanned in this manner.

3.3 Procedure for Solving P_2

First we separate the problem by warehouse. In the problem corresponding to warehouse j, either $Y_j = 0$ or $Y_j = 1$. If $Y_j = 0$, then $N_{jik} = 0$. If $Y_j = 1$, we should calculate the value of $Min \sum_{i=1}^{I}\sum_{k=1}^{K}\left[\theta_{jk} - \lambda_{ik} + d_{jik}\right]N_{jik} + f_j$. So

$$Z_2^j = \min\left\{ Min \sum_{i=1}^{I}\sum_{k=1}^{K}\left[\theta_{jk} - \lambda_{ik} + d_{jik}\right]N_{jik} + f_j, 0 \right\}.$$

To find $Min \sum_{i=1}^{I}\sum_{k=1}^{K}\left[\theta_{jk} - \lambda_{ik} + d_{jik}\right]N_{jik} + f_j$, we use an algorithm similar to the one for solving problem P_1. For each warehouse j, sort the $I \times K$ values $\overline{d_{i,k}} = \theta_{jk} - \lambda_{ik} + d_{jik}$. Starting with the smallest value of

$\overline{d_{i,k}}$, say $\overline{d_{i',k'}}$, if $\overline{d_{i',k'}} \geq 0$, then the solution is to ship none of products from this warehouse. That means $Y_j = 0$.

If $\overline{d_{i',k'}} < 0$, then ship as much of product k' as possible along arc (j,i') subject to satisfying warehouse capacity constraints (3). Then if the $w_{i'k'}$ has been completely fulfilled but the $\left\lfloor \frac{q_j}{s_{k'}} \right\rfloor$ has not been completely shipped, do the same for the next smallest value of customer/product combination, as long as it has negative cost value ($\overline{d_{i,k}}$). After having solved these, let π be a permutation of the number $1,2,\ldots,J$ such that $Z_2^{\pi(1)} \leq Z_2^{\pi(2)} \leq ..... \leq Z_2^{\pi(J)}$. The optimized solution to P_2 is to choose the p smallest values: $Z_2 = \sum_{j=1}^{p} Z_2^{\pi(j)}$

4.COMPUTATIONAL RESULTS

The solution procedures of finding the optimised value of the problem were coded in C++ as an integral part of the subgradient optimization algorithm. The problem sets were generated randomly but systematically to capture a wide range of problem structures. As all the customers' demands must be fulfilled, the total capacity of the selected warehouses and total production capacity of the plants must be greater than or equal to the total customer demand if the problem has a feasible solution. That means the two inequalities ($\sum_{j=1}^{p} q_j \geq \sum_{i=1}^{I}\sum_{k=1}^{K} s_k w_{ik}$ and $\sum_{l=1}^{L} v_{lk} \geq \sum_{i=1}^{I} w_{ik}$) are necessary conditions for the feasibility of the problem. We generated warehouse capacities and plant production capacities according to these two constraints. The warehouse capacity ratio (WCR), which is defined to be the ratio of the total demand of customers to the total capacity of the open warehouses, and plant capacity ratio (PCR), which is defined to be the ratio of the total demand of customers to the total production capacity of the L plants, were used as two input parameters of the problem.

The results of the performance of the algorithm are reported in Table 1-3. The warehouse capacity ratio (WCR) and plant capacity ratio (PCR) were also varied from 0.70 to 0.85 to present a different flavor to the problem under different operating considerations.

Table 1: Performance of Solution Algorithm (10 plants, 20 warehouses, 40 customers and 4 products)

WCR	PCR	CPU (sec)	WCR	PCR	CPU (sec)
0.70	0.70	0.114	0.80	0.70	0.151
0.70	0.75	0.41	0.80	0.75	0.278
0.70	0.80	0.291	0.80	0.80	0.123
0.70	0.85	0.214	0.80	0.85	0.256
0.75	0.70	0.117	0.85	0.70	0.130
0.75	0.75	0.138	0.85	0.75	0.116
0.75	0.80	0.130	0.85	0.80	0.126
0.75	0.85	0.132	0.85	0.85	0.133

Table 2: Performance of Solution Algorithm (20 plants, 40 warehouses, 80 customers and 8 products)

WCR	PCR	CPU (sec)	WCR	PCR	CPU (sec)
0.70	0.70	1.05	0.80	0.70	1.85
0.70	0.75	2.02	0.80	0.75	1.05
0.70	0.80	0.92	0.80	0.80	1.13
0.70	0.85	0.97	0.80	0.85	0.99
0.75	0.70	1.08	0.85	0.70	1.07
0.75	0.75	1.12	0.85	0.75	1.91
0.75	0.80	1.14	0.85	0.80	1.86
0.75	0.85	2.10	0.85	0.85	1.07

Table 3: Performance of Solution Algorithm (40 plants, 100 warehouses, 200 customers and 16 products)

WCR	PCR	CPU (sec)	WCR	PCR	CPU (sec)
0.70	0.70	16.81	0.80	0.70	16.15
0.70	0.75	32.57	0.80	0.75	16.39
0.70	0.80	33.57	0.80	0.80	15.28
0.70	0.85	26.67	0.80	0.85	45.20
0.75	0.70	15.45	0.85	0.70	26.07
0.75	0.75	29.54	0.85	0.75	14.71
0.75	0.80	15.72	0.85	0.80	17.57
0.75	0.85	15.41	0.85	0.85	27.81

Tables 1 to 3 indicate that the algorithm behaves well. For a problem with 40 plants, 100 warehouses, 200 customers and 16 product items, there are over 320,000 variables and over 5,000 functional constraints. Table 3 indicates that our algorithm was able to find solutions of this large mixed integer programming problem with low computational times varying between 14 and 46 seconds.

We also investigated the solution quality of our algorithm. We compared our algorithm with another two algorithms: "nearest" warehouse algorithm and exhaustive enumeration algorithm. The first algorithm assigns the "nearest" warehouse (with least transportation cost) to each customer/product combination and selects *p* "best" warehouses according to their total assigned transportation costs. The exhaustive enumeration algorithm searches all the possible assignments to find out the optimum solution. Obviously, the solution found by exhaustive enumeration algorithm is the best solution of the problem if we do not consider its computational time. Table 4 shows the algorithm solution qualities and computational times of these algorithms. As the solution got from exhaustive enumeration algorithm is the best solution of the problem, the relative differences between the exhaustive enumeration solution and the other two solutions were used to judge the qualities of the "nearest" warehouse and subgradient algorithms. The relative difference was expressed as a percentage of the exhaustive enumeration solution.

Table 4 Solution Qualities of the Algorithms (WCR=PCR=0.7, K =4)

L	*J*	*I*	Relative Difference of		CPU (sec) of		
			"Nearest" (%)	Subgradi-ant (%)	"Nearest"	Subgradi-ant	Enumeration
5	10	20	17.1	4.1	<0.01	0.03	0.06
5	15	40	26.3	5.3	<0.01	0.05	1.38
10	20	40	18.3	6.6	<0.01	0.12	181.52
10	22	40	30.5	6.8	<0.01	0.14	4918.27
10	25	40	25.6	7.1	<0.01	0.15	25070.38

From Table 4, we can see the computational time of exhaustive enumeration algorithm increased significantly while the problem scale increased. So this algorithm is suitable for only small-scale problem (for example, 5 plants, 10 warehouses and 20 customers). Although the computational time of "nearest" warehouse algorithm remained very low, its solution result was not satisfactory because its relative difference with exhaustive enumeration algorithm was much higher than our algorithm. These computational results indicate that our proposed algorithm consistently produced effective solutions and the computing times were well within the acceptable limits.

5. CONCLUTION

In this paper we present a practical supply chain network design model. This model, together with the effective solution procedures, can provide the decision makers with an efficient tool in the design of their supply chain network. The model aids the decision makers to best site warehouses and to determine the most efficient material flows among plants, warehouses and customers, in order to effectively satisfy customer demands. Computational results on a wide variety of the problems are reported and these results indicate that our algorithm can consistently provide stable solutions, regardless of the problem structure.

REFERENCES

Bramel and Julien, "The Logic of Logistics: Theory, Algorithms, and Applications for Logistics Management", Springer 1997

Fisher, M., "The Lagrangian Relaxation Multipliers Method for Solving Integer Programming Problems," *Management Science* 27, 1-18(1981), pp.1-18

Magnanti, T.L., Wong, R.T., "Decomposition Methods for Facility Location Problems", In: *Mirchandani P. B. Francis, R.L. (Eds), Discrete Location Theory, Wiley New York (1990),* pp.209-262

Melkote, S. and Daskin, M., "Capacitated facility Location/Network Design Problems," *European Journal of Operational Research* 129 (2001), pp.481-495

Pirkul, H. and Jayaraman, V., "Production, Transportation, and Distribution Planning in a Multi-Commodity Tri-Echelon System", *Transportation Science* Vol.30, No.4, (1996), pp.291-302

Simchi-Levi, D., "*Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies*", McGraw-Hill Companies, Inc. 2000

DIGITAL LIBRARIES AND INFORMATION RETRIEVAL

User and Resource Models Definition and Adaptation to Personalize the Multimedia Instructional Material in a Web-Based Distance Learning System

Antonella Carbonaro
Department of Computer Science
University of Bologna
Mura Anteo Zamboni 7, I-40127 Bologna, Italy
e_mail: carbonar@csr.unibo.it

KEYWORDS
User Modeling, Distance Learning, Information Filtering, Web Mining

ABSTRACT

This paper aims at overcoming the problem of information overload in an on-line course provided within a Web-based multimedia distance learning system that guarantees customized interaction and individualized instructional material to each different user. We propose a distance learning system that is capable of filtering interesting resources for a particular student, represented by the user profile, from a large amount of irrelevant information sources. After the refinement of the information coming from the visited resources, based on the correlative importance of different arguments, we present the technique we have used to automatically define the vector space model, that is the used information filtering model, and describe how to learn the user profile using learning and forgetting coefficients. Different experimental cases have been constructed and tested to verify the properties of proposed techniques; the suggested algorithms produce good results in a reasonable time also for Internet surfer students.

INTRODUCTION

The development of multimedia networked applications has been one of the most important factors for the success of the Internet and has changed the way many people study and work. To fully exploit the Web potentiality in the instructional field and to facilitate a student-based approach to the provided educational material, modern Web-based learning environments must be able to dynamically respond to the different and personal students' learning styles, goals, knowledge backgrounds and abilities. To this aim, Adaptive Hypermedia (AH) and User Modeling (UM) techniques have recently emerged as important technologies. Machine learning is one component of UM, a discipline which is concerned with both how information about users can be acquired and used by automated systems. In particular, the UM represents the system's belief about each learner's knowledge and updates it dynamically, based on the student's interactions with the system. But the enormous amount of information now available on the Web, makes it almost imperative to use some sort of automatic processing to adapt to student changing needs and interests, represented by the UM, locating relevant information and building effective AH. The selected information could be structured using ad hoc criteria and delivered to interested users to guarantee a personalized and technologically innovative service to the students. Most content retrieval methodologies use some type of similarity score to match a query describing the content, and then present the user with a ranked list of suggestions. These methodologies identify the information filtering (IF) problem [Belkin and Croft 1992]. In the application domain of IF, it is very difficult to generate and maintain user profiles using more traditional methods, like user interviews or user stereotypes. In order for the UM technique to be really effective it must build the user profile (that is, learn the student's interests and preferences) without explicit training from the user, maintain and update the profile by adapting to the changing interests of the user while exploring new domains that may be of interest to the user and adding them to the profile.

In this paper, we propose a system that is capable of filtering interesting resources for a particular student from a large amount of relevant and irrelevant information sources using asymmetric similarity function. After the refinement of the information coming from the visited resources, based on the correlative importance of different arguments, we present the technique we have used to automatically define the vector space model, that is the used information filtering model, and describe how to learn the user profile using learning and forgetting coefficients.

In particular, the paper is organized as follows. The next section presents the main functionality and the general organization of the system developed to assisting students who wish to improve their abilities using a distance learning environment. In the following, we introduce the information filtering module and some test cases on distance measures and, then, the technique we have used to automatically define the vector space model, that is the used IF model. Then, the sequent section describes and verifies how to learn the user profile using learning and forgetting coefficients. Finally, we present the system interface developed to propose to each different user a fitting environment corresponding to his preference in terms both of categories and web documents, and the conclusion of the paper.

SYSTEM DESCRIPTION

The mechanism that dynamically manages and maintains the multimedia hypertext typically consists of four different software modules that cooperate together in order to tailor both contents and presentation styles to the needs of each different student. Those components are i) the Domain Knowledge Module (DKM) that contains the domain knowledge representing the full educational material, ii) the User Model (UM) that represents the system's beliefs about the learner's knowledge, iii) the Hypermedia Management Module (HMM) that is responsible for the dynamical selection of the educational material to be presented to each student, on the basis of the information stored in both the DKM and the UM, and iv) the Run-Time Session Management Module (RTSM) that masters and controls the interaction between each student and the system. The internal organization of modern distance learning tools is shown in Figure 1.

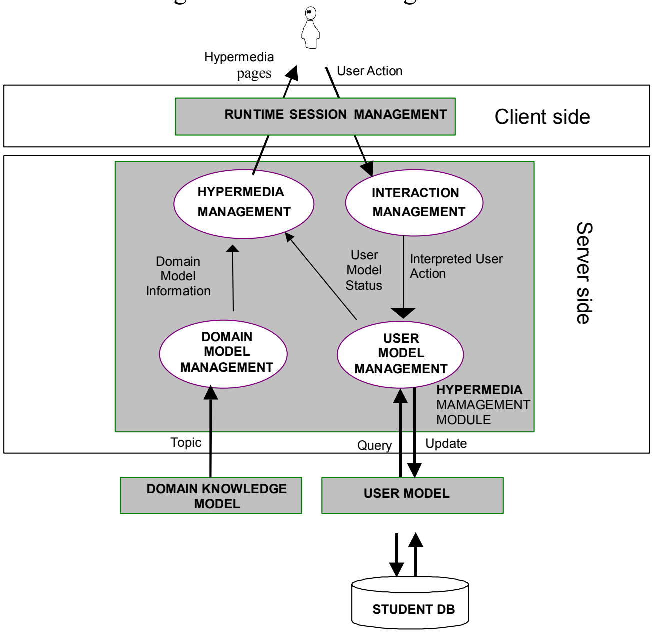


Figure 1: AH- and UM- based learning systems

Aim of this paper is to present the main features of an on-line computer science courses we have developed and integrated within a Web-based Multimedia Learning Environment. This system is based on the AH and UM technologies mentioned above, and was developed at the University of Bologna [Carbonaro et al 2001], [Roccetti and Salomoni 2001]. In particular, we report on how the internal organization of the learning environment needs to be augmented to automatically filter relevant information to be presented to each user on the basis of the particular UM. The filtering process may employ some degree of learning or adaptation to improve the quality of its assistance over time.

The learning system has been obtained as constructed out of the two following software applications that cooperate according to a typical Web-based client-server interaction scheme:

- the Client provides the integrated environment through which each student obtains the hypermedia pages constituting the domain lectures and interacts with the system during the problem solving activities.
- the Server provides each different student with an educational service for learning the domain content, which is tailored to the student's profile.

Within the Server application the automated provision of the domain lectures to each student is carried out by means of the three following cooperating software modules.

- A Specific User Model (SUM) that records the level of the student's knowledge. The SUM maintains a model for each user and changes it dynamically to follow the evolution of the learner's knowledge.
- A Specific Domain Knowledge Module (SDKM) that contains the domain knowledge information related to the selected course structured as a hypernetwork of knowledge items.
- A Specific Hypermedia Management Module (SHMM) that exploits both the SDKM and the SUM for dynamically producing the multimedia learning material whose content and presentation style are adapted to the student's needs and knowledge profile.

THE INFORMATION FILTERING SYSTEM

To filter information resources according to student interests we must have a common representation for both the users and the resources. This knowledge representation model must be expressive enough to synthetically and significantly describe the information content [Kurki et al. 1999]. In the proposed system we use the vector space model, a popular IF model for textual material. The use of one common model permits to update the user profile in according to consulted information resources.

The importance of the information filtering can only increase dramatically as more and more users begin utilizing the vast information resources available via electronic media. Suppose that we have *n* arguments by which the resources can be described and in which the user can be interested. More formally, let be $A = \{a_1, a_2,...,a_n\}$ the finite set of arguments, $U = \{u_1,u_2,...,u_q\}$ the user set and $R = \{r_1,r_2,...,r_s\}$ the resource set. Let us define $E =\{e_1,e_2,...,e_m\} = U \cup R$ the entity set, where $m=q+s$.

The function $\pi:ExA \rightarrow [0,1]$ defines for each entity-argument couple a real value belonging to [0,1] specifying the affinity between the actual entity and the argument. In particular, both the user profile and the resource profile are described as vectors in R^n. The *ith* element of π_j represents the proportion of the concept *i* in the entity e_j, or the proportion of user interest in that concept.

We have implemented and tested 11 different symmetric and asymmetric distance measures that consider how well the resource fits some part of the user profile and some of their combinations. Many other systems use symmetric distance measures, such as the cosine measure [Savia 1999]. In [Carbonaro 2001] we have showed the results obtained evaluating information according to the particular user's interest defining different test cases. From an accurate analysis of the values obtained comparing the different distance measure functions in the test cases we have evidenced a more accurate behaviour of *sat_n_sat_1* function. Then, we have considered this function for the definition of sequent techniques and experimental results. The obtained average execution times for *sat_n_sat_1* function have been 0.412 seconds, a satisfactory time for Web environment students.

THE VECTOR SPACE MODEL DEFINITION

To fully exploit the IF potentiality in the instruction field and to facilitate a student-centred approach to the provided educational material, modern learning environments must be able to automatically provide the vector space model, that is the used common IF model (for an in-depth introduction to the vector space model the reader is referred to [Salton, 1989]). To this aim, we have introduced a pre-processing step in the HMM. This processing also allows the system to automatically work in every knowledge domain environment, without specify the set of meaningful terms for each argument and their relevant values.

During this step, the system receives as input different documents that correspond to the fundamental contents of the current course as decided by the teacher. All the training documents are lexically analysed to obtain the attributes that describe the input. From this educational material the system extracts a set of meaningful terms, which is the attributes used in the following filtering phase. The terms are characterized by their frequency in the document, by the number of documents that contain the word and by their specificity. Furthermore, it is necessary to consider the role of the term inside the document, that is, for example, if it appears in the title or if it is bold or if it is a keyword. Interested reader can find more details regarding function definition in [Carbonaro, 2002].

To verify the terms extraction phase we have carried out tests using 100 Web documents for each class, obtained by the search engine Yahoo. For each test class we have realized a dictionary constructed out of all different words *w* extracted from the document test set.

It is possible to order each dictionary in a decrease manner respect to the importance of each different word *w* belonging to class *j,* suitable weighting words that are present in most of the considered documents even if their maximum frequency value in the document is not high, and the importance of the words not exclusively identifying a certain class.

To represent the different entity using the vector space model it is now necessary to associate relevant value to each extracted term. We remind that the *ith* element of π_j represents the proportion of the concept *i* in the entity e_j, belonging to the interval [0,1]. To transform the obtained relative frequency of each term in the document also considering its the role inside the resource structures (title, image, table, bold, …) we multiply the weights associate to the different structures in which the term appears and normalize the resulting value respect to 1. The extracted relevant and non relevant keywords are then used to generate 1000 test documents (500 related and 500 unrelated to the test arguments) used to evaluate the overall system.

We conclude this Section by specifying that the advantage in introducing an automatic vector space model definition module into our learning system has been twofold. On the one hand, this module has permitted an augmented personalization of the interaction during the use of the system. On the other hand, instead, the real instructor may be relieved from the two tasks of:

- designing and proposing to students, with different learning desire, lectures and documents for each new concept of the actual domain knowledge to be learnt, structured under the form required by the vector space model, that is specifying for each entity-argument couple a value representing the affinity between the actual entity and the argument,
- managing and evaluating each new resource that the Web environment could furnish continuously to the system.

The mechanism that dynamically filters and proposes documents to the students has been tested using different trial cases, in particular describing 200 relevant and 200 irrelevant resources. The algorithm has obtained the following classification results: 375 documents have been correctly classified with an error rate equal to 6.25%. Besides, analysing these erroneously classified resources we have observed that they represent edge cases not easily to categorize. The results indicate that the overall prediction accuracy of the proposed system was good and that it can learn user need for resource filtering.

USER MODEL DEFINITION AND MODIFICATION

Let us suppose that at some given time the student has already rated *n* documents. These resources don't describe the true student interest but only his actual preferences; in fact, in realistic system we cannot hope the user interest to stay constant but to vary slowly [Krulwich and Burkey 1997].

The user model updating algorithm uses a learning coefficient so that preference is given to resources that are rated more recently; the effect is that the user gradually forgets the resources that he has visited a long time ago. Moreover, we used a forgetting coefficient to decrease the relevance of some resources. A reasonable choice might be that the user forgets his previous interests as time passes by.

The mechanism that dynamically maintains and updates the student models has been tested using different test cases, describing an initial student model and hypothesizing the consultation of various resources, some of which with completely different profiles. Detailed information on the realized tests could be extracted from [Carbonaro 2001]; in this contest it is important to underline that the algorithm used to construct and update the student profile considering the information coming from the proposed resources is able to effectively adapt and personalize the instructional material to be presented to different students on the basis of personal interactions.

SYSTEM INTERFACE

Our web-based distance learning system is organized as follows: we have planned different subjects in the computer science field appealing for our students; these different categories are described in Table 1. In addition, each category is composed of a variety of web documents describing the actual topic. Aim of the system is to propose to each different user a fitting interface corresponding to his preference in terms both of categories and web documents. Thus, the user should select the desired argument and corresponding details without difficulty in the whole system. In addition, we want to suggest appropriate banner and e_mail in a specific interface's area to assist the student in his navigation. When the user selects some document corresponding to a specific category, we update his dynamic profile, maintaining the relevant information, for example, the particular category and document selected. Furthermore, we update the user profile when the student chooses some category, checks the corresponding documents and modifies his intention changing root category. In fact, in this case the user behaviour attests to disagree the proposed documents.

When the student concludes the opened session, we save his new profile taking into account the executed selections, that is, we combine in a suitable manner his past profile with the new dynamic one.

Now, we want to analyse some navigation paths of a generic user *u1* using the proposed web-based learning system and considering some specific starting profile. In particular, we want to evaluate how the system modifies both the proposed category list in successive work sessions and the student profile respect to different leaf categories when he consults some document. Additionally, we will examine how the system modifies the categories and document information memorized in the archives. Before to illustrate the evaluation tests made on the developed system, we want to underline that the system main page proposed to student *u1* respects his initial profile.

Now, let us suppose that the user selects the voice "games" from the system main menu and then the two leaf documents "The incredible machine" and "Pin ball". Afterwards, the obtained result representing the new state of the user *u1* is described in the last column of Table 1.

Table 1. Starting user profile of a just connected student and the obtained values after the choice of the two leaf documents of the category "Games"

CATEGORY	STRENGTH	STRENGTH
1 (sw)	0.20	0.17333
2 (hw)	0.15	0.13
3 (graphics)	0.15	0.13
4 (internet)	0.11	0.095333
5 (languages)	0.12	0.104
6 (AI)	0.08	0.069333
7 (training on line)	0.07	0.060667
8 (games)	0.05	0.176667
9 (electr.comp)	0.04	0.034667
10 (OS)	0.03	0.026

We could observe that the strength of the games category in the user profile is increased respect to the strengths of the other classes, neglected during the last session. Now, we want to analyse the modified interface that the system will propose to the same user *u1* during the sequent interaction. Clearly, the interface is influenced by the new profile presented in the third column of the Table 1 and by the characteristics of resources grant to the users. The environment interface is also modified both in the banner area and in the main menu section.

After the described navigation session of the user *u1,* the list of leaf categories associate to the main class "games" is modified; in particular, the system considers the new two preferred leaf documents with respect to the strength of the previous one. Now, we are interested in analysing what happen to the user profile during the same session. In other words, starting from an initial profile mainly centred on a certain argument class, we want to evaluate how it changes each time the user consults some document belonging to the same class. We hypothesise to test the system using the "Training on line" class and the following profile at the opening of the current session. Second column of the Table 2 shows the percentage of class interest (0.07) respect to the other classes and how this percentage is subdivided in the different documents belonging to the class.

Table 2. Current profile for the presented test case and new user profile after one visited document

NUM_CAT	STRENGTH	STRENGTH
7	0.07	
71	0.10	0.100
72	0.70	0.395
73	0.05	0.035
74	0.05	0.065
75	0.10	0.405

We would like to prove that, for each visited document, the user profile will progressively looks like the resource built on class 7 and allowed to the student during the actual session. Now, let us suppose that the user would like to see some document belonging to “Training on line” class. We will show (third column of the Table 2) how the profile will be modified after the first visited document, particularly respect to class 7. In the Tables 3 we show how the user profile is modified respect to class “Training on line”, after the consultation of 2, 3, 4 and 5 documents belonging to class 7.

From these examples we can deduce that the user profile in the leaf category level is modified obtaining a single profile as more similar as possible to the resource’s characteristics in the particular work session and increasing the opportunity to obtain that particular information as most desired.

Table 3. Values after 2, 3, 4 and 5 click

CAT	STRENGTH	STRENGTH	STRENGTH	STRENGTH
71	0.1000	0.10000	0.1000000	0.10000
72	0.2425	0.128125	0.1090625	0.16625
73	0.0275	0.021875	0.0209375	0.02375
74	0.0725	0.078125	0.0790625	0.07625
75	0.5575	0.671875	0.6909375	0.63375

CONCLUSION

This paper aims at overcoming the problem of information overload in an on-line course provided within a Web-based multimedia distance learning system that guarantees customized interaction and individualized instructional material to each different user. The intelligence needed to render the system adaptive to each different user’s profile is obtained by integrating together: 1) adaptive hypermedia techniques, 2) user modeling technologies and 3) machine learning strategies. In particular, we have proposed novel methods for information management, with special focus on the development of new algorithms for information filtering. The system is capable of filtering interesting resources for a particular student from a large amount of irrelevant information sources using asymmetric similarity functions. From this, the user profile is constructed and updated considering the information coming from the visited resources. Accordingly, we can deduce that our personalized filtering system satisfy the three main requirements of: i) specialization, in fact the system infers the habits of the user and specializes to them suggesting as many relevant documents and as few irrelevant documents as possible, ii) adaptation, in fact the system adapts its behaviour in response to the change of student interests, iii) exploration, in effect the system is able to explore newer information resources to find something of potential interesting to the user.

Different experimental cases have been constructed and tested to verify the properties of proposed techniques; the suggested algorithms produce good results in a reasonable time also for Internet surfer students. Currently, we are working to test the proposed architecture in a real long distance learning environment to collect and analyze feedback from teachers and students and to provide an on line model of the whole system.

References

Belkin, N. J., and Croft, W. B., 1992. “Information Filtering and Information Retrieval: Two Sides of the Same Coin.” *Commun. ACM 35*, 12, 29-38

Carbonaro A., “An Information Filtering Approach to Personalize the Multimedia Instructional Material in a Web-Based Multimedia Distance Learning System”, Proc. of 2002 Int. Conf. on Simulation and Multimedia in Engineering Education, San Antonio (Texas), January 2002

Carbonaro A., “A Comprehensive Approach to the Personalized Information Filtering Problem”, Proceedings of 2001 SCS Euromedia Conference, The Society for Computer Simulation International, Valencia (Spain), April 2001

Carbonaro A., Roccetti M., Salomoni P., “A Web-Based Didactical Environment for Learning Prolog Programming Abilities”, 2001 International Conference on Intelligent Multimedia and Distance Education, USA 2001

Krulwich, B. and Burkey, C., 1997. “The InfoFinder Agent: Learning User Interests through Heuristic Phrase Extraction.” *IEEE Expert*, September/October 1997

Kurki, T., Jokela, S. and Sulonen R., 1999. “Agents in Delivering Personalized Content Based on Semantic Metadata.” *Proc. of 1999 AAAI Symposium Workshop on Intelligent Agents in Cyberspace*, Stanford, USA

Roccetti M., Salomoni P., 2001 “A Web-based Synchronized Multimedia System for Distance Education”', Proc. of 2001 ACM Symposium on Applied Computing (ACM SAC`2001), ACM Press, Las Vegas.

Salton, G. 1989, *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, Reading, Mass.

Savia, E., 1999. “Mathematical Methods for a Personalized Information Service.” http://smartpush.cs.hut.fi web page

ANTONELLA CARBONARO received a degree in Computer Science from the University of Bologna, Italy, in 1992. In 1997 she finished her Ph.D. studies on Artificial Intelligent System, in particular, working on Machine Learning techniques. From 1997 to 1999 she received, beside the faculty of Computer Science of the University of Bologna, a research fellowship with theme of search "Artificial Intelligence". Now she is an Assistant Professor in the Department of Computer Science of the University of Bologna. Her current research interests concern soft computing, evolutionary algorithms and personalized learning environment. **Author’s Address**: Department of Computer Science, University of Bologna, Italy; e_mail: carbonar@csr.unibo.it; http://www.csr.unibo.it/~carbonar

dLibra — CONTENT MAINTENANCE FOR DIGITAL LIBRARIES

Paweł Gruszczyński
Cezary Mazurek
Stanisław Osiński
Andrzej Swędrzyński
Sebastian Szuber
Poznań Supercomputing and Networking Center,
ul. Noskowskiego 10, 61–704 Poznań, POLAND
phone: +48 61 858 20 30, fax: +48 61 852 59 54
e-mail: {grucha,mazurek,stachoo,kokosz,szuber}@man.poznan.pl

KEYWORDS

Digital library, content management, information society

ABSTRACT

In this paper issues of content management in digital libraries are addressed. We present three main factors that influence the quality of the digital library content organization and maintenance. We argue that, apart from sophisticated end-user tools, modern digital library systems must provide means for hierarchical content organization, document versioning and advanced access control. We also include a brief description of *dLibra* — a digital Library Framework developed by Poznan Supercomputing and Networking Center — with respect to the content management and maintenance facilities.

INTRODUCTION

With the development of the Internet, possibilities and needs for building digital libraries dramatically increased. One of the basic practical applications of such systems is using them as a building block of a global infractructure of services and applications for information society. National programme *Pionier: Polish Optical Internet* is an environment for building and making available those applications and services in Poland [1].

As the implementation work on digital library frameworks proceeds, end-users are delivered more and more sophisticated software tools for creating, viewing and searching electronic documents. Nevertheless, the constant growth of the number of available publications poses difficulties in several areas:

- content organization,
- duplicated, withdrawn and multiple document versions,
- access management and accounting,

Thus, apart from comprehensive end-user support, a digital library system must provide means for content management and maintenance. Such approach will greatly improve the quality and efficiency of the viewing and publication processes.

In the next three sections we discuss the above issues in general and in the last section we give an outline of the *dLibra* Digital Library Framework. To provide additional background information, we also briefly refer to similar functionality offered by the SCHOLNET Digital Library Testbed [2].

CONTENT MANAGEMENT

In authors' and readers' best interest is that documents are easy to find. However, in the world of digital publication a solution to this problem is not as easy to find. The difficulty obviously lies not only in the lack of good computer systems, which can assist readers in the process of searching, but also in our human nature. Authors are rarely concerned with supplying their work with the appropriate metadata description to make the publication easy to find. On the other hand, excessive medatata information (i.e. many irrelevant keywords) will much decrease the acuracy of searching.

The problem is obviously not solvable by means of a computer system. The only thing that can be done is to equip authors with a tool which makes inserting metadata to documents as painless as possible (e.g. automates the process to some extent), publishers with a tool which allows to easily control the authors' work and, of course, readers with a tool which allows to define queries easily and intuitively.

One solution is a very well-known and widely adopted hierachical catalogue, which has been in wide use long before the Internet was born. However, the question is: Are we using all the features which this structure offers in an electronic world? Consider the following issue. Should the author of fifty articles on neurobiology be allowed to publish a critical essay on Socrates philosophy? Probably not unless he can explain such a sudden change of his interests. This issue is discussed in more detail in section *Access Management* of this article.

The second problem can be best described by an example. If a reader is interested in one article about cat breeding then he is probably also interested in other works concerning this subject. Suppose that we have a branch in our catalogue (in *dLibra* such branch is called *directory* as a comparison to a computer filesystem) named "Cat breeding" and that we have only one article in it. A reader reads the article and then forgets about the library because he does not want to check every day if something new has appeared. We encounter a similar problem when a reader has found an article by running a search engine on "cat" & "breed" keywords.

The answer is a so-called *subscription service*. A reader can mark a directory as interesting to him, choose a form of notification and wait for information from the system if something new has appeared in this directory. Similarly, a reader can define a query which will be periodically performed by the system. If the query finds a new document, then the reader is automatically notified about the fact.

DOCUMENT VERSIONING

Documents and their Components

A document exists in many versions throughout its lifecycle. The first pre-print or draft version is replaced by its consecutive successors and eventually a final version is created. Some of the documents are under continous development, legislatives or software specifications being two of many examples. The problem of document versioning is not encountered in the world of well-known paper publications, because once printed on paper the document can never change. When the next version comes out, it has a different publication date, different ISBN number, probably different form and is in fact another document. This is, however, not true for electronic publications. An author or a publisher can switch to the next version available in electronic form (for example as an HTML web page) or remove the draft version completely until the next version is available without notice to anybody. A full coverage of this issue can be found in [3].

The answer to the problem is a system which assists an author and a publisher but also the readers in keeping track of the document's versions. This answer is not so obvious as it may seem, however. On the contrary to a paperwork, an electronic publication consists of many parts or *modules* [4], possibly independent to some extent. For example a web document can consist of many HTML files and a few graphic, sound, or video files as well. Each of these modules can be changed independently so they shall be tracked separately by the system.

A change made to a module does not however necessarily mean that the author wants to release a new version of the whole document. Consider an example document which is a book consisting of many long chapters. Each of the chapters is stored in a separate HTML file. It is fully understandable that during the process of preparation of the next version of this book, the author probably wants to improve more than one of the chapters. Only after changes have been incorporated into all of the chapters, does the author want to publish a new version of his work. Thus, there is a need for versioning on two levels. The first level is a module level when progress is tracked separately for every module (HTML file in our example) and on this level only the author can access all of the versions of modules. The second level is a document level when progress is tracked for the whole document. After a version of the document has been made available to public, it should not be taken away because it could create dead links and broken references from other documents. Consequently, when a new version of a document has been made available to the public, the old one is not removed, but is still in place so that the consistency with other documents is maintained.

Versions and Readers

A possibility of creating many versions of the same document is very attractive but it has its drawbacks. Let us consider the following scenario. A reader has found a very interesting article in a digital library. He has read it and made a research on his own, inspired by the article. Meanwhile, the authors of the article achieved new results and improved their article by adding new conclusions. These results and conclusions are probably of great interest to the original reader. The question is: how does he become aware that a new version of the article is available? He can of course check every day if there is a new version of a document but, given the number of documents available in electronic form nowadays, it may become very time consuming.

The answer is once again a subscription service. Using this service a reader can mark an interesting article to be tracked by the system. As soon as a new version of the document becomes available, the reader is notified of the fact by an e-mail or by different means.

ACCESS MANAGEMENT

A digital library or other electronic publishing system is a big infrastructure and needs proper access management. To achieve maximum flexibility, access restrictions should be applied on different levels of the library objects hierarchy. This section deals with three basic access management levels.

Let us consider the metadata by which authors describe their documents. The Dublin Core Metadata Element Set [5] is good for at least the majority of scientific publications but consider for example a set of fairy tales and a parent who wants to find a story which suits his needs. So he would like to search for a fairy tale designated for a child older than six and younger than ten. There is no possibility to prepare such a query using the Dublin Core Set mentioned above. There is an obvious need that a modern digital library system allows using more than one *metadata scheme*.

A question appears: who shall prepare such schemes in the system? Definitely not a developer or an administrator of the system because they lack the field knowledge and can only guess what is important and what is not for a given audience and type of documents. Leaving it to the authors is not a good choice either, because we would end up having one metadata scheme for each and every document in the library, which is even bigger mess, than not having metadata at all. The answer is that a person or a group of people should be chosen and given the rights to design and adjust metadata schemes for a given library. This is the first and most general level, which is the *library level* of right management.

Field knowledge is even more important as far as the hierarchical structure of library directories is concerned. Somebody has to make the subdirectories in the Biology directory and it apparently should not be an expert on atomic physics — the system should allow giving rights to modify the structure of each of the directories separately. The same applies to the permission to publish a document in a library directory. This is the *directory level* of access management.

Once a document is put in the library, the right to modify it is assigned to its creator. But there can be more than one author of the document. The system should then allow to grant rights to access and modify a document for more than one person. However, not everybody engaged in the document preparation process should be allowed to modify it. For example, people who are just reviewing the document and preparing comments or people who are accepting the document should not be allowed to modify it since they are not the authors. On the other hand, they should be allowed to access the document even before it was published because it is what their work is all about. All authors, reviewers and readers have their rights to the document but each right is different. This difference is meaningful only with regard to a specific document, so this access management level is the *document level*.

CONTENT MANAGEMENT IN *dLibra* DIGITAL LIBRARY FRAMEWORK

dLibra Digital Library Framework has been developed by Poznan Supercomputing Networking Center since 1999 [6][7]. *dLibra* facilitates all phases of a digital publishing process by supporting three basic groups of users: readers, writers and publishers.

Using a web-based interface the readers can easily browse the library and view selected publications. A search engine enables them to issue a query regarding various multilingual metadata attributes (e.g. using Dublin Core attribute scheme) such as the publication author, title, description, keywords, creation date and many more.

The editors are delivered intuitive GUI-based tools for placing new publications in the library and retrieving publications or some of their components for further editing. An advanced versioning system supports managing subsequent revisions of publication objects as well as branching.

The publishers receive tools for managing the whole library structure, in particular, putting out and hiding publications, managing access rights and other library resources.

Content organization

The whole library content is organized in a hierarchical structure of entities. A directory is an entity that groups any number of other items - subdirectories or publications. A publication is a unit of information (e.g. an article or a book) that consists of one or more basic objects of various types (e.g. HTML or image file). An example *dLibra* content structure is shown in figure 1.

The use of the hierarchical directory structure enables the publishers and library administrators to divide the whole library content into smaller areas, accordingly to e.g. the coverage or importance of the material. Additionally, with a comprehensive access management system, it is possible to assign readers and writers to particular parts of the library resources so that the search results are more accurate and new publications are placed only in appropriate directories. To support interdisciplinary documents, a system of links will be introduced that will make it possible to place a single publication in many directories of the library. The hierarchical structure of the publication itself enables the authors to logically group together modules of different types to make the digital material more attractive and comprehensive.

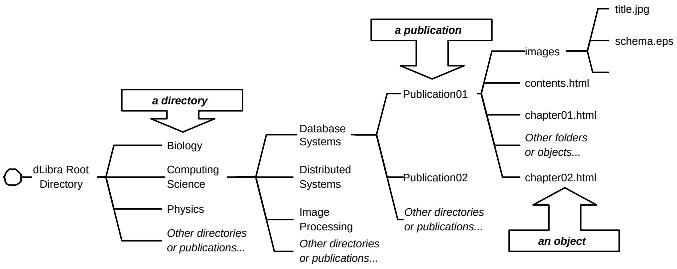


Figure 1: *dLibra* hierarchical directory structure

Every entity in the library — from the root directory down to a single publication object can be described by means of user-defined attribute schemes (e.g. Dublin Core). The values of attributes can be defined in several user-defined languages and are considered while searching the library.

Document versioning

dLibra provides support for both publication- and module-level versioning. A publication can be made available for public viewing by creating an edition, which is a set of certain versions of publication objects. Every publication can have an unlimited number of editions comprised of different versions of publication files or even different files. To explain the idea of object versioning and publication editions let us assume there is a publication consisting of only three files: `body.html`, `title.jpg` and `logo.gif`. Figure 2 shows how can subsequent versions of these files make publication editions.

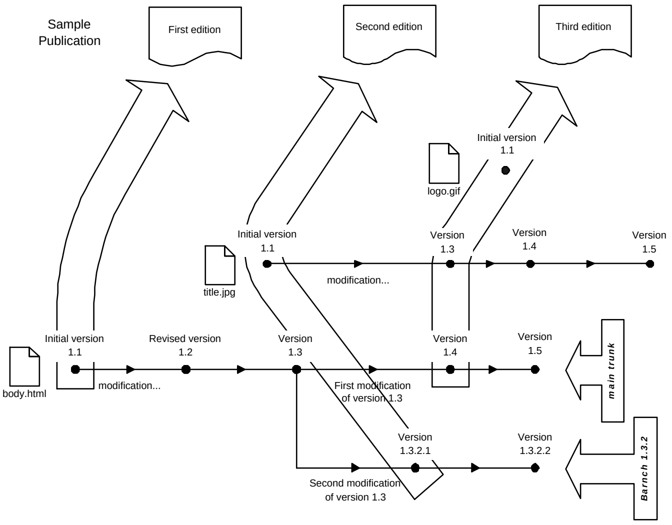


Figure 2: Object versioning and publication editions

In the figure, the sample publication starts from only one file — `body.html`. The first edition of the publication contains only this file, whereas the other files may or may not exist as well. Each publication object is versioned, which enables the editors to store subsequent revisions of the object and to place them in different publication editions. If there is a need for creating two or more object versions based on the same revision a branch can be created. The second edition of the sample publication, apart from newer reversion of `body.html` file, contains a new file: `title.jpg` in its initial version. The third edition comprises all the three files.

After creating an edition and making it accessible for public viewing, the author is unable to withdraw it from the library so that all references remain valid. It is, however, possible to create and publish another edition with modified content. Again, with the access management system, the roles of writers (content providers) and editors (reviewers) can be separated to increase the capabilities of the digital publication life cycle.

In the SHOLNET Digital Library Testbed a slighty different approach to publication structure and versioning is adopted. Apart from physical publication components, document view, which is a specific intellectual expression of a document instance, is used. This enables to present the document in a full text form, its abstract or e.g. metadata description. However, less emphasis is put on publication versioning (no branching support) and component reuse.

Access management

Access management in the *dLibra* library is based on a system of users and groups. A user can be made a member of any number of groups, which may be considered as an assignment of a specific role (the user inherits all the rights granted to the groups of which he is a member). Rights can be granted on a library, directory or publication basis. On each level, several access sublevels have been defined to enable a precise definition of user roles.

Library level access restrictions affect the functioning of the whole library:

- attribute scheme management
- file type hierarchy management
- user and group information management

Directory level access restrictions regard:

- directory visibility — some of the directories can be made invisible to particular users or groups
- permission to list a directory contents
- permission to read the contents of the publications contained in a directory
- permission to edit the directory structure (creating and removing empty subdirectories)
- permission to place publications in a directory
- right management for a directory and all its subdirectories

Publication level access restrictions regard:

- reading published editions a publication
- permission to edit the publication objects
- publication management (granting rights, branching, publishing)

In the SCHOLNET testbed, acces management tasks rest with system administrators. Every submission, withdrawal or replacement of a document involves a decision of an administrator of the appropriate part of the library. The advandate of such approach is its flexibility — every request can be handled individually. Nevertheless, the automated access control seems to be sufficient and less costly in most cases.

System model

dLibra Digital Library is implemented as a client-server system (figure 3). On the server side there are a number of independent modules connected via network intefaces. All modules are implemented using Java 2 techology, in particular RMI (Remote Method Invocation), JDBC (Java Database Connectivity) and JSP/Servlet technologies. Currently the data storage module utilizes the Oracle database system. Nonetheless, because of the use of the JDBC and SQL 92 standard *dLibra* can be easily ported to work with any other RDBMS. An event module built in the *dLibra* system provides a possibility of adding extension modules without modifying the already existing ones. On the client side GUI (JavaSwing) applications are provided that support publication creating and library management processes.

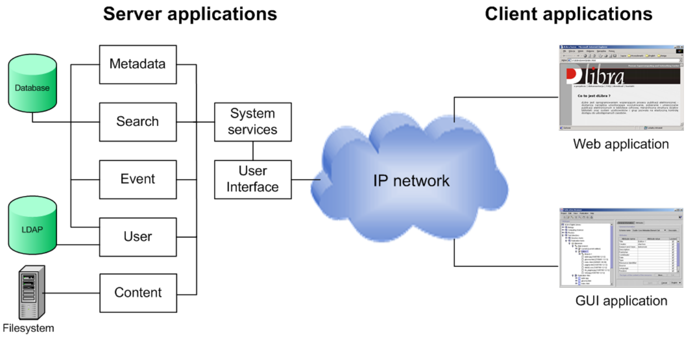


Figure 3: *dLibra* internal architecture

CONCLUSIONS AND FURTHER WORK

In this paper we have discussed some issues of content management in digital libraries. These observations are the result of work on system of digital libraries being conducted under the PIONIER programme. The system is one of the most important elements in general infrastructure of the programme. The system makes it possible to use achievements of the PIONIER programme like distance learning, groupwork applications etc. by information society.

We identified three crucial factors that contribute to the overall quality of publication and library maintenance processes: content organization, document versioning and access management. When attempting to build a large and well organized digital library, none of the elements can remain underestimated. During the design and implementation of the *dLibra* system we have put much emphasis on the library maintenance issues. The system proves to be an efficient tool for publishing and managing digital documents. Some aspects of the publishing process are currently investigated and have not been addressed in this article, e.g. groupwork and workflow management, document preparation process. Research and development will be continued to incorporate these ideas into the *dLibra* system.

Presently, *dLibra* is being put into practice in PSNC in order to facilitate publication, storage and access to company internal documents such as articles or reports. This will enable to evaluate and draw conclusions on the system's performance in an average workload environment.

References

[1] Jan Węglarz et al. *PIONIER — Optical Internet in Poland*. ISThmus, Poznań, April 2000.

[2] Donatella Castelli and Pasquale Pagano. *SCHOLNET — Global System Architecture Report*. CNR-IEI, Area di Ricerca di Pisa, 56124 Pisa, Italy, http://www.ercim.org/scholnet/del/D2.2.1-V2.pdf, January 2002.

[3] Joost G. Kircz. New practices for electronic publishing: how to maintain quality and guarantee integrity. In Dennis Shaw, editor, *Proceedings of the Second ICSU-UNESCO International Conference on Electronic Publishing in Science*. http://associnst.ox.ac.uk/∼icsuinfo/proc01fin.htm, 2001.

[4] Frédérique Harmsze. *A modular structure for scientific articles in an electronic environment*. PhD dissertation University of Amsterdam, http://www.science.uva.nl/projects/commphys/papers, 2000.

[5] Dublin core metadata element set, version 1.1: Reference description. Technical report.

[6] Cezary Mazurek and Sebastian Szuber. *Development of Digital Libraries at Poznań Supercomputing and Networking Center*. ISThmus 2000. Research and Development for the Information Society, Poznań (Poland), April 2000.

[7] Cezary Mazurek, Maciej Stroiński, and Sebastian Szuber. *Digital Library for Multimedia Content Management*. ERCIM 9th DELOS Workshop, Digital Libraries for Distance Learning, Brno, April 1999.

BIOGRAPHY

CEZARY MAZUREK, born in 1969, received his Master's Degree in Computer Science at the Poznan University of Technology in 1993. He worked for Poznan University of Technology between 1993 and 1994 and from 1993 for the Poznan Supercomputing and Networking Center (PSNC). He is currently the head of Network Services Department at PSNC. He is leading the development of services based on Internet technologies (e.g. Digital Library Framework: dLibra, Polish Educational Portal in co-operation with Interkl@sa, Multimedia City Guide in co-operation with Poznan City).

Multi-Agent Approach for Task Related Decision Supported Information Retrieval

T. ATANASOVA*, H.-J. NERN**, F. PAUTZKE***

*Institute of Information Technologies, BAS, Bulgaria, t.ata@dir.bg

** Aspasia-Systems
Technology Department Duesseldorf, Germany
info@aspasia-tech.net

*** Institute of Measurement and Control, University of Applied Science,
Bochum, Germany, friedbert.pautzke@fh-bochum.de

Abstract: In the paper an approach for building research models accessing distributed and heterogeneous knowledge pools using multi-agent information retrieval is presented. Accordingly task-oriented scenarios for intelligent information retrieval are discussed. Searching and browsing activities range from a well-defined search for a specific document to a non-specific task to estimate which kind of information is available. An algorithm for information retrieval is proposed that provides decision support to optimise the search results. The system uses metadata information schemes to provide services in refining user queries to focus a search, in automatically routing queries to relevant servers and in clustering related items.

Key-Words: multi-agent system, decision support, knowledge network, metadata description schemes, information retrieval, distributed and heterogeneous knowledge pools

1 Introduction

Nowadays intelligent software agents became popular research objects. In the past many attempts have been made towards the development of agents that assist in dealing with large information sources [1, 2]. Metacrawler is an agent that provides a common interface to a number of search engines. Webwatcher and Letizia agents are designed to assist and providing personalization to the user while browsing the WWW by performing a breadth-first search on the links ahead and provide navigation recommendations. Webcompass is directed towards off-line search and indexing. NewT is a system for WWW document filtering which utilized the weighted keyword vector representation. In terms of evolutionary filtering systems, NewT is a multi-agent system that uses evolution and relevance feedback for information filtering. NewT's application domain is structured newsgroups documents and the system is able to adapt successfully to such a dynamic environment. NewT employed only one kind of agents, namely specialized information filterers [3]. The Amalthaea system [4] introduces different types of agents, which base their relationships on a simple economic model. In [5] a multi-agent architecture for intelligent websites is presented and applied in insurance. The architecture has been designed and implemented using the compositional development method for multi-agent systems DESIRE. The agents within this architecture are based on a generic broker agent model. It is shown how it can be exploited to design an intelligent website for insurance [6].

The main difference of these acknowledged systems to the solution proposed in this paper is the use of knowledge representation schemes, the decision supported, task oriented information retrieval.

The paper is organized as follows: first the knowledge representation scheme is described, the task-based decision scenarios are proposed and the agent specialization is considered.

2 Knowledge representation

Knowledge databases (KDB) in the research knowledge network consist of theoretical and empirical knowledge distributed in heterogeneous sources.

The information in the KDB is accessed through:

- formulation of requests based on information demands;
- identification of potential sources of information;

- developing of successful search strategies.

In this investigation the decision models about the retrieved documents are constructed by using metadata for information objects. Current and previous problems are described in the KDB using a vocabulary of terms from the domain ontology by metadata. This is an extensible, scalable method of representation and it supports the requirements for decision-making. For the hypertext documents in the web the well known Dublin Core standard [7] has been developed. The Dublin Core (DC) is a 15-element metadata element set intended to facilitate discovery of electronic resources. DC serves as a core element set for resource discovery.

This representation provides a template for sets of objects. It also serves as a very important role for supporting distributed queries based on knowledge object type as a content-based document catalogue and search tool.

Thus, a knowledge object is represented by an object feature vector, whereby the attributes are summarised in an extended Dublin Core metadata set. The extension consists of metadata description for structured tasks and dynamically added optional attributes that reflect decision results of searching, for example, satisfying factor, relevance feedback factors and further classification items.

Within the prototype development and investigation phase of the proposed system the domain of control theory is chosen on the base of the Mathematics Subject Classification (MSC – 1991). Every object in the hierarchical structure receives primary and secondary classification numbers related to the relevant section classification subjects.

3. Decision scenarios for task oriented information retrieval

For the user given research task it is needed to recommend a list of approaches, methods, documents, software libraries and other relevant knowledge objects that satisfy some criteria of fitness. The task may be defined in different manner ranging from simple set of keywords to structured task description.

In this approach it is proposed that the decision about retrieving suitable knowledge objects is made according 3 possible scenarios (Fig.1):

1. searching by keywords;
2. tasks comparing - on the base of task description;
3. comparing of structured task - on the base of structured task descriptor

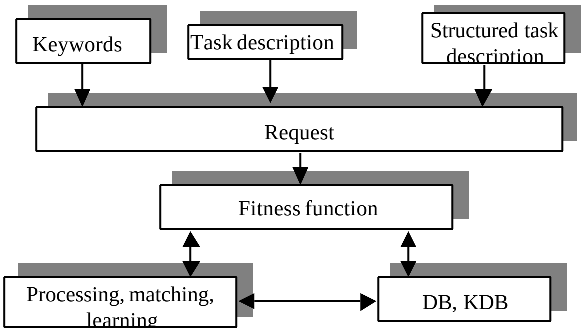


Fig. 1. Decision scenarios

3.1 Description by Keywords

The search by keywords is the most familiar case. The document may be found by identifying one or more fields of the given template given by the Dublin Core extension.

To enhance the usefulness of the results an optimisation is needed. The optimisation is proceeded on the base of clustering "Keyword/Value Notation" within the given ontology. Different clustering techniques in the feature space can be used for the ontological filtering engine. The cluster abstraction allows a large information space to be treated as a unit, without regard for the details of its contents. Clusters also provide convenient units for the partitioning of work and resource allocation among the distributed components of the system.

3.2 Task Description

Searching by general text description is a further facility. The user may describe his research task in free narrative text. In this case it is necessary first to provide pre-processing of this text in order to extract the description by keywords and after that to compare it with the existing clusters. The pre-processing may be done by some neural network techniques.

3.3 Structured task description

The space description of the structured task proposed in this approach is shown in Fig. 2. The metadata representation is used for describing this structured task. Some modifications are needed to adjust the set more suitable for the system's purposes. Similar to the knowledge representation, discussed in chapter 2, an extension of the Dublin Core is proposed for metadata description of the structured task (Table 1). In such a way it is allowed to embed structured content representation in the KDB.

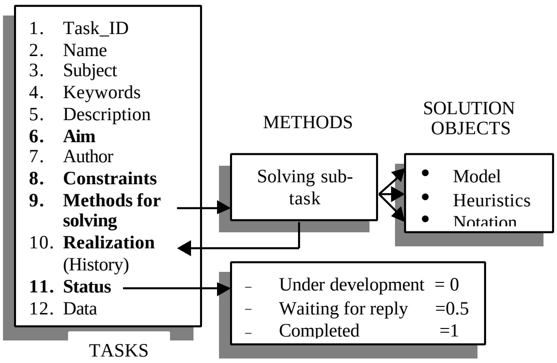


Fig. 2. Space of structured task

Thus the structured task descriptor vector can be described by Dublin Core metadata set plus the task extension:

1	Aim
2	Constraints
3	Methods (models) for solving
4	Realisation (history)
5	Status
6	Domain

Table 1. The extension of DC for structured task description

4. Algorithm for the decision support process

Searching of useful information is determined as an iterative feedback process. The target function is to minimize the quantity of trial and errors and to maximize the fitness function of the response to a given query.

Searching of a solution for user specified tasks is provided by database search utilities in case of the local knowledge base. If the local search gives insufficient results then the search in the web can be proceed. Web-agent search in the web, produces a metadata representation of the context in extended Dublin Core. The pre-filtered metadata (object vectors constant part) presents the document metadata instances. General user makes metadata query. Tools for realization are XML/RDF [8].

The system collects the knowledge objects using several autonomous web agents, which automatically analyse, index and assign the knowledge objects to subject classes and clusters. Ontological filtering engine based on clustering algorithms is used. Every cluster consists of the following detailed information:

- the list of members;
- summary description (set of keywords);
- related clusters.

The algorithm for the decision support can be summarised as follows (Fig. 3):

1. define the task;
2. define the fitness function - (set of keywords, or: matching parts of Structured Task Description Vector with metadata tags in DB);
3. search in DB (and web);
4. criteria for stopping (in internet-based searching);
5. show the results.

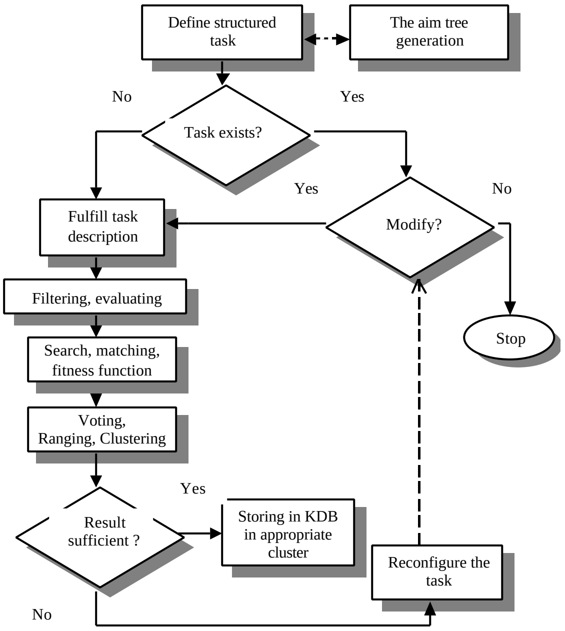


Fig. 3. An algorithm for decision support

The decision results received on the base of this approach can be tested and evaluated (Fig. 4). If their quality is good then the results are saved into knowledge base. The history of best results is archived and the context-based information is fixed and assigned in relation to the given task. Voting and feedback features allow a quality estimation of the proposed solving results. These systems are placed within the learning module of the Decision Support Module. Furthermore pure statistical data are included in the result vectors.

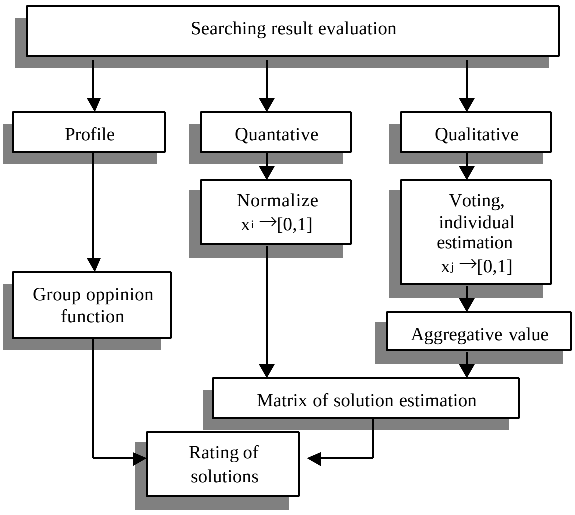


Fig. 4. Result evaluation

5. Agents for information retrieval, filtering and decision support

For realizing the proposed algorithm the agent technology is used. Instead of constructing a large, complex agent that have to solve the whole problem, it is more reasonable to create a group of smaller and simpler specialized agents that try to solve the problem collectively. Information filtering, discovery and estimation agents form a system of cooperative and competitive units.

Multiple agents work on a hierarchical structure identifying and grouping the knowledge objects. Agents execute tasks on behalf of a decision process, computer application or an individual user. Numerous agents in the multi-system can individually and concurrently look for groups of relevant objects at appropriate levels. More complex components are representing by hierarchy of agents. There are different groups of agents: info broker, DS agent, knowledge acquisition agent. One group of agents is searching in local DB and other group forwards the user query to all search engines, collects the results and returns a unified list according the object vector- metadata description.

An agent can perform only limited jobs: structured task edition, task managing, conversion of input data to meta data object vector, building decision model, producing relation reference vector by fitness function. Interface agent contains the object vector and sends it to the local DB or to the web agent. History agent corporate history or best practice and context based information. Other agents optimise the results, test them, and statistically estimate additional factors. For realising a higher goal, the activities of cooperative agents should be coordinated. Several coordination mechanisms [9] exist such as competitive coordination, cooperative coordination, temporal sequencing or no coordination.

5.1 Learning agents

The first stage of searching is the access of preliminary results. The second stage is their estimation and learning by user's feedback. This is done by decision agent - software unit that consists of hierarchical modules according Fig. 5. Agents are learned to create a user/task profile, to correct values of the relation matrix, to correct the list of objects or feedback from the user.

Agents allow inclusion of new categories and rules using historical practice agent, acquiring the knowledge.

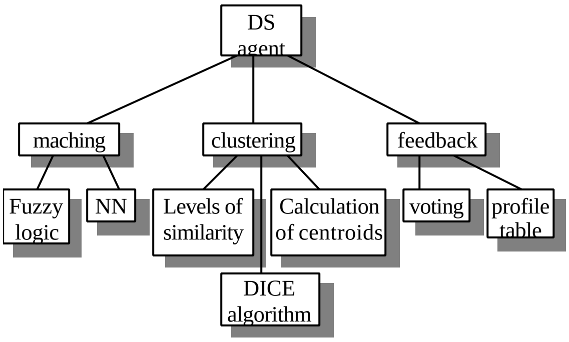


Fig. 5. Agents for decision support

5.2 Agent Architecture

The generic agent architecture [2] can be instantiated by adding specific types of knowledge to support functionalities and behaviour required. Depending on the choice of these requirements, an agent is created for a specific context by including the appropriate types of knowledge. The agent architecture supports its own modification due to the fact that basic functionalities are specified in an explicit declarative manner, in the form of knowledge. It is possible to dynamically modify the agent by adding or deleting some of its knowledge.

6. Conclusion

The amount of information available online and distributed among heterogeneous sources is constantly increasing. Users are interested in information that can be found in many different sources. Instead of constructing a large, complex agent that have to solve the whole problem, it is more reasonable to create a group of smaller and simpler specialized agents that try to solve the problem collectively. Information filtering and discovery agents form a system of cooperative and competitive units. The system adapts to the user's interests, which may change over time, while exploring new domains in local DB and the web that may be of interest to the user. In this paper an approach for searching and retrieval of relevant knowledge objects in distributed heterogeneous network environment is proposed. The search and decision support of results is done by agent technology on the base of extended DC metadata. Ontological filtering engine is used to response to simple user requests in the form of keywords. If the request is made in the form of free text then the intelligent processing is applied in order to obtain the weighted keywords that allow using the ontological filtering engine. Searching and browsing activities range from a well-defined search for a specific document to a non-specific task to realise which information is available. The structure of the given research task according to the metadata standard for hypertext documents is derived. The structured task description vector is used for searching in the local KDB and distributed information sources in the knowledge network. A query refinement is needed to overcome the problem of large result sets. This is provided by suggesting modifications to focus user queries. Query refinement is based on clustering algorithms. The organisation resp. classification of information into clusters of related items assists both the users and the system when dealing with large information spaces. The multi-agent system establishes related document retrieval, performs similar document identification and collaborative filtering, which helps scientific, technical and other professional Internet users building research models.

References

1. Honavar V., Les Miller, J. Wong. Distributed Knowledge Networks. http://www.cs.iastate.edu /~honavar/aigroup.html
2. Jonker, C.M., and Treur, J., (1999), A Re-usable Broker Agent Architecture with Dynamic Maintenance Capabilities. In: O. Etzioni, J.P. Mueller, J. Bradshaw (eds.), *Proc. of the Third Annual Conference on Autonomous Agents, Agents'99.* ACM Press, pp. 376-377.
3. Moukas A., (1996) Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem, Proceedings of the Conference on Practical Application of Intelligent Agents & Multi-Agent Technology, London.
4. Amalthaea: A Multi-Agent System that Discovers, Monitors and Filters Information Resources.http://belladonna.media.mit.edu/projects/amalthaea/
5. Jonker, C.M., Lam, R.A., and Treur, J., (1999) A Multi-Agent Architecture for an Intelligent Website in Insurance. In: *Proceedings of the Third International Workshop on Cooperative Information Agents, CIA'99.* Lecture Notes in AI, Springer Verlag.
6. Bjorn Hermans, (1996), Intelligent Software Agents on the Internet: an inventory of currently offered functionality in the information society and prediction of (near-) future developments, Tilburg University, Tilburg, The Netherlands, July.
7. Dublin Core metadata set recommendation (1999). http://purl.oclc.org/metadata/dublin_core/
8. Extensible Markup Language (XML), http://www.w3.org/XML/
9. Lesser, V.R. (1998), "Reflections on the Nature of Multi-Agent Coordination and Its Implications for an Agent Architecture", Autonomous Agents and Multi-Agent Systems, 1, pp.89-111, Kluwer Academic Publishers, Boston.

WebTrust
Agent Platform for Identifying Trustworthy Commercial Websites

ir. W.F. Stoutjesdijk
Accenture Technology Labs
BP 99, Les Genêts
449 route des Crêtes
06902 Sophia Antipolis, France

drs. dr. L.J.M. Rothkrantz
Delft University of Technology
Mekelweg 4
2628 BZ Delft, The Netherlands
L.J.M.Rothkrantz@cs.tudelft.nl

August 7, 2001

Abstract

In this document we present a blueprint for agents that can label websites according to a personalised view of trustworthiness. This trustworthiness goes beyond the technical trust that is implemented in security related issues and covers the whole 'feeling' of a trusted website selling a product. A model of site properties that contribute to this idea of a (dis)trusted site was designed. Agents will learn how their user reacts to these properties and calculate a trust-value upon it. With enough trained sites the agents can start labeling new websites. When the user is not acquainted with the topic of the site he can ask the agent of an expert or a consumer magazine to help: agents can negotiate with each other to get a more profound rating.

1 Introduction

On the ever growing Internet more and more products are being sold. To get the best price for an item one has to search a lot of sites. Only some of them look as if they will actually ship your order, handle your credit-card number properly and not pass your personal information around: not all sites can be *trusted.*

In this document we present agents that can help decide whether or not to order, by calculating a personal trust-label for websites. These agents can negotiate with other agents to come up with a shared trust rating.

2 Related work

2.1 Human trust

In his "Formalising Trust as a Computational Concept" [5], Marsh describes trust as a human phenomenon in 3 layers:

Basic trust; inherent in the trusting person. One can be an easy truster or a hard one. As with more things in life we can distinguish between optimists and pessimists: "...optimists are general trusters whose disposition to trust is relatively inflexible in a downward direction, despite past experience. Thus their trust in others can only increase, never decrease. The opposite is true for pessimists" (page 56). It is the way they look at life (from a trust perspective). They are gullible, suspicious or something in between.

General trust; measured from one subject to another. In other words how much person x trusts person y in general as a person.
Seen from a commercial perspective this is a customer-company relation. It is calculated regardless of the medium (shop, Internet or telephone). It is purely what the customer thinks of the company.

Situational trust; the most specific of the three describes trust in a specific context. The example Marsh gives is very clear: "while I would trust my brother to drive me to the airport, I most certainly would not trust him to fly the plane!".
In other words one can have General Trust, but not in every situation. For trust on the Web the site is the context: although you trusted this company, their site changed this perception.

Temporal issues are also discussed in his model, but will not be taken into account in WebTrust. From here on, trust factors can be seen from these three perspectives.

2.2 Trusting websites

Since the start of Internet commerce, company websites have been under scrutiny of trust. People are suspicious towards these sites. Cheskin did a survey on trust amongst Internet users in the Americas [1] [2]. Two of their main findings are:

1. Some people take more risk in trusting than others: they trust more. Cheskin actually confirms Marsh's idea here that people have different Basic Trust.
2. Respondents said to look at specific properties of the site. They looked for the presence of seals of approval, brandnames and layout. These seals of approval are logos from 3^{rd}-parties that guarantee security (VeriSign and BBB Online) or insure the transaction (American Express and Visa). These parties benefit from the quality of the transaction and associate themselves with 'trusted' sites. They are also called trusted 3^{rd}-parties.

Egger [3] makes use of this information to advise companies on how to make trusted websites. He makes the companies ask their customers what they think of the company's reputation, the content on the site, which 3^{rd}-party to choose and what the privacy policy should say.

2.3 Web of trust

Consumers do not have to depend on trusted parties at eOpinions [4]. On this website, where consumers review products and sites, everyone is part of the *Web of trust*. Users have a portfolio of people they trust. They do this either because they have read their reviews and concur with them, or because they trust the people *this* user has in his portfolio of trusted users. The idea behind this web of people is based on the assumptions that

- friends have a proven track-record. If a friend consistently gives you good advise, you are likely to trust him the next time too;
- you and your friend share interests. If you both like the same kind of films, you are more likely to trust his recommendations.

3 Feature Model

In order to get a more complete picture of the properties of sites mentioned by Cheskin, a survey was done. This survey asked 19 respondents[1] their trust opinion on 50 websites. On a 5-point scale they indicated how much they would trust sites (or the companies behind them) to do what they promised to do. These promises range from correct handling of payment and credit-card related issues to shipping the order and not spreading around personal information to others.
Aside from this, respondents had to write down anything that made them trust this site the way they did: the properties of the site and related issues (the fact that the company is public on the NASDAQ for example).

The reasons people submitted for their trust opinion were split up into 22 categories. Some of these categories are regarded as not trust-related: respondents indicating that they were not interested in the product or service, for example. From the others, the following 5 scored best (in order):

1. Design; the layout of the site was often mentioned positively and negatively. Sites that did not live up to the standards set by the 'professionals' and did not have anything else to offer (e.g. brandname) were rated negatively. Whereas websites with no track-record at all had benefit from their appearance.
2. Reputation; this category of answers was used positively in every case. When a company or its website is well-known, people have an opinion on it. This is often positive because respondents reason that a well-known company has to deliver good service. This category does not include 'experience' with the site, the product or the company.
3. Company information; respondents looked for real-world information on the website. They wanted to know where the company was based, what its phone number was etc. Contact information other than Internet related, gives more confidence in the company.
4. Privacy Policy; many sites have a privacy policy that states how the company will treat all personal information supplied to it. Respondents never mentioned any statements made inside this policy, just its existence.
5. Trusted Seals; as mentioned before, these seals are logos from organisations that are regarded as trusted.

Apart from the company information, no text-related issues are mentioned here. People do not seem to *read* the site, but merely *look* at it. These findings are confirmed by Cheskin's survey mentioned earlier, in which people looked for trusted seals, brandnamens (a form of reputation) and layout (design).

The ratings respondents gave to the sites were very conservative. Even though some of the sites could be regarded as awkward, the answers average on the 'Neutral' score. In Table 1 the ratings can be seen per scale-point.

To confirm the fact that trust is a subjective issue the agreement amongst respondents was measured. The 5-point ratings were mapped onto 'Negative', 'Neutral' and 'Positive'. For every site, the number of these ratings were counted. The agreement is then defined as the maximum group of ratings divided by the number of ratings. If 5 people said 'Negative', 1 said 'Neutral' and 2 said 'Positive' then the largest group is 'Negative'. The agreement would be $\frac{5}{5+1+2} = 63\%$.
This agreement measure strongly suggests that trust is subjective: only 6 out of 50 sites score an agreement of more than 75%. People do not seem to have the same

[1] 10 of which filled in the complete survey and are used in further analysis.

Table 1: Spread of the trust ratings

Not at All	*A little*	*Neutral*	*A lot*	*Absolute*
13%	21%	29%	24%	14%

trust rating for sites.

From work presented earlier and this survey, trust on the Web was understood as a process: people investigate the site for properties they find important and supply these properties with their personal weights. The 'sum' of these weighted properties will determine their judgment. The investigation of a site can involve the following:

Experience; the user will see if he knows the company or site already. If he does, he already has an idea about the trust he can attribute this site.

Layout; the design of the site is taken into account. Especially for unknown sites this can be very important. It is a sign of professionalism when it looks well. This will be reiterated later in this section.

Interaction; part of the layout are links and buttons. These are susceptible to errors. If the designer is not careful they may not work: the link does no longer exist or the button generates some kind of error. All these things can be discovered by interacting with the site: clicking on it.

Other people; finally the opinion of other people is important as shown by the eOpinions website, discussed before. In general these people can be friends, consumer magazines, newspapers or any other trusted party with an opinion. The trust rating of these others with regards to this particular site is of importance to the user.

All these questions asked and tests performed to assess trust are used to measure the *features* on a website. A feature is defined as a measurable entity present on, or related to a site or the organisation behind it.
It is the basis of the *Feature Model.* Except for the 'Other people' item which is part of the negotiating agents principle, explained later on.
The Feature Model consists of a list of all features identified by this research together with their metric, split up into 4 categories defined for it:

Reputation is an easy way for people to decide on a site. If they know it already because they read about it in a magazine, saw the company on television or heard about it from a friend they have an initial idea about its trustworthiness. Sometimes this idea is so strong that they will decide without really taking the site into account or even contradicting their feelings about other features. Reputation can be personal as well: the fact that they know it because they have used the site before (experience) is also included[2].
Reputation of a company and its site are not found on the site itself but on others indexing it. The features in this category are all external sources of information that give an impression on how well-known it is and what people think of it.
Examples: Number of references to this site (reversed search in AltaVista) and trust of partners (trust rating for outward pointing links found on the site).

Professionalism one seeks in both brick-and-mortar companies and online ones. It is typically a vague subject people refer to. Even the Cheskin report talks about "Clarity of purpose" and "Craftsmanship" (ibid. page 11).
Examples: Presence of 3^{rd}-party seals, number of pictures on the site and presence of a privacy policy.

Technical issues are noticed when they do not work, or when one wants the site to conform to a certain technical standard. This can be true for more categories, but here it is really obvious. The site is not working, not very fast etc. Or one really wants the site to be secure before ordering anything on it. Examples: Number of dead links found, use of encryption for site (indicated by the browser) and speed of the site (average roundtrip-time of packets).

Other features mentioned by people were hard to classify (i.e. implement) and ended up in a separate category. The modeling of them will take more research.

Using this model, sites can be represented as vectors: a vector of feature ratings. This will be called the *feature vector* of a site.

4 Trust agents

The Feature Model for trust is used to define the *trust agent.* This agent has 2 basic functions. It can

1. give a trust rating to a site based on what it perceives as trusted features;
2. communicate this rating to other agents to come up with a *negotiated* trust rating.

Both of these functions will be discussed below.

[2] As opposed to the survey category, where experience was not included.

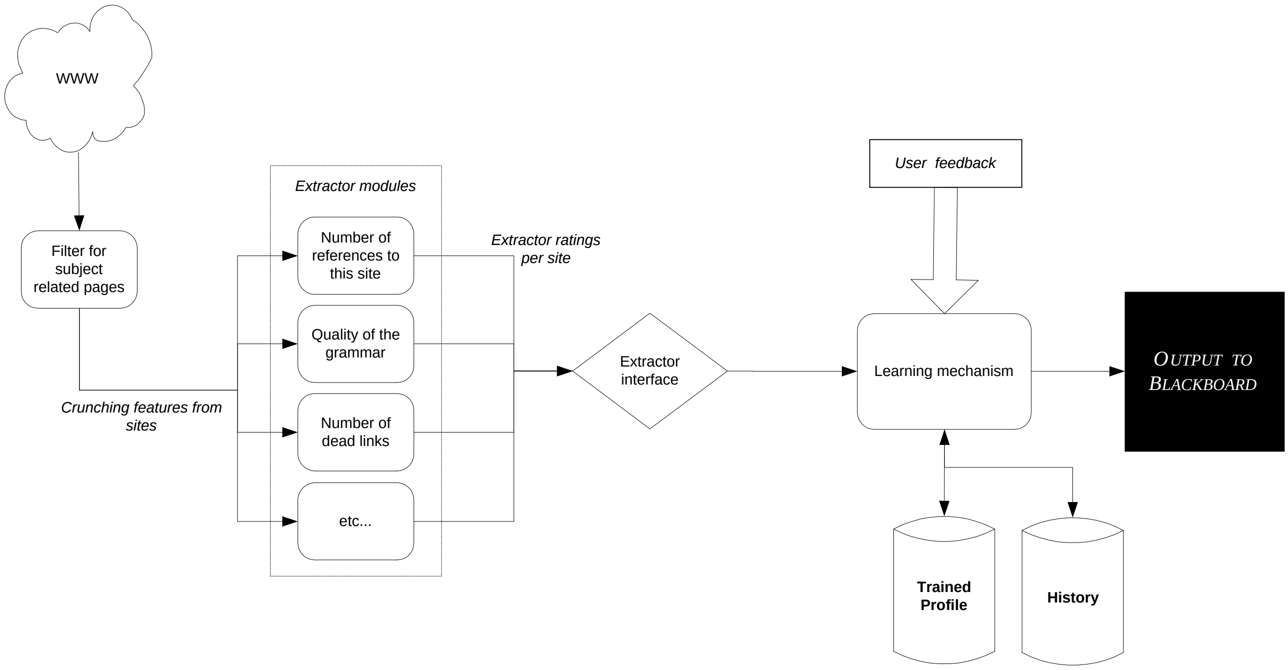


Figure 1: Architecture of a trust agent

4.1 A trust agent

The agent that was designed to rate sites is connected to a search engine that selects subject related websites. This is depicted in Figure 1. The list of websites that remain can now be labeled.

For every page the agent will extract features. This is done by extractors. Given the URL of the page, an extractor can calculate one feature rating (there is one extractor for every feature). For the 'quality of grammar' feature, for example, it will take all plain, visible text from the page and run it through a spell checker. The number of mistakes the checker finds (relative to the number of words) will be returned by this extractor. Some extractors use external sources. The 'number of references' for example, will go to AltaVista and do a reverse search for the website. This search will return the number of pages that point *towards* this site. This score will be returned as its rating. All these extractor ratings together make up the feature vector of the site.

The extractor interface is there to make the extraction process transparent to the rest of the application. This is done because in this way extractors can be added and removed in a modular way.

As stated before, trust is a subjective matter and the agents, therefore have to be user specific. The measuring of the feature vector was still done in an objective way, but the way the agent 'sees' these numbers is different for every user. In other words, the agent has to know which features contribute to the user's sense of trust and how much. It has to know the user's trust behaviour.

Calculating trust

Before the agent will make any recommendations it will show the user some selected sites and ask his trust rating for them. The assumption is, that in this way it will learn how the feature vector influences the final rating. For this learning process 3 methods have been implemented and tested:

- Feed-forward neural network; will take the feature vector as an input and train on the feedback given by the user.
- k-Nearest Neighbour; plots all trained sites in a space and will decide on the Euclidean k nearest ones.
- Decision Tree; takes every extractor as a leaf and calculates the order in which they are traveled down by looking at the most distinguishing extractors (algorithm from Russel & Norvig [6, page 534]). When, for example, the presence 3^{rd}-party seals always leads to a 'Absolute' trusted site, this will be a very distinguishing feature placed high in the tree.

When the feature vector is presented to one of these mechanisms, the agent will first look in the history file. If the URL is already present here, the user has already rated it and his own answer will be returned. If this is not the case, the mechanism will have to infer the answer from the trained user profile, which was generated by the first selected sites. If the user does not agree with the answer, he can supply the mechanism with feedback in order for it to learn more.

Together with this answer comes a confidence level. This level indicates how 'sure' the agent was given its inference mechanism and data. For the k-NN, for example, the confidence is partly based on the diversity in the 'neighbourhood' out of which the answer was generated. When $k = 5$, there are 5 sites close to the tested one that have ratings. These ratings can be very diverse, but one answer has to be chosen from them. The diversity is then defined as the number of answers in this set that were different from the one inferred. Confidence levels are expressed as percentages. Once the mechanism has inferred an answer and calculated its confidence, it will be written to the blackboard for other agents to see.

4.2 Negotiating on trust

Based on eOpinions' idea of the Web of trust, the user can select more agents to help him decide on trust. When looking for second-hand cars for example, he may want to ask a friend to take a look at the site or maybe even a consumer magazine. With WebTrust, these two parties can have their own agents. In this way, the user will profit from the trust knowledge put into the consumer magazine's agent together with his own to select sites that are trusted by both of them. The user's agent for example focuses on quality of grammar, whereas the consumer magazine's focuses on well-known cardealer names.

All agents selected by the user will write their ratings to the blackboard together with their confidence level. This data forms the input to the overall measurement, which is depicted in Figure 2.

This process models a basic form of human negotiation: when one enters a room with experts that all have an opinion one may ask them all for their answer and their confidence. The decision maker has to generate one single answer from all these experts.

In the current implementation all agents have equal weights, but for the negotiation to be more realistic, it is important to make these different. These weights can be used to turn the anonymous agents into agents that have a reputation for their field of expertise. Some agents will have a very low confidence level (below 30%), they are discarded. After this, the system looks for agents that are very confident (above 95%). This level is also equal to the agents that have found their rating in their history. Since this answer was supplied by the user himself, they are very confident. If confident agents are found then all the others are discarded. The weighted average (to confidence level) of the remaining agents will decide the final answer of the system.

5 Test results

The trust ratings yielded from the survey were used to test the quality of the agent mechanisms and the negotiation process. From the 50 sites rated by respondents a trainset was taken and the others had to be predicted. At the time of testing only 6 extractors had been implemented due to time constraints. This means that the predictions are not based on the full set of features that a respondent had taken into account when assessing his trust rating.

From the quality of this prediction a measurement could be made: taken over the average of all the predictions that had to be made, the quality measurement was defined as the Average error * (1 - Precision[3]). For a trainset of 40 sites and 10 predictions, the system's scores are displayed in Table 2. Presented in this table are the numbers for the 3 primitive methods and their 4 possible combinations. The lowest score is the best agent in this metric, the k-NN therefore scores best. The ANN combined with k-NN scores very well too. This is because the ANN most of the time, is not very confident and will not play an important role in negotiation, i.e. most of the answers are still determined by the k-NN method.

It can also be seen from this table that the combined methods do not score significantly better than the primitive ones. This is partly due to the calculation of the confidence level, which turned out not to work very well. The decision tree, for example, was always very confident, while its answers were very bad. All combinations with a decision tree have therefore been blurred by this.

Tests have also been done to see if the methods increase the quality of their predictions when the trainset was increased. Both the neural network and the k-NN managed to increase (lower quality measure), whereas the decision tree decreased when training more. For k-NN the average quality increase was 17% when doubling the trainset.

Finally benchmarks were setup to see if the agents actually scored better than (smart) random generators. We made an agent that gave scores uniformly and one that scored according to a normal distribution (which is a lot like the respondents of the survey did). They scored a quality measurement of 1.17 and 0.94 respectively. Even the worst method implemented, (combination of ANN and decision tree, 0.76) scored better than these benchmarks. Further improvements can be related to these scores.

6 Future Work

The quality measurements show that agents score better than the benchmark. However, they do need more extractors to have 'full vision'. The modular approach of the application facilitates adding extractors. This will therefore be relatively easy to do.

A more hybrid approach in which the extractors are measured by different methods may further improve results.

[3] relative number of correct answers

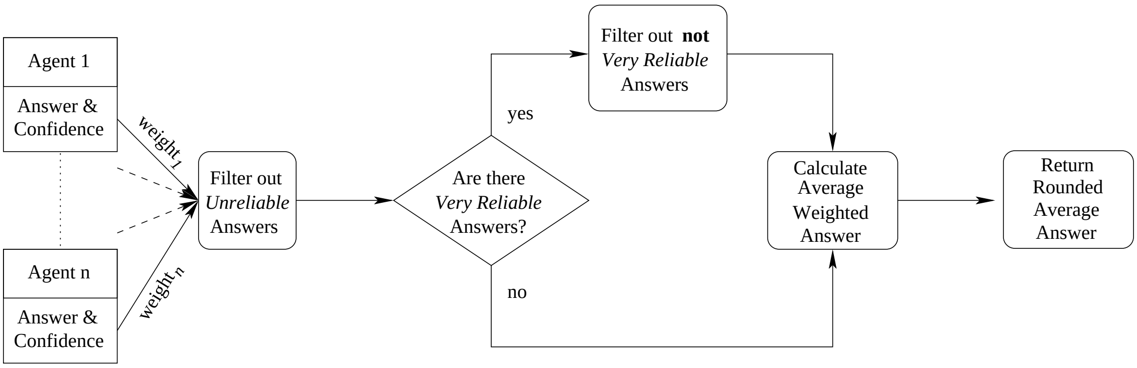

Agent 1
Answer & Confidence
Agent n
Answer & Confidence
weight 1
weight n
Filter out Unreliable Answers
Are there Very Reliable Answers?
yes
no
Filter out not Very Reliable Answers
Calculate Average Weighted Answer
Return Rounded Average Answer

Figure 2: The process of negotiating agents

Table 2: Quality measurement for prediction of 10 sites

ANN	*k-NN*	*Tree*	*A & k*	*A & T*	*k & T*	*All*
0.70	0.51	0.74	0.52	0.76	0.68	0.57

This means that methods specialise in certain features on a site. Neural networks, for example, are known to work well in visual recognition whereas the decision tree may work better when rule-based features are taken into account.
Furthermore, the negotiation process needs revision. Especially the decision tree needs a a confidence level measurement that better fits the actual quality of its rating. Other ways of inferring the answer can also be tried. The k-NN method can, for example, be 'upgraded' towards a full Case-based reasoning system with more knowledge.

7 Conclusions

We found strong support for our Feature Model describing properties of sites that influence trust behaviour. It presents a full listing of features that appear on, or are related to websites. Using this Feature Model, it proved possible to automate the labeling of commercial websites with respect to trust. The agents that were implemented, automated the way humans assess trustworthiness of a website by learning from their behaviour. k-Nearest Neighbour is the best method of the three tested to learn from human trust behaviour. Even combinations of these methods (by negotiation) did not perform better.

Acknowledgments

This work has been done as a Master thesis project for Delft University of Technology, the Netherlands in cooperation with Accenture Technology Labs, France. We would like to thank everyone who participated in the course of this project.

References

[1] Cheskin. eCommerce trust study. Technical report, Cheskin Research, Jan. 1999.

[2] Cheskin. Trust in the wired americas. Technical report, Cheskin Research, July 2000.

[3] F. Egger. Human factors in electronic commerce: Making systems appealing, usable & trustworthy. In *Graduate Students Consortium & Educational Symposium, 12th Bled International E-Commerce Conference*, June 1999.

[4] eOpinions.
http://www.eopinions.com.

[5] S.P. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, University of Sterling, UK, Apr. 1994.

[6] S. Russel and P. Norvig. *Artificial Intelligence, a modern approach.* Prentice Hall, 1995.

SIMULATION, MANAGEMENT EVALUATION AND CONTROL

NECTAR: Simulation and Visualization in a 3D Collaborative Environment

Law, Yee Wei
Facultaire der Informatica
Universiteit Twente
Postbus 217
7500AE Enschede, The Netherlands
Email: ywlaw@cs.utwente.nl

Chan, Kai Yun
Centre for Advanced Media Technology
Nanyang Technological University
Nanyang Avenue
639798 Singapore
Email: askychan@ntu.edu.sg

KEYWORDS
VRML 2.0 Standards, 3D Worlds, Synthetic Environments and Distributed Simulation (DIS).

ABSTRACT

For simulation and visualization in a 3D collaborative environment, an architecture called the Nanyang Experimental CollaboraTive ARchitecture (NECTAR) has been developed. The objective is to support multi-user collaboration in a virtual environment with an emphasis on cost-effectiveness and interoperability. The architecture is based on the modular decomposition of a virtual environment system into three basic components: the graphics engine, the user interaction engine, and the networking engine. The scene graph-based graphics engine, written in Java 3D, takes VRML models loaded through the NAnyang Vrml Elementary Loader (NAVEL), as input. For multi-user interaction, NECTAR calls for the Core Living Worlds (CoreLW) extension of VRML. Coupled with this content-level strategy, on the wire level, the Java Shared Data Toolkit (JSDT) is employed, to the benefits of monitored, multi-threaded, multi-sessioned and multi-channelled communication. Some thought has also been given to optimizing network communication. Lastly, the modular, loosely-coupled and layered architecture has been found to facilitate flexible incremental, component-wise development.

1. INTRODUCTION

Building a virtual environment (VE) is by no means an easy task, more so for networked virtual environments (net-VEs), as data now has an additional critical path: the network and the protocol stack. NECTAR offers a simple approach of modularizing the system into three loosely coupled components: the graphics engine, the user interaction engine and the networking engine. Right now, the targeted components are restricted to the graphics engine and the networking engine. The motivation behind the design and implementation of the architecture is that VE frameworks have traditionally been very expensive, based heavily on proprietary libraries and file formats, leading to a source base that is hard to maintain, limited upgrade path and strong vendor dependency. Also, the intention here is to work-around the traditional approach of creating VE application through low-level programming, by using file formats that provide high-level scripting functionality; in other words, a file format-centric approach is taken. The NECTAR approach emphasizes *component-wise development*, *extensibility*, *interoperability*, *cost-effectiveness* and *openness*.

The philosophy applied here is that to maintain the interaction of two or more independent users working together, a solution that adequately solves the problems for one person has to be in place first (Shaw et al. 1993). (Macedonia and Zyda 1997) outlines the main aspects of the problems as *bandwidth*, *latency*, *reliability* and *scalability*. To address these problems, a combination of architectural strategy and optimization techniques is utilized.

2. RELATED WORK

There are generally two ways for building a net-VE framework. First, mainly for older VEs which are not developed with networking in mind, it is customary to first build a standalone VE, then retrofit it with networking capabilities. Notable examples of this approach are demonstrated by MR Toolkit/Peer Package (Shaw et al. 1993), SVE/RAVEL (Kessler 1997, Kessler et al. 1998), WTK/World2World (Rahn 1998, Sense8 Corporation 1997), Maverik/Deva (Hubbold et al. 1996, Pettifer et al. 2000), CAVERN/CAVERNsoft (Park et al. 2000). The advantage of this approach is that developers can focus on the fundamental aspects before working towards the more advanced bits, delaying optimization to a later stage. Although the graphics engine and the networking engine can evolve separately, they are not completely uncoupled, as the internal representation of the graphics primitives indirectly affects the design decision of the networking engine. Such is the case with NECTAR. This is why NECTAR adopts a more “universal” representation of the graphics primitives, i.e. VRML, so that the networking engine is optimized for VRML. This is the biggest difference between NECTAR and the above-mentioned frameworks. Most of these frameworks maintain a low-level API to remain file format independent. The disadvantage is that building a VE application becomes a steep effort, typically entailing the writing of C/C++ programs.

The second way is with database/persistence and networking functionality integrated from the ground-up. Examples are

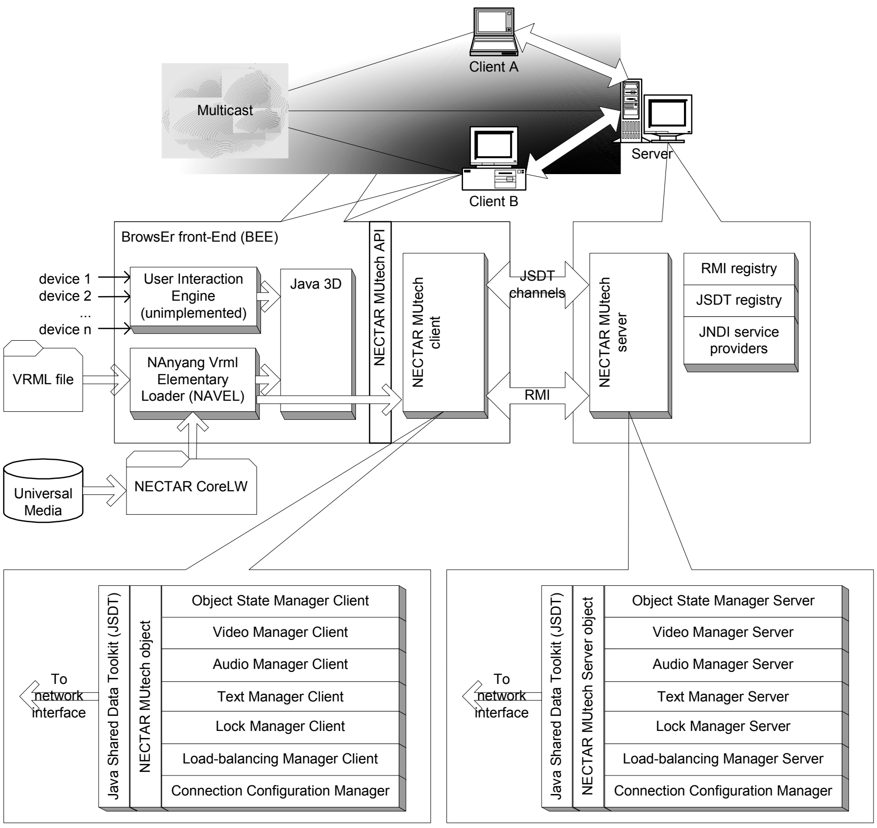


Figure 1: The Nanyang Experimental Collaborative Architecture (NECTAR)

VEOS (Bricken and Coco 1993), DIVE (Frécon and Stenius 1998, Hagsand 1996), Open Community/ISTP (Waters et al. 1996, Waters et al. 1997). This approach minimizes data conversion overhead as the graphics objects map readily to the distributed objects. However such tight integration also minimizes the degree of mutability. A slight change in any component might trigger corresponding or higher amount of change in the other components. NECTAR occupies the standpoint that before any optimization is attainable, it is more profitable to adhere to the loose coupling model. It is interesting to observe that the first type of (loosely coupled) frameworks seems to enjoy more development than the latter nowadays.

There is active research in the area of general-purpose networking framework. Net-Z (Proksim Software 2001a, Proksim Software 2001b) uses abstractions such as duplication objects, duplication space, match function, publisher and subscriber to improve scalability. In NECTAR, similar abstractions that are appropriate for VRML are used. InVerse (Singhal et al. 1997) emphasizes the fact that different information types are to be handled differently, e.g. video stream is handled differently than object states. NECTAR endeavours to support the basic information types as defined by InVerse. Also, the channel abstraction used in InVerse has important influence on the design of NECTAR.

3. THE GRAPHICS ENGINE

The discussion of NECTAR starts with the graphics engine. The core part of it comprises of the NAnyang Vrml Elementary Loader (NAVEL) and Java 3D (Sun Microsystems 2000) (Figure 1). Java 3D is a fourth-generation graphics API by Sun Microsystems. The decision to use Java 3D as the graphics API is based on careful evaluation of its features. Most notably, it is due to its scene graph abstraction and optimization, cross-platform support, interoperability with other Java APIs, hardware acceleration, VR inclination, and support for various advanced features like high-resolution locales and 3D sound.

While the heavy lifting of graphics rendering is taken care of by Java 3D, NAVEL is responsible for interpreting the VRML definition of objects into scene graph components. The similarity in scene graph structure between VRML and Java 3D adds to the advantage of using Java 3D and translates to straightforward and efficient implementation. NAVEL is based on the now defunct VRML-Java3D loader of the Web3D Consortium. Even though Xj3D (http://www.web3d.org/TaskGroups/source/xj3d.html) has succeeded the VRML-Java3D loader, NAVEL was started earlier than Xj3D to be able to take advantage of it. As a result, NAVEL only supports VRML97 instead of the latest X3D specification. But NAVEL does not implement the whole VRML97 specification either. Nevertheless, compared with the original loader, it has the following improvements (Law 2001):

1. Better support for prototypes (PROTO nodes). Among the more important fixes that have been incorporated are:
    a. Correct parsing of IS maps in PROTO definition.

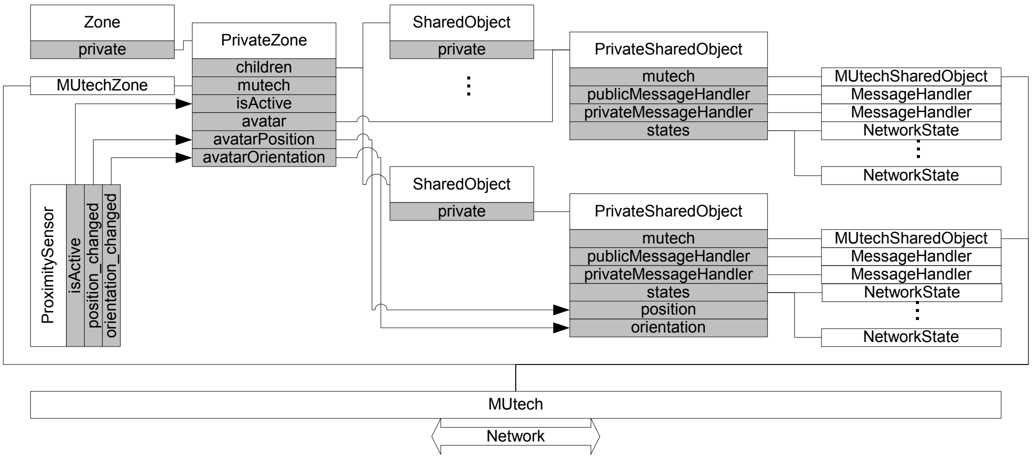


Figure 2: The Architecture of CoreLW

   b. Correct routing of field values.
   c. Correct resolution of DEF names.
   d. Correct recursive reference of PROTO nodes.
   e. Support for nested PROTOs.
   f. Correct instantiation of PROTO nodes.
   g. Sorted storage of field values.
2. Better and complete implementation of the Java scripting API as dictated by the VRML97 specification (Web3D Consortium 1997).
3. Full support for ECMAScript using Mozilla's Rhino library (http://www.mozilla.org/rhino).

Universal Media (http://www.web3d.org/WorkingGroups/media/) is considered an important feature of NAVEL. The benefits of Universal Media are substantial. First of all, these media building blocks have reusable value, saving content developers a bundle of development time. As the media is stored locally, download is avoided, and with the quality ensured by the Universal Media Working Group, users would have better browsing experience. Also, the feature is compatible with VRML97. The downside is that the browser has to handle URNs differently depending on the platform.

The BrowsEr front-End (BEE) is the GUI front-end for NAVEL. The core component of BEE is **`nectar.bee.BeePanel`**, which when plugged into **`nectar.bee.BeeViewer`**, **`nectar.bee.BeePlayer`**, **`nectar.bee.BeeScreen`**, **`nectar.bee.BeeApplet`** produces a static viewer, a fully dynamic browser, a full-screened application and an applet respectively.

4. THE NETWORKING ENGINE

NECTAR CoreLW

The networking engine consists of two major components: NECTAR CoreLW and NECTAR MUtech (Figure 1). Since the graphics engine is specialized in VRML, it becomes a natural requirement to be able to bridge the VRML contents to the networking engine. The Living Worlds Working Group of the Web3D Consortium has once worked on and came up with a solution. Even though the effort fizzled out at the stage of Draft 2 (Honda 1997), what remains of Living Worlds has converged to become Core Living Worlds (CoreLW) (http://www.web3d.org/WorkingGroups/living-worlds/CoreLW/spec.html). The purpose of CoreLW is to overcome the short-comings of Living Worlds in being over-ambitious. By being over-ambitious, Living Worlds is monolithic and too complicated. CoreLW is much more implementable, offers much more freedom, and since it arises out of the discussion among researchers and industrial specialists, though being unofficial, it is not one that is out of touch with the industry. What CoreLW offers is a first step to the multi-user extension of VRML.

Zone, SharedObject, NetworkState and MUtech

The essence of CoreLW lies in the concept of Zone, SharedObject, NetworkState and MUtech (Figure 2). Zone is the basis of area-of-interest filtering, whereby users in one Zone are only aware of the activities in that particular Zone. A Zone is also a container of SharedObjects. There may be a lot of objects in a Zone, but only SharedObjects are shared, i.e. whose states are maintained consistently across the network, among all the users. A SharedObject is also a container of NetworkStates. NetworkStates are the smallest units of shared information, e.g. a NetworkState node may be a NetworkSFBool, through which a boolean value is shared. PrivateZone and PrivateSharedObject are the internals of Zone and SharedObject respectively, that implements the actual funtionality. They exist on the reason of security.

In CoreLW, all the networking details are abstractized in a concept called MUtech ("Multi-User technology"). The MUtech object carries out all of the network related work. Associated with the Zone is a construct called MUtechZone (Figure 2). It is responsible for detecting when its associated PrivateZone becomes active; loading and initializing the user's avatar in the PrivateZone; facilitating the animation of the avatar's navigation; and attaching a MUtechSharedObject to every PrivateSharedObject in the PrivateZone.

Similarly, the MUtechSharedObject is associated with the SharedObject (Figure 2). It is responsible for facilitating state sharing between instances of the SharedObject;

facilitating inter-SharedObject communications; and managing the locking of SharedObjects.

It is important to note that MUtechs are not by definition interoperable. The reason is well given in (Honda 1997).

Separation of Functionality
The implementation of CoreLW in NECTAR is called NECTAR CoreLW, whereas the implementation of the MUtech object and all its associated machinery is called NECTAR MUtech. NECTAR MUtech is discussed later.

NECTAR CoreLW implements almost all of the CoreLW specification. To separate the functionality between NECTAR CoreLW and NECTAR MUtech, a distinct interface **`nectar.mutech.MUtech`** is created. In fact, NECTAR CoreLW does not refer to NECTAR MUtech directly. For the MUtechZone to get a handle of the MUtech object, the MUtechZone needs to supply an "implementation name" to the static method **`nectar.mutech.MUtechFactory.createMUtech()`**. The method constructs the class name of the MUtech object from: "nectar.mutech." + <implementation name> + ".MutechImpl". It then uses the class name to load the class of the MUtech object, and instantiates the MUtech object. This implementation name is specified by the virtual world author. Currently, as NECTAR MUtech is implemented on top of JSDT, the only valid implementation name is "jsdt". It can be easily seen that different implementations can be swapped in by supplying different implementation names.

Implementation
As mentioned, MUtechs are not meant to be interoperable, this is why the interface declaration of MUtechZone is deliberately left out by the specification. In NECTAR CoreLW, MUtechZone has the following interface:

```
EXTERNPROTO MUtechZone [
    #CoreLW/MUtechZone interface...
    eventIn SFNode set_privateZone

    eventOut SFString whichTechnology_changed

    eventIn MFNode childrenAddedToZone
    eventIn MFNode childrenAddedFromNet

    eventIn SFBool set_isActive

    eventIn SFVec3f set_avatarPosition
    eventIn SFRotation set_avatarOrientation

    eventIn SFNode valueToNet

    #MUtechZone/MUtech interface...
    field SFString implementation "jsdt"
    field MFString serverName ""
    field SFInt32 port 11976
]
"urn:web3d:nectar:include/nectar/corelw.wrl#MUtechZ
one"
```

As EXTERNPROTOs are used for all the CoreLW nodes (e.g. Zone, SharedObject etc.), to avoid download overhead, the importance of Universal Media is apparent.

NECTAR CoreLW is implemented in a mixture of VRML, ECMAScript and Java.

The structure of MUtechZone is shown in Figure 3. Data originating from the SharedObject to the network passes through the **`valueToNet`** field of MUtechZone, whereas data originating from the network to the SharedObject passes through the Java wrapper class **`nectar.mutech.ObjectWrapper`**.

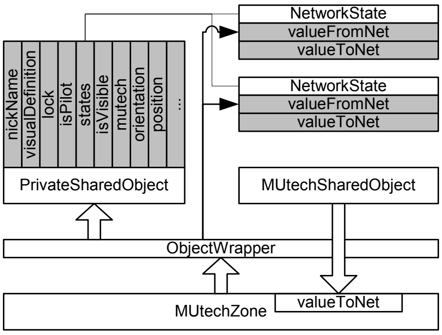


Figure 3: The Unsymmetrical Path of Values Updated to Net and from Net

The two main issues that the NECTAR CoreLW addresses are:

1. **The dynamic addition of SharedObjects to Zone**
   When a SharedObject is added to a Zone, it is added to the VRML scene graph *and* the object database. Depending on the type of SharedObject, different SharedObjects are added differently. *External objects* are objects that do not readily exist in the VRML scene graph. Currently, for simplicity and security, all external objects are obtained from the MUtech object. The API function is **`nectar.mutech.MUtech.listAddableObjects()`**. The external objects pass through the path in Figure 4 in the addition process. *Avatar objects* are a special type of external objects. They pass through the **`avatar`** eventIn of PrivateZone before going through the process in Figure 4. *Built-in objects* are those objects that are already present in the VRML file. Even though they are already present in the VRML scene graph, they have to be added to the object database. Their addition starts with the **`childrenAddedFromNet`** eventIn of MUtechZone, although they are not actually added from the net.

   Every SharedObject is assigned an object ID. However the job is relegated to the MUtech object, as the MUtech object relies on some networking-specific features to determine a unique ID. The API function is

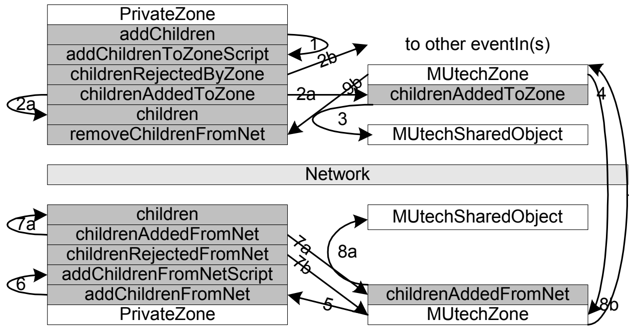


Figure 4: Dynamic Addition of SharedObjects to Zone

`nectar.mutech.MUtech.assignObjectID()`.

2. **The propagation of state updates**
   As mentioned, object states are contained in NetworkState nodes in the VRML scene graph, and the propagation of state changes follow the pattern in Figure 3. Object states are transmitted in serialized `nectar.mutech.ObjectState` objects. An `ObjectState` object contains the following information:
   a. Object ID: object identification number in string.
   b. Counter: a 16-bit sequence number which increments by one for every change in the object. The counter is managed by the class `nectar.mutech.CounterManager`.
   c. Owner ID: object's owner identification number in string. This is only set during full state updates.
   d. Array of field IDs: list of 8-bit field identification numbers corresponding to the fields that have changed. A field is encapsulated in `nectar.mutech.ObjectField`. The method `ObjectField.getID()` returns the field ID.
   e. Array of field vales: list of new values for the fields that have changed, ordered according to the list of field IDs above.
   f. Array of NetworkState tags: list of NetworkState tags in strings corresponding to the NetworkState nodes that have changed.
   g. Array of NetworkState values: list of NetworkState values for the NetworkState nodes that have changed, ordered according to the list of NetworkState tags above.

   Differential state updates are sent using the interface method `MUtech.sendDiffState()`, whereas full state updates are sent using `MUtech.sendFullState()`. An example of full state updates is when a client first joins a virtual world, or when a client introduces a new object into the virtual world.

CoreLW provides support for object locking explicitly, however the locking mechanism is currently not yet implemented.

NECTAR Mutech

As mentioned, the current implementation of NECTAR MUtech is based on JSDT. JSDT is an API designed solely by Rich Burridge of Sun Microsystems to support highly interactive, collaborative applications (Burridge 1999). The use of JSDT actually introduces another layer of abstraction because JSDT itself is a common interface beneath which a wide variety of implementation technologies can be employed. For instance, there are three default implementations that come with the toolkit, each being based on TCP/IP sockets, HTTP and the Light-weight Reliable Multicast Protocol (LRMP) (Liao 1998) respectively.

Strictly speaking, the current implementation of NECTAR MUtech is based on the TCP/IP socket implementation of JSDT, and is thus unable to support multicasting. LRMP is not used due to the reasons cited in (Waters et al. 1997). In either case, there is no perfect solution unless a custom implementation of JSDT is written. This is left as a future work.

The functionality of the MUtech server is exposed through the RMI interface `nectar.mutech.jsdt.server.MUtechServer` (Figure 5).

Zone and Multicasting

A Zone maps naturally to a multicasting group; in NECTAR MUtech, it maps to a channel. However there is not necessary only one channel for each Zone. In JSDT, a channel is uniquely identified by the host name, the port number, the connection type (e.g. “socket”, “http”, “lrmp” etc.), the client name, the session name and the channel name. In NECTAR, a client name is instead called client ID, session name called session ID and channel name called channel ID. Client IDs and session IDs are assigned incrementally, i.e. the numerical difference two consecutive IDs is 1. Channel ID takes the form: <zone ID> “.” <media category ID> “.” <channel ID suffix>. A zone ID is the ID assigned to a MUtechZone, but since each MUtechZone owns a separate MUtech object, a zone ID also uniquely corresponds to a MUtech object. Media category is an abstraction inspired by InVerse (Singhal et al. 1997). A media category can be *application-specific data*, *system-specific data*, *object states*, *byte stream*, or *text stream*, but right now only the object states media category is supported. Each media category is assigned a unique ID. Under one media category, channel ID suffixes are assigned. These channel ID suffixes are assigned by the method `nectar.mutech.MediaCategory.assignChannelIDSuffix()` . The assignment is again incremental. The whole assignment scheme ensures that a channel is zone-unique and media category-unique.

Multi-session

The NECTAR MUtech Server is a multi-sessioned server. What this means is that the server can host several sessions at a time. When a client connects, the client will be shown a list of existing sessions and allowed to either choose among the sessions to join, or create a new one. By this, several sessions can exist on the same server, virtually creating a string of parallel universes. However of course, as there is a performance limit as to how many sessions can be optimally supported at a time, the maximum number of sessions the server is willing to host has to be subject to administrative configuration.

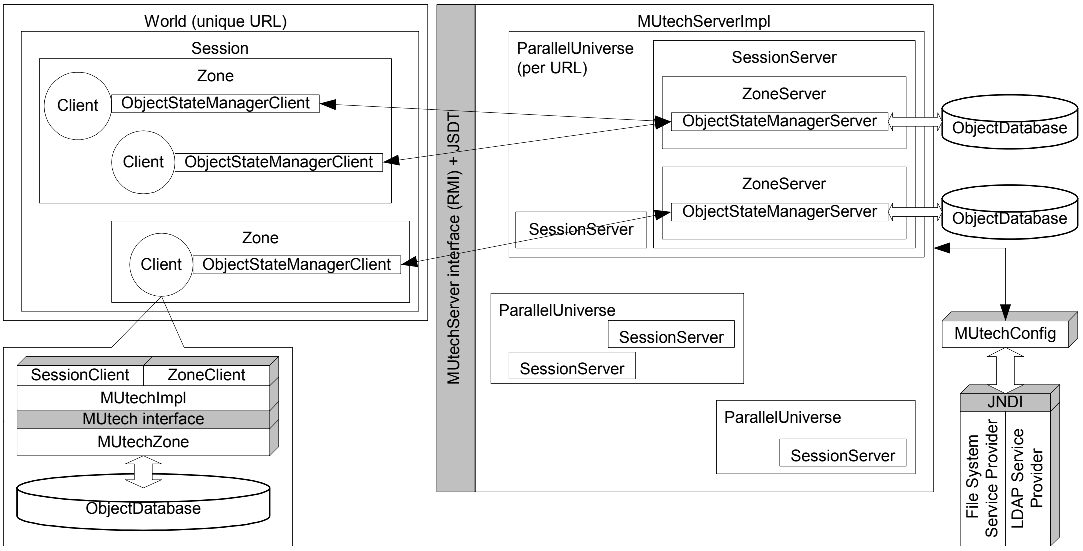


Figure 5: The Architecture of NECTAR MUtech

A virtual world is uniquely identified by a URL. For a URL, there maybe more than one (session name, session ID) pair (Figure 6). In the context of NECTAR, *session name* refers to the user-provided, user-friendly name of a session, whereas *session ID* refers to the system-assigned ID of a session. The MUtech server maintains a ParallelUniverse for each URL. Each ParallelUniverse maintains a hash table of (session name, session ID) pair. For each of such pairs, there exists a SessionServer for the pair. Under each SessionServer, there may exist more than one ZoneServer (Figure 5), with each ZoneServer corresponding to a single Zone.

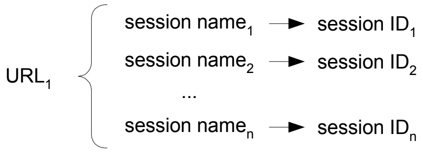


Figure 6: The Mapping of URL to Session Name and Session ID

Multi-channel

Media categories target different information types in a Zone. Channels target different sub-information types in a media category in a Zone. Channels are managed by managers; for example, object states are handled by the Object State Manager, text chat messages are handled by the Text Manager, and finally video and audio streams are handled by the Video Manager and Audio Manager respectively (Figure 1). Due to the limitation of the "socket" implementation of JSDT, it is unfortunate that video and audio streams cannot currently be supported. Currently only the Object State Manager is implemented. The beauty of JSDT is that, like DirectPlay, it supports reliable (guaranteed) or unreliable (unguaranteed), ordered (sequential) or unordered (non-sequential) data streams.

The interface method **`nectar.mutech.server.MUtechServer.listSupportedManagers()`** returns the list of supported managers. When the MUtech object starts up, it calls this function on the server through RMI. This way, the server and the client run the same set of managers, and this list has the flexibility to change from time to time, from application to application. A manager has two parts: one on the client that inherits from **`nectar.mutech.jsdt.ManagerClient`**, another on the server that inherits from **`nectar.mutech.jsdt.server.ManagerServer`**.

The Object State Manager has a reliable channel, and an unreliable channel which is or is not created depending on the result of the method **`decideConfig()`** of the Connection Configuration Manager (**`nectar.mutech.jsdt.ConnectionConfigManager`**). The Connection Configuration Manager runs a ClientSocketThread on the client, and a ServerSocketThread on the server. Periodically, the ClientSocketThread sends out a packet to the ServerSocketThread, and from the round-trip time estimates the latency and clock slew, following a methodology similar to that of http://www.codewhore.com/howto1.html. If the measured latency is higher than the latency threshold (returned by **`nectar.mutech.jsdt.MUtechConfig.getLatencyThreshold()`**, which is currently hard-coded as 50ms, but this is configurable), then the unreliable channel is created and designated as the *primary channel*. The primary channel is used for object states transmission. Otherwise, the unreliable channel is not created, and the reliable channel is designated as the primary channel instead. The rationale is that a latency that is higher than the threshold should not be worsened by employing TCP. In any case, the reliable channel always exists and serves as the channel for periodic synchronization, similar to Open Community/ISTP's 1-to-1 Connection.

Load-balancing

Net-Z (Proksim Software 2001a, Proksim Software 2001b) is an excellent source of idea on load-balancing. Most important is the idea of migrating duplication masters. It is easy to support this form of load balancing with CoreLW, because the CoreLW's notion of *pilot* and *drone* is equivalent to Net-Z's notion of *duplication master* and *duplica*. Inline with Net-Z's methodology, NECTAR supports the idea of distributing pilots according to a predefined set of load criteria. The load criteria currently only consists of the frame rate. Load-balancing is a responsibility of the Load-balancing Manager but it is currently not yet implemented.

Naming Service

As mentioned, `MUtechConfig.listAddableObjects()` is used to obtain a list of addable objects. The function accepts a `nectar.mutech.ObjectInfo` object which specifies the criterion for the objects asked for. The mechanism for supporting `MUtechConfig.listAddableObjects()` is provided by the Java Naming and Directory Interface (JNDI). JNDI is an industry standard interface to heterogeneous naming and directory services (Lee and Seligman 2000). There are two user-configurable parameters to get the job done: `jndiInitialContextFactory`, the class name of the *initial context factory*; and `defaultObjectsProviderURL`, the URL of the *objects provider*. First, these two parameters are used to obtain the *initial context*. The initial context is then searched recursively, i.e. the initial context yields a list of bindings, if any bound object in the list is a context, the bound object yields another list of bindings and so on. If any context satisfies the particular criterion, then all bound objects (non-context) under the context are considered addable objects for that particular criterion. Although JNDI allows transparent access to naming and directory services, currently the only supported service provider is the File System Service Provider.

This mechanism is intended to be extended to `MUtechConfig.listSupportedManagers()`, which currently is hard-coded to return only the Object States Manager.

Optimization

Learning from the experience of Open Community, it is imperative to minimize the cost of crossing between Zones. In CoreLW, it means the cost of Zone activation/deactivation has to be minimized. In the current implementation, it involves only adding or removing the client as a ChannelConsumer to the channel that is already created during MutechZone's initialization. According to observation, this operation is relatively lightweight. This means however that, resources are always allocated to the session and channels created at startup, even if the Zone is never activated. This represents a typical resource versus overhead trade-off.

Aggregation is performed on all network-going packets. The current technique is based on a hybrid timeout-based and quorum-based transmission policy (Singhal 1996). The timeout value is calculated as follows:

```
timeout = MUtechConfig.getLatencyThreshold()
        - ConnectionConfigManager.getLatency();
if (timeout <= 0) {
        timeout = DEF_TIMEOUT; //DEF_TIMEOUT=10
}
```

The quorum value is calculated as follows. Suppose there are *n* shared objects, emitting a state update every frame. If each individual update consists only of one translation and one rotation (a tuple of 3 floating point numbers and a tuple of 4 floating point numbers), then all these updates combined would consume $n \times (3 \times 4 + 4 \times 4)$ bytes. Although this estimate is crude, it can be quite effective if the scene is mainly populated by avatars. As an improvement, if the quorum is ever exceeded, the quorum is set to the size of the last update packet that exceeds the quorum. Furthermore, the fact that there are *n* shared objects does not mean that there are *n active* shared objects, or objects that are actively transmitting state updates. In other words, the quorum is now calculated from *n* active shared objects instead of *n* shared objects.

5. TEST RUN

The graphics engine and the networking engine are not yet integrated, mostly because the Browser API functions `Browser.createVrmlFromString()` and `Browser.createVrmlFromURL()` are not yet fully implemented. Therefore the graphics engine and networking engine are tested separately. For the graphics engine, while there is a lot of improvements made, there are still many non-conformance issues to be fixed.

For the testing of the networking engine, the graphics engine is "borrowed" from ParallelGraphics' Cortona VRML browser version 3.1 running on top of Microsoft Internet Explorer version 6. The test involves two users/clients. Figure 7 and 8 show the different perspectives of the users. The testing shows the successful addition of external objects and transmission of states, although the performance still leaves room for improvements.

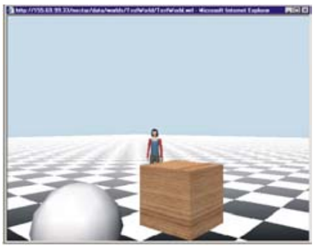


Figure 7: Perspective of the First User

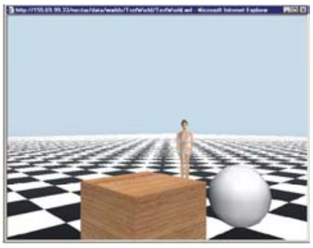


Figure 8: Perspective of the Second User

6. CONCLUSION

The initial design of NECTAR is meant to be simple in favour of short development time and low startup cost. In fact, using open standards and technology, and with very limited human resource (one full-time research student), it is extremely low-cost, in addition to having minimal vendor-dependency. Also, from an ease of maintenance point of view, this approach is certainly a big plus. Concerning its applications, as a foundation for future research in net-VE, it can be used for:

1. Experimentation with different graphics algorithms and simulation models.

2. Experimentation with sophisticated communication techniques, e.g. dead-reckoning, multicasting, locale-based connectivity etc.
3. Experimentation with different human-computer interaction techniques, e.g. two-handed interaction (Hinckley 1998), World in Miniature (WIM) (Stoakley 1995), See-Through Interface^{TM} (Bier et al. 1994) etc.
4. Generic visualizations, e.g. campus walkthrough, geographical information systems etc.
5. Exploring the application, strengths and weaknesses of VRML.

7. FUTURE WORK

Currently there is a lot of work in progress. Apart from completing the Browser API functions, there is plan to port NAVEL to using Xj3D, implement the entire VRML specification as much as possible, and submit to the NIST Conformance Suite (http://www.web3d.org/TaskGroups/x3d/translation/examples/Conformance/toc.html).

As mentioned, a custom implementation of JSDT is needed. The idea is to combine multicast UDP, TCP and RTP in a single implementation. Amongst the most important features, the Load-balancing Manager and Lock Manager are to be implemented.

Lastly, work on the user interaction engine has not started yet, but is planned.

8. ACKNOWLEDGEMENT

The project is supported by the NTU AcRF 04/98 research fund.

9. REFERENCES

Bier, E.A. et al. 1994. Toolglass and magic lenses: the see-through interface. In *Proceedings of the CHI '94 Conference Companion on Human Factors in Computing Systems*, 445-446.

Bricken, W. and Coco, G. 1993. “The VEOS Project”. Technical Report TR-93-3, Human Interface Technology Laboratory, University of Washington.

Burridge, R. 1999. “Java Shared Data Toolkit User Guide Version 2.0 (FCS)”. Sun Microsystems, Inc.

Frécon, E. and Stenius, M. 1998. “DIVE: A scaleable network architecture for distributed virtual environments”. *Distributed Systems Engineering Journal (Special Edition on Distributed Virtual Environments)*, 5(3):91-100.

Hagsand, O. 1996. “Interactive multi-user VEs in the DIVE system”. *IEEE Multimedia*, 3(1):30-39.

Hinckley, K. et al. 1998. “Two-handed virtual manipulation”. *ACM Transactions on Computer-Human Interaction*, 5(3):260-302.

Honda, Y. et al. 1997. “Living Worlds: Making VRML2.0 Applications Interpersonal and Interoperable”. Living Worlds Working Group, Web3D Consortium.

Hubbold, R. et al. 1996. “MAVERIK - The Manchester Virtual Environment Interface Kernel”. In *Proceedings of the 3rd Eurographics Workshop on Virtual Environments*, Monte-Carlo, Feb.

Kessler, G.D. 1997. *A Flexible Framework for the Development of Distributed, Multi-user Virtual Environment Applications*. PhD thesis, Georgia Institute of Technology.

Kessler, G.D. et al. 1998. "RAVEL, a Support System for the Development of Distributed, Multi-user VE Applications". *IEEE Virtual Reality Annual International Symposium (VRAIS) '98*, 260-267.

Law, Y.W. 2001. “Simulation and Visualization in a 3D Collaborative Environment”. Master's thesis, Centre for Advanced Media Technology, School of Computer Engineering, Nanyang Technological University. In review.

Lee, R. and Seligman, S. 2000. “JNDI API Tutorial and Reference”. The Java^{TM} Series. Addison Wesley, Jun. ISBN 0-201-70502-8.

Liao, T. 1998. “Light-weight Reliable Multicast Protocol”. Technical report, INRIA, France. http://webcanal.inria.fr/lrmp/lrmp_paper.ps

Macedonia, M.R. and Zyda, M.J. 1997. “A taxonomy for networked virtual environments”. *IEEE Multimedia*, 4(1):48-56.

Park, K. 2000. “CAVERNsoft G2: A Toolkit for High Performance Tele-Immersive Collaboration”. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, Seoul, Korea, Oct, 8-15.

Pettifer, S. et al. 2000. “DEVA3: Architecture for a Large Scale Virtual Reality System”. In *Proceedings of the ACM Symposium in Virtual Reality Software and Technology*, Seoul, Korea, Oct, 33-40.

Proksim Software, Inc. 2001. *Duplication Spaces*. White paper. http://www.proksim.com/documents/DuplicationSpaces.pdf

Proksim Software, Inc. 2001. *Net-Z 2.0 Technical Overview*. White paper. http://www.proksim.com/documents/technicaloverview.pdf

Rahn, S. 1998. *WorldToolKit Release 8 Technical Overview*. White paper, Sense8 Corporation, 100 Shoreline Highway, Suite 282, Mill Valley, CA 94941 USA. http://www.sense8.com/products/wtk_tech.pdf

Sense8 Corporation. 1997. *World2World Release 1 Technical Overview*. White paper, 100 Shoreline Highway, Suite 282, Mill Valley, CA 94941 USA. http://www.sense8.com/products/w2w_tech.pdf

Shaw, C. et al. 1993. “Decoupled simulation in virtual reality with the MR Toolkit”. *ACM Transactions on Information Systems*, 11(3):287-317.

Singhal, S.K. 1996. “Effective Remote Modelling in Large-Scale Distributed Simulation and Visualization Environments”. PhD thesis, Department of Computer Science, Stanford University.

Singhal, S.K. et al. 1997. “InVerse: Designing an Interactive Universe Architecture for Scalability and Extensibility”. In *Proceedings of the 6th IEEE International Symposium on High-Performance Distributed Computing (HPDC)*. IEEE Computer Society, Portland, USA.

Stoakley, R. et al. 1995. “Virtual reality on a WIM: interactive worlds in miniature”. In *Proceedings SIGCHI ‘95*, 265-272.

Sun Microsystems, Inc. 2000. “Java 3D^{TM} API Specification”.

Waters, R.C. et al. 1996. “Diamond Park and Spline: A Social Virtual Reality System with 3D Animation, Spoken Interaction, and Runtime Modifiability”. Technical report TR96-02a, Mitsubishi Electric Research Laboratory. http://www.merl.com/reports/docs/TR96-02a.pdf

Waters, R.C. et al. 1997. “The Interactive Sharing Transfer Protocol Version 1.0”. Technical report TR97-10, Mitsubishi Electric Research Laboratory. http://www.merl.com/reports/docs/TR97-10.pdf

Web3D Consortium. 1997. “The Virtual Reality Modeling Language. International Standard ISO/IEC 14772-1:1997”. http://www.web3d.org/technicalinfo/specifications/vrml97/index.htm

Layout Management for Cross-Platform Content Packaging

Patrick Brandmeier, Alexander Kröner, Thomas Rist

German Research Center for Artificial Intelligence (DFKI) GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{brandmei, kroener, rist}@dfki.de
http://www.dfki.de/~{brandmei, kroener, rist}

Abstract

In this contribution we report on our work towards a layout management system that supports cross-platform delivery of multimedia content. As input, the system expects content packages that have been compiled automatically or manually from repositories of existing media assets, such as text paragraphs and images. Since the authors of these assets may have specified layout preferences neither knowing in which package their assets will occur, nor knowing the customer's layout preferences, layout conflicts, such as incompatible style attributes, are preprogrammed during packaging. We present a constraint-based approach for resolving layout conflicts in automatically compiled content packages. Adopting the IMS standard for content packaging, the result of the layout computation is an IMS package enriched by annotations referring to style properties. A further aspect of our work is cross-platform delivery. With a focus on the customization of graphical representations for display on mobile devices with tiny screens, we have developed a module that uses techniques from the area of machine learning in order to chose from a set of available transformations the one which may produce the best result for a yet unseen image. The work has been conducted in the context of the EU funded project IMAGEN, which aims at the development of an integrated set of tools for the customized publication and distribution of multimedia content.

Keywords: Development and Design Tools for the Web, Interactive Web Publishing Tools

Introduction

Customized delivery of multimedia content is becoming a key factor for the operators of online portals in an increasing application field including E-Publishing of multimedia content, E-Learning, and electronic knowledge management. The term “customization” refers to both the choice of content according to a user’s information needs and interests, as well as to the presentation of content with respect to choices of media, style, layout, and rendering.

With regard to the “online world” one has to consider that content is provided from many different sources and most often already in the form of ready-to-present media assets. Rather than generating presentations from a rich semantic representation of domain knowledge,[1] selection and reuse of media assets are becoming the dominant tasks.

In this contribution we first sketch the IMAGEN[2] platform – an integrated set of tools for the customized publication and distribution of multimedia content. IMAGEN relies on the approach of customized content packaging, that is, in response to a user request, content is selected from repositories and assembled to a content package, which is delivered to the user.

Potential customers of IMAGEN are in the first place providers of repositories of rich content created by professional writers and artists. But a medium or large company's intranet may serve as such a repository as well.

For instance, DFKI's intranet offers –beneath a unified section of general information– a repository of homepages, which provide content about people and projects. In the following, we will use this example to demonstrate the benefits and the mechanics of the IMAGEN approach.

The IMAGEN Content Packaging Approach

Fig. 1 provides a rough sketch of the IMAGEN platform. IMAGEN can be conceived as a mediator between content providers on the one hand, and content users on the other hand.

The platform interfaces with a set of distributed repositories, which store so-called *content units*. A content unit is an XML file, which consists of a small composition of multimedia assets, such as texts, graphics, images, sound and video clips, and programmed interactive units, e.g., Java applets or Flash animations.

[1] For overviews of such tools see, e.g., [RFW97], [MW98].

[2] The platform is currently developed in the context of the EU-funded project IMAGEN IST-1999-13123. Project home page: *http://www.imagenweb.org/*

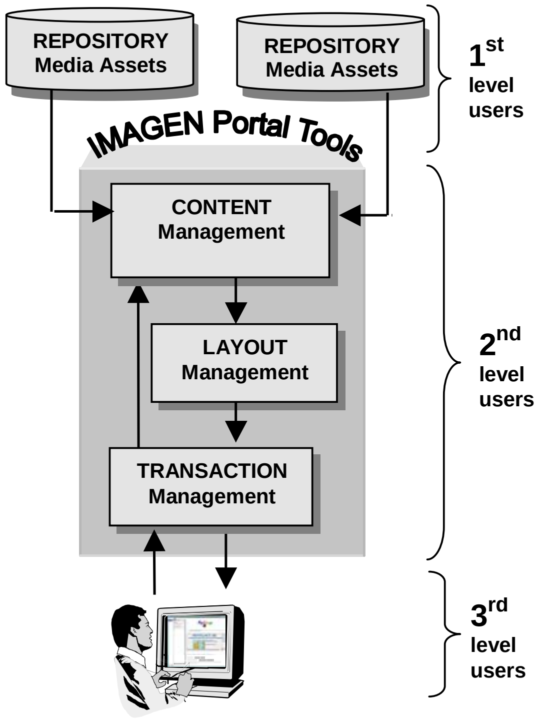


Fig. 1: Sketch of the IMAGEN Platform

The process of delivering content from the content provider to the content user is divided into three major steps – Content Management, Layout Management, and Transaction Management– which are implemented by separated components.

Content management comprises methods for content selection and content organization. The methods for content selection fall into two categories, methods that are based on profiles, which are learned from observing user behavior, and methods that take into account how humans express preferences and make judgments. Details on the deployed methods for content selection can be found in [Hal]. Content organization relies on a template-based approach. That is, different user queries are associated with different types of package structures that will be used to present retrieval and selection results.

The Transaction Manager basically keeps track of the content units that got compiled into packages and eventually were delivered to customers. Especially in an E-Commerce environment such a record is required in order to allow for the allocation of income from cleared transactions back to the rights owners of the media assets.

IMAGEN's Layout Manager receives as input a package comprising the selected media assets (or content units). The component's task is to find a "good" layout solution for the overall package.

One of the challenges for the Layout Manager is to meet the requirements that are imposed by the following three different user categories (see Fig. 1).

1st-level users are authors/creators of media assets that may be embedded in a content unit. They may use the LM for achieving a preview of if and how their layout preferences are integrated into a package's overall layout.

2nd-level users are usually publishers, content syndicators/ owners and administrators of web portals. A 2nd-level user may purchase media assets from a number of different 1st-level users in order to compile new multimedia products for his customers. As to the layout of a package, the portal administrators should be enabled to specify compulsory (required) constraints, which override 1st-level user preferences, and optional constraints, which enable the integration of 1st-level user preferences.

3rd-level users are end users or consumers of content. A driving force behind the IMAGEN project is the observation that 3rd-level users have a high, unsatisfied demand for customized presentations from 2nd-level users. In addition to a customized selection of content, the layout of a presentation need to be customized as well taking into account user preferences and available display resources.

In the next sections, we will discuss some of the methods applied by the Layout Manager to fulfill these requirements.

The Layout Manager

As already indicated, the Layout Manager aims at providing an appropriate layout for content packages. In IMAGEN, these packages are represented according to the IMS XML binding (cf. the IMS inset below). Such packages are provided as input to the Layout Manager. In turn, the output of the Layout Manager is again an IMS-compliant package, which may be further processed by the Transaction Manager.

The requirement of not changing a package's organizational had a strong impact on the design of the Layout Manager, since many of the current approaches to the generation of layout, e.g., [Krö00], require such a change in order to perform a virtual rendering process.

Other requirements include the demand for using standardized means of specifying and processing layout knowledge. Hence approaches like [BLM00] or [BKK+99] cannot be applied without modifications, since they rely on proprietary browsing software and/or specification languages.

Improving Layout

We identified several layout-related tasks, which may be addressed for improving the quality of layout.

Style Management

A frequent problem of content packaging is that the content units included in a package may have different style properties. A layout that integrates all of these properties in a straightforward manner will usually appear heterogeneous –or inconsistent. Following our intranet example, assembling the content of different homepages would result in a layout as shown on the left-hand side of Fig. 2. Therefore, a means of unifying style properties is required, which enables 2nd-level users to unify the appearance of the overall package layout as shown on the right-hand side of Fig. 2.

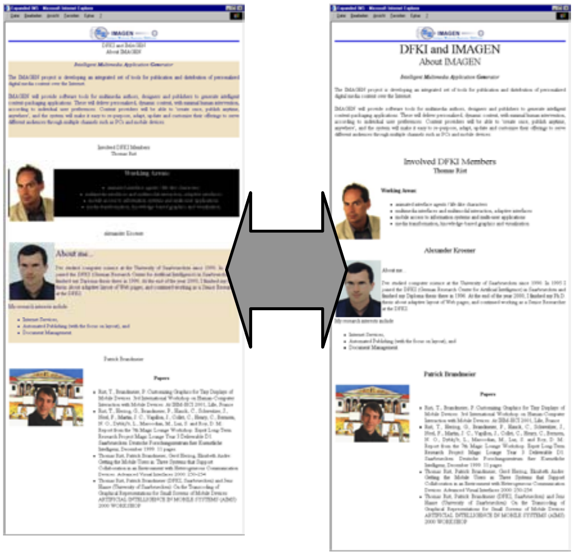


Fig. 2: Non-unified layout (left-hand side) vs. unified layout (right-hand side)

About IMS & IMAGEN. IMS Global Learning Consortium, Inc. (IMS) is developing and promoting distributed learning environments and content from multiple authors to work together.

Among the specifications provided by the IMS consortium are XML bindings for content packaging and meta data. Following these bindings, content packages are expressed as a manifest with three major sections: metadata, organization, and resources.

The metadata section is filled with information about the resources contained in the package. In IMAGEN, this section is primarily used to capture information about the 1st level user and about the transaction.

The section about the organization describes the logical organization of the resources contained in the content package. Originally intended for the learning community, it serves also in IMAGEN well as a description of the presentation logic.

Finally, a list of resources completes the package. Resources may be composed from other resources, and represent learning units. This idea was adopted for representing content units in IMAGEN.

Read more about IMS: *http://www.imsproject.org/*

Page Management

Packaging can result in very large documents which require extensive scrolling. Especially on devices with small screens this is cumbersome to handle. However, the page size depends on the 3rd-level user profile and is therefore usually unknown to the 2nd-level users. Hence they may appreciate a component that enables the expression of page breaking preferences, which are transformed into recommendations for page breaks before the document is forwarded to the rendering software.

Image Transformation

Another common problem is that images included in a package should meet constraints imposed by the 3rd-level user's output device. Thus images have to be transformed according to meta data provided by the 1st-level user as well as by the page management.

Architecture and Workflow

Fig. 3 shows the Layout Manager's architecture. As the overall structure of IMAGEN, the Layout Manager is organized in a modular fashion, which facilitates configuring it for new IMAGEN applications.

The modules, the Style Assistant, the Page Assistant,

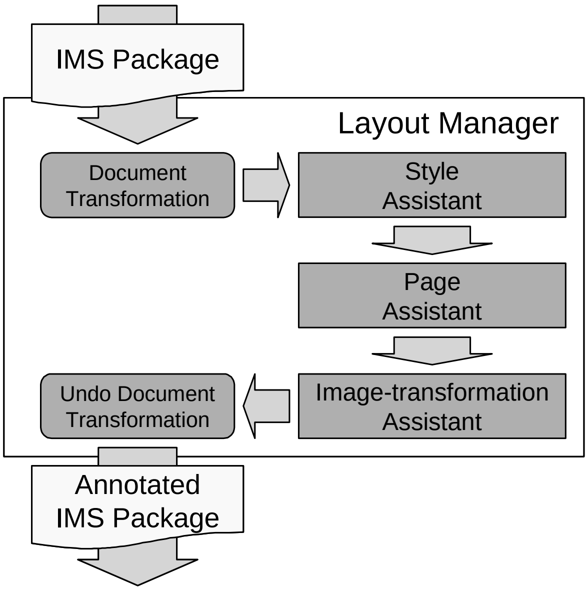


Fig. 3: Architecture and workflow of the IMAGEN Layout Manager

and the Image-transformation Assistant, implement the layout services we have proposed before. They share the idea that they would only add layout recommendations to the IMS package, rather than performing the actual layout rendering.

The advantage of this approach is that it decouples (spatially and temporarily) the process of specifying a uniform overall layout for a package from the actual rendering process, which may be performed by a standard rendering component, such as a WWW browser running on the PC of a 3rd-level user. It may even be the case that a package will be rendered only after it got downloaded and the 3rd-level user has disconnected from the portal server on which the layout specification has been computed.

The potential tradeoff –on the other hand– is that the specification process may not be able to consider all potential constraints and preferences that may be relevant at rendering time.

Processing runs through this architecture in three phases. In the first phase, the main input, an IMS package containing the documents selected by the Content Manager, may be optionally transformed. Such a transformation is recommended if the organization of layout differs considerably from the content organization of the IMS package. In that case, a transformation of the content organization into the layout organization may considerably simplify the specification of layout constraints.

In the second phase, the actual layout computation takes place. Here all layout assistants are allowed to add layout recommendations to the document. Fig. 3 suggests a workflow in which the computation starts with the style assignment, followed by the page break computation and finishes with the image transformation assistant. However, this is only one possible course of processing. In some situations the order may be changed, assistants may be disabled, and assistants may even suspend their work to request the support of another assistant.

Finally, in the last phase the computed layout recommendations are used to annotate the original document. During this process, any transformations performed in the first phase are removed. Thus the result of the layout computation is the IMS package that was provided as input but is now enriched by annotations.

The Style Manager

So far, most development work has been devoted to the Style Manager. In this section we will describe the underlying approach, refer to some implementation details, and finally work through a concrete example.

Approach

The Style Manager relies on constraints about style properties, such as font sizes and colors. We assume these constraints will be created and compiled by 2nd-level users. Once a constraint set has been specified, it may be applied to all documents sharing the structure of the document that was used to create the constraints.

The foundation of this approach is a representation of style properties as attributes of document nodes. Attributes represented this way are mapped to finite domain variables, which may be constrained by finite domain constraints. Here we distinguish the following two categories of constraints.

Required Constraints express layout requirements, which have to be met by a valid layout.

Rating Constraints are a means of specifying optional aspects of the computed layout. A rating constraint assigns a certain rating amount to a variable assignment, which may be further modified by a constraint weight. Such a weight serves as a measure of importance if rating constraints have to be compared.

The latter category of constraints helps a 2nd-level user expressing some kind of a quality measure for layout solutions. Thus the rating may serve as a selection criterion if there are several valid layouts available.

Technology

Major requirements to IMAGEN tools include the need for adopting standards, and the need for flexibility. To take both into account, we decided to implement the Style Manager using Java and XML.

The constraint solver consists of a kernel that relies on the JCL,[3] a set of wrapper classes that provide independence from the actual solver implementation, and several extensions. The most important of the latter ones include a module for processing rating constraints, and an XML binding for constraint satisfaction problems.

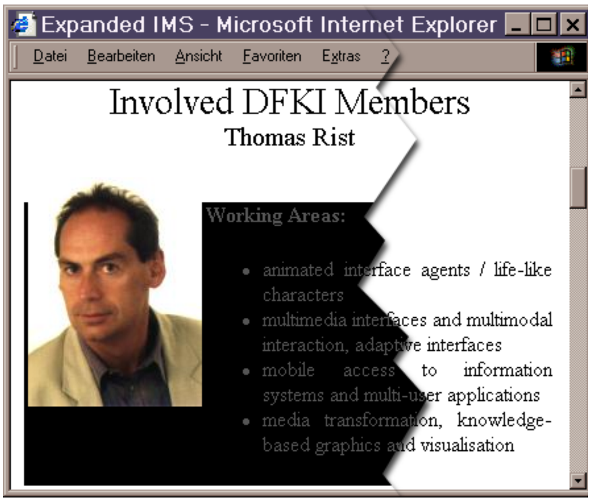
The XML binding serves as a standardized and flexible means of configuring the Style Manager. It facilitates not only integrating the Style Manager into existing frameworks, but provides also a notation for constraints and domains that is familiar to Web designers.

Example

In order to emphasize the advantages of the XML specification, and to provide some information about the details of our binding, we continue with a brief example.

A common problem when merging the styles of content units is that a unit's background color can provide a poor contrast to the font color specified for a subordinated unit.

This situation is demonstrated in Fig. 4, where the rendered fragment of a package is shown. On the left-hand side, the color selection, which is based on 1st-level user preferences, provides a weak color contrast. On the right-hand side, the color selection has been modified by a constraint about color contrast. Such a constraint can be specified as follows.



Involved DFKI Members

Thomas Rist

Working Areas:

- animated interface agents / life-like characters
- multimedia interfaces and multimodal interaction, adaptive interfaces
- mobile access to information systems and multi-user applications
- media transformation, knowledge-based graphics and visualisation

Fig. 4: Without (left-hand side) and with (right-hand side) constraining color contrast

```
<lm:constraint
  name="font-color-contrast"
  src="BC_COLOR_Contrast"
  target-var="lm:background-color"
  target-node="//ims:resource"
  source-var="lm:color"
  source-node="self::node()"
  min-contrast="250"
/>
```

This binary constraint establishes for each content unit a required constrained relationship between the variable carrying the foreground color and the variable carrying background color. The details of this constraint are as follows.

name is the constraint's name, a string that is unique with respect to the current constraint set.

src is the name of a Java class, which is applied as constraint operator. In our case, it is an operator that computes the contrast of two colors.

target-var is the variable that is constrained, here the background color.

target-node is an XPath expression, which describes the nodes of the document that are constrained. Here we choose all resource nodes, since these are containing the content units.

source-var is the constraining variable, here the foreground color.

source-node is an XPath expression, which describes the set of constraining document nodes. In our example these are the resource nodes itself.

[3] The Java Constraint Library (JCL) is developed at the Artificial Intelligence Laboratory at the Swiss Federal Institute of Technology in Lausanne, see also *http://liawww.epfl.ch/~akira/JCL/*

min-contrast is a threshold that is provided to the constraint operator as further input.

Cross-Platform Delivery of Content Packages

The IMAGEN approach of content packaging is not limited to a particular viewing platform. Hence the IMAGEN tools should support cross-platform delivery.

The Layout Manager addresses this issue in several ways. First of all, the finite domains used by the Style Manager may be set up according to the style that is appropriate for the 3rd-level user's platform. Similarly, the Page Assistant can be initialized with parameters such as the browser window dimension to produce pages, which fits into the 3rd-level users display device. Finally, the rendering itself may be addressed just by exchanging the style sheets used for processing the Layout Manager's outcome.

Of course, customizing layout alone cannot result in packages that are appropriate for each kind of device, since the customizations proposed before do not affect the content.

Especially for mobile devices, a customized content selection is required, since packages designed for a desktop PC will usually result in huge stacks of WML pages.

Partially, the Content Manager provides a solution to this problem, since it may consider device constraints such as display capabilities and bandwidth during content selection. Additionally, transcoding tools such as [IBM] may be applied to modify the selected content units before they are delivered to the client.

Nevertheless, there are some open issues. One of them are binaries such as images included in content units, which may exceed in byte and pixel size as well as color depth the output device's capabilities.

Transforming Images

In IMAGEN, the task of transforming images for a particular package is performed by the Image-transformation Assistant. This component is addressed by the Page Manager for fitting images into the computed pages. This task is realized straightforward by scaling the images.

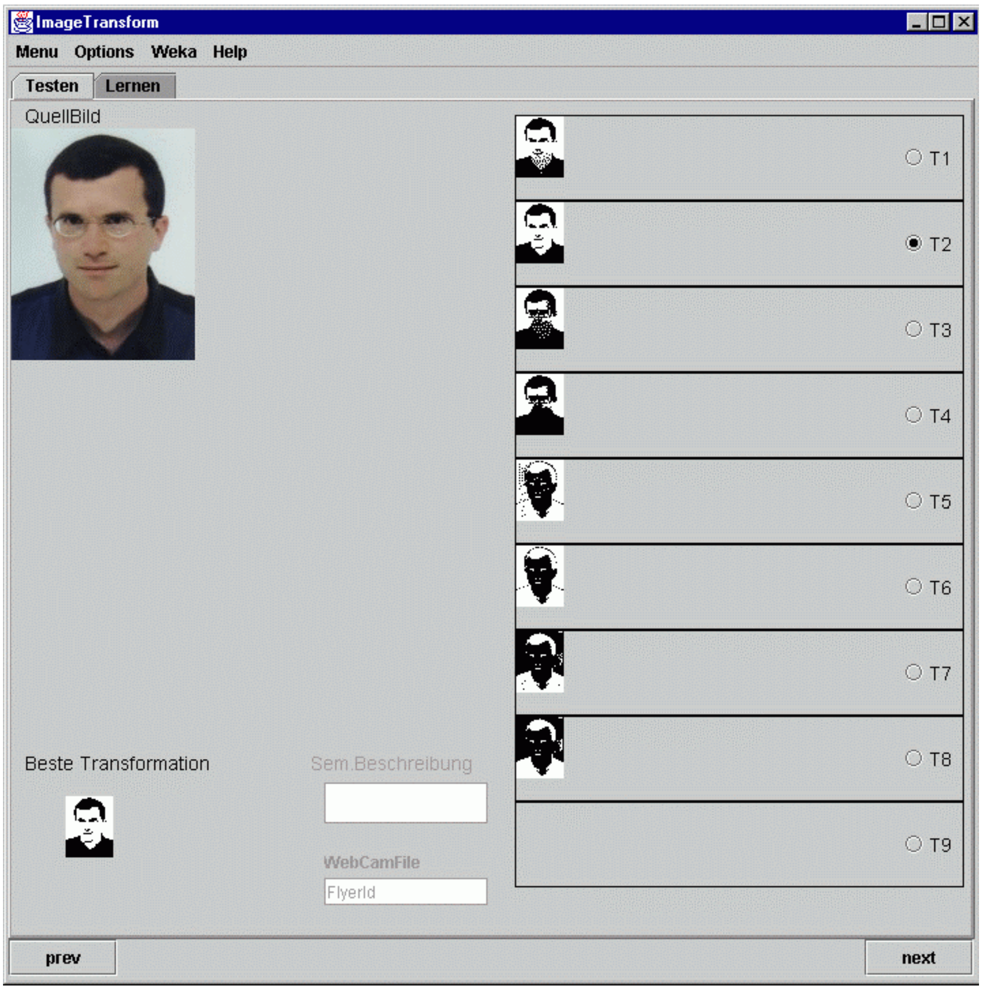
But for mobile devices, a deeper optimization may be required. Thus we started tackling the problem of how to transform images so that they can be displayed on very small displays, such as a 90x60 pixel display of a mobile phone. In particular, we are currently investigating different approaches to solve the graphical transformation problem: uniformed transformations, informed transformations, and re-generation of graphics.

Unfortunately, it is very difficult to find a general-purpose transformation that reliably produces suitable results for the large variety of graphics found on the web. A more promising approach starts with an analysis of the source graphics in order to inform the selection and adjustment of transformation parameters.

Basically the analysis phase performs a classification of the image source amongst syntactic or even semantic features. For instance, in our current work, the set of implemented semantic classifiers comprises classifiers that distinguish between portrait and non-portrait images, outdoor versus indoor images, outdoor images that show a scene with blue sky, clouds, sunset, water, forest or meadows, and snow-covered landscapes. In the ideal case, each image class can be associated with a certain transformation that produces acceptable results for the vast majority of instances of that class. While image classifiers provide a basis for selecting among different available transformations, it is still difficult to make an assignment between recognized features of an image on the one hand, and available transformations and their parameter adjustments on the other hand. We are currently investigating in how far this problem can be solved by deploying machine learning techniques. That is, in a training phase, a graphics design expert manually assigns transformations to source images and thereby allow the system to recognize and generalize correspondences between image features and transformation parameters.

A screenshot of our trainable image transformation system is shown in Fig. 5. When loading a source image into the system, the image gets displayed in the left part of the frame and an analysis is carried out to construct a feature vector that can be used to classify images. In our current test setting we use 430 features that result from regional color distributions of an image. In contrast, our repertoire of transformations comprises eight different transformations and is yet quite small. The result of applying each of the eight available transformations to a source image is displayed on the right-hand side of the screenshot.

By means of a software package for machine learning [WF00] we trained the system with a set of 130 images including portraits of persons, images of landscapes, in- and outdoor images of buildings, images of everyday ob-



○ T1
● T2
○ T3
○ T4
○ T5
○ T6
○ T7
○ T8
○ T9

WebCamFile: FlyerId

Fig. 5 A screenshot of the image transformation system used for testing an machine-learning approach

jects, and photographs of artworks.

In the learning or training phase, a graphics design expert would tell the system, which of the eight transformations yields the best result for a given source image.

In the test phase, the system is asked to recommend one of the eight transformations on its own for a yet unseen image. For the source image shown in Fig. 5, the system recommends T2 for transforming it to an image that can be displayed on a mobile WAP phone. In cases where the system remains undecided, which of the eight transformations to recommend, it may not recommend any of them but pick the nil option T9.

Summary

Automated content packaging provides a number of portal services with a means of providing customized content packages.

In this contribution we have addressed layout issues that arise when generating packages of multimedia content from pre-designed content units. A severe problem in this context concerns layout conflicts, which may emerge when trying to meet layout preferences as expressed by different authors.

Based on the assumption that automated content packaging will rely on an XML-compliant standard such as IMS, we propose an approach that takes as input an IMS package together with a declarative specification of constraints, and returns as output an IMS package whose structure is similar to the input package but which has been enriched by annotations concerning layout attributes and their values. Eventually, these layout attributes will guide the layout rendering process, which will be carried out in an IMS viewer or after applying XSLT transformations in a Web browser.

Furthermore, to support cross-platform delivery of content packages, we provide means of customizing layout using constraints imposed by the 3rd-level user's platform. Among these are methods of transforming images for display on mobile devices, which rely on a machine-learning approach.

Our work is a contribution to the IMAGEN platform for publication and distribution of personalized multimedia content over the Internet. While a prototype of the Layout Manger has already been developed, a first launch of an integrated IMAGEN pilot has been scheduled for the first half of 2002. The application will be a personalised WWW portal featuring contents from the Art-on-line collections of the 2nd largest Art Portal in Italy. The portal will provide clients (3rd-level users) with content packages that are automatically compiled according to a client's information request, interest profile, and display preferences.

By means of this field trial we expect feedback on a number of yet open questions. For example, it is not clear to which extend 3rd-level users will actually declare layout preferences, which they want to see taken into account, and how they would evaluate the solutions proposed by the Layout Manager. Feedback from 3rd-level users will also influence our further decisions on how to refine the proposed architecture of layout assistants.

Acknowledgements

We would like to thank our IMAGEN project partners for their valuable comments concerning all aspects of layout management. We are especially grateful to Niccolo Giaccone and Roberto Vaccaro for providing us with a rich corpus of content units.

References

[BJN00] Bergström, A., Jaksetic, P. and Nordin, P.: *Enhancing Information Retrieval by Automatic Acquisition of Textual Relations Using Genetic Programming*. In: Proceedings of Intelligent User Interfaces (IUI) 2000, ACM Press, 2000.

[BKK+99] J. Bateman, T. Kamps, J. Kleinz, and K. Reichenberger: *The DArt(bio) system: constructive text, diagram and layout generation for information presentation*. Association for Computational Linguistics, 1999.

[BLM00] A. Borning, R. Lin, and K. Marriott.: *Constraint-Based Document Layout for the Web*. Multimedia Systems **8.3**, pp. 177—189, 2000.

[Hal] M. Halamish: *Learning Users Interests for Providing Relevant Information*. MSc Thesis, Bar-Ilan Univ. Ramat-Gan, Israel. *To appear*.

[IBM] IBM Corp.: *WebSphere Transcoding Publisher*. Product description, 2001.

[Krö00] A. Kröner: *Adaptive Layout of Dynamic Web Pages*. In DISKI - Dissertationen zur künstlichen Intelligenz **248** (2001), infix, ISBN 3-89838-248-6.

[MW98] M.T. Maybury, W. Wahlster (eds.): *Readings in Intelligent User Interfaces*. Morgan Kaufmann, 1998.

[RFW97] T. Rist, G. Faconti, and M. Wilson (eds): *Intelligent Multimedia Presentation Systems*. Special Issue of the International Journal on the Development and Application of Standards for Computers, Data Communications and Interfaces **18**, No. 6 and 7, 1997.

[WF00] I.H. Witten and E. Frank: *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, 2000. Software: *http://www.cs.waikato.ac.nz/~ml/weka/*

HTTP TRAFFIC SIMULATION AND EVALUATION FOR MULTIPLE USERS IN AN INTRANET NETWORK

Elias Aravantinos
Petros Ganos
Research Academic Computer Technology Institute
Riga Feraiou 61
GR- 26221 Patras, Greece
E-mail: eliasara@cti.gr
E-mail: ganos@cti.gr

Christos Bouras
Panagiotis Kokkinos
Computer Engineering and Informatics Department
University of Patras
GR-26500, Patras, Greece
E-mail: bouras@cti.gr
E-mail: kokkinop@ceid.upatras.gr

KEYWORDS
IP Based Networks and Services, HTTP, Quality of Service, RMI, Web Server and Simulation.

ABSTRACT

In this paper we describe the architecture and design of a network-stressing tool developed to simulate and evaluate the HTTP requests and responses of hundreds of users. There are two major issues that have to be considered when designing and implementing such a simulation tool, the reliability evaluation and the user “friendly” environment. In our application we use a mechanism exploiting Intranet technology to stress the tested network and confirm the performance, proper operation and delay issues. We have also implemented an easy interaction with the user considering the creation of the scenarios and the presentation of the simulation results, pointing out the problems and the network’s status. Finally, we examined through a number of experiments using different simulation scenarios the behavior of the Intranet by load balancing the simulated users.

INTRODUCTION

The purpose of this paper is to evaluate network performance during the design of an intranet network, containing hundreds of terminals, for enterprise buildings, airplanes, ships etc. There are certain parameters like bandwidth usage, and delay defining network performance. The Web traffic is considered as the main component of Internet backbone traffic aiming to assess the Web behaviour of the network. In this paper we aim to measure the load of HTTP traffic using a new HTTP simulator called NHS (**N**etwork **H**TTP **S**imulator). The NHS was the software developed to measure network performance under different conditions.

There are several papers that studied the wide range of network problems, reliability and HTTP simulation (Davison 2001), (Floyd 1999), (Heidemann et al. 1997), (Judge et al. 1998), (L. Breslau et al. 2000), (Rosu et al. 2000), (Feldmann et al. 1999), (Krishnamurthy and Rexford 1998), (Badrinath et al. 2000). The main papers related to our work are described in the following paragraphs. The first paper (Davison 2001) describes the NCS HTTP simulator. This simulator tests HTTP 1.0/1.1 protocols in Ethernet, Fast Ethernet and modem environment. The NCS estimates client-side latencies and bandwidth usages. It also provides credibility; comparing simulation results to real world results and contains optional pre-fetching techniques. PROXIM is a caching and network effects simulator developed by researchers at AT&T Labs (Feldmann et al. 1999). It simulates a proxy cache using the HTTP trace as input. PROXIM can be used to simulate the three different scenarios using a proxy, without a proxy, where the bandwidth to the clients is the bottleneck, and without a proxy, where the bandwidth on the network connecting the clients to the Internet is the bottleneck. To assess the performance impact of proxy caching in a given environment, the simulation results are compared with-proxy to the without-proxy. In addition, the without-proxy simulation results are used to compare the simulation results against the original measured numbers. Network Simulator (NS) (Floyd 1999) is a discrete event simulator targeted at networking research. NS provides substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks. The NS simulator is likely the best-known networking simulator, but is not typically used for caching performance measurements, possibly because of slow simulation speeds. It uses detailed models of networking protocols to calculate performance metrics. This simulator has been validated by widespread user acceptance and extensive verification tests. NS is not a polished and finished product, but is the result of an on-going effort of research and development. In particular, significant bugs in the software are still being discovered and corrected.

NHS is described in this paper as a HTTP simulator, measuring HTTP traffic and presenting network performance. The Web Data in the simulated network are disseminated using the HTTP protocol. The NHS is a model independent of network technology and device configuration, flexible, scalable providing ‘smart’ Graphical User Interface (GUI); the user configures the simulation parameters easily in different scenarios. The terminals are informed automatically from a central point about the simulation parameters in a single process.

This paper is organized as follows: First, the simulation design of NHS is described. Next the phase of simulation and functionalities are presented in details. The next section describes the proposed simulation model, containing techniques and specifications. In the following part the simulation results are presented and discussed. Finally the concluding remarks are provided.

SIMULATION DESIGN

A general architecture of the simulation design is presented in Figure 1. The NHS user is located in the NHS server, configuring easily the simulation parameters from this central point. The simulation design is based on Remote Method Invocation (RMI) client server architecture.

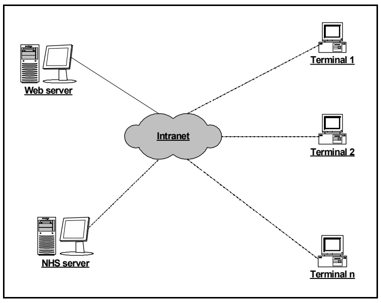


Figure 1: Network Architecture for HTTP Traffic

The NHS server hosts the server routine that communicates with the client routine hosted in the terminals. The NHS Server sends the parameters of simulation via RMI to the n terminals. A number of users are simulated in every terminal. During the simulation phase the terminals are connected to a Web Server via HTTP requesting Web pages. Each new Web page request creates a TCP connection. The Web Server responses to the Intranet, that consists of n terminals via HTTP. The NHS Server collects results of the simulation that are sent from the terminals and presents them to the user.

NHS FUNCTIONALITIES

The operation of the NHS consists of three phases (Figure2). The first is located at the NHS server; during this phase the user fills in various parameters and makes the selections involving the simulation scenarios. The second phase considers the NHS clients. Specifically during this phase HTTP traffic is created and several variables are measured. The third phase considers the NHS server and the presentation of simulation results to the user.

In the first phase the user defines various parameters and is guided to follow number of steps, like a wizard. In each step the user inserts values into the simulation parameters and configures the scenario according to the selections.

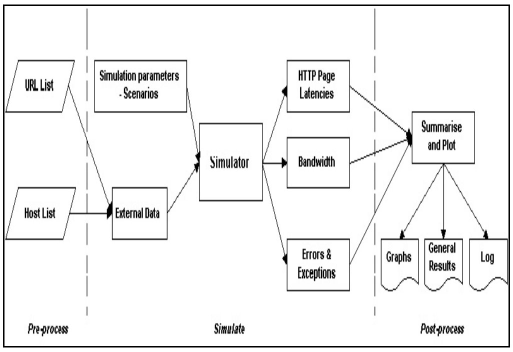


Figure 2: NHS Functionalities

The parameters and the selections are the following:

- The number of users simulated in all client terminals,
- The IP addresses of the client terminals
- The IP addresses can also be loaded from a text file in case of large networks. In each of these terminals 'runs' the client routine
- The Web Pages that will be used in the HTTP simulation. These Web pages can be loaded from a text file to make the step easier.

It is possible to configure NHS to repeat automatically an experiment in a specific IP address, with the same parameters, but each time with fewer users simulated. The user can choose to conduct an experiment with the same parameters in all IP addresses.

Then the user fills in the values of the parameters that define the simulation scenario like IP address, the number of simulated users, the number of HTTP requests that each simulated terminal will make, the waiting time, which is the time between the arrival of response and the next request and the timeout. This is a time limit referring to the time spent trying to get the hosts IP-address, establishing a connection with the server and also setting the timeout on the socket while reading the response. If this time is expired an exception is sent.

In the second phase, the NHS server routine sends to each of the IP address, specified by the user a message. This message contains some of the parameters that were pre-defined and are necessary for the simulation process (initialization). These parameters, which are the input for the client routine, are already mentioned like timeout, waiting time etc.

After a client routine receives this packet (message) creates and starts a number of threads. A thread is a process that runs concurrently and independently from the routine that created it. Each thread corresponds to a passenger simulated in the specific IP address. Each of these threads makes a number of HTTP requests to the Web Server, requesting several Web pages. One very important aspect of the simulation is the specific process that a thread requests a Web page and gets the corresponding response. The NHS uses the HTTP1.1 protocol to communicate with the Web server. The NHS takes advantage of some specifications of HTTP1.1 protocol, such as the persistent connection, in order to send the HTTP requests. The NHS uses the persistent connection to open only one TCP connection per

Web page, requesting through the same connection not only the html text, but also the images of the requested Web page. The advantage is to avoid opening a new TCP connection for each new HTTP request. During this whole process of sending HTTP requests and receiving the responses each thread stores some simulated data. Specifically the data that each thread stores are the following:

- The response code of each response, as defined in the HTTP1.1 protocol
- The request response time (page latency), for each request-response, includes the html page with the images
- The first to last byte time for each response, which is the duration between the arrival of the first and the last byte of the response
- The number and the kind of errors, such as timeout expiration
- The number of bytes that each thread received
- The html pages that each thread received.

When all the threads complete the requests and receive the responses, the client routine that created these threads is finished. The next step is to send to the NHS server routine the results of all simulated threads. In the third phase the NHS server routine collects the results from all the client routines of the various terminals. Then the routine processes them and presents the final results to the user through the GUI. These results include:

- Log table for all HTTP requests. This table contains the response code, the first to last byte time and the request response time (page latency) of each request.
- General results table. This table contains average, minimum, maximum values of the request response time (page latency) referring to each IP address. Also contains total values of errors, pages and bytes.
- The general results table contains the bandwidth used by each IP address. The bandwidth is calculated by dividing the total number of bytes received in an IP by the sum of the first to last byte time of all requests hosted by this IP address.
- Initial parameters table. This table contains the parameters that the user defined in the first phase.
- Graphs, performing any errors, bandwidth usage (bytes/sec), request response time and request response time (page latency) for each Web page. The results are saved in a file (html format), creating an archive of previous experiments.

IMPLEMENTATION ISSUES

NHS was implemented in Java. The NHS software consists of a number of classes, which are separated in to two different groups. The first group is the NHS server software, which runs in the NHS server, and its main task is the interaction with the user. The second group is the software that runs in each client terminal and its main task is to make the HTTP requests to the Web Server, to receive the responses, to keep a log of its operation and to measure various parameters involving simulation.

EXPERIMENTS

The NHS was tested in the network topology shown in Figure3. This is also the worst and most complicated case of our experiments described in details in this section. All the tested terminals are in different segments to increase the complexity of NHS experiments.

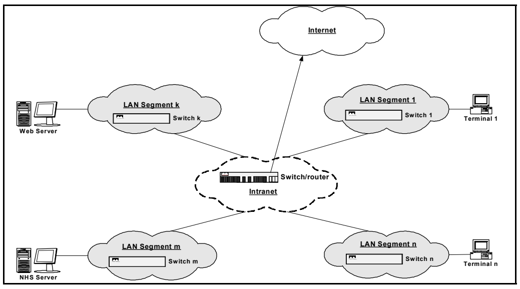


Figure 3: Network Topology

The tested network was an Intranet, supported by a Fast Ethernet network with star topology. The Fast Ethernet network cable is Unshielded Twisted Pair category 5 (UTP cat 5). The tested network consists of a certain number of terminals, servers and Local Area Network (LAN) switches. Each LAN segment is located in a different physical place in the Intranet. The Web Server responses to the client requests when download the Web pages. The NHS Server hosts the Simulator server routine that controls and communicates with the client-terminals. The terminals are single PCs. The specifications of the terminals and servers used in simulation tests are described in details in section 4 of this paper. The LAN switches are Fast Ethernet network devices that provide each sender/receiver pair with full 100 Mbps capacity. Each port on the switch gives full bandwidth to a single server or client station. Finally, there is a switch – router connected to the Internet to serve the Internet and Web based applications of the Intranet. The terminals and the servers are connected in different LAN segments. It was supposed that the tested terminals are located in different physical LANs instead of VLANs. All the terminals are connected to different switches instead of VLANs of the same switch. This view increases the complexity of NHS.

The terminals and the servers of the simulation had the following software specifications: The Web server supported Windows 2000 Server operating system and IIS 5.0, the NHS server contained Windows 2000 Server, IIS 5.0, JRE 1.3, NHS Server Routine and the terminals Windows 2000, JRE 1.3 package and the NHS Client Routine.

In the simulation phase several Web pages were used with different size (Kb), contained different number of images. The Web pages used, were 4 with total size between 38 and 140 KB, containing several number of pictures for example the first web page contains 25 pictures with total web page size 140 KB, the second 5 pictures with total size 38 KB, the third 16 pictures with 120 KB total size and the fourth 30 pictures with 121 KB size.

SIMULATION RESULTS

The results presented below are separated in two categories. The results that were measured in experiment and presented in NHS GUI and the results and values which were collected from NHS and presented in charts.
Firstly the results, as measured in experiment and presented in NHS GUI:
This experiment was conducted by setting the following values to the simulation parameters:

- Terminals: 4
- Used Users simulated per terminal: 70
- HTTP Requests Per simulated user: 3
- Web Pages used: 3
- Waiting Time: 1.0 sec,
- Timeout: 10 sec

The Table 1 contains samples of the simulation results, which are presented in the following charts, as exported after the simulation in the log file of NHS. The RRT symbolizes the Request Response Time and refers to the page latency, which is the time between HTTP request and response:

Table 1: General Results Table

Bytes	Bandwidth	Max RRT	Min RR T	Avg RRT
22632680	90587.889	4025	81	1351
22632680	98037.662	3625	50	1302
22632680	107639.012	4026	60	1152
22632680	107507.600	3104	50	1147

The chart (Figure 4) performs the simulation of 280 users in four terminals - 70 users per terminal. The graph represents the average RRT (Request Response Time (millisecond) for each tested Web page in relation with each IP address (terminal). In this experiment 3 different Web pages were tested (Simulation model Section).
The RRT is increased as the size of the web page is also increased. However there is a deviation between the RRT of the first and the third Web page, although their size (Kilobytes) is close. In addition the corresponding page latencies of the second and the third Web page are close, although the size of the two web pages is different. This is due to the fact that in both Web pages the size of the html page (text) is extremely small (10 and 7 Kilobytes), in contrast to the first that is 75Kbs. The first Web page contains more images, needing more HTTP requests and increasing RRT. The RRT difference between almost the same size Web pages is due to the size of the html page and to the number of images of each Web page.

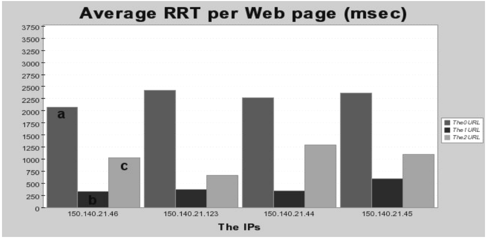


Figure 4: Results containing 4 terminals and 3 Web pages

Then we examined the results collected from the NHS processes, after a certain number of experiments: These experiments were conducted by setting the following values in the NHS parameters:

- Terminals: 4,
- Users simulated per terminal: 50/60/70/80/90,
- HTTP Requests Per simulated user: 3,
- Web Pages used: 3,
- Waiting Time: 1.0 sec,
- Timeout: 10 sec

The following charts (Figure 5) represent the bandwidth (bytes/sec) and the RRT in comparison to the number of users simulated. The several types of lines refer to the different IP addresses. In this chart are presented the measured results of 5 experiments - 50, 60, 70, 80 and 90 simulated users. We repeated the experiments, using the same parameters twice, in order to test the results.
The results lead to the conclusion, that the bandwidth of each terminal is decreased as the number of users simulated is increased. Certainly, this is a reasonable conclusion increasing the reliability of the NHS and the simulation process. From these charts we can conclude that the RRT is increased as the number of simulated users, in each IP, is also increased. It was also noted that the results, as measured, are very close.

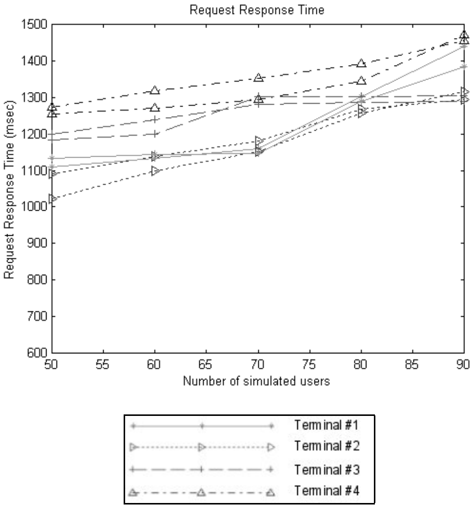


Figure 5: RRT Comparison of the Tests

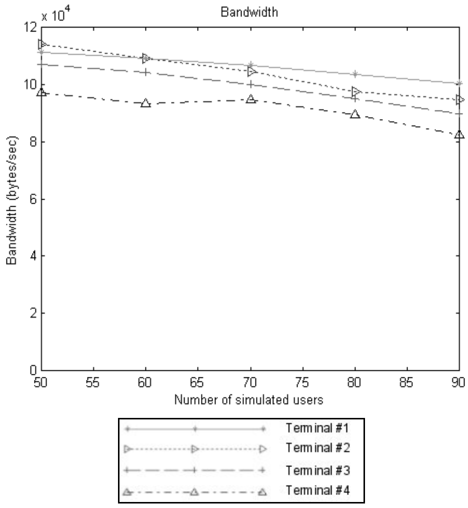


Figure 6: Bandwidth Comparison of the Tests

Finally we must remark that the previous experiments represent only some samples of the executed experiments. Also the results of these experiments depend from the way that NHS functions. NHS operates like a browser although there are differences; in the same way there are differences between known browsers like Explorer and Netscape.

SIMULATION CONCLUSIONS AND FUTURE WORK

NHS is a HTTP simulation tool, which can evaluate the reliability and verify the normal operation and performance of a LAN (intranet). The installation phase and operation of NHS are very easy. The components (servers and terminals) of the simulation process do not need to support any special hardware or software specifications.

The NHS user has a complete control during the simulation process, taking advantage of a very user-friendly GUI. So it is possible to set the values of several simulation parameters, creating different scenarios, to receive the results and assess the simulation process.

NHS is independent of the network topology where is executed. NHS can operate independently of the network topology and technology (ATM, Gigabit Ethernet) and the active network hardware and devices. The results of the simulation from all terminals are presented in a friendly and easy way in a central point, which is the NHS server (GUI).

In the future it is planned to conduct more experiments, using some hundreds of terminals and simulating even more users in each terminal. Also it is planned to test NHS in different intranets, various network topologies (bus, ring) and technologies (ATM, Gigabit Ethernet). In addition the purpose is to reduce even more the load of memory and CPU usage that the NHS creates in each terminal. Furthermore we will try different request distributions including Zipf-like distributions (Li 92). Finally one of the targets is to improve the GUI, importing new characteristics and new parameters creating even more complicated scenarios.

REFERENCES

Badrinath B. R. and P. Sudame. Gathercast, 2000: The design and implementation of a programmable aggregation mechanism for the Internet. In Proceedings of IEEE International Conference on Computer Communications and Networks (ICCCN), Oct. 2000.

Breslau L., D. Estron, K. Fall, S. Floyd, J. Heidemann, Helmy, P. Huang, S. McCanne, K. Varadhan, Y. Xu, and H. Yu, 2000. Advances in network simulation. IEEE Computer, 33(5): pp 59–67, May 2000.

Davison Brian D., 2001: HTTP Simulator validation using real measurements: A case study Published in Proceedings of Ninth International Symposium on Modeling, Analysis and Simulation on Computer and Telecommunication Systems (MASCOTS '01), Cincinatti, August 2001.

Feldmann A., R. C´aceres, F. Douglis, G. Glass, and M. Rabinovich, 1999: Performance of Web Proxy Caching in Heterogeneous Bandwidth Environments. In Proceedings of IEEE INFOCOM. pp. 106–116, New York, Mar. 1999.

Floyd S., 1999: Validation Experiences with the NS Simulator. In Proceedings of the DARPA/NIST Network Simulation Validation Workshop, Fairfax, VA, May 1999.

Heidemann J., K. Obraczka, and J. Touch, 1997. Modeling the performance of HTTP over several transport protocols. IEEE/ACM Transactions on Networking, 5(5): pp 616–630.

Judge J., H. Beadle, and J. Chicharo, 1998. Sampling HTTP Response Packets for Prediction of Web Traffic Volume Statistics.

Krishnamurthy B. and J. Rexford, 1998. Software Issues in Characterizing Web Server Logs. In World Wide Web Consortium Workshop on Web Characterization, Cambridge, MA, Nov.1998. Position paper.

Li W., 1992: Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38:1842–1845.

D. Rosu, A. Iyengar, and D. Dias, 2000: Hint-based Acceleration of Web Proxy Caches. In Proceedings of the 19th IEEE International Performance, Computing, and Communications Conference (IPCCC 2000), Phoenix, AZ, Feb. 2000.

Interaction Message Flow Control in Web Based Collaborative Virtual Environment

Qingping Lin, Weihua Wang, Jim Mee Ng, Chor Ping Low, Robert Gay
School of Electronic and Electrical Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Email: iqplin@ntu.edu.sg

Abstract

To achieve real-time natural interaction in the Collaborative Virtual Environment (CVE) with limited network bandwidth and computer processing power, the developments of efficient interaction model and system resource management mechanism are the key issues. In this paper, we propose a new behavior based active interaction model. The aim of this interaction model is to help achieving more efficient system resource management and better performance of the real-time natural interaction in large-scale CVE systems. The new approach enables the system to control the information flow from higher levels. Thus it gives a way to exploit the potential ability on managing the system resource and improving the real-time performance and scalability of the system.

1. Introduction

In the recent years, CVE has drawn significant research as indicated by the active work in this field, eg NPSNET by Macedonia et al [17], DIVE by Carlsson and Hagsand [6], MASSIVE by Benford and Greenhangh [11], NetEffect by Singn et al [7], the Virtual Society by Lea et al [14], and AVIARY by Snowdon et al [22], just to name a few. In a CVE system, multiple users can interact with each other in real-time, even though they may be physically located in different places around the world. This virtual environment provides users a sense of realism by incorporating realistic 3D graphics and stereo sound into the computer-human interface to create an immersive experience. CVE system gives the users a shared sense of space, and shared sense of time. It also provides users with the natural ways of communication and interaction. There are many scenarios that can significantly benefit from the solution in multi-user VR over simply non-collaborative VR or 3D workstation computer graphics. Most of them are military and industrial team training, collaborative design and engineering, collaborative scientific visualization, social activity simulation and entertainment like multi- user VR games. Envision future commercial applications include virtual shopping malls and showrooms, online tradeshows, conferences, and distance learning.

The main issues involved in the development of CVE include managing consistent distributed information, guaranteeing real-time interactivity, and contending with limited network bandwidth, computer processing and rendering resources.

These issues become especially important developing large-scale CVE systems that are large in the spatial extent, and large in the number of objects and users interacting within the virtual environment. For example, a CVE system might be deployed over the Internet where hundreds or even thousands of users will be participating in the system simultaneously. This requires an efficient way to manage large amounts of data for a large number of users to represent precise position and velocity information about objects in the virtual environment. Due to limited network bandwidth, the real time interaction performance becomes poor when the system is scaled up. Therefore, a new mechanism to manage system resources, such as network bandwidth and processor capacity, is required to achieve real-time interaction in large-scale CVE systems.

The main task of the system resource management is to reduce the requirement on network bandwidth and processor resources. The existing techniques include communication protocol optimization, flow destination control, user perception control, and system architecture refinement. To employ the flow destination control and user perception control in a CVE system, an appropriate interaction model must be created to provide the specification of the rules for its corresponding resource management algorithm. This is because that the interaction model will decide which users need to get the message and which need not. It also will decide the computing algorithm for the Level of Detail (LOD) and the update rate requirement of the information to the users. An efficient interaction model can improve the scalability and performance of real-time interaction of the CVE system dramatically. Thus development of an efficient interaction model becomes an important topic for CVE systems. That is the main focus of this research.

The paper begins with a literature review on the interaction models being used in the existing CVE systems. Then a new behavior based active interaction model is presented. This is followed by the implementation of a VRML based CVE system on the Internet which uses the behavior based active interaction model to manage the system resources.

2. Review of Existing Interaction Models

As discussed in Section 1, there are many CVE systems have been developed in the past few years. In this section, we will provide a literature review on some typical interaction models being used to save the network bandwidth and improve performance of real-time interaction in the CVE systems.

2.1 Distance Based Interaction Model

The simplest approach of the interaction model for CVE system is the distance-based interaction model. In this model, the distance between a user and the entities around him is used to decide the user's ability to know (see, smell, hear etc.) the entities. When the network bandwidth is limited, only the data from the entities nearest to the user are actually sent to that user's host.

There are two kinds of applications using this model. The first one is to use spatial distance to enable the interaction. For example, the spatial aura interaction model developed by University of Nottingham (Benford, Greenhalgh, [3]) employs the *aura* to enable the interactions among the entities in the virtual environment. Aura is an abstract sphere around the entity and move with the entity, like a magnetic field. When the aura collision is detected between two entities, the interaction between them is enabled. Aura is various for different mediums.

Another form of this kind of application is to enable the interaction through the *horizon count* method. The horizon count indicates the maximum number of entities that the user is prepared to receive updates from. The entities will be sorted by their distances from the virtual embodiment, the avatar of the user. A horizon count of 12 means that only the 12 closest entities will have their updates sent to the user. This is the approach of the interaction model used in Blaxxun's CyberHub [21].

The second kind of application is to use spatial distance to calculate the LOD. When the interaction is enabled, the distance between the two side of the interaction partner will be used as the parameter to calculate the LOD of the information sent between them.

2.2 Acuity Based Interaction Model

The distance-based approach is simple and easy to operate, but it may fail in some cases. The information flow in the virtual environment is enabled from two parties, the observer and the observed object. But the distance based interaction model ignores the observed objects' ability to attract observer's intention. For example, a large object that's far away may be more relevant than a small object nearby. So the acuity based interaction model [21] is introduced. Acuity is the ratio between size and distance. For example, if a user has a visual acuity of 0.1, an object 1 meter tall standing 10 meters away would be at the limits of relevance. Anything bigger or closer would be more relevant, anything smaller or farther away would be less relevant. In this model, every user has an acuity setting, which indicates the minimum size-over-distance ratio required in order for an entity to be know (see, smell, hear, etc.) to him. The system takes the size of an entity and divides it by the distance between the entity's location and the viewpoint of the user in the virtual environment. If the ratio is less than the acuity setting for the user, then the entity is too small and or too far away for the user to see. In such a case, no update from the entity is sent to him. In this model, the ability of the entity to attract others is also considered.

2.3 Region Based Interaction Model

In the region based interaction model, the virtual world is partitioned into smaller regions. The system will decide which regions are applicable to each particular user. Only the updating information from these regions is transmitted to the corresponding user. The division of the regions is not apparent to the user. A user can see several regions at once--generally the region containing the user's viewpoint and those neighboring it. But he does not see any seam between the regions or any abrupt change as the view point moves from one region to another.

For example, considering the apartment showed in Figure1, a user could see the washroom if he is standing in the kitchen. But he won't be able to see other users in the master room or the two common rooms. If the user moved into the hall, he could see others in the three bedrooms, but could not see the washroom now. Thus the graphical updating data from the rooms which could not be seen will not be transmitted to the user. The division of the region is also medium dependent. For example, a user may not be able to see others in the master room from the kitchen, but he might still hear the song from the master room.

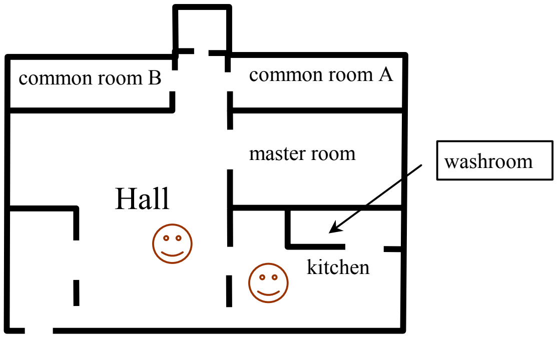


Figure 1. Region Division in an Apartment

The regions are predefined while the users are moving among the regions. When this kind of migration occurred, the LOD of the information between the user and other entities are dynamically changing accordingly. Bounding boxes and Binary Space Partitioning trees can be used to define the region partition in the virtual environment.

Region division is widely used in the CVE systems, as the *locale* in Spline [2], the *zone* in VRClass [8], the *hexagonal cell* in NPSNET [17], and the closed community and building/room in NetEffect [7]. It also appears as third party objects in MASSIVE II (Model, Architecture and System for Spatial Interaction in Virtual Environments) [11].

2.4 Peer/Group Based Interaction Model

When a user does not like to receive the information from some entities, those entities could be "ignored". In a similar way, by designating an entity as a peer, certain data streams are sent only to that peer. This provides a form of private communication, like whispering.

Grouping is another variation of this idea. By designating certain entities as part of a group, private conversation could be set up among the group members. The users out of the group will not receive the updating information of the entities in the group. Different from the region division model, which partitions the virtual environment spatially, this model partitions among the entities that could be involved in the interactions. The CyberSockets used in CyberHub [21], for example, supports the concepts of peers and groups.

Another good example uses the group based interaction model is the NPSNET. NPSNET (Macedonea, Zyda, Pratt, [17]) is developed by the Computer Science Department of the Naval Postgraduate School, USA. It incorporates the IEEE 1278 DIS application protocol and the IP Multicast network protocol for multi-player military simulation over the Internet. The virtual environment of the system is logically parted by associating spatial (hexagonal cell region division), temporal, and functionally related entity classes to form network multicast groups. Each user has a local Area of Interest Manager (AOIM) which is used to identify the multicast groups which are potentially of interest to them, and to restrict network traffics to these groups.

Besides the spatial group by the region division, entities may also belong to a functional class in which an entity may communicate with a subset of entities. An example of a functional class in the military domain could be a VE "air control" group. The group world includes entities that are primarily concerned with entities or events occurring in the air. Therefore, air defense and aircraft entities would comprise the majority of the group. Aircraft and air defense systems are relatively sparse in the whole compared to other combat systems such as tanks. And the entities can also belong to a temporal class. The updating information among the group members is distributed once in a while. NPSNET exploits wide area multicast communications and entity relationships to partition the virtual world to enable the scalability of the CVE system for military applications.

2.5 Spatial Aura Interaction Model

It is an interaction model used in several CVE systems, such as DIVE (Distributed Interactive Virtual Environment) (Carlsson and Hagsand, [6]) developed by the Swedish Institute of Computer Science, Virtual Society Project by Sony CSL (Lea, Honda, [14]). The spatial aura interaction model is designed by Unversity of Nottingham as the interaction model used in MASSIVE conferencing system (Benford, Greenhalgh, [3]). It is supported by this system to restrict the network communication for user's awareness maintaining in the virtual environment.

The key concepts of the spatial model include *medium*, *aura*, *awareness*, *focus*, *nimbus* and *adapters*. Medium is a communication type such as audio, visual or text. Aura is an object- and medium-specific subspace in which interaction may occur. Awareness is used to quantify one object's significance to another object in a particular medium, which depends on focus and nimbus. Focus represents an observing object's interests in a particular medium. Nimbus represents an observing object's wish to be seen (or heard, smelt, etc) in a given medium. Adapters are objects that can modify other object's auras, foci and nimbi in order to customize or modify its interaction with other objects. After the interaction is enabled by the aura collision, awareness is negotiated through combining the observer's focus and the observed entity's nimbus.

3. Analysis of the Review Findings

The review reveals that the interaction model is the core of the system resource management mechanism of CVE systems. The spatial aura interaction model [3] is one of the most completed interaction models and been used in many existing CVE systems. It has been noticed from analysis on the spatial aura interaction model that its basic modeling idea is passive. The passiveness is directly inherited from the earlier 2D cyberspace: file browser, sharing database, etc. By using the spatial aura model of interaction, we are not directly facing to the collaborative behaviors which should be supported in the CVE systems, but the passive awareness, observing. By this way, the intention of the behavior is ignored. To overcome the shortcoming of this restrict passive service, the adapter, third party object was extended into the spatial model to get some flexibility to the social feature, context reflection [12]. But the basic concepts are still passive. The problem can not be resolved from the root. For example, when two clients are discussing face to face, the third client goes through between them by chance. The former two clients' awareness to the third one could be ignored in reality, but may be set to the highest level by the spatial aura interaction model. It is because that this model only deals with the awareness level according to the spatial distance among the entities and the observing direction of the clients, but not the real intention of the clients.

To achieve more efficient system resource management and better performance of the real-time natural interaction for the large-scale CVE system, there are a number of fundamental issues that need to be addressed. These include: How to develop the interaction model more according to the user's interest/intention? How to make the interaction model adaptive to the dynamically changing of the user's interest/intention? How to support collaborative working more efficiently? And how to develop an integrated interaction model that is more flexible for different CVE applications?

At least some of the answers to these questions could be found if research in this area is carried out. In this research study, we try to develop a behavior based active interaction model, which takes into account the user's interest/intention, and employs a layered structure to integrate different information flow controls into an integrated interaction model.

4. Design the Interaction Model in CVE: An Information Filter

Analysis on the available interaction models in CVE systems reveals that the core of the interaction model is "information flow control", as illustrated in Figure 2. To improve the real-time performance of CVE that support large number of users, it is important that the data transmission among the entities or users be minimized. Thus the control of the information flow with the reasonable and acceptable realism becomes the key issue.

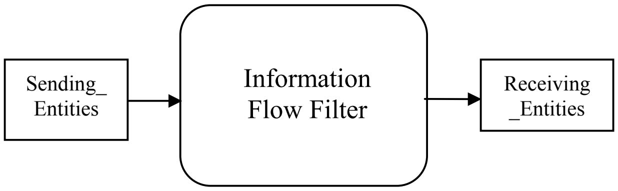


Figure 2. The Key Concept of the CVE Interaction Model: Information Flow Control

For controlling the information flow, there are two common methods: to control the destination of the flow, and to control the Level of Detail of the information. These two methods are normally used to define an information filter. The information flow output from the filter has a form of: from x *(Sending_Entities)* send *i (Information*) to *y (Receiving_Entities)* at *l (level number)* LOD. The main component of the filter is the grouping algorithms that are used to classify the Sending_Entities, Information, and Receiving_Entities of the information flows.

5. Behavior Based Active Interaction Model

To provide more natural and effective information flow control, we proposed two-layered filter's approach. The active interaction model includes two components: the low-level direct flow-control, and the high-level behavior management. The high level control is derived from the type of the interactions, while the low-level control is derived from the real time interactions.

5.1 Low-level Direct Flow-control

Based on the user's intention/interest, the direct flow-control could be adopted from two directions: from the intention of the Sending_Entity and the interest of the Receiving_Entity. This makes the two kinds of controls: the input information flow control and the output information flow control.

5.1.1 Input Information Flow Control

Input information flow control affects the LOD of the information according to Receiving_Entity's intention to get that type of information. The interest of the user will decide his awareness level on other entities in the virtual environment.

The static AOI model is a kind of input information flow control in which the AOIs of the clients do not change with the time. Before a user subscribes into a CVE system, he will define the AOI by choosing a series of keywords among a set of selections. The system will transfer the information to the clients using different LODs according to the matching of information's property (the property of the source entity) with the clients' AOI.

Besides the static AOI model, we also use some dynamic (adaptive) AOI modes. With the static AOI model, we can determine the information's detail level based on client's predefined interest. But in many cases, user's interest is dynamically changing. And the content of the scenario is also dynamically changing, not limited in some named areas which could be predefined. As such, it is important to have dynamic AOI models to meet the changing requirements.

Our adaptive model includes four sub_models: Time Based Sub_model, Frequency Based Sub_model, Topic Based Sub_model, and User_input Based Sub_model, as illustrated in Figure 3.

Different interactions have different properties. We classified them as the continued interaction, and discrete interaction. The former will keep on for some time and then terminated by the either of two interaction parties. The lasting time is not known before the interaction's occurrence. In this kind of interaction, time based adaptive sub_model should be used to control the information detail level between the two interaction parties. While for the discrete interaction, the occurrence of the interaction is more important than the lasting time as in many cases the interaction time is predefined. For example: a client selects to watch the same video clip for many times. The time length of the video clip is predefined, while the frequency of the interaction shows the client's interest to the video. For this type of interaction, the frequency based adaptive sub_model would be more suitable.

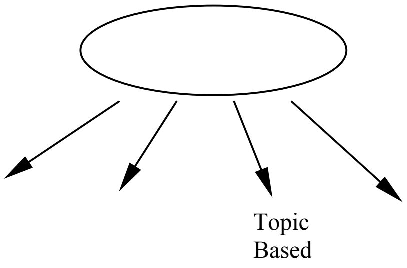


Figure 3. Dynamic AOI Model

In some scenarios, we can deduce the client's AOI from his "interest path". Two clients who have interacted with the similar entities in a certain period of time show similar AOI. They will be considered as having high topic_based information detail level.

With the user real time input model. Clients can input their interests in real time and the information LOD will be changed accordingly.

5.1.2 Output Information Flow Control

Another party of the interaction flow is the sending_entity. Output information flow control works on the control of the destinations to which the sending_entity want to send the information.

The working mode of the output information flow control could be two styles:

(i). *Entity A* **send** info **to** *Entity B, Entity C, …*

(ii). *Entity A* **send** info **to** the *Entities* (which are **under condition** *c*)

For the first style, the destination of the information is clearly identified by the source of the information (Entity A). For the second style, the destination of the information is decided by a condition judgement. The controlled information from Entity A will be sent to each client who satisfies the condition.

The key issues of this model are the user identification and the condition judgement. For different interaction and different information, the condition will be different. So we need a more general interaction model to combine them into an integrated definition, model operation, and the implementation. That will be discussed in the next section.

5.2 High-Level Behavior Management

The behavior based interaction model is designed to provide users a high level management of the entities' behavior in the VE. It incorporates different information flow controls into it's behavior control algorithms in different levels, and provide an integrate interface for different applications. The model will provide more abundant interactions for the VE. More importantly, this new approach enables the system to control the information flow from higher levels and thus provides us a way to exploit the potential ability to manage the limited system resource.

The basic concept of the behavior based interaction model is illustrated in Figure 4. The interaction model is based on the behaviors among the entities in the VE.

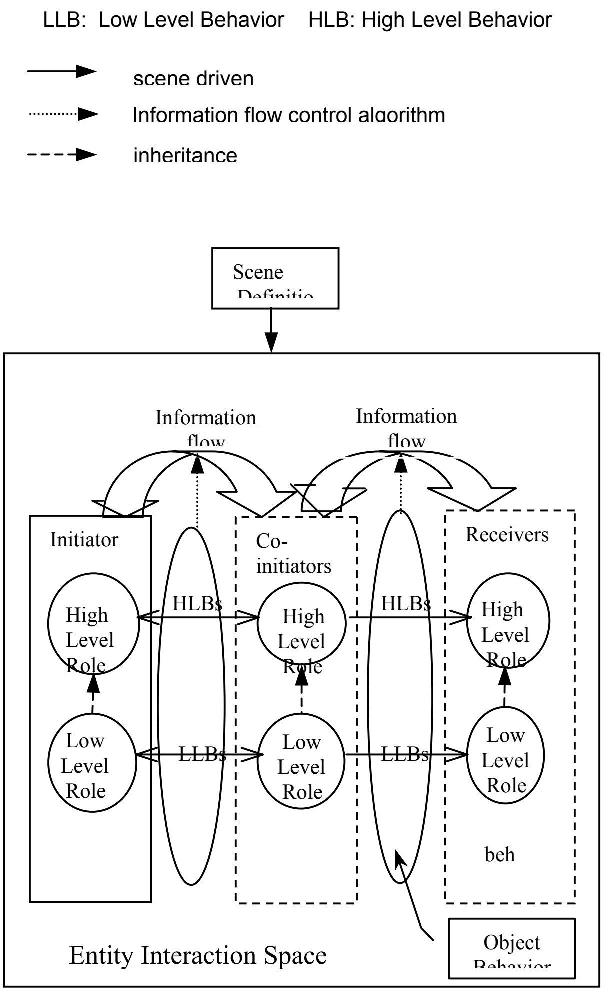


Figure 4. The Behavior Based Interaction Model

The scene definition outlines the functions of the scene and thus, provides a guideline to control the information flows from a higher level.

Under the definition of scene function, some high-level roles could be identified for participants of the scene. A user could get a better feeling of reality by acting as a participant of the scene (role-playing). That brings us a way of mixed reality, which breaks down the bound between the digital reality and the physical reality. It should be the kernel principle of virtual society. For example: in a virtual shopping mall (the scene), a participant can work as an assistant (role). That's a high level role required by the scene, compared with the low-level " person" which is the only role for a participant in the spatial aura model of interaction.

The role-relevant behavior is introduced into the model as the layered behavior. For example, a "teacher" can "give a lecture" (high level behavior) to the "students", while a "person" could only "chat with" (low level behavior) the others. The behavior-based model will support interactions far beyond walking and chatting. And we can control the information flow in the interaction space from the high-level behaviors' point of view. With this, Artificial Intelligence can be introduced to control a virtual, synthetic entity CVE.

Initiator is the entity that initiates an interaction. *Co-initiator* refers to the middle entity in the interaction, while the *receiver* is the objective entity of the interaction. The system will provide control on the information flow according to the algorithm defined for that interaction. The model directly supports the collaborative working by the definition of the *Initiator*, *Co-initiator* and *receiver*.

5. Implementation Work

To implement and evaluate the new approach of the interaction model, a VRML (Virtual Reality Modeling Language) based CVE system has been developed. For achieving an integrated system resource management mechanism, a scaleable server architecture is designed to improve the scalability of the CVE system from the system structure point of view. Virtual Reality Modeling Language (VRML) is used to create the VE in the system. The External Authoring Interface (EAI) of VRML is employed to achieve interaction among the users. The behavior based interaction model is adapted into the system as an information filter to achieve the more natural interaction and optimize the usage of the system resource.

5.1 Server Architecture of Experimental CVE System

For improving the scalability of the CVE system and providing more efficient packet delivery, the basic client-server architecture is expanded to include a sever federation. Multiple Acting Servers (AS) are working simultaneously as a cluster. The AS' communicate with each other in a peer-to-peer manner. At the same time, these AS' themselves behave as clients in a hierarchical client-server relationship. A Master Server (MS) works as the higher Server and is in charge of the total traffic-monitoring task. The clients are allocated and reallocated among the AS' according to the network traffic situation in the system by the MS. Through the multi-server (AS') workload sharing mechanism, the system can support more simultaneous clients with the real-time interaction. Thus the scalability of the system is improved. The server architecture is described in Figure 5.

When a new client logon to the VE, the MS will check the workload of the AS', and allocate the client to a less loaded AS and referent the initial position of the user in the VE. The MS maintains the user list of every AS, and monitor on the traffic load of the AS'. When an AS is overloaded, the MS will reallocate a part of its clients to another free AS. The server processes are distributed among different physical machines to share the traffic load in the system.

Every AS is in charge of the consistence maintaining of its clients. It propagates the updating information of the VE among its directly connected clients. It also shares the updating information with all other peer AS' if they are serving the same VE. As the database of the VE is replicated in the AS', compared with the centralized database architecture, the failure tolerance of the system is improved.

The AS and MS are implemented in JAVA. JAVA socket connection is used as the networking interface to realize the multi-user real time interaction, and to maintain the consistency in the multi-user VE. VRML files store the description of the VE and the virtual embodiment of the clients: Avatar.

Monitoring on the workload in the AS end, the random data of the traffic load in the system is sampled. The traffic load of 100 concurrent user scenario is illustrated in Figure 7 and Figure 8.

5.2 Implementation of the Behavior Based Interaction Model

A small script language: Collaborative Behavior Description Language (CBDL) is designed to work as the interface for integrating various control methods for different interactions in a shared scene. CBDL is designed to support the behavior based interaction model in the CVE system. The implementation of the interaction model in the CVE system is in fact the implementation of an information filter. Through the filter, the information flows in the VE are controlled according to the descriptions in the CBDL files.

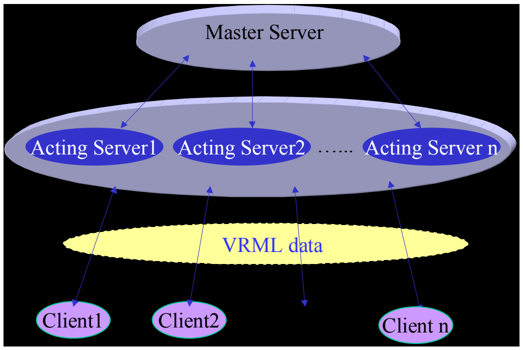
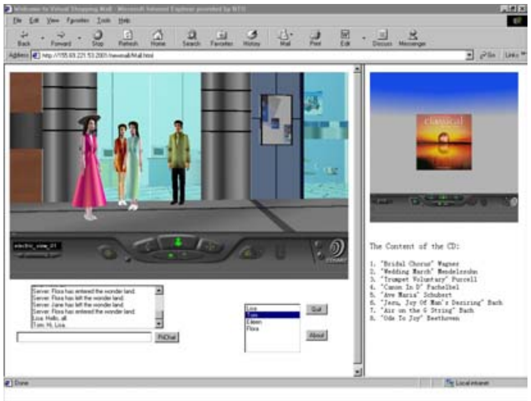


Figure 5. Server Architecture to Support VRML Based Large-scale CVE System on the Internet.



Server: Flora has entered the wonder land
Server: Flora has left the wonder land
Server: Jane has left the wonder land
Server: Flora has entered the wonder land
Lisa: Hello, all
Tom: Hi, Lisa

The Content of the CD:

1. "Bridal Chorus" Wagner
2. "Wedding March" Mendelssohn
3. "Trumpet Voluntary" Purcell
4. "Canon In D" Pachelbel
5. "Ave Maria" Schubert
6. "Jesu, Joy Of Man's Desiring" Bach
7. "Air on the G String" Bach
8. "Ode To Joy" Beethoven

Figure 6. screen capture of the system from the client side interface.

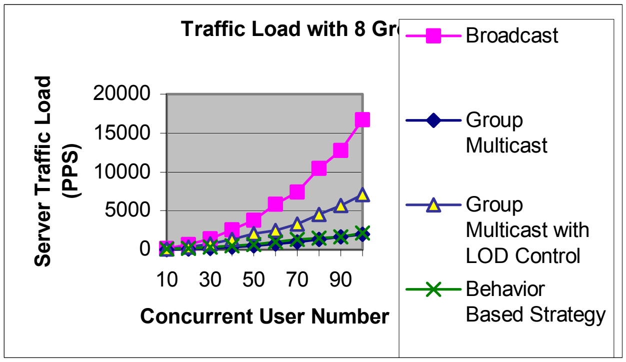


Figure 7. Server Traffic Load comparison with existing interaction message control approaches with 100 concurrent users in 8 groups.

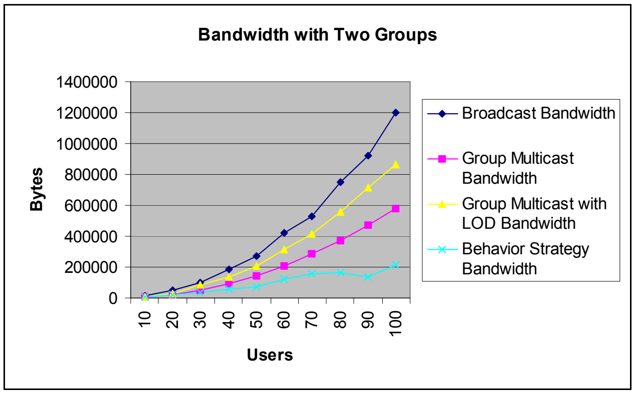


Figure 8. Server Traffic Load comparison with existing interaction message control approaches with 100 concurrent users in 2 groups.

For every application, a series of VRML files are given to describe the entities in the VE. And a series of CBDL script files are also attached to define the collaborative behaviors of the entities in the VE and the collaborative information flow controls for these behaviors. Because that the CBDLworks as an enhancement to VRML for multi-user application, the filter works as enhancement to VRML browser for the same purpose. For the single user application, VRML employs an event model to describe the behaviors of the entities in the VE. The filter works on an event driven model in accordance with the existing event model of VRML.

The events in the VE are triggered and propagated by the entities in the VE. The client side implementation will be in charge of the event monitoring for event/information trigger/propagation of the client. While the server will be responsible for monitoring on the event and triggered/propagated the event/information from/to other roles, for example, the interactive objects.

6. Sample Application and Concluding Remarks

A prototype of a Virtual Shopping Mall (VSM) CVE application has been developed based on the proposed approach. It contains three shops: an electric shop, a design studio and a musical instrument shop. Figure 9 is a screen capture of the system from the client side interface. Through the client interface of the VSM, the users could log into the system, interact with the 3D interactive objects demonstrating the products, communicate with other simultaneous shoppers from different physical places, and get help from the shop assistants. The behavior-based interaction management mechanism is used in the application to manage multimedia transmission in the VE and improve the efficiency of the communications. It has demonstrated good performance with this prototyped application. The performance of the proposed interaction message flow control approach has also been demonstrated in the server traffic load comparison with existing interaction message control strategies as shown in Figure 7 and Figure 8.

Bibliography

[1] A. L. Ames, D. R. Nadeal, and J. L. Moreland, "VRML 2.0 Sourcebook "

[2] J.W.Barrus, R. C. Waters and D. B. Anderson, " Locales: Supporting Large Multiuser Virtual Environments", *IEEE Computer Graphics and Applications*, November 1996, pp 50-57,

[3] S.Benford, L.Fahlen, " A Spatial Model of Interaction in Large Virtual Environments", *ECSCW'93*, Milan, September, 1993 [4] S. Benford, J. Bowers, L. Fahlen, C. Greenhalph,"Managing mutual awareness in collaborative virtual environments", *Proc. of the ACMconference on Virtual Reality Software and Technology (VRST'94),*Singapore, August 1994, ACM Press. 1994

[5] S. Benford, J. Bowers, L. Fahlen, and C. Greenhalph, " Networked Virtual Reality and Cooperative Work ", *Presence,* Vol.4. No.4. Fall 1995, pp 364-386

[6] C. Carlsson and O. Hagsand, " DIVE - a Platform for Multi-user Virtual Environments", *Computer&Graphics*, Vol.17. No. 6., 1993, pp. 663-669

[7] T. Das, G. Singh, A. Mitchell and others, " NetEffect: A Network Architecture for Large-Scale Multi-user Virtual Worlds", *Proceeding of VRST'97*, 1997

[8] P. Garcia and others, " VR-CLASS: A multi-user VRML teaching environment ", *EuroMedia99,* 1999

[9] R.Gossweiler et " An Introductory Tutorial for Developing Multiuser Virtual Environments", *Presence*, Vol.3. No.4. Fall 1994. pp. 255-264

[10] C.Greenhalgh, " An Experimental Implementation of the Spatial Model", in *Proc. 6th ERCIM workshops*, Stockhom, June 1-3, 1994

[11] C.Greenhalgh and S.Benford, "MASSIVE: A Distributed Virtual Reality System Incorporating Spatial Trading", in *Proceeding of the 15th International Conference on Distributed Computing Systems(DCS'95),* Vancouver, Canada,1995, pp.27-34

[12] C.Greenhalgh, "Dynamic, embodied multicast groups in MASSIVE-2", *Technical Report NOTTCS-TR-96-8,* Department of Computer Science, The University of Nottingham, UK, December, 1996

[13] C.Greenhalgh, "Supporting Complexity in Distributed Virtual Reality Systems*", Technical Report NOTTCS-TR-96-6,* Department of Computer Science, The University of Nottingham, September, 1996

[14] R. Lea, Y. Honda and K. Matsuda, " Virtual Society: Collaboration in 3D Spaces on the Internet", Couputer Supported Cooperative Work*: The Jounal of Collaborative Computing* 6. 1997, pp 227-250

[15] J.Leigh et. "CAVERN: A Distributed Architecture for Supporting Scaleable Persistence and Interoperability in Collaborative Virtual Environments", *Virtual Reality*, Vol.2, No. 2, 1996, pp217-237

[16] R. Macedonia, M. J. Zyda, D. Pratt and others, "Exploiting Reality with Multicast Groups: A Network Architecture for Large-scale Virtual Environments", *in Proc. 1995, IEEE Virtual Reality Annual International Symposium (VRAIS'95*), March, 1995, RTP, North Carolina

[17]M. Macedonia, and others, "NPSNET: A Network Software Architecture For Large-Scale Virtual Environments", *Presence*, 3(4), Fall, 1993, pp. 265 – 287

[18] J. Maxfield, "A Distributed Virtual Environment for Collaborative Engineering", *Presence,* Vol. 7. No. 3. June 1998, pp241-261

[19] D. Ratenko and B.Kirsch, " Sharing Virtual Environments over a Transatlantic ATM Network in Support of Distant Collaboration in Vehicle Design", *Proc. Virtual Environments 98 Conference, Stuttgart, Germany*, 1998, pp 1-48

[20] M. Reddy, " Managing Level of Detail in Virtual Environments: A Perceptual Framework*", Forum Short Paper in PRESENCE*, Vol. 6, No. 6, December 1997, pp 658-666

[21] B. Roehl, J. Couch, and others, " Late Night VRML 2.0 with Java "

[22] D. Snowdon and J. Adrian, " AVIARY: Design Issues for Future Large-Scale Virtual Environments ", *Presence*, Vol. 3, No. 4, Fall 1994. pp.288-308

[23] D.Snowdon, S.Benford, C.Greenhalgh, R.Ingram, C.Brown, L, Fahlen and M.Stenius," A 3D Collaborative Virtual Environment for Web Browsing", *Virtual Reality WorldWide'97*, Westin Santa, CA, April 1997

[24] M.Wray, "Distributed Virtual Environments and VRML: an Event-based Architecture", *Proceedings of the 7th International WWW Conferece*, Brisbane, Australia, 1998

COMTEC

TELE-COMMUNICATION SECURITY

INSIDER THREAT RE-EMERGES THROUGH IMPROVED INTRANET SECURITY

Jorma Kajava
University of Oulu,
Department of Information Processing Science, Linnanmaa,
FIN-90014 Oulu University, Box 3000, FINLAND,
E-mail: Jorma.Kajava@ oulu.fi

KEYWORDS

Information Security, Intranets, Insider Threats

ABSTRACT

This paper discusses the three dimensions of information security: confidentiality, integrity and availability. The key to understanding Intranet security involves recognizing the crucial differences between Intranets and the Internet and the various co-operation possibilities that virtual networks offer. The specific security threats of Intranets can be found in communications, software, data and operations security. When protection technologies and tools against criminals outside organizations have improved, the new emerging threat is insiders. In this paper, we shall put forward some ideas concerning protection against such threats.

INTRODUCTION

The use of Intranets as internal information transmission channels within organizations serves to emphasize the importance of their secure realization. Information security threats between Intranets and other networks and information systems are rather similar. The technology used in Intranets and the way that technology is used comprises a new threat source. Information security solutions in Intranets are based both on experiences gained from the Internet and on new solutions designed particularly for Intranets. Special areas of interest within Intranet security are communications, software, data and operations security. By researching these four areas, we hope to find usage differences in information security solutions between Intranets and the Internet.

Information security seeks to protect data and information, systems and services under both normal and exceptional conditions. Such protection creates data and information confidentiality, integrity and availability.

The first period of this research was conducted in 1995 - 97. The goal was to understand the possibilities offered by Intranets and to establish their information security vulnerabilities. The first results were published by Remes (1997).

We have researched artificial intelligence and its applications since the 1980s (Kajava & al, 1988a, b). The second stage of the research reported here aimed at introducing more intelligence to management information systems, such as executive information systems (Massa, 1993). The third stage involved the construction of a management information system on an Intranet platform (Kajava & al, 1997).

In the light of our research findings from the past years, we are in a position to say that information in organizational networks is now better protected against criminals at the technical level. In the future, however, we must be prepared to deal with a new threat: insiders.

Previously, insiders were in many cases people who used inside information in their stock market dealings. This insider concept has now been redefined by the likes of Dorothy Dennings, who argues for a broader meaning of inside crimes in her influential book 'Information Warfare and Security' (1998). In this book, she discusses various kinds of attacks or misuses reported in the 1998 CSI/FBI (Computer Security Institute / Federal Bureau of Investigation) computer crime and security survey. According to the survey, inside attacks were somewhat more common than outside attacks, the actual figures being 36 % and 28 %, respectively.

As our earlier work includes a presentation of basic technical solutions for Intranets (Kajava & Remes, 2000), we concentrate in this paper on the re-emergence of insider threats and its implications on technology and software environment changes. Having built their first firewall a number of organizations continued along the same path and constructed others with the result that the networks of these organizations are now divided into several zones. This, in turn, has brought about a change in roles of their employees. This change is reflected in such questions as who has the right to read information in a particular zone, who has the right to add to or change that information, create or delete information, and so on. Where are the lines of responsibility to be drawn in this new situation?

Information security used to be technologically biased, but now we might say that technical solutions constitute the basic platform on which to build secure systems including human activities. Technical security solutions will still have a role to play, but the front stage may soon be occupied by personnel security.

INTRANETS VERSUS THE INTERNET

Intranets have been defined as internal communication systems of organizations based on the standards of the Internet and the World Wide Web (WWW) (Telleen, 1996). Intranets are based on Internet technology, i.e., technologies that all together generate the Internet (Hinrichs, 1997). For example, Internet routers and their communications connections constitute a part of this technology. Intranets could also be defined as the use of Internet technology in organizational networks (Levitt, 1996). In yet another sense, Intranets could be viewed as private networks based on WWW servers (Pal, 1996).

What makes Intranets particularly attractive targets for external and internal attacks is the internal information of the organization. The use of Intranets as internal information systems emphasizes the importance of threats against Intranets from within organizations. It goes without saying that the connection to other networks outside an organization must be designed very carefully during the Intranet construction process. However, the same careful design and construction work is necessary even when implementing connections to other networks within the organization. A potential attacker working inside the organization should not find it any easier to gain access to confidential information. Most networks are protected against external attacks, while failing to address insider attacks adequately. Such internal attacks tend to be underestimated, although can be extremely harmful to the organization.

The main idea of Intranets is the same as that of the Internet, i.e., to facilitate the flow of information. Intranets enable users to keep in touch with physically remote sites of their organizations by using the Internet or other communication networks, which simplifies the transmission of information between the different sites. When using the Internet, the information security threats posed by the Internet to the organization must be borne in mind at all times. For example, the Internet service could be shut down for a while constituting a severe threat for the availability of information.

THE ADVANTAGES OF USING INTRANETS

One significant advantage of using Intranets in comparison with other solutions based on corresponding network systems is the hardware independence of the Internet technology. Internet technology has successfully solved the problem of incompatible hardware and software and permits organizations to connect all their computers, operating systems and databases to a single independent system (Cortese 1996). There is, however, an unresolved compatibility problem in browser technology.

Another major advantage of Intranets in comparison with other corresponding technologies is the simplicity and versatility of the Intranet user interface and the ease with which voice and images can be transmitted. The WWW-based user interface is easy to learn and use. By clicking on hyperlinks it is easy to move from one document to another and to fetch files to workstations. The technology also allows the real time transmission of voice and images.

Intranets also enable one user interface to contain functions belonging to a number of different applications. Reports, announcements, notices, memorandums, phone and address books could all be accessed by a single browser. In addition, the maintenance and use of databases can also be carried out via the Intranet.

Moreover, Intranets make it possible to distribute real time information to all employees, provided that every information producer puts their documents immediately on a proper platform. The use of electronic platforms decreases the distribution of paper documents, as everyone can read and print only such documents as s/he deems important. The distribution of information over an Intranet makes it easier for everybody to follow important news.

The transfer of multienvironment applications to a single Intranet-based system significantly facilitates the use and maintenance of the system. As every Intranet application is based on the same principles of use, users are no longer required to learn the characteristic details of every system. Also maintenance personnel has to deal with one working environment instead of several disparate systems.

INTRANETS AND THE PERSONNEL

The way an organizational Intranet is used has great effect on how the personnel perceives the system. Thus, the following arrangements in the personnel policy must be carefully considered:

•How to organize the technical maintenance of the Intranet
•How to produce information from systems
•How to publish information on the Intranet
•How to use the Intranet.

As far as the ease of use in information maintenance is concerned, the optimum situation would include that all information producers on the Intranet would transfer their documents to the right place in the system and delete them when they become obsolete. This would enable maintenance personnel to concentrate their efforts on the technical administration of the Intranet. However, the free presentation of documents on Intranets is not a good option from the point of view of information security. Lack of content control could result in the publication of incorrect information either intentionally or unintentionally. This, in turn, could have an adverse effect on the organization, for example, in making decisions concerning the development of new products.

The lines of responsibility for the logical design of the system must be defined. Those in charge should make administrative decisions concerning the Intranet and make plans for dividing the activities of the Intranet to accommodate different user groups.Intranet users comprise the fourth group of people involved. Users do not produce new materials, they only use the information produced by others.

Information security threats based on Intranets

There are several ways of classifying information security threats. In this research, we use the following classification: threats based on technology (WWW technology, software, software code under process, telecommunications and viruses), threats based on human activities and natural phenomena. Our earlier work includes a presentation of threats based on technology (Kajava and Remes, 2000).

Threats based on human activities

Hacking comprises the most serious threat for the modern society in the next millennium. All other types of harm may be derived to hacking. Thus, hackers may be behind virus attacks, software piracy, theft, information misuse and sabotage. (Kajava, 2000 a)

Hackers exploit human illiteracy and weaknesses in computer systems to gain access to these systems. Hackers may study the system they attack, damage it or steal information held on it (Simonds, 1996) causing harm to the owner of the system. Even if a hacker does not cause any damage, there will be expenses for the owners as they have to study whether any damage has occurred or not.

Using the Internet as a part of an Intranet poses a serious threat, because the Internet is inherently nonsecure. As a result, users must be very careful particularly in encrypting their communications. Imitation (spoofing), reply (rapid fire), alteration of message contents (superzapping), prevention of service availability and active and passive wiretapping are among the most malicious threats. Wiretapping, for example, could lead to a situation where strategic knowledge regarding an organization gets in the hands of outsiders, if communication encryption is not implemented by means of strong encryption methods.

Hacker tools, although developed for the Internet, are also usable on Intranets. They can be software or hardware based or a combination of both. Their authorized use includes finding and correcting information security weaknesses on Intranets. However, they also enable insiders to hack such communication systems and access information which they are not authorized to access.

Threats caused by people, employees in particular, can be much more serious on Intranets than on the Internet. Personnel invariably constitutes the most severe information security threat. Intranets have made it easier for employees to access information, but they have also made it easier to misuse this information.

Dishonesty among personnel is always an information security threat. After all, it is easier for corporate personnel to gain access to sensitive information than outsiders. Authorized users may be tempted to misuse vital information. Even unauthorized members of staff may access sensitive information, if, for example, they know the weaknesses of the system.

Also carelessness and negligence among personnel may result in sensitive information landing in the hands of unauthorized persons. Papers left lying on a table are easy to read and copy. Printing a document into a wrong address may also have an undesirable outcome.

A low level of knowledge concerning security among personnel is a clear vulnerability. This is partly a result of the current employment situation in which it might in some cases be a problem finding educated personnel to recruit. Short-comings in training and education combined with the use of uneducated personnel in the realization and maintenance of Intranets may seriously undermine the security of the information held on Intranets.

The use of outsiders in the construction and maintenance of Intranets could also be detrimental to information confiden-tiality. Therefore, security concerns are of utmost importance in designing software and systems outsourcing (Kajava and Viiru, 1996) and in drawing outsourcing contracts (Kajava and Jurvelin, 1998).

PROTECTION METHODS

In an attempt to provide protection against Intranet security threats a number of governments have adopted the model of the Canadian Mounted Police (1981) based on eight security levels. The most important levels in Intranets are those of communications, software, data and operations security. These are the areas in which Intranet security solutions differ most from their Internet counterparts.

Communications protection

Communications security is very important, particularly when the Internet is employed as a communication channel between the different sites of an organization. The same solutions are applicable for isolating Intranets from external networks than in protecting internal networks. Intranets may be protected against external hacking and other information security threats by firewall hardware. Firewalls are used to control traffic in communication networks (Siyan and Hare, 1995), and they do it by examining the communication passing through them and by imposing certain restrictions on it. Such communications as fail to follow the restrictions are filtered out.

The various parts of an Intranet can be isolated with firewalls thus increasing information security against internal misuse. Such firewalls also prevent external intruders from moving on the Intranet.

An organization's network addresses can be encrypted for outsiders using address translation. In address translation, all data packages have an identical network address when they leave the network of the organization. Address conversion disables outsiders from making decisions based on real network addresses and the number of addresses within the organization (Bernstein et al, 1996).

Encryption precludes external parties from examining the contents of messages in networks and information in systems. Encryption is realized by means of encryption methods based on encrypting algorithms. The most common methods are DES (Data Encryption Standard), IDEA (International Data Encryption Algorithm) and RSA (Rivest Shamir Adleman). One popular method on the Internet is SSL (Secure Sockets Layer), which is used in WWW contacts (Stallings, 1997, 2000). In Intranet communications, encryption works with application software that uses some encryption algorithm. It is of paramount importance that the encryption software fulfils the requirements of the organization and is applicable to the intended application area. The strength of the encryption is particularly important, if the organization uses the Internet as a communication channel.

Another method of protecting communication on the Internet is known as tunnelling. In this method, the data area of a communication packet contains another communication packet in its entirety. To enhance the degree of protection, a tunnelling communication contact may also be encrypted.

Software protection

Development is faster within software security than in any other area of Intranet security. The application software used on Intranets constitutes the most salient difference between Intranets and the Internet. In practice, Internet software security equals browser and server software security and security of software for producing WWW applications. Also Intranet software security is based on application programmes in browsers. This fact has far-reaching consequences, because many Intranet applications are tailor-made for one single organization.

All critical Intranet applications must require user identification. Together with password controls, user identification provides protection against unauthorized use of applications. From the organizational point of view, the right to use critical applications could be enhanced by one-time passwords.

Audit trails record exceptional incidents and other security relevant events in the use of Intranet applications. Audit trails should be produced and stored for an agreed period of time to assist in eventual investigations and to monitor access control by recording, for example, all dates and times for logon and logoff.

Virus protection is an important aspect of Intranet security. Intranets make it easy to transfer documents and files between the different sites of organizations. Moreover, e-mail attachments can also be used as a vehicle for transferring computer viruses and infecting new computer systems and networks.

Data protection

Appropriate data protection decreases the probability of internal information security threats. Personnel should have guidelines concerning all materials that can be published on a corporate Intranet. A simple, yet effective way of providing a security classification on an Intranet is to divide all corporate data into data that can be published and data that cannot be published on the Intranet. When publishing confidential data on the Intranet, the limitations of user rights are an essential consideration. There must also be clear guidelines about the deletion of data on the Intranet. In addition, all data published on the Intranet must be backed up using appropriate back-up media. Finally, user rights for every Intranet directory and file must also be defined carefully.

Operations protection

Intranet operations security consists of activities which advance security without influencing practices. Thus, security threats posed by corporate personnel should be prevented in a simple and efficient manner without compromising the efficiency of the system as perceived by the users. The right to use the different parts of an Intranet and the right to access each data directory must be defined for every employee in accordance with to his/her tasks (Code of Practice, 1993). Remote access must also be carefully regulated to ensure security.

The design and implementation of effective user rights management is in many ways a demanding process, but a successful solution significantly decreases information security threats posed by corporate personnel.

TECHNOLOGY CHANGE

Information security education is closely connected with changes in technology. The past decade witnessed a remarkable improvement in security in connections between two adjacent servers. However, the advent of the Internet changed all that by interconnecting all those small networks, thereby enabling large numbers of people to access the systems. The rapid growth of telecommunications adds to the problem as most of it takes place via the Internet (Kajava & Varonen, 2000).

During the 1990s, interconnected partners could negotiate an agreement on the secure use of their networks, but this 1:1 relationship has changed dramatically to a N:N relationship involving an enormous number of interconnected servers. This paradigm change that affects large computer systems such as those used by public administrations brings other problems in its train. For example, the question of who owns a particular piece of information may arise, along with its corollaries such as who has the right to alter it, delete it, append to it, and so on. These questions may be solved in a particular government office or company, but even that may result in loss of compatibility with other organizations. If technology changes produce unresolved questions of this kind, the ensuing effects include loss of flexibility in interconnectivity and reduction in security control.

Researchers have debated at recent international conferences over the issue of whether a run-through of the information security features of the Internet protocol is enough to educate end-users. If the development of networks had stayed at the level it had reached when EDI-based (Electronic Data Interchange) solutions were designed in the 1980s, the answer would be a resounding yes. After all, we can be fairly sure that encryption between two interconnected partners, be they administrative offices or commercial companies, is secure. But the mere use of an encryption algorithm may pose problems when a third server is included, let alone one which has to be contacted through a network of other servers. A case in point is the identification of the various co-operations partners. Moreover, how can we ascertain the integrity of a service we buy through an open network? And whose responsibility is it to answer for it?

Education and training are a means of directing people's attention to these matters and, gradually, to act accordingly. Therefore, the emphasis in security education must be diverted from predominantly technical issues to including a strong people component, by raising end-users' awareness level, enhancing their security skills and simply by respecting and paying more attention to them.

CONCLUSION

Intranets have gained in popularity during the past few years. They started out as private communication channels, but their use has been extended to include DSS (Decision Support System), CSCW (Computer Supported Cooperative Work), expert systems, database maintenance tools, private phone catalogues and user guidelines. Unfortunately, also the number of security problems has increased. From the organizational point of view, these problems call for particular Intranet solutions.

The solutions offered are similar to those provided for the Internet. However, as the usage area of Intranets differs from that of the Internet, it is important to re-examine well-known security threats on the Internet and try to find ways of protecting Intranets against these threats. Intranets make organizations more vulnerable to internal threats.

In 1995, Internet security came under discussion. Owing to poor security, a lot of organizations started seeking better security solutions for their private networks. One measure that is widely used today is firewalls. Efficient as they might be, firewalls are no panacea. As a result, Internet security remains relatively poor, although there are a number of successful software and hardware protection solutions in use. At the same time, also protocols have improved their security-related definitions. Consequently, if we exclude "professional" criminals, amateur criminals find it exceedingly hard to gain access to organizational networks.

It is a sad fact that the difference between good and bad escapes a lot of people. During the mainframe period 20 years ago, insiders posed the worst threat to organizations. And today, barring professional criminals, insiders have regained their position as the most common threat to organizations.

So, after all those years, we are back in square one. The question is, what can be done to remedy the situation. In my view, we need more education and training inside organizations, and we need information security awareness programmes. In addition, novel network solutions enable us to understand the importance of the non-technical implications of information security.

After every developmental stage in information and communication, people seem to return to their old ways. Perhaps we should not talk so much about the intelligence of the human race, if people are incapable of learning the basic rules of society.

REFERENCES

Bernstein, T.; Bhimani, A. B.; Schults, E.; and Siegel, C. A. 1996, *Internet security for business*. J. Wiley & Sons Inc.

A Code of Practice for Information Security Management. 1993. Department of Trade and Industry. GBIS. PD 0003, England.

Cortese, A. 1996. "Here comes the Intranet". Business Week. February 26.

Denning, D.E. 1998. *Information Warfare and Security.* Addison Wesley Longman, Inc. Reading, MA.

Hinrichs, R. J. 1997. *Intranets: What´s The Bottom Line?* Prentice Hall.

Kajava J.; Paani, H. 1988. "Building Software Systems by Using Conventional and Knowledge Based Approaches". University of Oulu,. Working Papers Series B 13. Oulu.

Kajava, J.; Turunen, J.; and Junno, S. 1988. "Organization of Optimization Models of Decision Support Systems". University of Oulu, Working Papers Series B 14. Oulu.

Kajava, J. and Viiru, T. 1996. "Delineation of Responsibili-ties regarding Information Security during an Outsour -cing Process from the Client's Point of View". In *Notes on Information Security Management 1996*. R. von Solms(Ed.) International Federation for Information Processing. Port Elizabeth, South Africa.

Kajava, J.; Hilden, K.; Helander, S.; Mustajärvi, K.; and Valtanen, J. 1997. " Ebony Intra Web - Product Description". Elektrobit Ldt, Oulu.

Kajava, J. and Jurvelin, P. 1998. "Information Systems Outsourcing: Contract and Requirements for the Vendor". In *Databases and Information Systems*. Janis Barzdins (ed.). Proceedings of the Third International Baltic Workshop, Riga, Latvia.

Kajava, J. and Remes, T. 2000. "Intranet Security Threats from Organizational Point of View". In *Databases and Information Systems*. Proceedings of the Fourth IEEE International Baltic Workshop. Albertas Caplinskas (ed.). Vilnius, Lithuania.

Kajava, J. 2000 a. "Hackers – the Most Insidious Threat to the Information Society". In *Nätets juridik*. Nordisk årsbok i rättsinformatik 1999. Red. Ari Koivumaa. Stockholm: Jure AB.

Kajava, J. 2000 b. "Intranet Security from the Organizational Point of View - The Re-emerging Insider Threat". In Steven Furnell (ed.): Proceedings of the Second Interna-tional Network Conference (INC'2000). Plymouth, UK.

Kajava, J. and Varonen, R. 2000. "Information Security Education: From the End-User Perspective to Public Administration Applications". 3. Internationale Fachtagungen *Verwaltungsinformatik in Theorie, Anwendung und Hochschulausbildung* (FTVI HBS 2000). Magdeburg, Halberstadt, Germany.

Levitt, L. 1996. Intranets: Internet Technologies Deployed Behind the Firewall for Corporate Productivity.

Massa, A. I. 1993). "Executive Information Systems (EIS) - The Possibilities in Steel Industry". University of Oulu, Department of Information Processing Science, Oulu.

Pal, A.; Ring, K.; and Downes, V. 1996. *Intranets for Business Applications; User and Supplier Opportunities*. Ovum Ltd.

Remes, T. 1997. "Intranet Information Security Threats and the Protection against them". University of Oulu, Department of Information Processing Science, Oulu.

Royal Canadian Mounted Police. (1981), Security in the EDP Environment. Security Information Publication, 2nd Ed. Gendarmere Royale du Canada. Canada, October.

Stallings, W. 1997. *Cryptography and Network Security: Principles and Practice*. Second Edition. Upper Saddle River, New Jersey. Prentice Hall.

Stallings, W. 2000. *Network Security Essentials: Applications and Standards*. Upper Saddle River, New Jersey. Prentice Hall.

Siyan, S. and Hare, C. 1995. *Internet Firewalls and Network Security*. Indianapolis, USA: New Riders Publishing.

Telleen, S. T. 1996. The Intranet Architecture: Managing information in the new paradigm. Sunnyvale, CA. Amdahl

A RESPONSE-ORIENTED TAXONOMY OF IT SYSTEM INTRUSIONS

M.Papadaki[†], S.M.Furnell[†], B.M.Lines[†] and P.L.Reynolds[‡]

[†] Network Research Group, Department of Communication and Electronic Engineering, University of Plymouth, Plymouth, United Kingdom
[‡] Orange Personal Communications Services Ltd, St James Court, Great Park Road, Bradley Stoke, Bristol, United Kingdom.
e-mail: nrg@plymouth.ac.uk
Web: http://www.plymouth.ac.uk/nrg

KEYWORDS

Intrusion Detection Systems, Intrusion Taxonomy, Intrusion Response.

ABSTRACT

The ability to select and initiate appropriate response(s) is an issue that is often neglected in Intrusion Detection Systems (IDS). In order to address the problem, a means is required to consider different potential security breaches, the differing contexts in which they may occur, and the differing potential consequences. Current intrusion taxonomies have limited application in this regard, considering categories of intrusions that could not be detected by an IDS, or representing potential results in too few dimensions to enable any fine-grain selection of response options. This paper presents an overview of a new taxonomy, which is specifically targeted towards enabling the consideration of responses. A number of generic incident and target categories are identified, encompassing the most common forms of intrusion/attack and the contexts in which they may occur. An assessment of the likely results is then presented in each case, considering the security impacts, the time available to respond, and further potential attacks that may be initiated as a result. By encompassing alternative targets, and considering multi-dimensional results, the taxonomy provides a means of differentiating the incidents on the basis of the responses they require, rather than by characteristics of the attack method or their security impacts alone.

INTRODUCTION

Intrusion Detection has been an active research area within the computer security domain for more than 15 years. The challenges associated with this area have so far been largely concentrated on the process of detecting an intrusion. However, automation of the next stage after detection, the response to an incident, is a significant issue that has not been adequately addressed and therefore requires further research in its own right.

Intrusion response is defined as the process of counteracting the effects of an intrusion. It includes the series of actions taken by an Intrusion Detection System (IDS), following the detection of a security-related event. The justification for advancing the automated response capability of IDS is twofold: firstly, to reduce the significant overhead that manual response poses to the administration of increasingly large and complicated IT infrastructures, and secondly, to cope with the widespread use of automated scripts that can generate attacks of distributed nature.

In order to select appropriate responses, it is necessary to know more than just the type of incident that has occurred, or the basic security impact that has resulted. However, many current intrusion classification taxonomies provide little understanding beyond this level. As such, a new taxonomy has been developed as the basis for studying the issue of response, aiming to consider incidents and identify their different results in different contexts. It is intended that this taxonomy will give insight into the process of selecting appropriate responses and forming the basis of decision-making in an automated responder system (Furnell and Dowland 2000)

The discussion begins by summarizing previous work that has been conducted in relation to intrusion and attack taxonomies, before proceeding to present details of the new approach. The concept of the response-oriented taxonomy builds upon previous ideas, originally introduced by Furnell et al (2001).

CURRENT INTRUSION TAXONOMIES

Previous research has given rise to a number of intrusion taxonomies, each of which presents an alternative view of the situation. Brief summaries of a number of notable approaches are given below.

A common method of classifying security incidents is according to the impacts or outcomes resulting from their occurrence. This has led to a number of result-based taxonomies of incidents and attacks. In such approaches, all attacks are grouped into basic categories according to their result, aiming to give more insight into their severity. An example is a taxonomy devised by Cohen (1995) that includes result categories such as *Corruption, Leakage,* and *Denial. Corruption* is defined as the unauthorised modification of information, *leakage* is when information ends up where it should not be, and *denial* is when computer or network services are not available for use. Another result-based taxonomy is specified by Russell and Gangemi (1991), who define similar outcome categories, but use a different set of terms (i.e. *secrecy* and *confidentiality* instead of *leakage; accuracy, integrity*, and *authenticity* instead of *corruption;* and *availability* instead of *denial*).

Although result-based taxonomies can be useful in providing a meaningful association between different types of attacks, the end result of an attack is not the only significant characteristic and thus it represents only one aspect of the problem. In order to detect, respond, and specify protection, it is necessary to have some classification of the incidents that lead to the results. In this respect, there are also a number of prior works that can be considered.

Cheswick and Bellovin (1994) classify attacks into the seven categories listed below:

- Stealing passwords - methods used to obtain other users' passwords
- Social engineering - talking your way into information that you should not have
- Bugs and backdoors - taking advantage of systems that do not meet their specifications, or replacing software with compromised versions
- Authentication failures2 - defeating of mechanisms used for authentication
- Protocol failures - protocols themselves are improperly designed or implemented
- Information leakage - using systems such as *finger* or the *DNS* to obtain information that is necessary to administrators and the proper operation of the network, but could also be used by attackers
- Denial-of-service - efforts to prevent users from being able to use their systems.

Although this approach provides a general overview, including the main categories of intrusions, it is not specified in any further detail, and thus is too general to provide any insight to the relationship among different classes of attacks or their different characteristics.

Neumann and Parker (1989) developed an intrusion taxonomy based on a large number of incidents reported to the Internet risks forum. The taxonomy classifies intrusions into nine categories, according to key elements that might indicate a particular type of incident. Table 1 below summarises the overall scheme.

NP 1 EXTERNAL MISUSE	Nontechnical, physically separate intrusions
NP 2 HARDWARE MISUSE	Passive or active hardware security problems
NP 3 MASQUERADING	Spoofs and Identity changes
NP 4 SUBSEQUENT MISUSE	Setting up intrusion via plants, bugs
NP 5 CONTROL BYPASS	Going around authorised protections/controls
NP 6 ACTIVE RESOURCE MISUSE	Unauthorised changing of resources
NP 7 PASSIVE RESOURCE MISUSE	Unauthorised reading of resources
NP 8 MISUSE VIA INACTION	Neglect of failure to protect a resource
NP 9 INDIRECT AID	Planning tools for misuse

Table 1: SRI Neumann-Parker taxonomy

An extension of the Neumann-Parker taxonomy was produced by Lindqvist and Jonsson (1997), which further refines security incidents into intrusions, attacks and breaches. It examines these issues from a system-owner perspective, based on a number of laboratory experiments. The results of these experiments indicated a need for further subdivision of the Neumann-Parker classes 5, 6 and 7, as shown in Table 2 below. Their work provides further insight into the process of spotting aspects of system elements that might indicate an intrusion.

Extended NP5 CONTROL BYPASS	Password attacks, spoofing privileged programs, utilizing weak authentication
Extended NP6 ACTIVE RESOURCE MISUSE	Exploitation of write permissions, resource exhaustion
Extended NP7 PASSIVE RESOURCE MISUSE	Manual browsing, automated browsing

Table 2: Lindqvist and Jonssen extension of the Neumann-Parker taxonomy

A final example is provided by Howard (1997), who follows a different approach by focusing on the process of an attack, rather than classification categories. Howard's taxonomy establishes a link through the different potential *attackers* (classified as hackers, spies, terrorists, corporate raiders, professional criminals and vandals) and the *tools* and *access methods* that they may utilise, leading to the *results* that enable the attackers to achieve their objectives. This taxonomy was based on the analysis of real incidents, as reported to the CERT/CC from 1989 to 1995, and thus represents a very valuable tool for systematically studying attacks. Having said this, it does not present a comprehensive top-level classification of intrusion incidents, or yield an appropriate classification that could be used to determine the required response – a criticism that could also be levelled at the other examples considered here.

Although most of the existing taxonomies succeed in contributing to the systematic study of intrusions, they are not immediately applicable to the domain of automated intrusion detection and response systems. From a detection perspective, it is clear that a number of the incident classifications identified (e.g. social engineering, physical tampering), and issues such as the objectives of attackers, could not be detected or determined by an automated system. In addition, they do not give any insight into the issue of response. A taxonomy that would serve this purpose ought to give consideration to the classification criteria, which will include aspects such as incident type, target, and/or potential impact. This will lead to indication of generic response categories, considering what can be done to halt an attack in progress, reduce its impact and/or prevent reoccurrence. The discussion of such a taxonomy is the focus of the next section.

A RESPONSE-ORIENTED TAXONOMY

The aim of the new taxonomy is to determine the effect an incident has on specific targets, and demonstrate how that may influence the response decision process. In order to demonstrate that concept a set of incidents have been used and are listed below:

1. Information gathering (Probe / Scan, Sniff)
2. Authentication failure (Masquerade / Spoof, Bypass)
3. Software compromise (Buffer Overflow, Flood / Denial of Service (DoS)
4. Malware (Trojan Horse, Virus / Worm)
5. Misuse (Unauthorised Alteration, Unauthorised Access)

As with the previous taxonomies, the selection of incidents is by no means exhaustive, but the five top-level categories aim to encompass the most significant set of incidents that affect current systems. Also, the description of the incidents used in the taxonomy aims to preserve a high level of abstraction, in order to include as many cases of incidents as possible. So, for example, although there are many different methods of launching Denial of Service attacks (such as SYN Flooding, SMURF attacks, Ping of Death, Trin00, and others), their ultimate effect on a system is similar, and it is this that will be the main determinant of the desired response(s). The five incident categories, and example incidents, are described more fully later in this section, following discussion of the other elements of the taxonomy.

Another important characteristic that can influence response is the *Target* of the intrusion, since the same incident can have different impacts upon different targets. The target groups considered in the new taxonomy are as follows:

- *External server:* Public-facing servers that are accessible from external networks and represent the public image of the host organization (e.g. web, email, DNS, FTP servers). Ideally, if configured correctly, external servers should not contain or facilitate access to confidential information, but ought to provide uninterrupted service to clients.
- *Internal server:* A server accessible only within the internal network of the organization (e.g. intranet web and file servers).
- *User workstation:* Computing units used by average users, likely to contain information specific to a particular user and their role within the organisation.
- *Network Component:* Networking equipment such as routers, switches, firewalls, which may be targeted as a means of accessing other systems or subverting operations.

This is by no means a detailed or exhaustive list, but it is sufficient to give a high level abstraction of the different elements that might be targeted in a typical organisation.

As well as the incident type and the target, the other significant characteristic that must be considered in order to select a response is the likely *result(s)* of an intrusion. However, this aspect cannot be represented in only one dimension, and the taxonomy presented here considers it to be comprised of *urgency*, *severity*, *impact(s)* and *potential incidents* arising from an incident.

The *Urgency* relates to the need for timely response, and partially reflects the speed of the attack. Since some attacks can evolve more rapidly than others, it is important to consider how much time is available to respond in each case. A Denial of Service attack, launched with the use of automated scripts is an example of a rapidly evolving attack, while sniffing traffic in a Local Area Network (LAN) allows a greater window of opportunity for response, as it is likely to evolve in a longer period of time. Another dimension of the result is the *Severity* of the intrusion, which relates to the magnitude or extent of the attack. The more severe an intrusion is, the sooner it needs to be contained, in order to eliminate its impacts and the threat introduced in the system. In the taxonomy, both urgency and severity are rated on a scale of Low, Medium, High for each incident / target combination.

Apart from the urgency and severity, another aspect of the result is the consideration of the *Impact(s)* of an intrusion upon a system. The *Impact(s)* relate(s) to the asset(s) of the system that have been compromised by the intrusion and may be observed and measured in relation to the *Confidentiality, Integrity* and / or the *Availability* of systems and data. Although in scenarios such as conventional risk analysis (Davey 1991) it is normal to rate these impacts on a sliding scale to indicate their severity, the taxonomy in the table that follows simply indicates whether there is a potential impact or not, as assignment of values would be too subjective.

The final element of the result relates to whether any further incidents are likely to be facilitated as a consequence of the initial attack. This is expressed in the taxonomy as *Potential Incidents*. For example, when sniffer software is used to capture network traffic, it is likely that the information obtained (e.g. user names and passwords) will enable attackers to log in as legitimate users at a later date and thus succeed in the masquerade. In other words, the potential incidents indicate the threat that has been introduced in the system after the occurrence of the original incident.

Having introduced the top-level elements of the taxonomy, the focus will now move to the incident categories identified earlier, as well as justifications to accompany the various ratings included in Table 3.

Information Gathering

The main characteristic of there intrusions is that they aim to collect information about a target and identify exploitable vulnerabilities. Although information gathering does not have significant impact upon a system, it carries the danger of the knowledge gained subsequently being used for launching other attacks with higher severity. Probe, Scan and Sniff are intrusions that fall into that category and will be described below.

Probe / Scan

Probe is used to access a target in order to determine its characteristics. Scan, on the contrary is used to access a set of targets in order to determine which of them have a specific characteristic. The characteristics in question aim to identify the architecture of targeted systems and networks, and usually relate to network configuration, as well as specific versions of services, operating systems and other types of software. The information obtained can subsequently enable the occurrence of incidents, such as spoofing, exploiting vulnerabilities and thus bypassing authentication, compromising software and introducing malware. The impacts relate to breach of confidentiality, as information is obtained without authorisation. Probing and especially scanning can also degrade availability, by producing large amounts of traffic when probing / scanning multiple targets. External servers as well as network components can be affected in this manner, as in both cases availability is highly important and it is those targets that are more likely to deal with that traffic.

The severity of scans / probes varies, depending on which target it is directed to. In the case of external servers and network components, which are genuinely subjected to unknown and thus untrustworthy users, they should be designed to be more tolerant with attacks of this nature. After all, within their normal activity they often provide the same nature of information anyway. Thus the severity of probing / scanning is not significant in those two cases. The urgency to respond is equally low, as apart from having low severity, probing / scanning is not likely to escalate rapidly. On the contrary, probing or scanning an internal server is not usual and thus it raises higher level of suspicion. Bearing in mind the importance of preserving confidentiality in internal servers, the level of high severity is more appropriate. The urgency to respond is medium, due to the high level of severity on one hand and its slow nature, in terms of escalating on the other. As for user workstations, although probing / scanning a user workstation is even more rare and thus raises higher level of suspicion, its impact is not as severe, as the threat to confidentiality in this case is significantly lower. Thus the severity can be regarded as ‘medium’. However, the occurrence of such an incident could mean prior breach of another target (e.g. DNS server), and thus a medium level of urgency to respond is considered appropriate.

Sniff

Sniffing consists of the interception of traffic while it travels across the network. It is achieved with the use of software tools that can capture network packets either locally or remotely. The sort of information obtained with sniffing could be anything that travels across the network, such as user name and password combinations, data files, and system or network information. After obtaining information with sniffers, the potential incidents likely to follow can mainly be masquerading, bypassing, and software compromise.

The impacts of sniffing mainly involve loss of confidentiality, however its severity and urgency depend on the type of targets subjected to it. In external servers the severity is low, since again the nature of information disclosed cannot be significant enough to raise the level of severity. Similarly with probing / scanning, the need for timely response is low, since the severity of the incident and the chance of escalating are low. In the case of internal servers, the severity is again high, however the need to respond is high as well, since the nature of information that can be disclosed in this case is more significant and thus requires a more urgent issue of response. As for user workstations, the nature of information exposed is not significant enough to increase the level of severity and urgency, so as in the case of probing / scanning, both are considered as medium.

INCIDENT	TARGET	RESULT					
		URGENCY	SEVERITY	IMPACT			POTENTIAL INCIDENTS
				C	I	A	
1.Information gathering							
Probe / Scan	External server	Low	Low	✓		✓	Spoof, Bypass, S/w compromise, Malware
	Internal server	Medium	High	✓			
	User workstation	Medium	Medium	✓			
	Net. component	Low	Low	✓		✓	
Sniff	External server	Low	Low	✓			Masquerade, Bypass, S/w compromise
	Internal server	High	High	✓			
	User workstation	Medium	Medium	✓			
	Net.component	Medium	Medium	✓			
2. Authentication failure							
Masquerade / Spoof	External server	High	High	✓		✓	Misuse, Malware, Software compromise
	Internal server	High	High	✓			
	User workstation	Medium	Medium	✓			
	Net. component	High	High	✓		✓	
Bypass	External server	High	Medium	✓			Misuse, Malware
	Internal server	High	High	✓			
	User workstation	High	Medium	✓			
	Net. component	High	Medium	✓			
3. Software Compromise							
Buffer Overflow	External server	High	High		✓	✓	Bypass, DoS, Misuse, Malware
	Internal server	High	High		✓	✓	
	User workstation	High	Medium		✓	✓	
	Net. component	High	Medium		✓	✓	
Flood / DoS	External server	High	High			✓	Spoof
	Internal server	High	High			✓	
	User workstation	Medium	Medium			✓	
	Net. component	High	High			✓	
4. Malware							
Trojan Horse	External server	High	High	✓	✓	✓	Bypass, Misuse, Malware, S/w compr., Info. gathering
	Internal server	High	High	✓	✓	✓	
	User workstation	High	High	✓	✓	✓	
	Net. component	High	High	✓	✓	✓	
Virus / Worm	External server	High	High	✓	✓	✓	Misuse, Malware, S/w compr., Info. gathering
	Internal server	High	High	✓	✓	✓	
	User workstation	High	High	✓	✓	✓	
	Net. component	High	High	✓	✓	✓	
5. Misuse							
Unauthorised Alteration	External server	High	High		✓	✓	Malware
	Internal server	High	High		✓	✓	
	User workstation	High	Medium		✓	✓	
	Net. component	High	High		✓	✓	
Unauthorised Access	External server	High	Low	✓			Malware, Unauthorised Alteration
	Internal server	High	High	✓			
	User workstation	High	Medium	✓			
	Net. component	High	Low	✓			

Table 3: Response–oriented Intrusion Taxonomy

Finally, in the case of network components, the severity of sniffing is medium, since the nature of information exposed in this case (e.g. Access Control Lists, administrator user account details) is significant enough to raise the level of severity. The urgency to respond is also medium, since network components represent single points of failure and a possible compromise could affect multiple hosts.

Authentication failure

Users and processes need to identify and authenticate themselves quite often in order to obtain specific access privileges. As a result, defeating the authentication process is very common objective for attackers, and can be summarised in three main ways, namely Masquerading, Spoofing and Bypassing.

Masquerade / Spoof
Masquerade is the action in which valid identification and verification information that belongs to legitimate users is obtained and used by an impostor. For example, an attacker might use a sniffer to capture user name, password and IP address combinations that are sent across the network, and then use this information to log into accounts that belong to other users. Spoofing, by contrast, involves the provision of false information. In network communications, each packet of information traveling on a network contains source and destination addresses either in the form of MAC, IP addresses, TCP connection IDs, or port numbers. Supplying accurate information is often assumed, however it is possible that incorrect information is entered into these communications, in order to accept an impostor address as original and either trick other machines into sending it data or to allow it to receive and alter data. Examples include IP spoofing, email spoofing and DNS spoofing.

Masquerading and spoofing are mainly a threat to the confidentiality of systems, since they most often provide unauthorised increased access to attackers. However, in the case of external servers and network components, it is possible to cause loss of availability as well, if used as a technique to enable the occurrence of DoS attacks. The potential incidents that can follow masquerading and spoofing are obviously misuse (unauthorised access and alteration of information), malware (introduction of Trojan horses, viruses / worms) and software compromise (Buffer overflow, DoS).

The severity of masquerading and spoofing is considered high in external servers, as it may result in loss of availability. The urgency to respond is high as well, since IP spoofing can very soon escalate to a DoS incident. However, even in the case of masquerading, once unauthorised access is achieved to external servers, it is possible to alter information that can harm the public image of the organisation and thus cause further embarrassment and disruption of operation. In the case of internal servers, even if services are not accessed externally, the danger of disclosing confidential information is considerably high, resulting in severe embarrassment to the organisation, and disruption of its operation. So, the level of severity and the urgency to respond in this case are high as well. As for user workstations, the severity is less significant, as in many cases the nature of information or access level obtained will not pose a great level of threat to the system (although some users will always be exceptions). The level of urgency is medium as well, since the workstation is probably used as a step to achieve increased access into a more significant component of the system (either internal or external server). Obtaining unauthorised access in network components, as well as making them unavailable by achieving DoS attacks is highly severe, as it can affect multiple hosts or even the entire internal network, depending on the scale of the problem. The urgency to respond is thus high as well.

Bypass
Bypass is an action taken to avoid the authentication process by using an alternative method to access a target. For example, some operating systems have vulnerabilities that could be exploited by an attacker to gain privileges without actually logging into any privileged account. Bypass is usually a result of software compromise (e.g. buffer overlow) or malware (e.g. if a trojan horse is used instead of the original authentication process). The issue is again a threat to confidentiality, as increased unauthorised access is achieved. The potential incidents that can follow are misuse (unauthorised access and alteration of information) and malware.

The severity is medium in the case of external servers, since their availability is not threatened directly. However a rapid response is needed to avoid further escalation of the incident, so the urgency in that case is high. In internal servers both severity and urgency are high, as the direct threat is higher, so is the need to avoid escalation of the incident. Although the severity in the case of user workstations is lower, and thus can be considered as medium, the need to respond is equally high, since bypassing authentication is an indication of an already compromised system, so further action should be taken as soon as possible. Finally, bypassing authentication in network components is of medium severity, since the threat to confidentiality is not as severe as in the case of internal servers, but again the need to respond and eliminate any chances of escalating the problem is high.

Software compromise

Intrusions that involve the exploitation of software vulnerabilities fall into this category. There are three main categories of vulnerabilities within a system, namely design, implementation or configuration vulnerabilities (Howard 1997). The main categories of intrusions that fall into this category are Buffer Overflow and Denial of Service; they are presented below.

Buffer Overflow
Buffer overflow is a result of deficient software implementation that allows the assignment of data in a buffer without checking in advance if its size is sufficient to 'host' that data. So in the case of someone sending larger amounts of data, the targeted system will allow the input of data in the buffer anyway, with the result of either crashing the system or overwriting part of memory adjacent to the buffer. As a result of the latter, unauthorised access could be obtained by modifying the flow of program execution, and allowing the execution of arbitrary code with the same access rights granted to the compromised program (Aleph1 1996).

Such incidents can compromise the integrity and availability of the targeted system, and can lead to further incidents such as bypassing authentication, denial of service, misuse or execution of malware. In all cases, the amount of time elapsing before that happens is usually small, as in many cases it even happens almost simultaneously.

Buffer Overflows are more commonly exploited in server software (web, ftp, email, file) since they are easily accessible from external sites and often run under root/administrator privileges. Thus high potential severity can exist for external servers, as well as internal (intranet) servers in some organisations. The urgency to respond is high as well, since apart from the significant severity of the incident, the likelihood of escalation is significant as well, so an urgent response is needed.

In the case of user workstations the severity is medium, since the chances of being subjected to attacks of this nature is less substantial. Also, even if targeted (e.g. server software is running, probably by default) the number of hosts affected are limited (probably only one), so the scale of the problem is less significant. However, the urgency to respond is still high, in order to avoid execution of malware or further compromise of other systems.

The chance of exploiting buffer overflows in network components is even less significant, but the potential impacts of doing so are more substantial than in the case of workstations, since a greater number of hosts can be affected. Thus the severity of buffer overflow is medium in this case. The urgency to respond is again high, for the same reason.

Flood / Denial of Service
Denial of Service (DoS) attacks aim to overload (flood) the capacity of a target by accessing it repeatedly. The result of such action is to make the target unable to respond to any other events / requests and thus become inaccessible to legitimate clients. Subsequent occurrences could include another party assuming the role of the target, resulting in spoofing.

The impact of Denial of Service attacks clearly relates to the availability of the targets. Since these attacks are most often conducted with the use of automated scripts, the need to respond immediately is crucial in most cases. In the case of an external server, the severity is likely to be high, given that a site may represent a public interface of the organization. Inaccessibility could result in embarrassment and loss of custom. The urgency to respond is also high, since usually the time available to prevent either the occurrence of the incident, or subsequent escalation, is very limited. Although DoS to internal servers and network components does not risk causing embarrassment to the organisation, their failure to provide services could have impact on multiple hosts, or even the entire internal network of the organisation, so the severity is also high, as is the urgency to respond. In the case of user workstations, the likelihood of being subjected to a DoS attack is rather small, simply because the impact of doing so is not as significant. User workstations are mostly used as (potentially unwitting) tools to conduct DoS attacks in order to achieve maximum level of effectiveness, but are not the targets. However, it is possible, and it can result in either degradation of performance, or total loss of legitimate usability. Thus the severity in that case is medium. The urgency to respond is medium as well, as the impacts of the attack are of medium severity and the time available to encounter the attack or avoid escalation is usually more.

Malware

Malicious software, also known as malware, characterises the classes of intrusions that are conducted under complete software control. Intrusions falling into this category differentiate from automated software tools used to launch other classes of attacks (e.g. DoS attacks), in the sense that humans are not involved in the escalation of malware attacks; after the initial human involvement to begin the distribution of malware, individual attacks can subsequently occur without the need for the instigator's further involvement. Thus malware can constitute an attack in its own right. There are three main types of malware, namely Trojan horses, viruses and worms and will be discussed below.

The impacts of malware can differ significantly from case to case, since the code in the payload can do nearly everything that is feasible under software control. For example, it is possible to initiate posting of legitimate users' working documents to all the members of his/her address book, resulting to breach of confidentiality (SARC 1999). Alternatively, it is possible to delete or modify files in the system, achieving a breach of integrity. Finally system or network resources can be consumed at the execution of the payload, resulting to either degradation of performance or entire inaccessibility of targets for legitimate use.

The potential incidents that can follow the execution of malware can also be nearly anything. Misuse, other forms of malware, software compromise and information gathering are examples of potential results of malware. Thus the severity of malware varies according to the specific incidents. However, if considering the execution of malware in general, the severity is high in all types of targets, since such a great variety of functionality can potentially be included in the payload. In addition, the risk of spreading to additional targets is extremely high, so the urgency to respond and contain the execution of malware is high as well in all cases.

Misuse

Misuse relates to unauthorised or unacceptable use of system resources. In this sense, it is a quite general term that can actually include all the incidents described so far, since all of them are somehow a form of misusing system resources. However, incidents falling into this category mainly take place after unauthorised access has been obtained in a target and include cases that mainly involve misuse of files and data within a system. It is important to mention at this point that the occurrence of incidents from this category indicates that the targeted system may have already been in a compromised state, unless the activity is being perpetrated by a legitimate user. Hence any response issued might be affected by this factor as well.

Unauthorised alteration

Unauthorised alteration includes actions such as creating, modifying, deleting system or data files. This will affect the integrity and / or availability of resources and represents an important issue that needs to be addressed.

The severity in the case of external servers is high, as information or services might be altered in such a way as to cause embarrassment to an organisation and further disruption to its normal operation. For example, web site defacements (Alldas.de 2001) represent a highly important incident that can immediately attract the interest of media and put the organisation into a difficult situation. Also the modification of information or services could potentially mislead or cheat customers, and result in making the organisation liable for those actions. Although the urgency to respond in such case is high, the feasibility of doing so might be another issue. Certainly the current state of the system needs to be considered in order to determine the effectiveness or selection of an appropriate response.

Unauthorised alteration is highly severe in the case of internal servers and network components as well, since it can result in misleading internal users to make decisions based on inaccurate information or disrupting their operation. Even if the likelihood for rapid escalation of the incident is very small, the need for timely response is high again, since the severity of the incident is so significant.

Finally in the case of user workstations, the importance of the target is typically lower, as it can affect only a limited number of users. The severity is therefore medium. Still, the urgency to respond is high, mainly because the current state of the targeted system should be assessed and any potential risks minimised.

Unauthorised Access

Unauthorised access includes actions that involve disclosure of information to unauthorised parties. As a result of their occurrence, incidents such as unauthorised alteration or execution of malware might follow. Thus the severity of unauthorised access can vary according to the target (and whether confidentiality is at high risk) but the urgency to respond in all cases should be high. That is to firstly assess the current state of the system and prevent further escalation of the incident and occurrence of unauthorised alteration or execution of malware as well.

When external servers or network components are subjected to unauthorised access, the severity is low, since no confidential information should be at risk and no modification has taken place. On the other hand the current state of the system is unknown and needs to be assessed. By contrast, unauthorised access to internal servers has high severity, because there is more important information available for attackers. In the case of user workstations the severity is medium, as there is risk to confidentiality, but it is less substantial.

CONCLUSIONS

In this taxonomy, several incidents have been considered, aiming to illustrate the effect of different types of targets on the results of an intrusion. The ultimate intention is to give insight into the main intrusion characteristics that can influence intrusion response, and subsequently lead to the indication of generic classes of response. Although the response-oriented taxonomy is quite generic and cannot depict the complexity of the response decision process, it can still serve as a basic tool that will enable the research to progress towards that direction. After looking into the results of different intrusions on various targets, it seems that intrusions directed towards internal servers always have the most significant results, mainly due to their importance in the operation of an organisation. By contrast, user workstations have the least significant results, as their role within the organisation is less important and the consequences after the occurrence of an intrusion can more easily be addressed. Finally, network components and external servers seem to depend on the type of intrusion to a greater extent, as some classes of intrusions have more significant results than others.

In terms of response and how different intrusion characteristics can influence the response process, it can be argued that the more severe an intrusion is, the more important it is for the response to focus on the prevention of its occurrence, or its containment. In classes of intrusions with low or medium severity and high urgency, the risk for rapid escalation is significant, and so the response process should focus on the prevention of further escalation (prevent the occurrence of potential incidents). Finally, the severity and urgency can affect the transparency of the initiated response. It seems that there should be a trade-off between them, as the more severe the intrusions, the less transparent responses can apply.

It should be noted that there are several limitations in this taxonomy. For example, apart from the type of target, the number of systems targeted could also be considered, as the scale of an incident will certainly influence its severity. For example, a virus that infects a small number of user workstations is not as severe as one that infects all of them. However, the omission of this factor does not prevent the taxonomy from fulfilling its objective of demonstrating that the same category of incident can demand different responses in different contexts.

As regards the responses themselves, it may appear curious that they have been omitted from the taxonomy presented here. The basic reason is that the taxonomy is intended to provide the foundation for an automated decision mechanism within a software agent. The specific response options available could vary depending upon the environment in which the agent is deployed, and thus the classification taxonomy is independent of any particular mapping. In the context of such an agent, the decision-making process could also be more complex. Although incident and target related characteristics are the main determinant of the likely result of the incident, various other contextual factors could be measured when an incident is detected in order to better inform the response decision process. For example, the account in use, the current alert level of the IDS, and the nature of any responses already issued could all influence the choice of response that is likely to be the most effective. Further consideration of this issue is presented in (Papadaki et al. 2002), and the issue represents the focus of ongoing research by the authors.

REFERENCES

Aleph1 (1996), “Smashing The Stack For Fun And Profit”, Phrack online journal, vol. 7, issue 49, 8 November 1996.

Alldas.de (2001), “Defacement Archive”, http://defaced.alldas.de/, 26 October 2001.

Cheswick W.R., and Bellovin S.M. (1994), 'Firewalls and Internet Security: Repelling the Wily Hacker', Addison-Wesley Publishing Company, 1994.

Cohen F.B. (1995), *Protection and Security on the Information Superhighway*, John Wiley & Sons.

Davey J. (1991), “The CCTA risk analysis and management methodology (CRAMM)”, Current Perspectives in Healthcare Computing, pp. 360 – 365.

Furnell S.M. and Dowland P.S. (2000), "A conceptual architecture for real-time intrusion monitoring", *Information Management & Computer Security*, Vol. 8, No. 2, pp65-74.

Furnell, S.M, Magklaras, G.B, Papadaki, M. and Dowland, P.S. (2001), "A Generic Taxonomy for Intrusion Specification and Response", in *Proceedings of Euromedia 2001*, Valencia, Spain, 18-20 April 2001: 125-131

Howard J.D. (1997), PhD thesis 'An Analysis of Security Incidents on the Internet 1989 - 1995', Carnegie Melon University, 7 April 1997, http://www.cert.org/nav/reports.html

Lindqvist U., and Jonsson E. (1997), “How to Systematically Classify Computer Security Intrusions”, in Proceedings of the 1997 IEEE Symposium on Security and Privacy, May 4-7, 1997, IEEE Computer Society Press.

Neumann P.G., and Parker D.B. (1989), “A summary of computer misuse techniques”, in Proceedings of the 12th National Computer Security Conference, Balitimore, USA, 10-13 Oct 1989, pp. 396-407.

Papadaki, M., Furnell, S.M., Lee, S.J., Lines, B.M. and Reynolds, P.L. (2002), “Enhancing response in intrusion detection systems”, submitted to *Journal of Information Warfare*.

Russell D. and Gangemi G. T. (1991), "Computer Security Basics", O’Reilly & Associates, Inc., Sebastopol, CA, 1991.

SARC. 1999. “W97.Melissa.A virus overview”. Symantec AntiVirus Research Center. http://service1.symantec.com/sarc/sarc.nsf/html/W97.Melissa.A.htm

ACKNOWLEDGEMENTS

The research presented in this paper has been supported by funding from the State Scholarships Foundation (SSF) of Greece.

NETWORK MANAGEMENT ENSURING QoS ACCORDING TO CONTENTS POLICIES

Kazunori Ueda
Osaka School of International Public Policy
Osaka University
1–31 Machikaneyama, Toyonaka, Osaka, JAPAN
E-mail: ueda@osipp.osaka-u.ac.jp

Takao Shimayoshi
Toshihiko Hata
Industrial Electronics & Systems Laboratory
Mitsubishi Electric Corporation
8–1–1 Tsukaguchi-Honmachi, Amagasaki, Hyogo, JAPAN
E-mail: {simayosi, hata}@img.sdl.melco.co.jp

Shinji Shimojo
Cybermedia Center
Osaka University
5–1 Mihogaoka, Ibaraki, Osaka, JAPAN
E-mail: shimojo@cmc.osaka-u.ac.jp

Manzoor Hashmani
NS Solutions Corporation
7–20–1 Fukushima, Osaka, JAPAN
E-mail: hashmani.manzoor@osk.ns-sol.co.jp

Kazutoshi Fujikawa
Media Center
Osaka City University
3–3–138 Sugimoto, Sumiyoshi-ku, Osaka, JAPAN
E-mail: fujikawa@media.osaka-cu.ac.jp

Hideo Miyahara
Graduate School of Engineering Science
Osaka University
1–3 Machikaneyama, Toyonaka, Osaka, JAPAN
E-mail: miyahara@ics.es.osaka-u.ac.jp

KEYWORDS

DiffServ, Bandwidth Broker (BB), Admission Control, Resource Allocation, Contents.

ABSTRACT

A lot of architectures assuring QoS (quality of service) have been proposed. To apply these architectures to a network, rules regarding the allocation of resources is needed. These rules are called “policies.” Since each network has its own policy, when data go through several networks, policy conflicts may arise. Therefore, it is very hard to assure the contents level QoS. Furthermore, a human operator has to input how resources are allocated to data stream. In this paper, to solve these problems, we propose a new management framework and new policy definitions. Our proposed framework includes the new mechanisms that can notify how to classify data stream automatically by using only policies of network devices. Moreover, by using our policy definitions the new mechanisms can solve conflicts among policies.

INTRODUCTION

A number of users and companies are able to access high-speed networks, because of rapid progress of the Internet technology. On high-speed networks a lot of services with computer networks provide for example, Video on Demand (VoD) and IP telephony (VoIP) are available on the Internet. However, a network has to consider many requirements of applications. One of such requirements is to assure quality of service (QoS). Although there are several definitions of QoS assurance, in this paper we define QoS as to classify a data stream and to treat a data stream according to its class.

Some architectures such as IntServ (Shenker et al. 1997), DiffServ (Blake et al. 1998; Nichols et al. 2001), etc. have been proposed to assure QoS. In addition to these architectures, Bandwidth Broker (BB) has also been proposed. A BB manages and allocates resources such as a bandwidth within a network domain (Shirahase et al. 1999; Tajima et al. 2000; Neilson et al. 1999).

These architectures provide functions that ensure bandwidth or classify data streams. However, QoS of user applications is not assured only by these architectures. To assure QoS of user applications, it is necessary to notify these architectures how to classify data stream and to apply classifications to network devices. So far, to notify how to classify data stream a human operator had to input or notify the information when request came. In this paper, we propose a new mechanism is used to notify how to classify data stream automatically by using only policies of network devices.

To apply our QoS framework to available networks, the following issues must be considered. Firstly, in the case that a source host and a destination host exist in different network domains, BBs must communicate with each other. Secondly, each network domain has its own policies for resource allocation and resource control in the domain. Since each policy is defined independently by each domain, conflicts among policies may arise. We propose new policy definitions and a mechanism that can solve conflicts among policies by using the policy definitions.

Our proposed framework includes the new mechanisms that can notify how to classify data stream automatically by using only policies of network devices. Moreover, by using our policy definitions the new mechanisms can solve conflicts among policies. Our framework enables an automatic resource allocation.

QoS ARCHITECTURE

To assure end-to-end QoS, we assume that DiffServ architecture and a resource management mechanism are available on the network. The main functions of the resource management mechanism are “Policy system”, “Admission control” and “Policy negotiation” (e.g. Bandwidth Broker). Since most proposed architectures are able to manage only bandwidth, we manage bandwidth as a network resource.

In the next subsection, we discuss related works concerned with a DiffServ architecture, a policy system, an admission control and a BB.

Differentiated Services (DiffServ)

DiffServ architecture provides priority control for data packets on the IP network. The concept of the DiffServ architecture is shown in Figure 1.

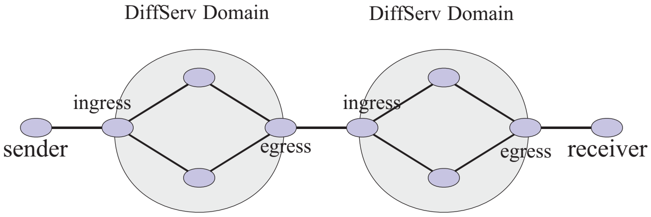


Figure 1: DiffServ components

DiffServ Domain (DS Domain) is a set of nodes that have DiffServ functions. Nodes in DS Domain are classified into boundary nodes and inside nodes. Boundary nodes connect own domain with other DS Domains or non DS Domains that don’t have DiffServ functions. Inside nodes are connected with nodes in the same domain. In DS Domain, boundary nodes that accept packets from other domains (DS Domain or non DS Domain) are called ingress nodes, and ones that send packet to other domains are called egress nodes. Data packets which are classified in the same class are aggregated at boundary nodes. Since the classification is carried out at only boundary nodes, DiffServ architecture is considered to be scalable for a number of data stream.

In DS domain, nodes treat data packets according to only those classes. Per Hop Behaviors (PHBs) (Brim et al. 2000) are descriptions of packet treatment per hop. Assured Forwarding PHB (Heinanen et al. 1999), Expedited Forwarding PHB (Jacobson et al. 1999), etc. have already been proposed as one of PHBs. AF PHBs provide network environments have various qualities with classification and priority of classes. We can provide the airline-model services (services are first, business, and economy) with AF PHBs. EF PHB provides a high quality network environment, in which the bandwidth for aggregated data packets is assured. We can provide the virtual leased line service (available bandwidth is constant) with EF PHB.

Policy System and Admission Control

The policy system is an application that manages a network with policies (Moore, Ellesson et al. 2001; Snir et al. 2001; Moore, Rafalow et al. 2001; Westerinen et al. 2001). A policy is a set (or sets) of conditions and actions. In the policy it is specified when and how decisions are made. Some policies have a hierarchal architecture, i.e. policies contain other policy(s). It is thus possible to build complex policies by composition of simple policies.

Admission control is a main function of a resource management system. Admission control accepts or rejects requests for resource allocation according to policies (Yavatkar et al. 2000). To accept requests from senders, a management system has to recognize the network state regarding available resources. Therefore, a measurement of data packets on the network and a judgment for resource reservation are necessary.

Bandwidth Broker (BB)

Bandwidth Broker (BB) architecture is proposed as an architecture supporting policy negotiation. One of BB’s functions is a negotiation about bandwidth allocation with other BBs. A BB architecture is shown in Figure 2.

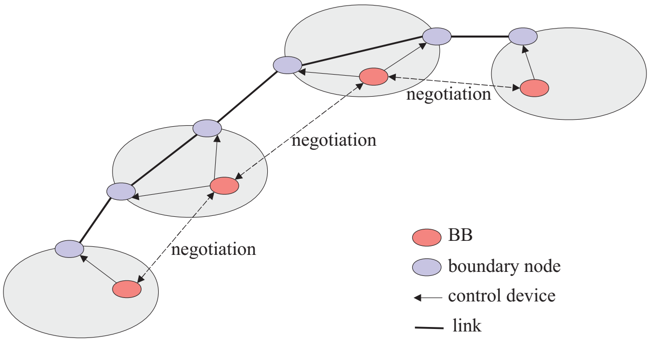


Figure 2: BB architecture

If a BB receives request message for data transfer, the BB decides whether bandwidth, which are on the path from the source host to the destination host, are available or not. That is, a BB asks other BBs whether bandwidth is available. If bandwidth is available in each domain on the path, BBs reserve the bandwidth and control network devices to reflect the reservation.

MANAGEMENT FRAMEWORK

In this paper, we will call a software that sends data from a host to another host on the network the "network application", e.g. Video on Demand. Some network applications require QoS guarantee. For example, there are real-time movies, data transfer applications and so on. Real-time movies need much bandwidth, short delay and small jitter, but it does not require all data packets to be received. On the other hand, data transfer applications need all data packets to be received, but does not need short delay or small jitter. Thus, several QoS requirements vary from a network application to a network application.

In most network applications that require QoS, the data pass through more than two networks. In order to assure QoS we assume that all networks are managed by BBs. We call each network which is managed by a single BB a "domain." In the case where a sender and a receiver are in the same domain, bandwidth reservation is easy, because a single BB can manage all resources (bandwidth) on the path. On the other hand, in the case where a sender and a receiver do not belong to the same domain, each BB on the path not only manage its own resources , but also has to cooperate with each other (Nicolouzou et al. 2001), because, generally, each domain has its own policies and policies may conflict with each other. Due to policy conflicts, it is therefore necessary to adjust their own policies and adopt a new mechanism by which BBs can cooperate with each other.

We propose "contents selection function" as an additional function of a BB so that a BB can select contents to be allocated bandwidth when policy conflicts arise. Also, we propose a new framework including the contents selection function. Figure 3 shows our proposed framework.

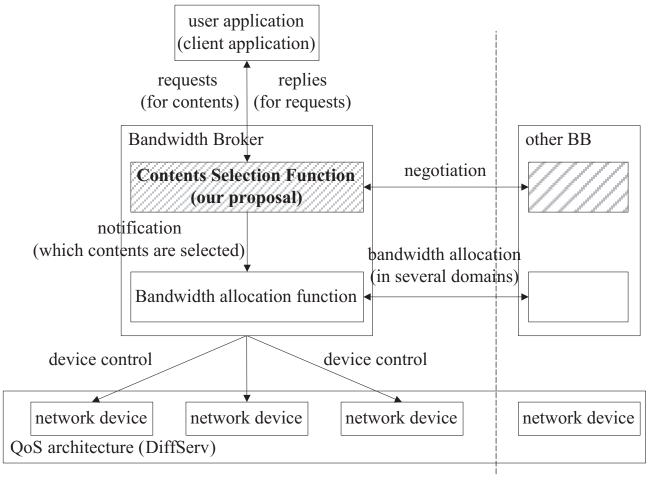


Figure 3: our proposed framework

When contents' requests that users want to transfer data packets on a network are sent, a BB receives those requests. If the BB decides to allocate bandwidth to some contents, the BB attempts to reserve bandwidth for the requests. In the case where a sender and a receiver do not belong to the same domain, BBs in neighboring network domains on the path of the contents communicate with each other. Since each BB controls network devices in its own domain, as a result, bandwidth is reserved on the path from a sender to a receiver. If bandwidth is reserved, the BB replies that the requests for contents are accepted. A user application sends data stream after it receives the reply. In this way, bandwidth is allocated to contents automatically.

To assure QoS of contents, domains decide how they treat data of contents according to their own policies. On the other hand, contents have policies that show how they wish to be treated. We call a policy for domain "domain-policy" and a policy of contents "contents-policy." A domain-policy relates to other domains, and hosts in the domain. A contents-policy relates to network applications, bandwidth, and so on. Now, we define "priority value", to decide which contents are allocated bandwidth. When there are many contents that the domain cannot assign all contents enough resources, the domain assigns resources to the contents that have high priority values.

POLICY DEFINITIONS

In this section, policy definitions and contents selection system for bandwidth allocation are described.

Domain-Policy

Parameters of a domain-policy are as follows:

- priority value for domain
- priority value for host
- priority value for service type
- DiffServ class bandwidth (under DiffServ environment, the bandwidth is assigned for each DiffServ class)

A priority value of domain shows how important each neighboring domain is. A priority value of host shows how important each host, each user or each device is. For example, a video camera that is watching an important place or a computer that is used for an important task has high priority value. A priority value of service type shows how important each service is in the domain. DiffServ class bandwidth shows the amount of bandwidth that the domain can use for the class. These values are determined individually in each domain and their size have only relative meaning. The domain-policy examples are shown in Table 1. Figure 4 shows network topology in this example.

Contents-Policy

Contents-policies are generated when users send requests for contents. Contents-policies include information of a source host, a destination host, and policy parameters requested by

Table 1: examples of domain-policy

Domain name	**A**	**B**	**C**	**D**
for domain	C:100	A:2	A:50	C:1000
	B:50	C:10	B:30	
			D:100	
for host	A1:100	B1:10	C1:100	D1:1000
	A2:70	B2:5	C2:80	D2:900
	A3:60	B3:5	C3:80	D3:800
	A4:90			
for service type	monitor:70	monitor:8	monitor:90	monitor:800
	machine:100	machine:10	machine:80	machine:800
	archive:50	archive:5	archive:50	archive:500
	voice:90	voice:9	voice:90	voice:900
DS Class & bandwidth	EF:40MB	EF:50MB	EF:70MB	EF:90
	AF:30MB	AF:20MB	AF60MB	

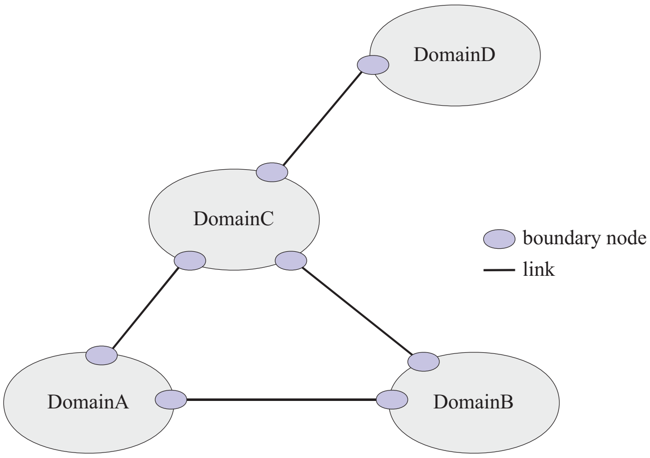


Figure 4: network topology (in this example)

the clients. A contents-policy consists of the following two components:

- application-policy
- QoS description

Application-Policy
Parameters of an application-policy are as follows:

- service type (ex. machine control)
- information about the source host
- information about the destination host

Service type in a application-policy and the service type in a domain-policy are used together to calculate priority values of contents. Information about both hosts also is used to calculate priority values of contents.

QoS Description
Parameters of QoS description are as follows:

- media type (ex. MPEG2)
- bandwidth (fixed, discrete, variable)

QoS description includes media type in order to take into consideration characteristics of data format. Bandwidth represents how much bandwidth to the contents require and also how the bandwidth changes. Contents-policy examples are shown in Table 2.

Table 2: examples of contents-policy

	Contents1	**Contents2**	**Contents3**
source host	A1	A2	A3
destination host	D2	D2	D2
service type	machine	machine	machine
bandwidth	20MB	20MB	20MB

	Contents4	**Contents5**	**Contents6**
source host	B2	B3	D1
destination host	C2	C3	A4
service type	monitor	monitor	monitor
bandwidth	10MB	10MB	10MB

Contents Selection System

In the Contents Selection System there are three functions: (1) Calculation of a priority value (2) decision of contents to be allocated bandwidth (3) sending request that some certain contents are allocated bandwidth to other BBs.

Priority Value Calculation
Priority values of contents are calculated by binding domain-policy with contents-policy.

Priority values of contents P are defined by the following formula.

$$P = \begin{cases} \frac{1}{3}(PD + PH + PT) \\ \frac{1}{2}(PD + PT)\,(when\,PH\,is\,none) \end{cases}$$

In this formula, PD, PH, and PT are priority values of a domain, a host, and a service type respectively. The sum of these value (that is P) is used to decide which contents are allocated resources.

PD is calculated according to the domains that contents go to and come from. If the source host is in the domain in which the priority value calculation takes place, the source domain is not taken into consideration. Likewise, if the destination host is in the domain making the calculation, it is not taken into consideration. PD is defined by the following formula.

$$PD = \begin{cases} \frac{1}{2}(PDsrc + PDdst) \\ PDdst\ (src\,host\,is\,in\,domain) \\ PDsrc\ (dst\,host\,is\,in\,domain) \end{cases}$$

$PDsrc$ is the priority of the domain from which the contents come, $PDdst$ is the priority of the domain to which the contents go.

PH is likewise defined by the following formula.

$$PH = \begin{cases} \frac{1}{2}\left(PHsrc + PHdst\right) \\ \quad (both\,host\,are\,in\,domain) \\ PHdst \;\; (src\,host\,is\,in\,domain) \\ PHsrc \;\; (dst\,host\,is\,in\,domain) \end{cases}$$

$PHsrc$ and $PHdst$ are the priority of the source and destination hosts.

Contents selection
To select which contents are allocated resources, all the contents' priority values are compared. Then, the contents having the highest priority is selected. The contents selection function then judges whether unused resources can be allocated to the contents. If possible, resources are allocated to the contents. Otherwise, the contents with the second highest priority value are selected and the same operations are performed. When several contents with the same priority value, the contents have the earliest timestamp are selected. Moreover, according to the result of the priority judgment, it is also able to release resources that have already been allocated.

Bandwidth Reservation Request
On our proposed system, domains send and receive requests from other domains. Moreover, our proposed system plays the role of an interface between connected domains. As BB sends and receives requests, there are resource allocation requests and resource release requests. Each request is delivered between connected domains and finally the requests are transferred from the domain in which a source host exists to the domain in which a destination host does.

Contents selection examples are shown in Table 3 by using domain-policies of Table 1 and contents-policies of Table 2.

SIMULATIONS

This section describes the results of simulations using the policy and bandwidth allocation methods described. In this simulation, we assumed a network topology with a hierarchical structure as shown in Figure 5.

There are four contents types, live-movies (6Mbps), movie-archives (6Mbps), 64kbps data transfers (64kbps) and 128kbps data transfers (128kbps). The priority value of 64kbps data transfer is the highest among these types, and the priority value of live-movies is higher than the priority value of movie-archives.

In the simulations, requests for contents are sent about once a second. In these simulations, we have pre-defined 108 data transfers. A sender, A receiver, and a service type is determined for each data transfer. One of data transfers will be selected at random. About half of the data transfers go through between Domain M and Domain Y. Therefore, Link M-Y is a bottleneck link.

Table 3: examples of contents selection

DomainA	**1**	**2**	**3**
PD	100	100	100
PH	100	70	60
PT	100	100	100
P	100	90	86.67
Class:BW	EF:20M	EF:20M	AF:20M

DomainB	**4**	**5**	**6**
PD	10	10	10
PH	10	5	5
PT	8	8	8
P	9.33	7.67	7.67
Class:BW	EF:10M	EF:10M	EF:10M

DomainC	**1**	**2**	**3**	**4**	**5**	**6**
PD	75	75	75	30	30	30
PH				80	80	80
PT	100	100	100	80	80	80
P	87.5	87.5	87.5	63.3	63.3	63.3
Class:BW	EF:20M	EF:20M	AF:20M	EF:10M	EF:10M	EF:10M

DomainD	**1**	**2**	**3**
PD	1000	1000	1000
PH	900	900	900
PT	800	800	800
P	900	900	900
Class:BW	EF:20M	EF:20M	AF:20M

Simulations are performed considering the "static reservation." "Static reservation" means that bandwidth allocation requests for contents have "static reservation" is always accepted. That is, bandwidth for contents that have "static reservation" is assured.

First, Figure 6 and Figure 7 show the relationship between priority and request rejection rate. In this figures, priority values are normalized so that the maximum value becomes 100. The higher the priority becomes, the lower the rejection rate. Only priority:85-90 in Figure 6 has low rejection rate. The result comes from that data transfers of this priority value require small amount of bandwidth.

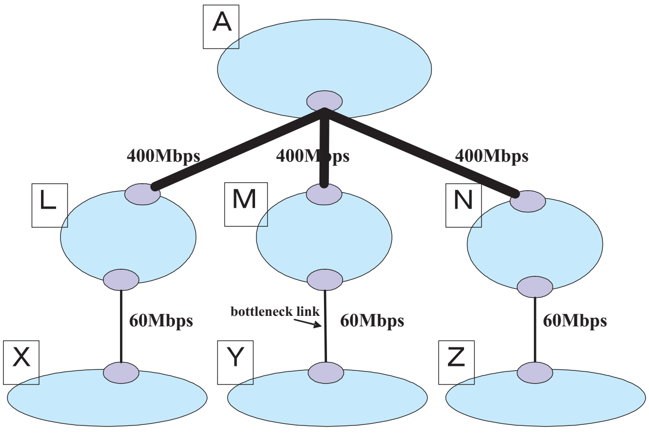


Figure 5: network topology

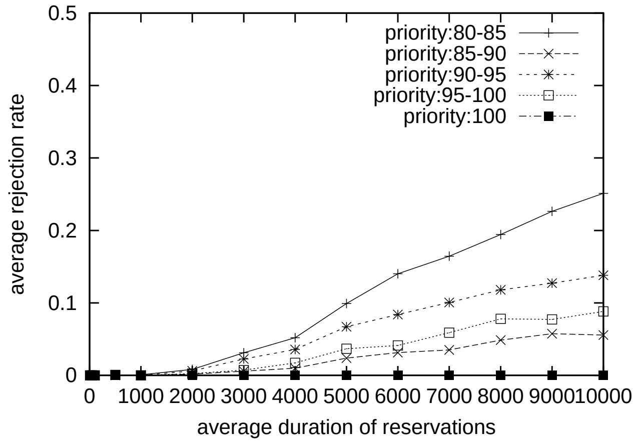


Figure 6: the rejection rate based on priority ("static-reservation" is not used)

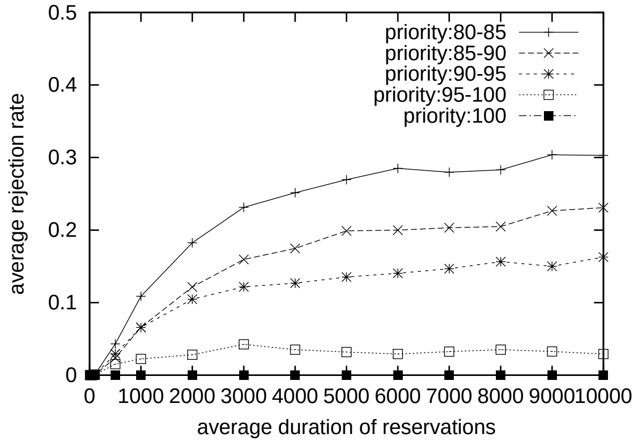


Figure 7: the rejection rate based on priority ("static-reservation" is used)

Next, Figure 8 shows the relationship between duration of reservation and request rejection rate. In this figure, the priority isn't considered. This figure shows that the longer the duration of reservation is, the higher the request rejection rate.

Figure 9 and Figure 10 show the relationship between duration of reservation and bandwidth utilization. When "static reservation" is permitted, statically reserved bandwidth may not be requested and the bandwidth utilization drops.

In Figure 8, when average duration of reservation is less than about 4000 seconds, in the case that the rejection rate 0.05 is acceptable, it is better not to use "static reservation." In this way, we can decide how we should design our network (e.g. to use "static reservation" or not) according to the network environment or other requirements for the network.

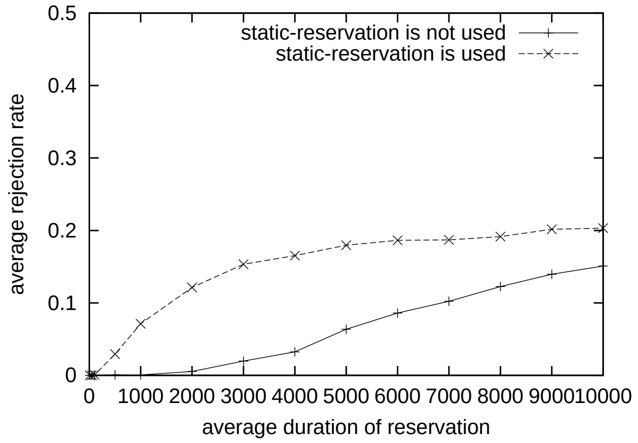


Figure 8: the rejection rate

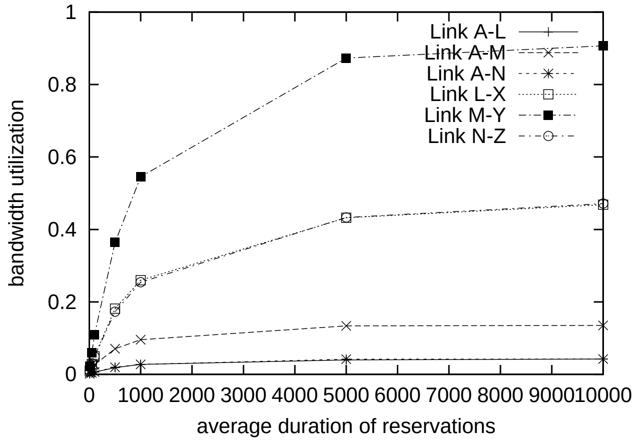


Figure 9: the bandwidth utilization ("static-reservation" is not used)

SUMMARY AND CONCLUSION

We proposed a resource allocation framework and policy definitions that can solve the difficult problem of selecting the contents that are allocated resources. In this framework, a priority based resource allocation function is added to BB functions. We performed some experimental simulations of resource allocation and showed the results with sample policies. We found that our proposed framework can be both advantageous and disadvantageous depending on the network design.

We focus on the method to select contents is based on the priority. We will continue this research by considering other mechanisms such as optimization of bandwidth utilization. Furthermore, we will analyze the relationship between contents characteristics (ex. duration) and the resulting allocations.

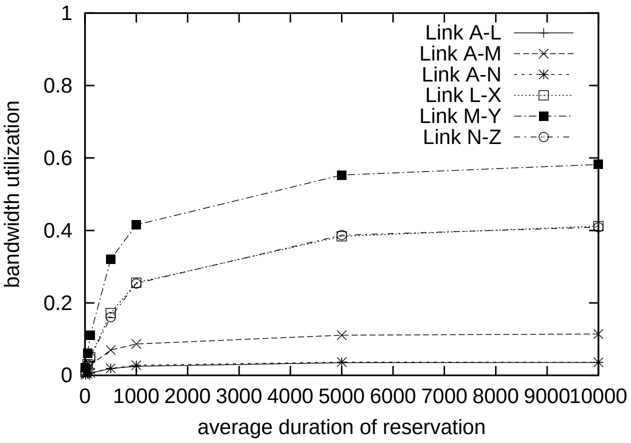


Figure 10: the bandwidth utilization ("static-reservation" is used)

We are currently implementing a prototype of the bandwidth allocation system. Using this prototype, we will verify the framework, the system, and the policies proposed in this paper.

ACKNOWLEDGEMENT

This work was partly supported by Special Coordination Funds for promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology of Japan.
This work was supported in part by Research for the Future Program of Japan Society for the Promotion of Science under the Project "Integrated Network Architecture for Advanced Multimedia Application Systems" (JSPSRFTF97R16301) and "Advanced Multimedia Contents Processing" (JSPS-RFTF97P00501).

REFERENCES

Blake, S.; D. Black; M. Carlson; E. Davies; Z. Wang; and W. Weiss. 1998. "An Architecture for Differentiated Services," *Request for Comments* 2475.

Brim, S.; B. Carpenter; and F. L. Faucheur. 2000. "Per Hop Behavior Identification Codes," *Request for Comments* 2836.

Heinanen, J.; F. Baker; W. Weiss; and J. Wroclawski. 1999. "Assured Forwarding PHB Group," *Request for Comments* 2597.

Jacobson, V.; K. Nichols; and K. Poduri. 1999. "An Expedited Forwarding PHB," *Request for Comments* 2598.

Moore, B.; E. Ellesson; J. Strassner; and A. Westerinen. 2001. "Policy Core Information Model – Version 1 Specification," *Request for Comments* 3060.

Moore, B.; L. Rafalow; Y. Ramberg; Y. Snir; J. Strassner; A. Westerinen; R. Chadha; M. Brunner; and R. Cohen. 2001. "Policy core information model extensions," *IETF Internet draft* <http://www.ietf.org/internet-drafts/draft-ietf-policy-pcim-ext-01.txt > *,work in progress*.

Neilson, R.; J. Wheeler; F. Reichmeyer; and S. Hares. 1999. "A Discussion of Bandwidth Broker Requirements for Internet2 Qbone Deployment Version0.6," Available at http://www.merit.edu/working.groups/i2-qbone-bb/doc/BB_Req6.doc.

Nicholouzou, E.; and G. Politis; and P. Sampatakos; and L. S. Venieris. 2001. "An Adaptive Algorithm for Resource Management in a Differentiated Services Network," *In Proceedings of the IEEE International Conference on Communications* (Helsinki, Finland, Jun.11-16).

Nichols, K.; and B. Carpenter. 2001. "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification," *Request for Comments* 3086.

Shenker, S.; and J. Wroclawski. 1997. "General Characterization Parameters for Integrated Service Network Elements," *Request for Comments* 2215.

Shirahase, A.; M. Hashmani; M. Yoshida; and S. Shimojo. 1999. "Design and deployment of qos enabled network for contents businesses," *International Conference on Computer Communication 1999*.

Snir, Y.; Y. Ramberg; J. Strassner; and R. Cohen. 2001. "Policy qos information model," *IETF Internet draft* <http://www.ietf.org/internet-drafts/draft-ietf-policy-qos-info-model-03.txt> *,work in progress*.

Tajima, K.; M. Hashmani; and M. Yoshida. 2000. "A resource management architecture over differentiated services domains for guarantee of bandwidth, delay and jitter," *EuroComm 2000*.

Westerinen, A.; J. Schnizlein; et al. 2001. "Terminology for Policy-Based Management," *Request for Commnets* 3198.

Yavatkar, R.; D. Pendarakis; and R. Guerin. 2000. "A Framework for Policy-based Admission Control," *Request for Comments* 2753.

SYSTEM AND KNOWLEDGE MANAGEMENT

A knowledge management instrument to support learning of highly interactive systems development

Charles van der Mast
Dept. of mediamatics
Delft University of Technology
Delft, The Netherlands
c.a.p.g.vandermast@its.tudelft.nl

Jan-Willem van Aalst
Atos Origin, eBusiness solutions
P.O. Box 8348
Utrecht, The Netherlands 3503 RH
jan-willem.vanaalst@atosorigin.com

ABSTRACT

The main interest of universities can be seen as maximizing the quality of the transfer of knowledge from teacher to learner. Web-based environments increasingly facilitate the possibilities to do this. In this paper we present Performer, an interactive web-based instrument for teachers and students that supports knowledge management, in order to maximize the quality of the transfer of knowledge from teacher to student, and from student to student. We report on experiences with the use of Performer by various groups of university students in producing a multimedia CD-ROM for a real customer. The conclusions of the research are that students may benefit (1) by structuring their learning process using the objectives provided by Performer; (2) by using the knowledge elements (examples and hints on content) stored in Performer; and (3) by perceiving that sharing experience and knowledge within and between student teams may enhance personal learning and collaboration. We have thus shown that the transfer of knowledge about realizing highly interactive systems using a web-enabled knowledge management instrument can be significantly enriched.

Keywords

Design of highly interactive systems; multimedia; teacher-learner relationship; knowledge management

INTRODUCTION

Knowledge intensive organizations and institutes (e.g., universities, research centers) are constantly looking for ways to improve the quality of their educational material. This is because the student of today may be the teacher of tomorrow, and students should be facilitated in their learning process as well as possible. In other words, an important part of the main interest of universities and research centers can be seen as maximizing the quality of the transfer of knowledge from teacher to learner. Traditional instructional methods and didactic instruments have become insufficient to support this due to the ever-faster pace of (technological) developments and the increased specialist level that is required these days [7].

In recent years, two important new instruments have evolved that can significantly aid teachers in doing just that, namely (a) the Internet and (b) evolving theory of knowledge management. However, in the field of highly interactive systems (in which human-computer interaction plays a key role), the research focus has up to now mostly been targeted towards maximizing the quality of the system itself (the final product, or deliverable) for the end-user, and not so much on improving the quality of transfer of knowledge about the design of highly interactive systems from teacher to learner. This is chiefly because the research field of Human-Computer Interaction has experienced such an enormous pull from dissatisfied computer users who demand more attention to human factors in design. As a consequence, the HCI research field has focused more on the end user than on transferring HCI knowledge itself. Only very recently has interest in this research area emerged (see for example [4]). Such interest focuses mainly on offering tools to teachers that can aid them in the transfer of their knowledge to students.

In this paper, we present an instrument based on knowledge management theory (which we briefly discuss later) that facilitates the teacher in improving the knowledge transfer about designing highly interactive systems for students in that area. To this end, we start by providing an overview of the design process of highly interactive systems that we teach in a course on multimedia CD-ROM production. We then investigate the issues that the teacher faces when thinking about improving the quality of the transfer of knowledge to students; we then present the actual instrument that we have devised to aid the teacher in doing just that. We do this not only from a technical point of view, but we also take into account much less technical matters such as the gathering of relevant content (knowledge) and the assessment of the quality of content; the issues about getting students enthusiastic and keeping them motivated to use such an instrument; and, finally, we discuss issues about keeping the instrument itself organized and “alive”.

We also present our experiences in the first year of having used this instrument and the applicability of this instrument to related research fields. On a more abstract level, we also briefly discuss how the use of our instrument relates to the overall level of professionalism of the instructional (teacher-learner) process (See [2]). Finally, we also report on comparisons of the effectiveness of our instrument in a commercial company environment versus a university environment.

The development process

At Delft University of Tehnology, a course on multimedia CD-ROM production is presented to senior students from all programs of the university, e.g. computer science, architecture, engineering studies, management and business, etc., see [8]. This course is project-based, with interdisciplinary teams of six students each from different programs (disciplines) who are to produce a CD-ROM for external commissioners from local institutes and companies such as art centers, museums, theatres, small companies and businesses.

The content of the CD-ROMs may be for instructional, for public relations, or for marketing purposes. The main objectives of the course are (1) to learn to design and realize an attractive (persuasive) multimedia CD-ROM using professional methods and tools (Director, Premiere, Photoshop, 3D Studio, etc), and (2) to learn to communicate and collaborate effectively with colleagues from other disciplines.

The student teams work according to a project framework of multimedia roles (director, administrator, graphic designer, AV-specialist, text writer, programmer, etc.), phases (analysis, design, realization) and required deliverables (documents, audiovisual assets, storyboard, CD-ROM). About ten teams are working simultaneously during a four-month period, each student spending about two hundred working hours.

The student groups use a digital multimedia studio and servers to store their work. They can also access their data over the 100 Megabit campus network to do work from their own department or from their apartments on campus. They are faced with mastering many different kinds of skills: writing quality assurance plans, scripting and storyboarding, photo and movie shooting, programming, etc. Also, they must master available tools in a very short time. During the project, we give a series of lectures and workshops to provide the fundamental knowledge and skills they need.

The campus-wide "Blackboard" system [12] for distance learning contains a hundred-page paper "project book" to guide student groups. For the sharing of documents and other assets, the project teams use the BSCW system [11] for shared workspaces. At the end of the project, each student group gives a presentation to the teacher and the commissioner. This presentation covers both the multimedia product and the process that was followed.

Problem analysis

The multimedia course is now seven years old. According to the standard university course evaluation, teachers and students were overall satisfied about the course results over the years. However, from detailed evaluations organized by the teachers we observed that many students are hesitant to start full collaboration in the multidisciplinary teams they must work with for the first time. It takes several weeks to know and accept each other's different styles of thinking and working.

Every time the course is given, structuring the group process and sharing insights in methods requires a lot of effort from teachers and students alike. This is chiefly caused by the different cultural backgrounds of students from the different programs, but it is also caused by the different ways that students use to structure the development process of their own disciplines (e.g., to architecture students, "development" does not mean what it does to computer science students). Another problem we observed is that the required skills to use multimedia development tools are not easy to acquire in a few days. This is also caused by the rapid sequence of new releases of programming and editing environments and tools. The senior students (who finished the course earlier) who are assisting the teams do not have the time to be sufficiently skilled in using the newest releases of all tools: this typically takes months of intensively using these tools. For example, typical issues with teams are: how to implement metallic-style buttons with special effects; what kind of scripting is needed to add special effects to animations. Finding practical hints and tricks is important with the kind of developing process the students are performing.

Offering a knowledge management instrument to the student teams may partly solve these and many similar problems. Two goals are particularly important: first, to support the structuring of the development process followed by the students from various disciplines into a more professional level as soon as possible, and, second, to provide a means to find expertise on technical features of the tools they need for design and production and to share expertise with students from all groups. In both cases, the transfer of knowledge from teacher to student and from student to student was our aim.

A brief overview of knowledge management theory

According to a large majority of researchers on knowledge management (cf. [3], [10]), the two basic components of knowledge management are flow (tacit) and stock (explicit). Tacit knowledge is transferred informally in teacher-learner discussions, while explicit knowledge is usually transferred through books or other media. Weggeman holds that these are the two basic distinct but not separable components of knowledge management. Each of these requires a unique approach. Recently these two types have been theoretically connected to three levels of formality in the organizing of groups, or teams:

1. Communities of interest. These are groups of experts with a shared interest who meet regularly on an informal basis. This is typically about flow knowledge, because knowledge is disseminated through (informal) discussions.
2. Communities of purpose. This is more formal than the previous type of community. Here, we

see formal networks of practitioners take shape, who meet on a regular basis with a defined group of experts. This is a mix of flow knowledge management and stock knowledge management.

3. Communities of practice. On this level of formality, the work process is explicitly defined and the sharing of knowledge is organized around this defined process. Knowledge is made explicit wherever this is feasible and sensible, and categorized and disseminated for a wide range of experts involved in the same area of interest.

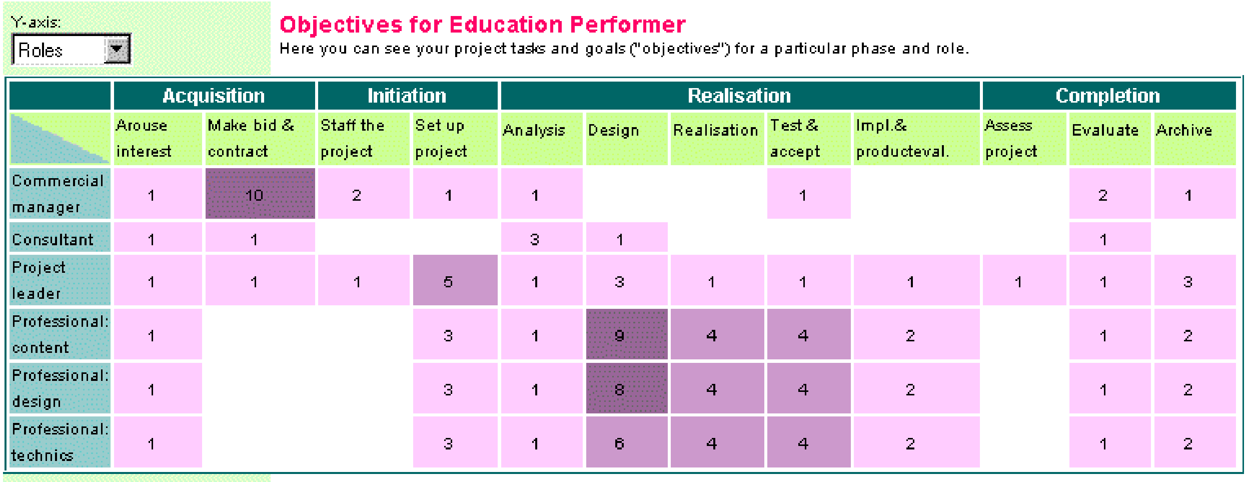
In the *knowledge value chain* [10], knowledge is created or identified, then gathered, categorized, disseminated, and finally archived or dismissed.

Both types of knowledge are equally important to the learning process, and knowledge management therefore should pay attention to both components. In this view we see college lectures as flow knowledge management. Until now, stock knowledge management was hardly available at the university. To improve the balance, we have developed an instrument called Performer.

The Performer knowledge management instrument

From a knowledge management perspective, Performer is an instrument to facilitate stock knowledge management. This means that Performer is targeted towards the categorization and dissemination of formal knowledge that can be made explicit relatively easy. This knowledge can be made explicit using media such as presentations, text documents, spreadsheets, but also references to websites or existing paper literature. Our way of positioning Performer excludes flow knowledge management, in which knowledge is mainly transferred through informal discussions. Performer is typically useful for communities of practice, which is what our students are (for the four-month duration of the course).

From a maturity point of view, Performer is an instrument that aids teams and organizations in achieving a more professional way of working. Pertaining to the Capability Maturity Model, a widely used framework for measuring the maturity of primary organizational processes, Performer helps to achieve CMM level 3 (for a comprehensive overview of CMM, see [2]). Performer does this by providing a visual representation of the way of working of project teams in terms of project phases (or main process steps), project roles, and objectives to be achieved by each role in each phase. Performer makes this visible by presenting a matrix that shows the way of working of a project team in a single view. We show this in figure 1. This English version of Performer is made for use by professional teams. We derived a simplified version for use by students in the multimedia course, see figure 5.



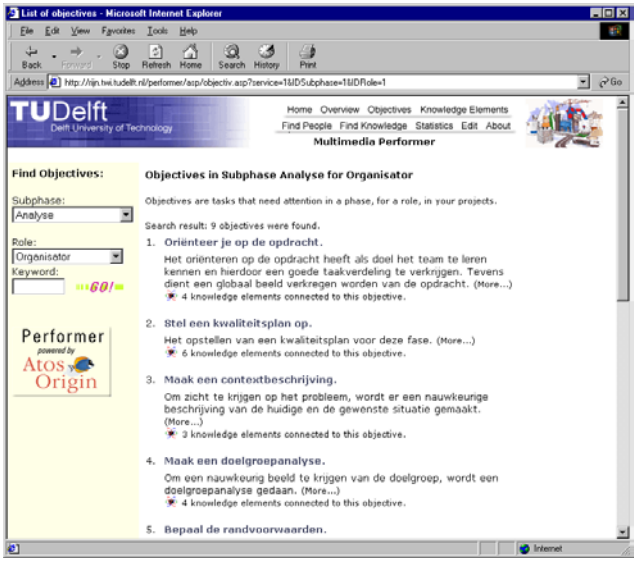
Objectives for Education Performer

Here you can see your project tasks and goals ("objectives") for a particular phase and role.

	Acquisition		Initiation		Realisation					Completion		
	Arouse interest	Make bid & contract	Staff the project	Set up project	Analysis	Design	Realisation	Test & accept	Impl.& producteval.	Assess project	Evaluate	Archive
Commercial manager	1	10	2	1	1			1			2	1
Consultant	1	1			3	1					1	
Project leader	1	1	1	5	1	3	1	1	1	1	1	3
Professional: content	1			3	1	9	4	4	2		1	2
Professional: design	1			3	1	8	4	4	2		1	2
Professional: technics	1			3	1	6	4	4	2		1	2

Figure 1. The Performer objectives matrix, which shows the relative denseness of objectives for each role in all sub-phases of a project or process.

In figure 1, we show the way of working for project teams that are involved in producing educational software. We therefore call this Performer view the "Education Performer" (one may envision Performers for other knowledge domains in a similar way). In the Education Performer, we define six main roles that together produce Educational Software in twelve project phases. Each role has certain (sub) goals to achieve in each project phase. We call these (sub) goals *objectives*. These are shown in the cells. The number in each cell corresponds directly with the deepness of the color and represents the number of objectives to achieve. When clicking one of the cells, we get a list of objectives for a particular role in a particular phase. We show this in figure 2. Each objective contains a brief summary. When clicking on the title of an objective, we see an *objectives details* screen, where activities, preconditions, deliverables, and other attributes of objectives are shown. In this way, we have unambiguously defined our way of producing educational software. The objectives matrix provides a quick and easy-to-understand view of this development process, for example for customers or new employees. For the students of the course this view teaches them implicitly the way of working they have to follow during the project.


List of objectives - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search History Print
Address
Go
TUDelft
Delft University of Technology
Home Overview Objectives Knowledge Elements
Find People Find Knowledge Statistics Edit About
Performer
powered by
Atos Origin
GO!
Internet

Address: http://rijn.twi.tudelft.nl/performer/asp/objectiv.asp?service=1&IDSubphase=1&IDRole=1

Multimedia Performer

Find Objectives:

Subphase: Analyse

Role: Organisator

Keyword:

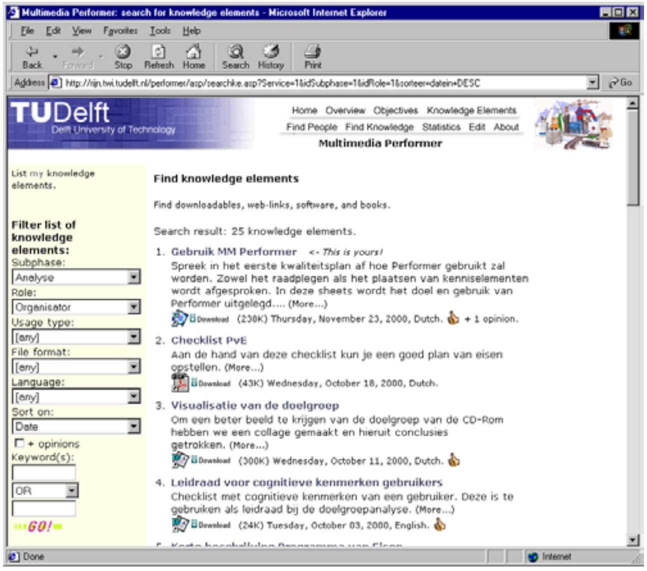
Objectives in Subphase Analyse for Organisator

Objectives are tasks that need attention in a phase, for a role, in your projects.

Search result: 9 objectives were found.

1. Oriënteer je op de opdracht.
   Het oriënteren op de opdracht heeft als doel het team te leren kennen en hierdoor een goede taakverdeling te verkrijgen. Tevens dient een globaal beeld verkregen worden van de opdracht. (More...)
   4 knowledge elements connected to this objective.
2. Stel een kwaliteitsplan op.
   Het opstellen van een kwaliteitsplan voor deze fase. (More...)
   6 knowledge elements connected to this objective.
3. Maak een contextbeschrijving.
   Om zicht te krijgen op het probleem, wordt er een nauwkeurige beschrijving van de huidige en de gewenste situatie gemaakt. (More...)
   3 knowledge elements connected to this objective.
4. Maak een doelgroepanalyse.
   Om een nauwkeurig beeld te krijgen van de doelgroep, wordt een doelgroepanalyse gedaan. (More...)
   4 knowledge elements connected to this objective.
5. Bepaal de randvoorwaarden.

Figure 2. A listing of objectives, obtained by clicking on a cell in the Performer objectives matrix. Each objective is summarized in one sentence. The number of available knowledge elements about this objective is mentioned. Clicking on the title of an objective brings up the description, preconditions, activities, and roles involved. Clicking on the number of knowledge elements gives direct access to them.


Multimedia Performer: search for knowledge elements - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search History Print
Address
Go
TUDelft
Delft University of Technology
Home Overview Objectives Knowledge Elements
Find People Find Knowledge Statistics Edit About
Download
GO!
Done
Internet

Address: http://rijn.twi.tudelft.nl/performer/asp/searchke.asp?Service=1&idSubphase=1&idRole=1&sorteer=datein+DESC

Multimedia Performer

List my knowledge elements.

Filter list of knowledge elements:

Subphase: Analyse

Role: Organisator

Usage type: [any]

File format: [any]

Language: [any]

Sort on: Date

□ + opinions

Keyword(s):

OR

Find knowledge elements

Find downloadables, web-links, software, and books.

Search result: 25 knowledge elements.

1. Gebruik MM Performer <- *This is yours!*
   Spreek in het eerste kwaliteitsplan af hoe Performer gebruikt zal worden. Zowel het raadplegen als het plaatsen van kenniselementen wordt afgesproken. In deze sheets wordt het doel en gebruik van Performer uitgelegd.... (More...)
   Download (238K) Thursday, November 23, 2000, Dutch. + 1 opinion.
2. Checklist PvE
   Aan de hand van deze checklist kun je een goed plan van eisen opstellen. (More...)
   Download (43K) Wednesday, October 18, 2000, Dutch.
3. Visualisatie van de doelgroep
   Om een beter beeld te krijgen van de doelgroep van de CD-Rom hebben we een collage gemaakt en hieruit conclusies getrokken. (More...)
   Download (300K) Wednesday, October 11, 2000, Dutch.
4. Leidraad voor cognitieve kenmerken gebruikers
   Checklist met cognitieve kenmerken van een gebruiker. Deze is te gebruiken als leidraad bij de doelgroepanalyse. (More...)
   Download (24K) Tuesday, October 03, 2000, English.

Figure 3. Knowledge element search filtering mechanism. Various filters may be set to zoom in on a desired set of knowledge elements. The knowledge elements are specified in a standard text format including description, available since, submitted by, approved, related objectives, roles, sub-phases, access hits. Download by clicking the icon.

Perhaps the most important entity connected to these objectives are the knowledge elements that we defined earlier. Since a piece of knowledge that is relevant to developing educational software always pertains to one or more roles in a certain project phase, we can locate the spot in the matrix where each knowledge element belongs. Within each cell, we can even define to which objective(s) a particular knowledge element should be connected. In this way we cannot only merely collect all knowledge elements that are relevant to developing educational systems, but we can also categorize them, and thus make them easily accessible to everyone involved in such projects. Performer further facilitates the quick finding of knowledge through a filtering mechanism. We show this in figure 3. Using the filters shown on the left hand side of the screen, project team members can define the project phase and project role for which they want to find knowledge elements (thereby effectively choosing one cell in the matrix). Additionally, they can define the usage type of the knowledge element (is it a template? A best practice? Literature?), the file format (for example, Microsoft Word), and the language of the knowledge element.) Finally, a general keyword mechanism (A and B, A or B, A and not B) completes the filtering mechanism. This potentially greatly decreases the time needed to find knowledge.

Additionally, Performer features some statistics tools, for example a listing of who has contributed the largest amount of valuable knowledge elements to Performer. Other statistics include a top 20 of most popular knowledge elements, and who has contributed last month. Every employee involved in a certain type of project (educational projects in this case) is encouraged to contribute to the knowledge collection in Performer. This requires some form of quality control; otherwise, the value of the available knowledge will quickly diminish. To this end, each Performer has its own content owner. This is typically a senior employee who regularly monitors newly added knowledge, and who has the authority to accept or reject knowledge elements.

Implementing Performer: setting it up and keeping it alive

The design and realization of the knowledge management instrument itself is not sufficient to make it work. For example, we have seen that a content owner is required to monitor the overall quality of the body of knowledge in Performer. In a similar way, we might argue that we also require a *Performer champion* whose task it is to constantly motivate employees and keep them enthusiastic about knowledge management. This requires certain pre-conditions, such as an available critical mass of knowledge elements when the application is first introduced. This is because employees will quickly abandon the use of such an application if they cannot find what they want the first time they try it. In other words, there is a wide range of non-technical activities that must be carefully organized in order to make the technical application work. This in itself is not a new insight, but the way in which these various aspects should be modeled is not generally accepted. A simple visual representation of the most important aspects is given in figure 4 (see [6]).

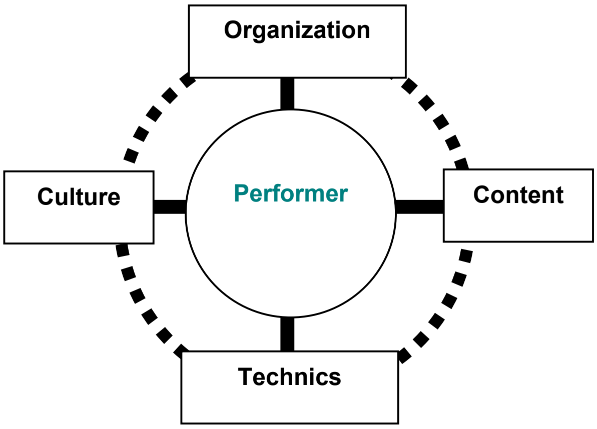


Figure 4. The KnowledgeWorks model for successful knowledge management implementations.

Figure 4 shows that the technical aspects of Performer comprise only about a quarter of the required effort. The following aspects require at least as much attention:

1. Organization. This includes setting up procedures for keeping alive the knowledge management organization; also, it includes a deployment process with roles, tasks, and responsibilities to guarantee a minimum amount of quality.
2. Content. The content aspect includes all knowledge, expertise, and individual estimates of the employees; a certain amount (the "critical" amount) of knowledge must be ready in the system by the time it is used by employees.
3. Culture. This is about getting people motivated to share knowledge, and to keep people enthusiastic. The strengths of the knowledge management instrument must repeatedly be highlighted and the weaknesses identified, considered, and improved.

Our research setting involved groups of students designing and realizing multimedia CD-ROMs, which requires a wide range of disciplines. Their knowledge sharing interests can be defined on two levels:

1. To learn from other disciplines in order to expand knowledge about multimedia system development in general (knowledge sharing in breadth);
2. To learn from his or her own particular discipline from previous-year students, to expand knowledge about particular aspects of multimedia system development (knowledge sharing in depth).

For both types of knowledge sharing, we wanted to offer students an instrument that supports a stock type of knowledge management because we found that a flow type of knowledge sharing was already done informally outside college hours. Thus we defined the main direction for the characterization of the tool Performer.

For the first type of knowledge sharing, we had to make explicit the entire process of designing and realizing multimedia CD-ROMs as taught in the Delft University course. In this definition, it should be immediately clear in a visible way what activities were assigned to each role during the entire process. In this way it is easier for discipline A to get an overview of the activities of discipline B, which in turn makes it easier to find relevant knowledge for each activity. The Objectives matrix, shown in figure 1, supports this. The Performer champion supervises this process (the teacher in this case) and two graduate students guarded the quality of the content.

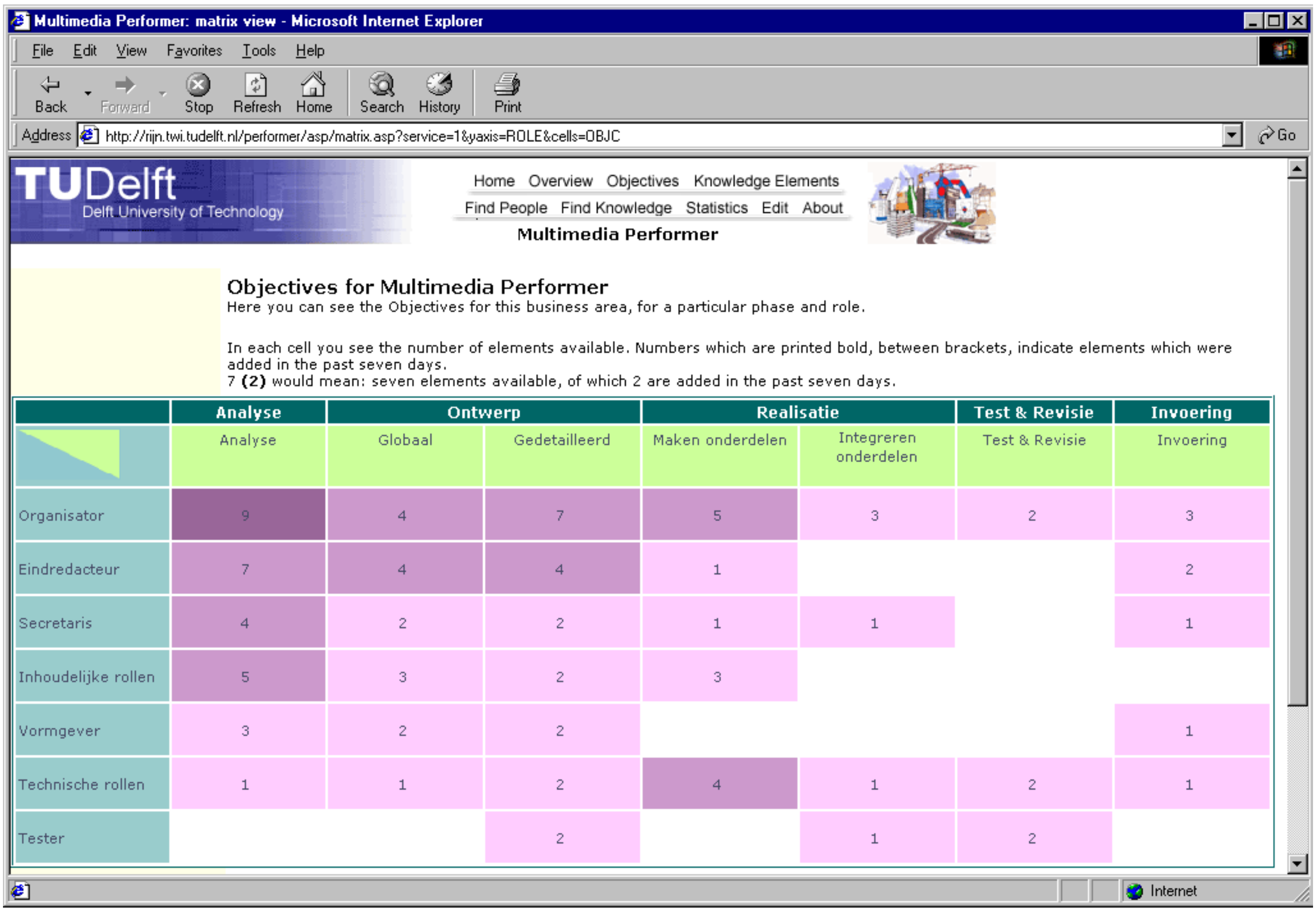
For the second type of knowledge sharing, we had to make sure that students can easily find relevant knowledge that had been submitted by previous-year students; that the knowledge was of a high quality and immediately usable; and that it encouraged our subjects to contribute new knowledge for that particular discipline too. To this end we offered the Knowledge Element Search filters, shown in figure 3. Using the filters it becomes easy to select a usage type, file type, project phase and knowledge element keyword to quickly locate a particular knowledge element.

The use of Performer by students

For the academic year 2000-2001 we decided to offer a slightly adjusted version of Performer to the fifty students following the multimedia course. A standard "empty" Performer for professionals was taken as the starting point, and some features that were not relevant for our students were excluded from the user interface. All data (such as the definition of our objectives matrix) was entered into the "Multimedia Performer" by using the standard editing features of Performer itself.

First, the roles and the phases were clearly specified and documented. This was not too difficult, because the teacher of the course was involved in the Performer research project [6]. The sentences could be derived from the descriptions in the "project book" that was already available for the course, in a straightforward manner. Next the objectives for the cells in the Performer matrix had to be described carefully. The teacher did this, assisted by some students who participated in the course in the past. The matrix containing the roles, phases and objectives was a means to convey to students the activities they should perform for this multimedia course, see figure 5. It was offered as a framework to structure the way of working.

The roles and phases were different from the more extensive set as described in the Performers for the professionals (see figure 1). The initial commercial phases and the phases for deployment and maintenance of the product were omitted. The objectives were completely different because they were described in educational terms for students. We connected to the objectives a set of forty initial knowledge elements, specified from teacher and student experiences of the past years. The knowledge elements are documents included as examples (e.g. selected user impressions, scripts, storyboards made by students in the past) and hints included as tricks (e.g. how to make metallic buttons, how to make some kind of generic components with some tools).


Multimedia Performer: matrix view - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search History Print
Address http://rijn.twi.tudelft.nl/performer/asp/matrix.asp?service=1&yaxis=ROLE&cells=OBJC
Go
TUDelft
Delft University of Technology
Home Overview Objectives Knowledge Elements
Find People Find Knowledge Statistics Edit About
Multimedia Performer
Internet

Objectives for Multimedia Performer

Here you can see the Objectives for this business area, for a particular phase and role.

In each cell you see the number of elements available. Numbers which are printed bold, between brackets, indicate elements which were added in the past seven days.
7 **(2)** would mean: seven elements available, of which 2 are added in the past seven days.

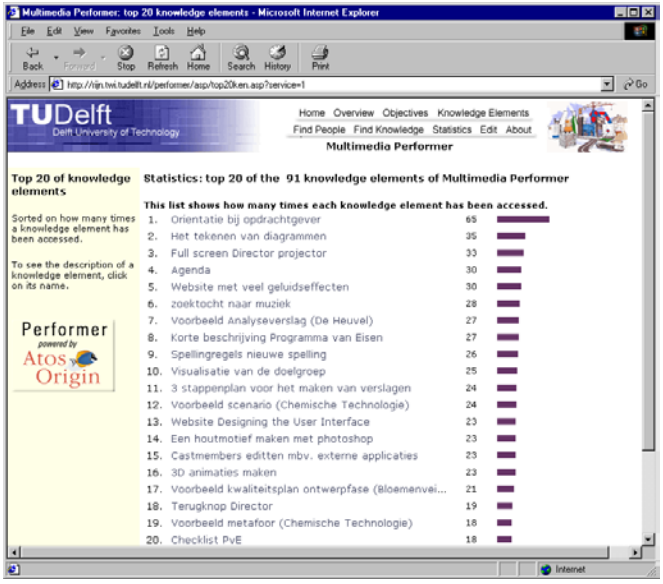
	Analyse	Ontwerp		Realisatie		Test & Revisie	Invoering
	Analyse	Globaal	Gedetailleerd	Maken onderdelen	Integreren onderdelen	Test & Revisie	Invoering
Organisator	9	4	7	5	3	2	3
Eindredacteur	7	4	4	1			2
Secretaris	4	2	2	1	1		1
Inhoudelijke rollen	5	3	2	3			
Vormgever	3	2	2				1
Technische rollen	1	1	2	4	1	2	1
Tester			2		1	2	

Figure 5. The objectives matrix of the Multimedia Performer (adapted version of Performer with contents in Dutch for the students of the course; the roles and phases are also different), which shows the relative denseness of objectives for each role in all sub-phases of a project or process. Seven roles are defined (see y-axis on the left) and five phases (x-axis on the top), resulting into seven sub-phases.

At the start of the Multimedia project, the students got a one-hour workshop on the use of Performer. In the workshop all the features of Multimedia Performer were presented in a tutorial with exercises. This workshop was finished with a questionnaire on what the students know on Multimedia Performer at that time. The results of the initial questionaire showed that the students were confident how to use all the features offered by Multimedia Performer, and were aware of the purpose and possible benefits it might bring them. We conclude that the workshop was effective.

The Multimedia Performer was always online during the multimedia course. In order to provide a positive cultural and organizational attitude (see figure 4) some students were appointed as *Champions* to stimulate and support their fellow students in using the Multimedia Performer and sharing knowledge. At the end of the course the students had added 25 new approved knowledge elements. Some added knowledge elements were not approved by the content owner because they would not be usable for other students, or because they offered the same content as other existing knowledge elements. One student added five approved knowledge elements; eleven other students added the next twenty knowledge elements.

During the project, usage statistics of the Multimedia Performer were recorded automatically. The students could follow the access hits for all objectives and knowledge elements in a matrix view. According to these statistics, the number of hits to the objectives was 650 and to the knowledge elements was 514. These recordings were anonymous. Most hits were recorded during the first weeks of the three project phases (analysis, design and realization). The objectives most frequently accessed were: write a quality assurance plan (129 accesses), analyze the assignment (57), design a metaphor (53), perform task analysis (44), make interaction design (30), perform user analysis (27), construct skeleton in Director (26), etc. The most frequently accessed knowledge elements were: orientation with customer (65 accesses), making diagrams (35), creating a full screen Director projector (33), website with audio samples (30), finding music (28), example of a final report on analysis phase (27), visualization of user characteristics (25), etc. The students could follow the sharing of knowledge week by week in different ways, e.g. by the list of top 20 knowledge elements, see figure 6.


Multimedia Performer: top 20 knowledge elements - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search History Print
Address http://rijn.twi.tudelft.nl/performer/asp/top20ken.asp?service=1
Go
TUDelft
Delft University of Technology
Home Overview Objectives Knowledge Elements
Find People Find Knowledge Statistics Edit About
Multimedia Performer
Performer
powered by
Atos Origin
Internet

Top 20 of knowledge elements

Sorted on how many times a knowledge element has been accessed.

To see the description of a knowledge element, click on its name.

Statistics: top 20 of the 91 knowledge elements of Multimedia Performer

This list shows how many times each knowledge element has been accessed.

1.	Orientatie bij opdrachtgever	65
2.	Het tekenen van diagrammen	35
3.	Full screen Director projector	33
4.	Agenda	30
5.	Website met veel geluidseffecten	30
6.	zoektocht naar muziek	28
7.	Voorbeeld Analyseverslag (De Heuvel)	27
8.	Korte beschrijving Programma van Eisen	27
9.	Spellingregels nieuwe spelling	26
10.	Visualisatie van de doelgroep	25
11.	3 stappenplan voor het maken van verslagen	24
12.	Voorbeeld scenario (Chemische Technologie)	24
13.	Website Designing the User Interface	23
14.	Een houtmotief maken met photoshop	23
15.	Castmembers editten mbv. externe applicaties	23
16.	3D animaties maken	23
17.	Voorbeeld kwaliteitsplan ontwerpfase (Bloemenvei...	21
18.	Terugknop Director	19
19.	Voorbeeld metafoor (Chemische Technologie)	18
20.	Checklist PvE	18

Figure 6. The list of top 20 knowledge elements that were most frequently accessed by the students. This list was provided to gain insight into the behavior of sharing knowledge within the group and to let students become curious about the knowledge that was most often shared.

Conclusions and discussion

The instrument Performer was developed for stock knowledge management in teams of professionals. The first use of a slightly adjusted version for senior students was quite successful. The students frequently consulted the matrix with objectives for all roles and phases. They reported that the objectives helped them to structure their way of working in the multidisciplinary teams. On the value of the knowledge elements they were hesitating; the average response was *neutral,* with a high standard deviation. The students reported that they prefer getting technical advice from the senior students assigned to assist them (*flow knowledge management*). However, 30% of the students consulted the available knowledge elements frequently up and until the very end of the project. They also contributed their own new knowledge elements. This may be caused by cultural differences in the heterogeneous group of students from rather different programs.

Recently, in another course not reported here [9], some 120 junior computer science students used Performer. They were enthusiastic about the use of the objectives matrix for a course on human computer interaction. Students had to develop a user interface in teams of four. In this course for junior students almost no new knowledge elements were added and their opinion about the use of knowledge elements was not positive at all. They preferred the assistance by senior students strongly (in other words, a preference for flow knowledge management instead of stock knowledge management).

We observed important differences in the use of Performer by university students on the one hand, and practitioners from commercial companies on the other hand. In particular, each of these target audiences requires a wholly different approach.

The value of the effect of using Performer is also different. In the case of university students, the value of using Performer lies especially in stimulating the group process, team co-operation, and the clarification of the working process using the objectives matrix. Peer-to-peer student advice is often preferred to using existing knowledge elements. Due to time pressure, students were not able to consult Performer as often and as thorough as some of them would have liked. The motivational aspect was the most difficult problem: a student has little personal benefit of adding knowledge elements for other, future students. It required a set of small awards (gifts) for students who were willing to take this time anyway. In a professional commercial environment, on the other hand, the motivational aspect is less an issue because each practitioner sees the value of sharing his or her own knowledge with other colleagues, particularly for new colleagues. The objectives matrix is not used as often here because the way of working is already generally known.

Another important observation is that a stock knowledge management solution such as Performer is probably fit for many other knowledge domains as well, as long as it is used in a project-oriented environment. We are currently investigating the use of Performer in other types of university courses.

References

[1] Dobson, W.D., and C.K. Riesbeck, "Tools for incremental development of educational software interfaces". In: *Proceedings CHI 98,* Los Angeles, 384-391, ACM Press.

[2] Johnson, D.L., and J.G. Brodman (2000), "*Applying CMM project planning practices to diverse environments*". In: IEEE software, Vol 17, nr. 4, special issue on process diversity, pp. 40-47.

[3] Nonaka, I., and H. Taceuchi (1995), "*The knowledge-creating company*". Oxford Press, New york, U.S.A.

[4] Philips, B.H., M. Rahman, and J. Järvinen, "Building a human factors "knowledge shelf" as a collaborative information tool for designers". In: *Proceedings CHI2001,* Seattle, 98-103, ACM Press, 2001.

[5] Van Aalst, J.W., and C.A.P.G. van der Mast (1998), "Creating the multimedia experience database". In: Sutcliffe, Ziegler, and Johnson (Eds) "*Designing Effective and Usable Multimedia Systems*". Proceedings of the IFIP Working Group 13.2 Conference, Stuttgart, Germany, September 1998. Kluwer Academic publishers, pp. 117-129.

[6] Van Aalst, J.W. (2001), "*Knowledge Management in Courseware Development*". Delft University Press, The Netherlands. ISBN 90-407-2161-0

[7] Van der Mast, C. (1995) “Professional Development of Multimedia Courseware”. In: *Machine-Mediated Learning*, vol. 5 (3&4), 269-292.

[8] Van der Mast, C., C. Verwijs, P. Fisser (2000) “Collaborative distance learning using MESH workstations”. In (Eds. F. Broeckx, L. Pauwels) *Conference Proceedings Euromedia 2000*, May 8-10, 2000, Antwerp Belgium, Society for Computer Simulation International, San Diego, ISBN 1-56555-202-4, 207-211.

[9] Koppelman, H., van Dijk, E.M.A.G., van der Mast, C.A.P.G., van der Veer, G.C., Team Projects in Distance Education: a Case in HCI Design, in: *5th annual ACM SIGCSE/SIGCUE conference on Innovation and Technology in Computer Science Education,* July 11 - 13, 2000, Helsinki, Finland, pp. 97-100.

[10] Weggeman, M. (2000), “*Kennismanagement: de praktijk (Knowledge management in practice)*”. Scriptum books, Netherlands (In Dutch).

[11] BSCW: Basic System for Co-operative Work, see http://bscw.gmd.de

[12] Blackboard: see http://www.blackboard.com

An Integrated Network and System Management Framework based on Adapted Software Components

Martin H. Knahl[❢], Udo Bleimann[⊕], Steven M. Furnell[?], Holger D. Hofmann[❊]

e-mail: martink@soc.plym.ac.uk

Keywords

Integrated Network and System Management, Reusability, Componentware, Computer Networks

Abstract.

Componentware represents an evolutionary step in software development which adds a new packaging granularity to object-oriented systems: software components. Such as components in other sciences or technical domains, software components facilitate reuse even across application domains and allow the realisation of nearly any distribution scenario. In Integrated Network and Systems Management (INSM), the subjects to management are distributed. Furthermore, the heterogeneity of managed components often requires specific adaptations to the management software. The INSMware framework approach proposed throughout this paper achieves a maximum level of manageability by combining INSM with componentware. Software components by default are immutable which would even hinder minor adaptations to the management system or management semantics. To overcome this architecture-inherent limitation of componentware, we integrate the Component Adapter approach to INSMware, which allows the integration of even semantically incompatible software components.

1 Introduction

Nowadays, distribution represents a system-inherent criterion for software and hardware systems. Hardware as well as software components are to collaborate in large-scale environments such as the Internet and thus create new challenges in managing both. We develop the concept of integrated network and system management (INSM) and propose INSMware, a framework which provides component-based INSM, managing both hardware and software components [8].

Traditional management software is monolithic: stand alone applications performing all necessary functions. This paradigm is reaching its limits in terms of code reusability, extensibility, configurability and especially concerning the speed and size development. One reason for this is that traditional development approaches require the application to contain the entire functionality even if much of it is only required for very specific tasks. This paradigm also requires extensions to the design to be carried out and integrated by an application developer familiar with the system (typically the original developer). However, it is the user, not the developer, who is often the one most familiar with the application domain and aware of required extensions. The natural consequence of this would be to provide an interface that allows the user to add new functionality to the monolithic block and that defines a migration path towards compound applications.

Software components provide a different approach to the development process. Like a child with Lego$^{\text{TM}}$ building blocks, one can build a compound software application by using several elemental blocks – software components – rather than a monolithic entity. Once the application has been built, it has similar functionality but has the advantages of being distributable, scalable, configurable and can be modified by adding or replacing components. Configurability in this context means to use components or component-based applications in more than one application domain by setting parameters affecting the component's behaviour. Component software aims to provide an architecture to tailor individual needs easily and at low cost. The Latin proverb 'Divide et Impera' can now be changed to 'Build and Manage'.

2 Software Components and Adaptation

The concept of software components goes back to 1969 when McIlroy envisioned an industry of reusable software components and introduced the concept of formal reuse though the *software factory* concept [9]. The idea behind the concept of software components was and is to use self-contained, pre-fabricated and pre-

[❢] Network Research Group / School of Computing, University of Plymouth, Plymouth, United Kingdom

[⊕] Fachbereich Informatik, University of Applied Sciences, Darmstadt, Germany

[?] Network Research Group / Department of Communication and Electronic Engineering, University of Plymouth, United Kingdom

[❊] ABB Research Centre, ABB, Heidelberg, Germany

tested units in order to build more complex units or entire applications.

As there is no agreed formal definition of a software component [13], we will first present the definition that forms the basis for our work: a software component is a piece of software with one or more well-defined interfaces that are configurable, integrable and immutable [4]. This definition highlights the immutability of software components. Though it guarantees black-box reuse [2], immutability can lead to static caller relationships between software components, a lack of system configurability, and poor support of object-oriented reuse mechanisms such as implementation inheritance [12]. Thus, the physical shape of software components highly influences their reusability.

The driving force behind the use of pre-fabricated components, be it in computer science or other domains, has always been reuse. Reuse relies a great deal on the availability of descriptions of entities to be reused.

Figure 1 shows a taxonomy of reuse potentials with respect to adaptability and modifiability.

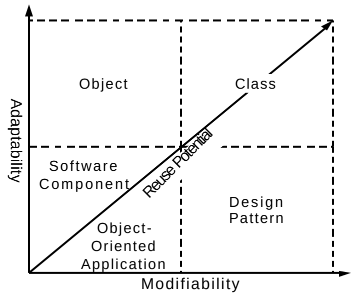


Figure 1: Taxonomy of Reuse Potentials

At one extreme we have classes, which exist in the form of modifiable source code. The reuse potential of a class is very high as one can always change the class' source code to meet the requirements of the developer. Objects, on the other hand, are adaptable, not modifiable, entities. An object exists only at runtime as an instance of a class and cannot be modified. However, an object's state and its relationships with other objects can be changed at runtime. Design Patterns [2] exist in graphical and/or textual format, and like classes, can be modified. Because a Design Pattern is dedicated to a specific use, it has very limited adaptability.

Object-oriented applications, i.e., applications that have been developed using object-oriented techniques, also have limited adaptability. Moreover, they are not modifiable since they exist in an immutable physical shape. It may come as a surprise to find software components at the lower left end of our taxonomy of reuse potentials. But a closer scrutiny reveals that software components share the same characteristics as object-oriented applications. They come in an immutable physical shape, having been implemented on distributed object-oriented architectures such as CORBA [11] and DCOM [1]. They provide, at most, limited support for the differential reuse mechanisms: inheritance, aggregation, and delegation, and are therefore not very adaptable either.

Software components are reused by composition [10], which means the grouping of a set of components to form a new component or even an application. The components are integrated on a common *composition layer* and communicate by exchanging messages. The interaction of these components is controlled by a *control logic*. Software component composition requires component compatibility. If components that are to be composed are not compatible, i.e., they are not able to collaborate because of interface or semantic reasons, the composition language can be used to realise component compatibility. We call this concept *adaptation*.

Let C_{org} be a component to be adapted that contains the implementations I_j, $)I_j$ the adapting implementation, and operator ρ an operation to combine I_j with $)I_j$. Then the adapted component C_{adapt} can be defined as:

$$C_{adapt} = C_{org} \; \rho \;)I \text{ with })I = \{I_1 \; \rho \;)I_1, ..., I_n \; \rho \;)I_n\}$$

The following criteria for software component reuse and adaptation mechanisms have been identified: transparency of use, black-box characteristics, configurability, reusability and architecture independence and efficiency [6]. As adaptation functionality at the level of a composition language is not reusable, it is not an appropriate mechanism for software component adaptation.

To cope with this problem, we apply the concept of Component Adapters [5]. Component Adapters are software components which represent a specific view of one or more software components to client components and which act as surrogates for these.

The incoming interfaces of a Component Adapter represent a view required by client components while its outgoing interfaces are connected to the incoming interfaces of server components to be used to realise the Component Adapter's functionality. The interface members of the Component Adapter can be mapped to one or more implementations provided by the server components. Depending on the implementation of the Component Adapter, these caller relationships can be changed at runtime or can be statically assigned at the time of development.

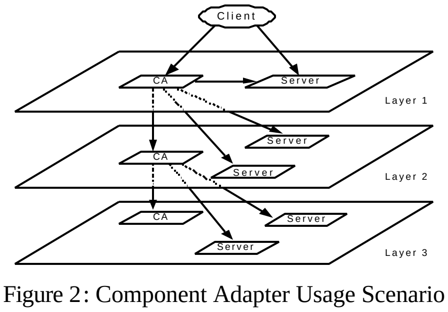


Figure 2: Component Adapter Usage Scenario

The mapping of incoming to outgoing interface members is installed by configuration and includes possible parameter and return type conversions. Though it is possible to do an automatic data conversion for elemental data types such as integer or float, the mapping of complex data types has to be configured by the user. Without any data conversion facilities only server components forming part of a subtype/supertype relation with client components could be used by the Component Adapter. Figure 2 shows a usage scenario of Component Adapters in which the approach is used recursively.

The design of Component Adapters is based on a combination the structural Design Patterns [15] [2]: *adapter* (adaptation of object interfaces), *decorator* (addition of functionality to existing implementations), *facade* (provision of high-level interfaces to sub-systems), and the behavioural Design Pattern *mediator* (centralisation of control). This makes it possible to change communication paths between software components, to define new interfaces to existing implementations, and to centralise the required implementation for the adaptation (control logic) in one place without the necessity of modifying the concerned components.

The integration of the Component Adapter approach with a component management architecture such as discussed in [5] provides flexibility in integrating self-developed software components with pre-fabricated components and supports systems evolution. This means that even software components that were unknown at the time of the management system's development can be slightly integrated with it. The major benefit of using a management system together with Component Adapters is that software components to be managed neither have to provide specific management functionality nor specific management interfaces.

3 Component-Based Integrated Network and Systems Management

The terms network, systems and applications management stand for all precautions and actions taken to guarantee the effective and efficient use of hardware and software resources of distributed systems and their underlying communication networks [3]. Several management architectures have been proposed and standardised as a basis for integrated management (e.g., SNMP based TCP/IP management, Telecommunications Management Network by the ITU). Management platforms that integrate different management applications are available on the market (e.g., IBM Tivoli TME/10 Management Framework). However, the proposed management architectures and platforms do not provide integrated management solutions for network management of LANs and WANs and for their different requirements on systems management. They generally represent management toolboxes with differences in middleware, management protocols or incompatible management applications. Given the diverse technologies and vendor specific implementations in today's heterogeneous networks, the management architecture itself becomes a heterogeneous mixture comprising different management applications and platforms.

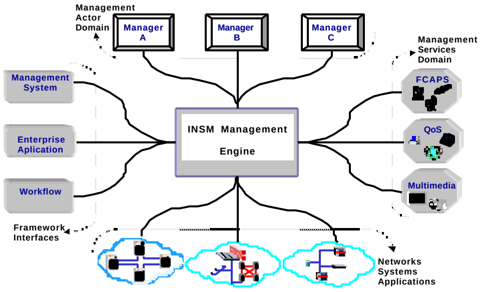


Figure 3: Management Framework

One attempt to address this problem could be to provide a single system capable of providing all Network and Systems Management services. Such a system is referred to as an Integrated Network and Management System. The problem with these is that they are very complex and therefore expensive and processor intensive. The Management Framework itself must be a distributed system with open interfaces, where the required management services are put together to provide the required management functionality (see Figure 3). Furthermore, open interfaces are required to provide interoperability to other management and enterprise applications and to expand the management framework to meet future requirements. Therefore, we propose a new component-based Management Framework [7].

The impact and leverage of distributed systems technology is prevailing not only for design and implementation of applications but also for the deployment of management systems. Thus far, management systems have typically been of two categories. Either specialised along one dimension, e.g., vertically targeting one or a few management aspects, e.g., configuration, billing or security or horizontally dedicated to management of a specific layer, e.g., network elements. Alternatively, they have resembled monolithic "main frames" based to a large extent on proprietary solutions. The presented research uses contemporary componentware technology to leverage a modular approach to the design of management systems, thus facilitating openness and extensibility on one hand and adaptability, i.e., customisation of management services, on the other.

The distribution of component-based systems mirrors the distribution of managed hardware/software components. We have already argued that a component-based approach does not inevitably guarantee reusability. We argue that this deficiency can be overcome by using Component Adapters that also allow for the integration of legacy components/applications into the management system.

As mentioned before, the conventional Management Systems cannot meet the needs of today's rapidly changing network systems. To make the system flexible to change, we propose a new style of the network management system, in both the development and operation phases. We call it "Componentware Integrated Network and System Management" (INSMware), as it uses the component-oriented approach for building and running the Integrated Management System [7] [8]. Our component-oriented Management System solves the problems of heterogeneous management protocols and can help reducing development costs.

The component-oriented software approach frees the developer from cumbersome coding, as the componentware based approach provides integration and customisation of Software components. This enables rapid and efficient system development, since tool software handles and validates much of the integrity of the system, but not a human.

4 INSMware

Limitations and restrictions of existing Network and System Management frameworks such as distribution of the management services and adaptation and integration of new management services can be overcome by providing a component-based approach [7] [8].

INSMware is a componentware-based framework for Integrated Network and System Management. We provide a component-based development approach meeting requirements for integrated management services. There are two versions of INSMware: one implementation is using CORBA [11], the other is based on DCOM[1]. This allowed us to study both middleware architectures in detail.

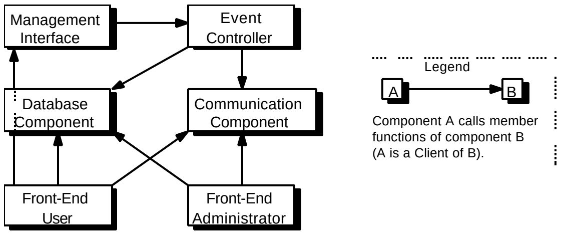


Figure 4: INSMware Components and Connectivity

The design of the individual INSMware components (see Figure 4) is based on a domain specification that subdivides the entire application domain into subdomains. First, the data processing system requires a connection to a data source. This is realised by the Management Interface component that exists in several different forms, similar to the device drivers of an operating system. It is configurable as required for different data sources.

The Management Interface component interprets the received data. It is filtered and analysed, and the component notifies the event controller when particular pre-defined exceptional states occur. Data storage is accomplished by a call to the database component and user notification is effected via communication components. It should be emphasised that all information about the users that need to be notified, e.g., access to user, user's role regarding the monitored processes, are stored within the system. The communication component itself consists of a set of several sub-components that again implement sub-domains, for example, faxes, voice mails, e-mails. The user can visualise system states using the front-end user component and can maintain the system by using the front-end administrator component.

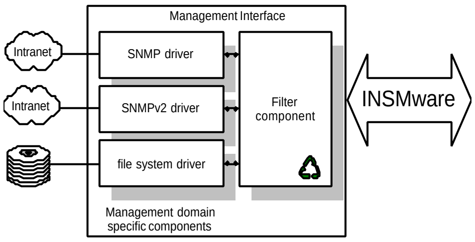


Figure 5: Management Interface Component

In our approach, a new management protocol can be installed by merely adding a new management domain specific component that represents the protocol into the Management Interface. This means that the protocol-handling part of the Management System is usually encapsulated to one component. This strengthens the adaptability of the component-oriented NMS to new management protocols, as the developer only has to develop the communication component specific to the new protocol. Therefore, we use a driver concept for the Management Interface that consists of the specific management domain and generic Filter component as shown in Figure 5.

4.1 Integration of Security Component

In any application domain, all data in the INSMware system is stored using the database component. To prevent unauthorised access to security-relevant data, clients must encrypt data before being transmitted to a database component that then decrypts it. Similarly, the database component must encrypt the data to be transmitted while client components have to decrypt the data. It was decided to add this functionality to the INSMware system, which means that four existing components, namely, the database component, the event controller component, the front-end user component, and the front-end administrator component have to be modified and re-built. These changes are not required if the additional functionality is integrated into the INSMware system in the form of Component Adapters via surrogate substitution. We demonstrate this using the DCOM version of INSMware.

Two approaches are possible:

i. realisation of the required security functionality inside a Component Adapter
ii. realisation of the required security functionality by a separate software component that can be used by a Component Adapter.

Approach (i) implies that the security functionality has to be realised in the programming language in which the Component Adapter is implemented. This restricts the reusability of the security functionality to one particular programming language. Approach (ii) allows to possibly reuse the provided security functionality not only by the Component Adapter, but also by several other software components. Coming from these two scenarios, we chose to realise approach (ii) because of reusability issues.

When realising approach (ii) either a security component can be developed from scratch or a software component can be purchased on the market that meets the requirements. We decided to develop our own security component to be able to reuse it in our various management domains. One demand for the realisation of a security component was the integrability with the development tools used. As these tools were ActiveX-capable, i.e., Microsoft ActiveX components could be integrated with these, we decided to implement the security component as an ActiveX component. ActiveX is a part of the Microsoft DCOM architecture that allows the realisation of scripting-capable software components.

Figure 6 shows the integration of a security component into INSMware. A client component accesses a Component Adapter that encrypts the data and sends it to a Component Adapter located on the same host as the database component. The second Component Adapter decrypts the data and sends it to the database component. After processing the client request, the database component transmits the results to the local Component Adapter that encrypts the data and sends it to the client-located Component Adapter. Finally, the results are decrypted and transmitted to the client component.

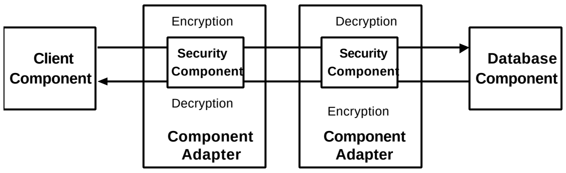


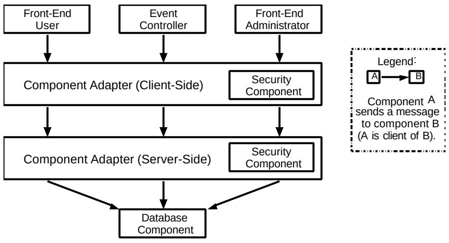
Figure 6: Component Adapter Integration

Both Component Adapters are located locally to the adapted software components. This location constraint is obvious as data transmission between the adapted components and their associated Component Adapter is still insecure. This does not represent a problem to local inter-component communication while this is not acceptable for remote communication.

The integration of Component Adapters with the INSMware system implies that every client of the database component, i.e., the event controller component, the front-end user component, and the front-end administrator component, accesses a client-located Component Adapter instead of directly accessing the database component. This integration is transparent to all client components and can be dropped if necessary.

Implementing a separate security component optimises the reusability of implemented algorithms, but its integration to a system may adversely affect the system's performance since the number of inter-component communications necessarily raises.

The Component Adapters must implement the interface of the software component to be adapted and expose this interface to client components. In the case of the INSMware system, this is the interface of the database component.



Front-End User
Event Controller
Front-End Administrator
Component Adapter (Client-Side)
Security Component
Component Adapter (Server-Side)
Security Component
Database Component
Legend:
A
B
Component A sends a message to component B (A is client of B).

Figure 7: INSMware Component Adapter Integration

Two Component Adapters, one for the client and one for the server side, are required. On the client side, data has to be encrypted before being sent to the server-located Component Adapter and data has to be decrypted after results being received from the server-side Component Adapter. The reverse behaviour is required on the server side.

The Component Adapters are integrated into the INSMware system by surrogate substitution and thus are transparent to client components. Each component of the INSMware system that accesses the database component does so through a client-side Component Adapter that communicated with a server-side Component Adapter. This is shown in Figure 7.

4.2 Management of SW Components

Component monitoring provides data about the run-time state of a set of software components. Data such as server host utilisation and memory usage of software components can be monitored. Without the use of Component Adapters, software components would have to be especially designed and implemented for the use within the management framework. This would far limit the components that can be used within our management environment. With Component Adapters, software components, with or without individual monitoring facilities, can be integrated and managed.

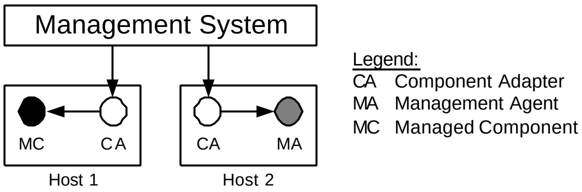


Figure 8: Component Adapters - Component Monitoring

Figure 8 shows two types of software component monitoring. In Host 1, a Component Adapter directly monitors a managed software component (MC). This approach can be applied if a MC supports monitoring functionality or if the Component Adapter gathers information about a component by making calls to its operational interface. In Host 2, the Component Adapter monitors a management agent (MA) that gathers information about a software component's environment, for example, the number of running processes or software component instances on a particular host, the operational status of hardware components (e.g., on/off/stand-by).

Using the Component Adapter approach, monitoring software components can be easily integrated with new or existing management systems. If, for example, Component Adapters implement a Simple Network Management Protocol (SNMP) interface [14], they can be integrated with any SNMP-compliant management system.

Along with the distribution of management components comes a distribution of management knowledge. In particular scenarios, it may be required that elemental management tasks are performed by local management components while complex tasks that might require user interaction may be co-ordinated by a central authority.

5 Conclusions and Outlook

Our work with INSMware has shown that a distribution of management software and thus of management knowledge can help in mastering the inherent complexity of distributed hardware/software systems. The realisation of management systems as componentware, i.e., software entirely composed of immutable pieces of software called "software components", can significantly support software reuse in this domain. In a scenario where normally the management software would have to be adpated to changing requirements, we propose the use of the Component Adapter concept. The latter helps to integrate and configure even incompatible software components and supports a common denominator on the software level.

For particular application domains, pre-fabricated Component Adapters may exist providing specific interfaces to managed components or to management systems. Other scenarios might require custom adapters whose development could be supported by libraries or code skeletons providing basic functionality.

6 References

[1] N. Brown, C. Kindel. *Distributed Component Object Model Protocol - DCOM/1.0*. Microsoft Corporation, Network Working Group, 1996.

[2] E. Gamma, R. Helm, R. Johnson, J. Vlissides. *Design Patterns — Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1994.

[3] H. Hegering, S. Abeck, B. Neumair. *Integrated Management of Networked Systems*. Morgan Kaufmann, 1999

[4] H.D. Hofmann. *Componentware — Integration of Software Components in Distributed Computing Environments*. M.Sc. Thesis, Cork Institute of Technology/Ireland, 1997.

[5] H.D. Hofmann, J. Stynes. *Implementation Reuse and Inheritance in Distributed Component Systems*. Proceedings of Twenty-Second Annual International Computer Software and Applications Conference (COMPSAC'98), Vienna/Austria, 1998.

[6] H.D. Hofmann, J. Stynes, G. Turetschek. Software Reuse by Adaptation. *Proceedings of the Second International Network Conference (INC 2000)*. University of Plymouth: Plymouth, UK.

[7] M. H. Knahl, H.D. Hofmann, A. D. Phippen. A Distributed Component Framework for Integrated Network and System Management. Information Management and Computer Security, 7(5), pp. 254-260, MCB University Press, Bradford/UK, 1999.

[8] M. H. Knahl, U. Bleimann, H.D. Hofmann, S. M. Furnell. 2000. An Integrated Management Architecture for Heterogeneous Networks: INSMware. In: Jajszyczyk, Andrzej (Editor), Proceedings of the IEEE Workshop on IP-oriented Operations and Management (IPOM '2000). September 2000. pp. 111-118.

[9] M.D. McIlroy. *Mass-produced software components*. In J.M. Buxton, P, Naur, and B. Randell (editors), *Software Engineering Concepts and Techniques*, pp. 88-98, 1968 NATO Conference on Software Engineering, 1976.

[10] O. Nierstrasz, S. Gibbs, D. Tsichritzis. *Component-Oriented Software Development*. Communications of the ACM, 35(9), pp. 160-165, 1992.

[11] OMG. *The Common Object Request Broker: Architecture and Specification*, Revision 2.2. OMG Document 98-07-01, Object Management Group, Inc., 1998.

[12] M. Sakkinen. *Inheritance and Other Main Principles of C++ and Other Object-oriented Languages*. PhD thesis, University of Jyvaeskylae/Finland, 1992.

[13] J. Sametinger. *Software Engineering with Reusable Components*. Springer, 1997.

[14] Stallings, William. SNMP and SNMPv2: The Infrastructure for Network Management. IEEE Communications: Management of Heterogeneous Networks, 36(3), 1998.

[15] W. Zimmer. *Relationships between Design Patterns*. Proceedings of PLoP '94 - Pattern Languages of Programs, Addison-Wesley, 1995.

MEDIATEC

SOUND AND SPEECH SYNTHESIS

DSP IMPLEMENTATION OF REALTIME 3D SOUND SYNTHESIS ALGORITHM FOR MONAURAL SOUND SOURCE

Noriaki Sakamoto
Isao Shirakawa
Dept. Information Systems Eng.
Osaka University
2-1 Yamada-Oka, Suita, Osaka, Japan
sakamoto@ise.eng.osaka-u.ac.jp

Wataru Kobayashi
Arnis Sound Technologies, Co., Ltd.,
2-7-9 Kita-Senzoku, Ota-Ku, Tokyo
Japan

Takao Onoye
Dept. Communications & Computer
Eng., Kyoto University
Yoshida-Honmachi, Sakyo-Ku, Kyoto
Japan

KEYWORDS
3D sound synthesis, DSP implementation, memory size, calculation accuracy, computational costs.

ABSTRACT

This paper describes a single DSP implementation of a realtime 3D sound synthesis algorithm, which is distinctive in that the audible frequency band of a head-related transfer function (HRTF) is divided into three subbands in such a way that the portion of the functional behavior proper to each subband can be realized efficiently by a distinct digital filtering structure. The whole digital filters are implemented on a commercially available 16-bit fixed-point DSP, from the aspect of memory size, calculation accuracy, and computational complexity. As a result, the realtime 3D sound synthesis is performed by employing a single DSP with low computational labor of 51.3 MIPS and small memory size of 4.6k words, which demonstrates its applicability for mobile computing.

1. INTRODUCTION

According as 3D sound effects are widely used in the fields of entertainment and mobile computing, the 3D sound synthesis has been investigated for portable applications like CD and MD players. Thus a number of different approaches have been attempted for realizing 3D sound effects in such a way as to construct a virtual sound source with 2-channel stereo.

In the 3D sound synthesis, there are three types of approaches, one to transform 5.1-channel surround signals into 2-channel stereo signals (Kawano et al. 1998), the second to reproduce 3D sound signals from a 2-channel stereo-sound source (Kim et al. 1997) and the third to synthesize 3D sound effects from a monaural sound source.

However, the more realistic 3D sound synthesis from 2-channel stereo and monaural sound sources can be attained by exploiting head-related transfer functions (HRTFs). The key technology of this 3D sound synthesis consists in digital filtering to realize HRTFs defined for a given sound-source position at left and right external ears. The functional characteristics are closely related to the sound reflections and diffractions via head, shoulders, external ears, etc., which are different according to directions and distances from a given sound source to human ears (Wenzel et al. 1993).

Since conventional methods intend to realize the overall characteristics of HRTFs throughout the audible frequency band, the process of approximating precisely the whole HRTF behaviors involves too heavy computational labors to perform a single DSP implementation for mobile applications (Okamoto et al. 1997).

In addition, some conventional approaches with the use of FIR filter (Brown and Duda 1998) and IIR filter (McCabe and Furlong 1991) have been attempted, aiming at the accurate HRTF realization. However, all of these conventional methods suffer from the impractical computational cost.

On the other hand, among other conventional methods, there exists an interesting one from the aspect of the computational cost. This is distinctive in that the HRTF frequency band is divided into three subbands and the input signal level into eight, such that 600 DSP instructions per channel can be achieved for the 3D sound synthesis (Kim et al. 1997). However, although the computational cost and memory size can be much saved, the high precision HRTF approximation can be hardly achieved, since only the average value of the signal in each subband is used for the synthesis.

Thus to cope with these technical difficulties of the 3D sound synthesis, a novel realtime algorithm is devised dedicatedly for mobile applications (Kobayashi et al. 2001). In this algorithm the audible frequency band of a given HRTF is divided into three subbands such that the 3D sound synthesis in each subband can be efficiently achieved by employing a distinct optimal filtering structure in a different way from the conventional methods.

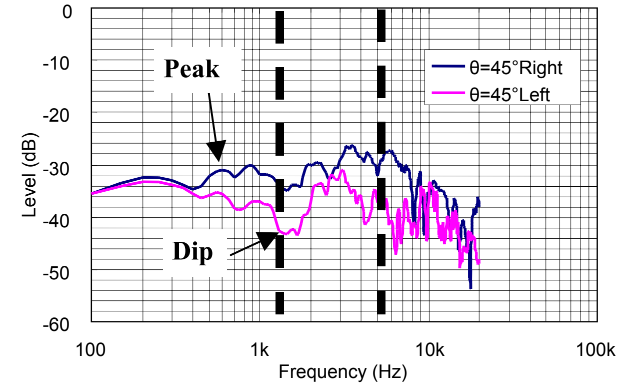


Figure 1: HRTF Frequency Response Characteristic Measured by Binaural Recording

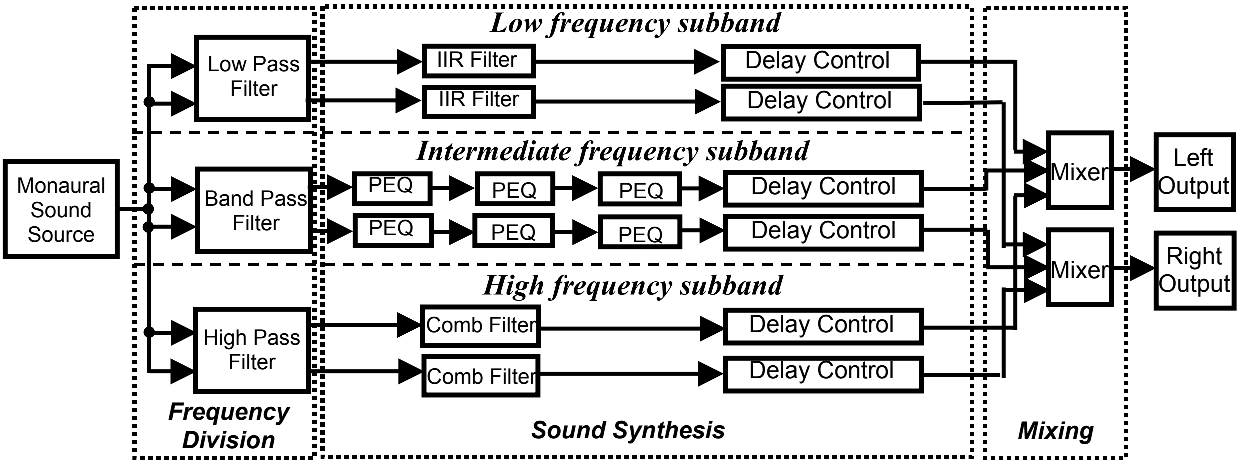


Figure 2: Block Diagram of 3D Sound Synthesis Algorithm

Fig. 1 depicts an HRTF frequency response measured by binaural recording under the condition of setting a sound source at the rightward angle of 45˚, where the characteristic curve in the low frequency subband changes mildly up and down, that in the intermediate frequency subband suddenly turns bumpy, and that in the high frequency subband quakes intensely. In what follows, a specific modeling of the diffraction characteristic proper to each subband is attempted as shown in Fig. 2. An optimal implementation of 3D sound synthesis by a single DSP is devised from the aspect of memory size, calculation accuracy, and computational complexity and, in addition, a hardware implementation with digital input/output interface is practically described.

2. PROPOSED FILTER STRUCTURE

As depicted in Fig. 2, the implementation process of this algorithm is divided into three stages, (i) Frequency Division Stage, (ii) Sound Synthesis Stage, and (iii) Mixing Stage. The digital filters to be used at each stage for DSP implementation are discussed in what follows.

2.1 Frequency Division Stage

The proposed 3D sound synthesis algorithm begins with the frequency division of a given monaural audio input. At this stage, a linear FIR filter is used for the frequency division, since it has a characteristic that the phase is proportional to the frequency, and hence the output can be easily adjusted to the prescribed phase. The order of the FIR filter is set to the 33rd, which can keep 40 dB/oct slope. Thus the sound signals to be processed in the low, intermediate, and high frequency subbands are divided through the use of the 33rd order low pass, band pass, and high pass FIR filters, respectively. This frequency division process is applied to each channel (left/right).

2.2 Sound Synthesis Stage

The 3D sound synthesis algorithm is invoked in each of these three frequency subbands. At this stage, the direction of a sound source and the distance from the source to both ears are taken into account.

In the low frequency subband, we have only to consider the parameters of delay and level differences between the left and right channels. The delay control buffers in the left and right channels of Fig. 2 are used for the delay difference. Therefore, the level difference between the left and right channels is taken into account. As can be seen from Fig. 1, in the low subband, the functional behaviors of HRTFs in both channels are of almost the same structure, only with the level differences increasing with the frequency. Thus these two characteristics can be realized by a pair of first-order Shelving IIR filters. Fig. 3 depicts their filter structure, where x[n], y[n], d[n] indicate the input, output, and delay signals, respectively. At this stage, the data in the low subband are processed with the use of three parameters of A, B×C and C, which indicate gains of the IIR filter. From Fig. 3, we have the following equation for each channel

$$y[n] = (x[n] - A \times d[n-1]) \times C + B \times C \times d[n-1].$$

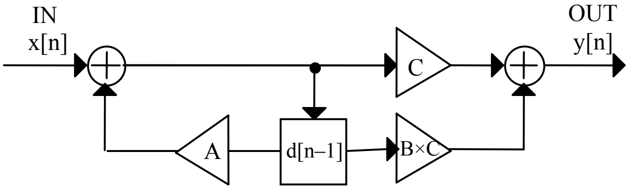


Figure 3: Structure of Shelving IIR Filter

In the intermediate subband, the frequency response characteristic of an HRTF is much more complicated, as can been seen from Fig. 1. In this case, parameter equalizers (PEQs) are used, which can easily adjust filter coefficients to control peaks and dips. These PEQs are generally constructed by using IIR and FIR filters. However, an IIR filter has fewer parameters than a FIR filter, and hence the memory can be much saved in the 3D sound synthesis. Thus in the intermediate subband, IIR filters are adopted, and moreover, in the HRTF characteristic any peak or dip is not extremely steep or deep, respectively, hence the order of the IIR filter can be set to the 2nd, the lowest order to realize a PEQ. Fig. 4 illustrates a 2nd order IIR filter, where x[n], y[n], and d[n] indicate the input, output and delay signals, respectively, and a0, a1, a2, b0, and b1 indicate gains. Thus each PEQ is represented by the following equation

$$y[n] = (x[n] + b1 \times d[n-1] + b2 \times d[n-2]) \times a0$$
$$+ a1 \times d[n-1] + a2 \times d[n-2].$$

In this intermediate subband, there are three peaks and dips in most cases, as can be seen from Fig. 1, the number of IIR filters is set to 3 per channel and hence totally 15 parameters per channel are necessary.

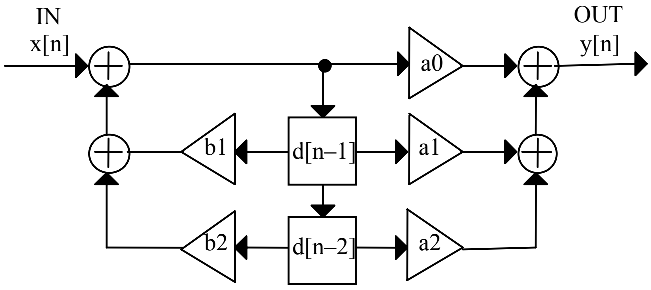


Figure 4: Structure of 2nd-order IIR Filter

In the high subband, the characteristic behaves like a comb, as can be seen in Fig. 1. Thus we adopt a comb filter. Fig. 5 depicts a structure of a comb filter, where totally 4 parameters are necessary, that is, direct gain (Gd), effect gain (Ge), feedback gain (Gf), and delay (d). For input x[n], output y[n], and delay z[n], we have for each channel

$$y[n] = Gd\times x[n] + Ge \times z[n-d],$$

$$z[n] = x[n] + Gf \times z[n-d-1].$$

As for the fundamental frequency of a comb filter, it is derived from the inverse of the delay value (d).

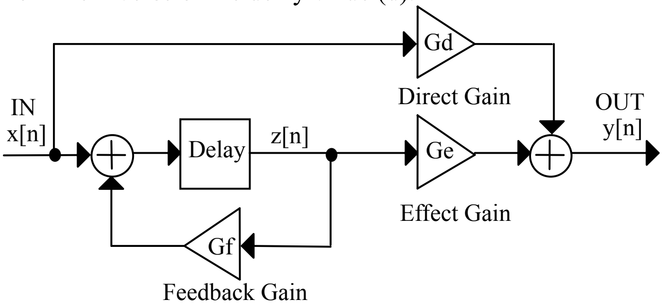


Figure 5: Structure of Comb Filter.

After the sound synthesis, the delay difference can be controlled by delay control buffers of Fig. 2. The delay control buffers conceal the delay difference between sound diffraction paths in three subbands, the delay difference caused by filter processings in three subbands, and the arrival time difference between left and right ears from a sound source.

2.3 Mixing Stage

After the sound synthesis, audio data in the left and right channels in three subbands are mixed. At this stage, the sound level in each frequency subband can be adjusted by multiplying the input data, x1[n], x2[n], and x3[n] by gains, a0, a1, and a2, respectively. Thus the output y[n] is represented by the following equation

$$y[n] = a0 \times x1[n] + a1 \times x2[n] + a2 \times x3[n].$$

3. DSP IMPLEMENTATION OF 3D SOUND SYNTHESIS

According to the proposed filtering structure, a scheme of DSP implementation of the 3D sound synthesis is devised with the use of a Texas Instruments low-power 16bit fixed-point DSP, TMS320C54x.

In what follows, for the simplicity, assume that the sampling frequency of the input and output signals is fixed at 48 kHz, since the same discussion can be attained for other sampling frequencies. In addition, the input signal is specified as a monaural channel, and the output signals are obtained in the left and right channels. For realtime processing, the whole processing of 3D sound synthesis for both channels is executed within a sampling period.

Now, consider the following three technical factors for the 3D sound synthesis.

- Memory size
- Calculation accuracy
- Computational cost

3.1 Memory Size

It is obvious that the total memory size necessary for this scheme is smaller than the internal memory size of the DSP. Thus, let us estimate the memory size necessary for our implementation. The total memory size is determined by the program size, the memory size for filter parameters, the delay buffer size in each filter processing, and the memory size for calculated data and address pointers. Additionally, as shown in Fig. 2, it is also dependent on the delay control buffer size for the delay difference between left and right channels.

There are three kinds of delay differences. First, we take account of the delay difference which depends on sound diffraction paths in three subbands. Now, suppose that a human head is of spherical shape with diameter of d = 200 mm. If the half of a sound wavelength is greater than the head diameter, the effect of sound diffraction through the head is negligible small. Thus a sound in the low subband can be regarded as arriving directly at both ears. On the other hand, a sound in the high subband reaches both ears through the diffraction along the head surface. Hence the maximum path difference between left and right ears is a head diameter for a sound in the low subband and a half circumference for that in the high subband. This means that the maximum delay difference between sound diffraction paths is $(\pi d/2-d)/v$ = 336μs, where v denotes the sound velocity set to v = 340 m/s. Secondly, consider the delay difference caused by filter processings in three subbands. This delay difference is due to FIR filters at the frequency division stage. In the intermediate subband, the FIR filters are constructed by a series of 33rd high pass and low pass FIR filters, while a 33rd FIR filter is used both in the low and high subbands. Therefore, in the low and high subbands, the delay of a 33rd FIR filter, 688μs, must be added to adjust the output timing of the processed data. Finally, consider the arrival time difference between left and right ears. This maximum difference is set to d/v = 588μs, which is due to the difference of a head diameter.

As a result, three kinds of the delay differences should be taken into account, and totally the maximum time difference can be estimated at 1.61 ms. Therefore, a delay control buffer size of Fig. 2 is determined by this delay of 1.61 ms.

Fig. 6 shows a memory map for the 3D sound synthesis. Program and filter parameters are assigned to ROM, and the data buffer and the data area are assigned to RAM. Note that the memory size for filter parameters per sound-source position corresponds to 320 words in ROM.

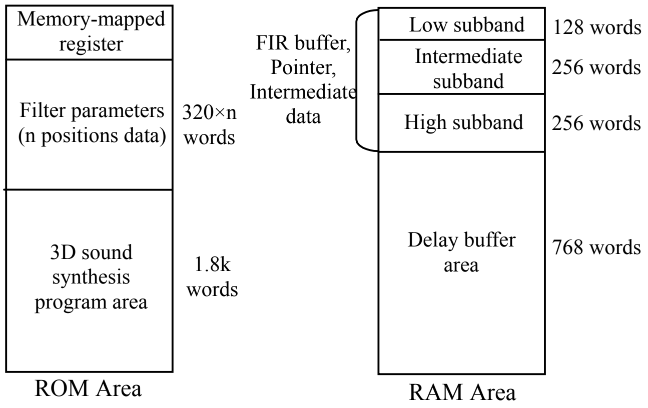


Figure 6: Memory Map of 3D Sound Synthesis

For example, in the case of implementing the 3D sound synthesis for four sound-source positions, only 3.2k and 1.4k words are assigned to ROM and RAM, respectively. Therefore, the memory size for the 3D sound synthesis is much smaller than that of a DSP TMS320C5409, which has 16K and 32K words of ROM and RAM, respectively, and thus the condition on the memory size can be satisfied.

3.2 Calculation Accuracy

The monaural input signals are given in the form of 16-bit integers. However, the intermediate data are calculated in the 32-bit precision including the fractional part, to keep the calculation accuracy. Finally, the output data after processing are rounded off in the form of 16-bit integers for both the left and right channels.

In addition, the frequency division and the sound synthesis of Fig. 2 are processed in series for each subband, since the 32-bit data after the frequency division can be utilized as the input of the succeeding sound synthesis, maintaining the high precision.

To implement the band pass filter at the frequency division stage, the sequential process of a high pass FIR filter and a low pass FIR filter is performed. In comparison with a band pass FIR filter, two chained FIR filters have a steeper characteristic at both of the cut-off frequencies. Therefore, the higher precision HRTF approximation can be attained by PEQs in a border frequency band.

3.3 Computational Cost

The total computational cost can be reduced by the optimal filtering structure. In addition to the advantage of the calculation accuracy, the sequential processing of the frequency division stage and the sound synthesis stage can reduce the overhead, since this processing can transfer the intermediate data between these stages without storing temporarily, and can save load/store instructions.

Consequently, the 3D sound synthesis can be implemented by a single DSP, and the same algorithm is applied to a monaural sound source regardless of direction of a sound source, that is, the computational cost and the memory size are the same, regardless of the direction.

4. RESULTS OF HARDWARE IMPLEMENTATION

To evaluate the DSP implementation of the realtime 3D sound synthesis, we have developed an evaluation system, as illustrated in Fig. 7. Fig. 8 depicts a photograph of this hardware board. A monaural input signal is captured through a digital audio interface. For this purpose, we adopt Cirrus Logic CDB8415A as the evaluation board of a receiver and a transmitter with digital audio interfaces. In addition, an evaluation board with crystal oscillators is designed to control the data transfer rate of input and output for sampling frequencies of 48 kHz, 44.1 kHz and 32 kHz. On this board, the input signals are processed by a Texas Instruments DSP

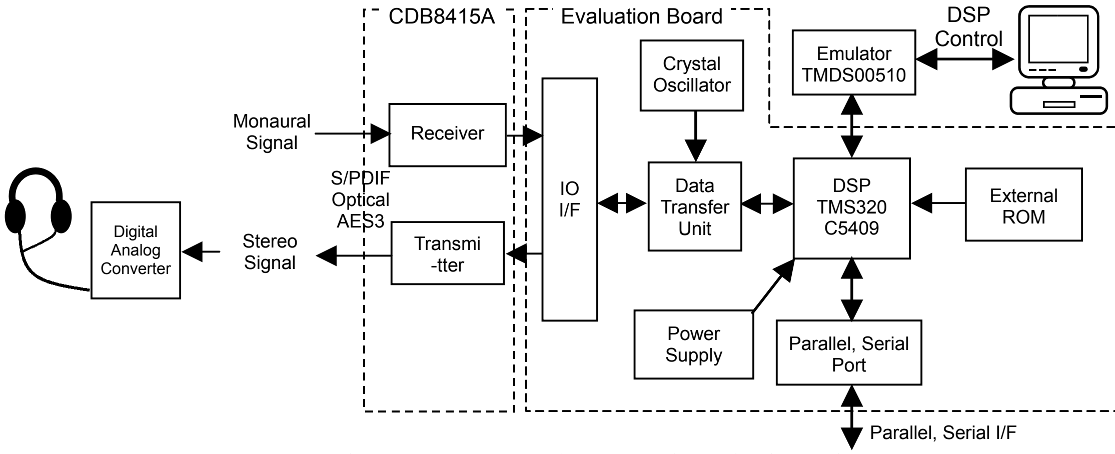


Figure 7: Structure of 3D Sound Synthesis Evaluation

TMS320C5409, which is monitored by a DSP emulator Texas Instruments TMDS00510 connected to PC. This board also involves an external ROM for a boot program and parallel and serial ports to evaluate the system. The synthesized data through the DSP are output from a transmitter of CDB8415A, and are finally reproduced by a D/A converter through headphones.

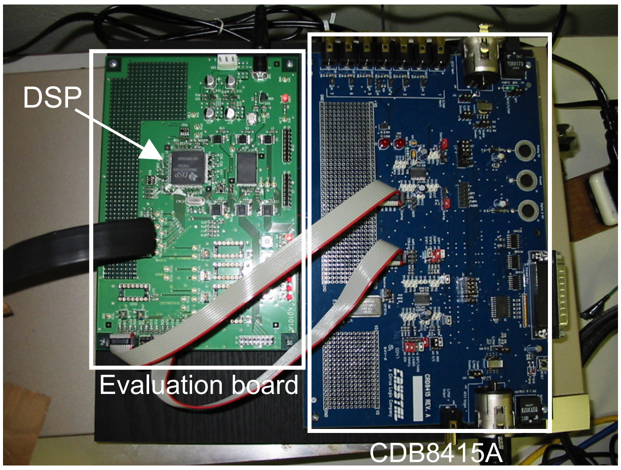


Figure 8: Photograph of 3D Sound Synthesis Evaluation System (Left: 3D Sound Synthesis Evaluation Board, Right: Cirrus Logic CDB8415A)

Table 1 shows computational costs of the 3D sound synthesis implementation, where it can be seen that the computational costs are 19.8 MIPS (Mega Instructions Per Second) at the frequency division stage, 27.0 MIPS at the 3D sound synthesis stage, 4.51 MIPS at the mixing stage, and totally 51.3 MIPS. Consequently, 530 DSP instructions per channel are required for the 3D sound synthesis, which are comparable to (Kim et al. 1997).

Table 1: Computational Costs at Each Stage (Unit: Mega instructions per second)

	Frequency division	Sound synthesis	Mixing	
Low subband	5.42	3.94	4.51	
Intermediate subband	9.31	17.7		
High subband	5.09	5.38		
Total	19.8	27.0	4.51	51.3

The power consumption of this DSP implementation of the 3D sound synthesis is evaluated as 83 mW, which includes 0.9 mA/MIPS of the DSP core consumption in FIR processing, the dominant part in this implementation, with 1.8 V supply voltage (due to the datasheet of TMS320C54x (Texas Instruments Inc. 1997), (Texas Instruments Inc. 2000)). The power consumption of a commercial available portable game system is 600 mW, and hence the addition of this 3D sound synthesis system to it becomes 14 % increase. Considering that a portable game system generally uses a specific chip, an ASIC implementation instead of a DSP one enables much more reduction of power consumption.

As for the practical implementation of this 3D sound system, noting that both input and output signals are processed at the same sampling frequency, we can see that the crystal oscillators used for generating sampling frequencies on the evaluation board can be eliminated. Similarly, the parallel and serial ports can also be eliminated. In addition, the external ROM is not necessary, if the data stored in the external ROM are transferred to the masked ROM in a DSP in advance, and therefore a small size implementation of a single DSP can be achieved.

5. CONCLUSION

This paper has described a single DSP implementation of a realtime 3D sound synthesis algorithm, which is distinctive in that the audible frequency is divided into three so that an optimal filtering structure can be constructed differently for each subband to realize accurately HRTFs at low computational costs and with small memory size.

Specifically, by employing a Texas Instruments 16-bit fixed-point DSP, TMS320C54x, the realtime 3D sound synthesis for a given monaural sound source can be performed at low computational costs of 51.3 MIPS, with small memory area of 4.6k words, and at low power consumption of 83 mW with 1.8 V supply voltage.

Consequently, our new approach can effectively provide a listener with the 3D sound effects through headphones at low cost and low power consumption by a single DSP. This 3D sound synthesis implementation can be applied to mobile applications such as a handy-phone with the use of an embedded DSP.

REFERENCES

Brown, C. P. and R. O. Duda. 1998. “A structural model for binaural sound synthesis.” *IEEE Transactions of Speech Audio Processing* 6, No. 5 (Sept), 476-488.

Kawano, S.; M. Taira; M. Matsudaira; and Y. Abe. 1998. “Development of the virtual sound algorithm.” *IEEE Transactions of Consumer Electronics* 44, No. 3 (Aug), 1189-1193.

Kim, T. S.; S. W. Jeong; B. C. Park; and S. I. Park. 1997. “New real-time implementation of 3D-sound system using TLA algorithm.” *IEEE Transactions of Consumer Electronics* 43, No. 3 (Aug), 671-678.

Kobayashi, W.; N. Sakamoto; T. Onoye; and I. Shirakawa. 2001 “3D acoustic image localization algorithm by embedded DSP.” *IEICE Transactions of Fundamentals* E84-A, No. 6 (June), 1423-1430.

McCabe, C. J. and D. J. Furlong. 1991. “Spectral stereo surround sound pan-pot.” In *90th Convention Audio Engineering Society.* preprint 3067.

Okamoto, M.; I. Kinoshita; S. Aoki; and H. Matsui. 1997. “Sound image rendering system for headphones.” *IEEE Transactions of Consumer Electronics* 43, No. 3 (Aug), 689-693.

Texas Instruments Inc. 1997. “Calculation of TMS320LC54x Power Dissipation, Application Report.” (June).

Texas Instruments Inc. 2000. “TMS320VC5402 Datasheet.” (Aug)

Wenzel, E. M.; M. Arruda; D. J. Kistler; and F. L. Wightman. 1993. “Localization using nonindividualized head-related transfer functions.” *Journal of the Acoustical Society of America* 94, No.1 (July), 111-123.

A New Fractal Word Synthesis

S. Fekkai, M. Al-Akaidi
Faculty of Computing Sc. & Engineering,
De Montfort University, Leicester, LE1 9BH, UK.
Email: mma@dmu.ac.uk

Abstract

Synthesis or artificial speech has been developed steadily during the last decades. Especially, the intelligibility has reached an adequate level for most applications, for example communication-impaired people and automatic generation of speech waveforms [1,2]. Recent progress in speech synthesis has produced synthesisers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications.

The goal of this work was to develop a new speech synthesis system, which is based mainly on the fractal dimension to create natural sounding speech. Our initial work in this area showed that by careful use of the fractal dimension together with the phase of the speech signal to ensure consistent intonation contours, natural –sounding speech synthesis was achievable with word level speech. In order to extend the flexibility of this framework, we focused on the filtering and compression of the phase to maintain and produce natural sounding speech.

Introduction

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms has been under development for several decades [1,2]. In an ideal world, a speech synthesizer should be able to synthesize any arbitrary word sequence with complete intelligibility and naturalness. The trade-off schematic in Figure 1 illustrates how current synthesizers have tended to strive for flexibility of vocabulary and sentences at the expense of naturalness (i.e., arbitrary words can be synthesized, but do not sound very natural). This applies to articulatory, rule-based and concatenative methods of speech synthesis [3,4,5,6].

An alternative strategy is one, which seeks to maintain naturalness by operating in a constrained domain. There are potentially many applications where this mode of operation is perfectly suitable. In conversational systems for example, the domain of operation is often quite limited, and is known ahead of time [7].

Past work by others have examined how unit selection algorithms can be formulated, and what constraints must be maintained [3,5,6].

In this work, we develop a framework for natural-sounding speech synthesis using fractal dimension. The developmental philosophy that we have adhered to throughout the work, places naturalness as a paramount goal. In our preliminary work involving word fractal dimension, the vocabulary size is relatively small, but naturalness is very high. Our research follows the bottom curve of Figure 1 where we view naturalness as the highest priority.

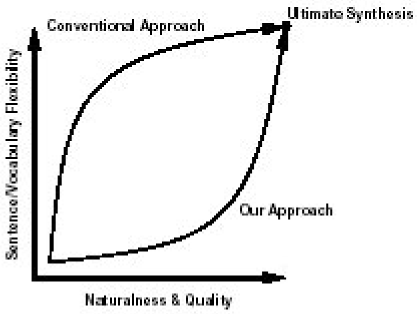


Fig. 1. Synthesis development trade-off schematic.

Fractal Dimension

Fractal dimension as parameter is important because it can be defined in terms of real-world data, and can be measured approximately by means of experiment [8-10]. Fractal dimension is a real number that in general, falls between the limits of 1 and 5 and can be calculated in a number of ways.

For the case of fractal speech signals and curves the fractal dimension lie between 1 and 2. The Power Spectrum Method (PSM) [11] has been used as an application of the Fourier power spectrum technique to calculate the fractal dimension of speech phonemes. The speech signal is Fourier Transformed by means of an FFT and the power spectrum is computed, $P_i = \mathrm{Re}(k_i)^2 + \mathrm{Im}(k_i)^2$. Assume that P_i is the measured power spectrum then $\hat{P}_i$ is the expected form of the fractal power spectrum, $\hat{P}_i = \mathrm{c}\ |k_i|^{-\beta}$, where c is a positive constant and β the positive spectral exponent [12].

Applying the Least Square approach to calculate the spectral exponent β and c yields to the following equation:

$$\beta = \frac{N \sum_{i=1}^{N} (\ln P_i)(\ln |k_i| - (\sum_{i=1}^{N} \ln P_i)(\sum_{i=1}^{N} \ln |k_i|)}{(\sum_{i=1}^{N} \ln |k_i|)^2 - N \sum_{i=1}^{N} (\ln |k_i|)^2} \qquad (1)$$

&

$$C = \frac{\sum_{i=1}^{N} \ln P_i - \beta \sum_{i=1}^{N} \ln |k_i|}{N} \qquad (2)$$

Where C= ln c. Using the relationship:

$$D = \frac{5-\beta}{2} \qquad (3)$$

Provides a simple formula for computing the fractal dimension from the power spectrum of a signal.

The implementation of the PSM consists of applying the FFT to the speech signal in order to obtain a spectral representation of the phoneme. A pre-filter step is then used to adjust the estimated values of the fractal dimension to fit within the range 1 and 2. The power spectrum of the pre-filtered signal is computed then the least square approach is applied to calculate the power exponent β (Eq. 1). Hence the fractal dimension D (Eq. 3) is obtained.

It is important to mention that without the pre-filtering step, the values of the fractal dimension were not satisfying the range of the fractal model. However the use of the Pre-filter ($\frac{1}{w}$) has the effect of confirming the speech data to fit the range of the fractal dimension for speech signal which lies between the range 1 and 2.

Non-Stationary algorithms for speech Synthesis

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies. The methods are usually classified into three groups:

1- Articulatory synthesis, which attempts to model the human speech production system directly.
2- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
3- Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

Articulatory synthesis typically involves models of the human articulators and vocal cords. The articulators are usually modeled with a set of area functions between glottis and mouth. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment [13]. For rule-based synthesis the articulatory control parameters may be for example lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture. Phonatory or excitation parameters may be glottal aperture, cord tension, and lung pressure [14].

When speaking, the vocal tract muscles cause articulators to move and change shape of the vocal tract, which causes different sounds. The data for articulatory model is usually derived from X-ray analysis of natural speech. However, this data is usually only 2-D when the real vocal tract is naturally 3-D, so the rule-based articulatory synthesis is very difficult to optimize due to the unavailability of sufficient data of the motions of the articulators during speech. Other deficiency with articulatory synthesis is that X-ray data do not characterize the masses or degrees of freedom of the articulators [13]. Also, the movements of tongue are so complicated that it is almost impossible to model them precisely.

Advantages of articulatory synthesis are that the vocal tract models allow accurate modeling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior [15].

In the next section we will introduce the synthesis of speech using fractal. A new technique, which develop a framework for natural sounding speech synthesis

Synthesising Speech with Fractals

In the previous section we have discussed how the power spectrum of a signal's Fourier transform can be used to extract the fractal dimension. This followed from assuming the power spectrum, $\hat{P}_i$, was related to the dimension in the following form,

$\hat{P}_i = \text{c} \, |k_i|^{-\beta}$, where β=5-2 D

To create a synthetic fractal is then the process of filtering white noise of the required size with a low pass filter, q whose Fourier transform is:

$$Q(k) = |k|^{-\beta/2}, \text{ where } \beta = 5-2D$$

Using this principle we will start by creating a fractal signal. The process is consist of four stages explained bellow:

***Step 1**:* Compute a random Gaussian distributed array $G_i, i = 0,1,...,N-1$ using a conventional Gaussian random number generator, with zero mean and unit variance. Compute a random number sequence of uniform distributed numbers $U_i, i = 0,1,...,N-1$ in the range zero to one.

***Step 2**:* Calculate the real and imaginary parts; $N_i = G_i \cos 2\pi U_i$ and $M_i = G_i \sin 2\pi U_i$. This defines G_i as the amplitude and U_i as the phase.

***Step 3**:*Filter N_i, M_i with $W_i = \frac{1}{K_i^{\beta/2}}$ to create N' and M'.

***Step 4**:* Inverse DFT the result using a FFT to obtain

$$n_i = \text{Re}(\hat{F}^{-1}\{N'+iM'\}).$$

The exponent is $\beta/2$ to ensure that the power spectrum, P_k, satisfies

$$P_k = (N'_k)^2 + (M'_k)^2 \propto k^{-\beta}$$

By using the same random noise for *U* and *G*, we can see how changes to *D* affect the signal.

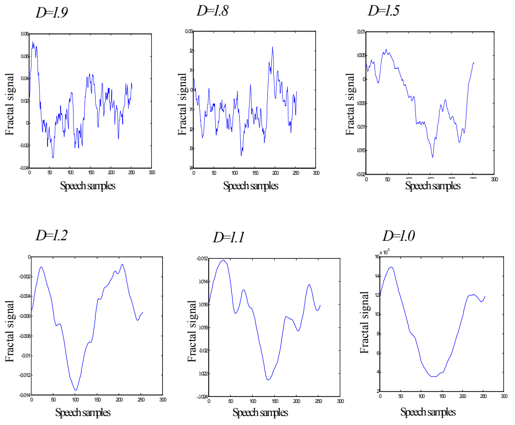


Fig. 2. Fractal Signals.

The next section will illustrate the new algorithm used to reproduce a natural sounding speech synthesis system, which make use of the fractal dimension of the word and its unwrapped phase.

Simulation & Results

The main objective is to create natural sounding speech and to ensure consistent intonation contours. The work carried out was based on two hypotheses:

- First, that the phase of the speech signal carries important information, hence intelligibility of the synthesis speech.
- Second, that fractal mathematics has the property of generating natural sounding speech.

1. Phase Compression Method

A real valued signal can be represented in terms of the so-called analytical signal, which is important because it is from that signal that the amplitude, phase and frequency modulations of the original real valued signal can be determined.

The phase compression method makes use of two parameters related to the speech signal namely the fractal dimension D and the phase ϕ.

The fractal dimension of the speech word is computed using the power spectrum method (as seen previously), then it is used to generate the fractal signal *f*, the aim is to use the fractal properties of the word in the calculation of the amplitude envelope in order to obtain more natural synthesis speech. To avoid the problem of choosing a principal range to compute the phase ϕ, we use the function unwrap phase and ϕ_1 is obtained. This latter is then compressed and used with the amplitude envelope within a specific algorithm to produce the synthesis speech. This method is illustrated in Figure 3.

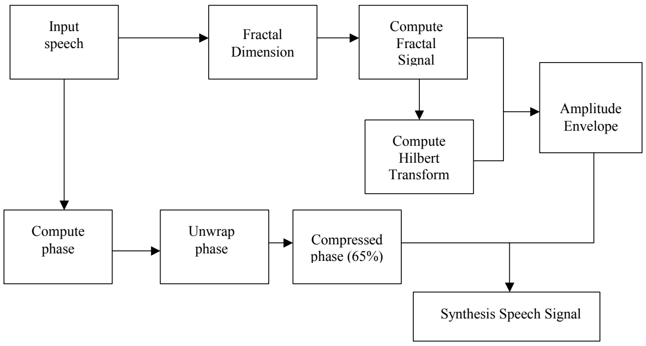


Fig.3 Phase compression algorithm

The simulation process consists of the following steps:

Step 1:Compute the fractal dimension D of the speech word $S_i(t)$.

Step 2: Compute the fractal signal $F(t)$ of the speech signal $S_i(t)$, using the power spectrum method as explained in chapter 3.

Step 3: Compute the Hilbert transform $H(t)$ of the speech signal S_i, (*i=1,2...N. N* is the length of S_i) as follows:

Take The Fourier Transform $F(k)$ of the signal $S_i(t)$.

Multiply the result by $[-i\,\text{sgn}(k)]$.

Compute the inverse Fourier Transform to obtain the Hilbert transform as follows:

$$H(t) = F^{-1}\{(-i\,\text{sgn}(k)F(k))\}$$

Step 4: Compute the amplitude envelop

$$A(t) = \sqrt{F(t)^2 + H(t)^2}$$

Step 5: Compute the phase $\phi(t) = \tan^{-1}\left(\frac{H(t)}{F(t)}\right)$ of the speech signal.

Step6: Compute the unwrapped phase $\phi 1(t) = unwrap(\phi(t))$.

Step 7: The synthesis speech signal is then reproduced as follow: $S_o(t) = A(t)\cos(\phi 1(t))$.

Step 8: Compress $\phi 1(t)$ by 65% from it's original with the use of the Discrete Cosine Transform (DCT), which has for role to keep the low frequencies and transform the high frequencies to zeros.

2. Phase Compression Method with Fractal Synthesis

This method adds novelty to the synthesis of speech words. In fact it doesn't use the fractal dimension D to compute the fractal signal of the input word but it also compresses the amplitude envelop by using a low pass filter then add white Guassian to the remaining signal. The algorithm used is illustrated in Figure 4.

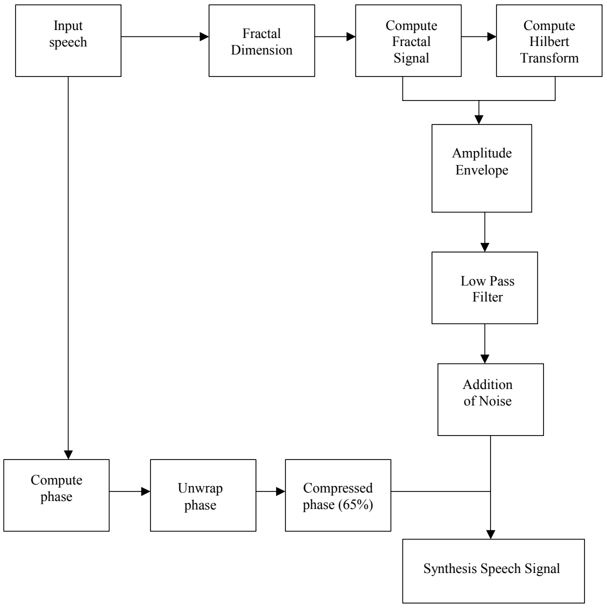


Fig.4 Fractal Synthesis algorithm

Three experiments have been conducted in the simulation process involving four different words namely "test", "best", "Open" and "zone", all spoken by different speakers.

- In the first experiment, only the unwrapped phase of the word has been used along with the fractal signal.
- In the second one, the phase is 65%compressed from the original
- In the third experiment the amplitude envelope is low pass filtered and the remaining signal is replaced by a white Guassian noise

The three experiments gave good quality and intelligibility of the synthesised words and they all sounded very natural, however, among the three the best natural sounding speech was enhanced when the phase of the speech signal was low pass filtered and white noise added.

It is important to mention here, the novelty of the use of the fractal signal in the energy of the reconstructed speech signal, which has for effect to control the naturalness sounding of speech synthesis.

The results are illustrated in figure 5 and 6, respectively where the unwrapped phase; the amplitude envelope, the original word and its reconstructed one are plotted. We can clearly notice the similarities between the waveforms of the input speech word with the reconstructed one, which confirm the accuracy and efficiency of the new used algorithm in the synthesis of speech word

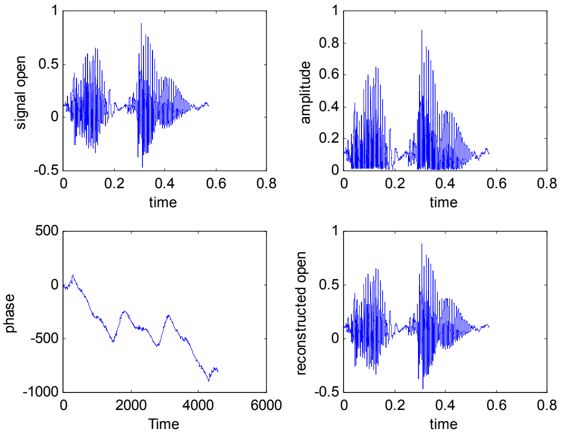


Fig.5 Synthesis results of word "open".

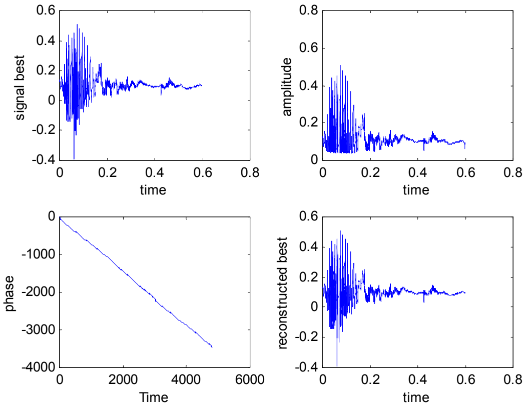


Fig.6 Synthesis results of word "best".

Results Evaluation

To test the validity of our results 10 people have been listening to the synthetic speech and their evaluation is elaborated in Figure 7.

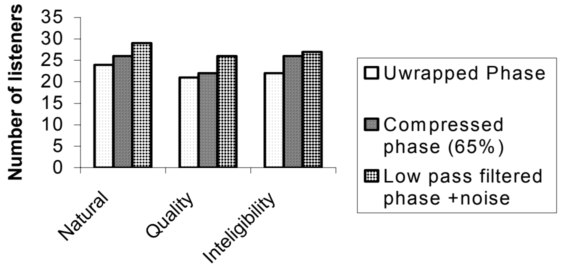


Fig.7 Evaluation of the synthesis

The statistical evaluation of the results of Figure 7 can be summarised in the following table 1:

	Quality	Naturalness	Intelligibility
Unwrapped phase	Good	Good	Good
Unwrapped compressed phase (65%)	Good	Very Good	Good
Unwrapped compressed phase with fractal synthesis	Very Good	Excellent	Very Good

Table 7.1 statistical evaluations

Conclusion

As seen in this paper the results obtained are very satisfactory. They highlight the importance of the use of the fractal dimension in generating very natural sounding speech synthesis. In fact we can notice from Figure 7 and Table 1 respectively that for the case of unwrapped phase and compressed phase methods simultaneously the quality and the intelligibility of the words is good and the difference occurs only in the naturalness where it is found to be more natural when the phase is 65% compressed from its original. On the other hand, it is clearly shown (Table 1) that the compression method with fractal synthesis, which we claim to be our novel contribution, gave the best results with a high quality and intelligibility of the word and very natural sounding words synthesis.

A new algorithm based on fractals has been used for the synthesis of speech words. The naturalness level we placed as paramount in our work was highly achieved as a result of the fractal characteristic used in the synthesis process. Despite the small size of vocabulary we used, the naturalness is very high an as the pursuit of naturalness dominates, human listening provided the best feedback.

References

[1] Kleijn K., Paliwal K., “Speech coding and synthesis”, Elsevier Science B.V., The Netherlands, 1998.

[2] Santen J., Sproat R., Olive J., Hirschberg J., “Progress in speech synthesis”, Springer-Verlag, New York Inc., 1997.

[3] Campbell N., “CHATR: A high-definition speech re-sequencing system,” *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, Dec. 1996.

[4] Huang X., Acero A., Adcock J., Hon H., Goldsmith J., Liu J., and Plumpe M., “Whistler: A trainable text-to-speech system,” in *Proc. ICSLP*, Philadelphia, PA, pp. 2387–2390, Oct. 1996.

[5] Hunt A. J. and Black A.W., “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. ICASSP*, Atlanta, GA, pp. 373–376, May 1996.

[6] Sagisaka Y., “Speech synthesis by rule using an optimal selection of nonuniform synthesis units,” in *Proc. ICASSP*, New York, NY, pp. 679–682, Apr. 1988.

[7] Jon R.W. Yi.and James R. Glass, “ Natural-sounding speech synthesis using variable-length units”, ICLSP98.

[8] McDowell P.S and Datta S. “ A fractal approach to the characterisation of speech ”, Acoustic Letters, Vol.17, No.1, 1993.

[9] Maragos P. “ Fractal Aspects of Speech Signals: Dimension and Interpolation ”, Proceedings *IEEE* international Conference on Acoustic Speech and Signal Processing (ICASSP), Vol.1, pp417-424, 1991.

[10] Fekkai S. and Al-Akaidi M. “ A Novel approach to measure fractal dimension of speech phonemes ”, Euromedia, Belgium, 1999.

[11] Fekkai S.,Al-Akaidi M., “ Word recognition based on fractal techniques ”, Proceeding of international conferences on image, science, system and technology, Las Vegas, USA, pp589-595, 1999.

[12] Srinivas J. “ Fractals classification, generation and application ”, University of Texas, *IEEE*, vol. 2, pp.1024-1027 (1992).

[13] Klatt D., “Review of Text-to-Speech Conversion for English”, Journal of theAcoustical Society of America, JASA vol. 82 (3), pp.737-793, 1987.

[14] Krger B., “Minimal Rules for Articulatory Speech Synthesis”, Proceedings of EUSIPCO92 (1): 331-334, 1992.

[15] O'Saughnessy D., “ Speech Communication - Human and Machine”, Addison-Wesley, 1987.

DEVELOPMENT OF A SPEECH RECOGNIZER FOR THE DUTCH LANGUAGE

Pascal Wiggers, Jacek Wojdel, Leon Rothkrantz
Data and Knowledge engineering,
Delft University of Technology,
Mekelweg 4, 2628 CD Delft, The Netherlands
Email: P.Wiggers@its.tudelft.nl,
J.C.Wojdel@its.tudelft.nl,
L.J.M.Rothkrantz@its.tudelft.nl

KEYWORDS
Speech recognition, Hidden Markov Models

ABSTRACT

This paper describes the development of a large vocabulary speaker independent speech recognizer for the Dutch language. The recognizer was build using Hidden Markov Toolkit and the Polyphone database of recorded Dutch speech. A number of systems have been build ranging from a simple monophone recognizer to a sophisticated system that uses backed-off triphones. The system has been tested using audio from different acoustic environments to test its robustness. The design and the test results will be presented.

INTRODUCTION

This paper describes the design and construction of a speech recognizer for the Dutch language. The system was created as a baseline system for further research on the subject of speech recognition within our group; in particular on the integration of information from multiple modalities in a Hidden Markov based speech recognizer in order to create a natural and robust human computer interface. Although a number of well-performing speech recognizers is now available, we decided to create our own system to ensure that it fits our needs and that it can easily be altered and extended.

The system was designed to recognize sounds recorded by an ordinary desktop computer microphone. To make the system as general as possible it was decided to create a speaker independent large vocabulary continuous speech recognizer and to make the system easy adaptable to new vocabularies it was decided to create a phoneme based recognizer.

The final system uses context dependent models, but as this system was build an refined incrementally, actually a whole set of recognizers has been created ranging form a simple monophone recognizer to a sophisticated multiple mixture triphone system.

This paper first describes the tools and data that were used to create the system it then gives a brief description of the development process and in the last section experimental results concerning the performance of the system are presented.

DEVELOPMENT OF THE RECOGNIZER

The system was developed using a variety of tools, programs and scripts. By far the largest set of tools and software libraries was taken from the Hidden Markov Toolkit (HTK). This is a portable software toolkit for building and manipulating systems that use continuous density Hidden Markov Models. It has been developed by the Speech Group at Cambridge University Engineering Department (Young et al. 1995). HTK provides the means to create and manipulate Hidden Markov models in general but it is primarily designed for building HMM based speech-processing tools.

The toolkit comprises a number of script-driven tools supported by a set of software libraries. Furthermore a number of programs was written that filled the gaps in the development process not covered by any of the tools. This includes a program for data selection, some tools for creating an initial acoustic model set and a tool for creating scripts for triphone clustering.

The system was built in four stages. First data for training and testing was selected and prepared. Then a simple monophone system was created. This system was subsequently refined and finally a number of evaluation tests were conducted to measure the performance of the system. Each of these steps will be described in the following sections.

Data preparation

The training and testing data for this project was taken from the Dutch Polyphone database (Damhuis et al 1994; Boogaart et al. 1994). This is a rather large corpus containing telephone speech from 5050 different speakers in 222075 speech files, based on 44 or in some cases 43 items per speaker. The speakers were selected from all dialect regions in the Netherlands and the ratio between male and female speakers is almost fifty-fifty. The utterances contain all Dutch phonemes in as many phonetic contexts as the designers of the database could find.

As the Polyphone database was recorded with automatic voice-interactive telephone services in mind most speech files contain examples of phrases useful for this kind of applications, this includes street names, bank-accounts, numbers and answers to yes-no questions. Training a large vocabulary recognizer on these samples may result in a recognizer that performs well on recognizing numbers and 'yes' and 'no' but which generalizes very poor to other words. To avoid these problems only nine items per person were used. All sentences were selected from newspaper articles. Five of these sentences belonged to the group of phonetically rich sentences that were selected to cover as many phonetic contexts as possible. The other four sentences were selected because they contained common frequently used application words.

The telephone recordings contained many samples of poor quality or contained background noise or even background speech. To ensure well-trained models that can be used for recognizing speech recorded for example using a PC microphone only utterances that were spoken by native speakers and which did not contain any background speech, background noise, no stuttering or disturbing hesitations and no mispronunciations or foreign pronunciations were selected. Mouth noises, like smacking, sniffing, loud breaths and verbal hesitations however were allowed.

Selection of the utterances that adhered to this profile was done automatically using meta-information provided by the Polyphone database. For each selected utterance a word level transcription was created.

Three different data sets were extracted from the Polyphone database this way. A training set containing 22626 utterances, a development set containing 2673 utterances, which was used for testing and fine tuning during development of the system and an evaluation test set comprising 2885 utterances, to evaluate the final performance of the system. The development set contained persons and sentences that did not occur in the training set. And the evaluation test set contained persons that did neither occur in the training set nor in the development set, but its phonetically rich sentences did also occur in the training set.

Bigram language models for testing and evaluation were calculated using the training data. The models were smoothed to incorporate words that did not occur in the training data using a backing-off scheme. Part of the probability mass from other words was transferred to these words.

To test how well the final system would adapt to other environments we recorded a small data set using a digital video camera (Wojdel, 2001). This set contains data from 5 different persons, all of them computer science students at TU Delft, four male students and one female student. From each person four or five recording sessions were used. Each session contained 23 sentences, ten of which were phonetically rich sentences, similar to those in the Polyphone database. The other sentences contained either a sequence of short words, a sequence of numbers, a spelled word or a command from a telebanking application.

Since the Polyphone utterances were encoded in 8-bit A-law wave format we converted our recordings, that were originally in 16-bit 44 kHz stereo wave format, to A-law format. Subsequently, all utterances were encoded to Mel-frequency cepstral coefficient vectors. Each vector contains twelve cepstral coefficients with log energy and delta and acceleration coefficients added, all scaled around zero by subtracting the cepstral mean from all vectors. This resulted in 39 dimensional feature vectors. A sampling rate of 10 ms was used and each vector was calculated over a segment of 25 ms.

The phonemes from the SAMPA set (Boogaart et al. 1994) were adopted as phoneme set for the recognizer. Three special purpose phonemes were added. The first, *sil*, models (longer periods of) silence that occur between sentences or when a person is not speaking at all. The second phoneme, *sp*, also represents silence, but only optional periods of short duration, like the silences that occur between words. The thirth phoneme that was added *mn*, models all kinds of mouth noise and verbal hesitations. The idea behind the inclusion of this phone was that real, natural speech always contains mouth noise and modeling this may improve the results in real-life environments. Verbal hesitations like ‘uh’ and ‘ehm’ were modeled by including them in the dictionary like any other word. The resulting set contained 45 phones.

For each of these phones a Hidden Markov Model was created. All models except the silence models shared the same topology, which consisted of non-emitting start and end states and three emitting states using single Gaussian density functions. The states are connected in a left-to-right way, with no skip transitions. Each state has a transition to itself. The model is shown in figure 1.

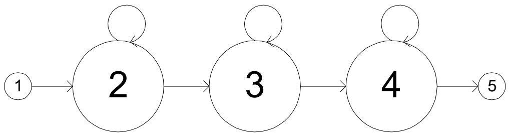


Figure 1: Acoustic HMM Topology

This is a rather simple model topology, but earlier experiments with three and five state models showed that the three state model performs as well, and in some cases even slightly better than the five state model, despite the fact that it has fewer parameters (Wiggers, 2001).

Training

As only unsegmented training data was available we initially set the mean and variance of all the Gaussians of all the models to the global mean and variance of the complete data set. These models were then trained using embedded Baum-Welch re-estimation. In the embedded re-estimation algorithm for each utterance the transcription is used to create a composite HMM which spans the whole utterance by concatenating instances of the phone HMMs corresponding to each label in its transcription. The

Forward-Backward algorithm is then applied. When all of the training files have been processed, the new parameter estimates are formed and the updated HMM set is created. This way no boundary information for the acoustic models is needed.

To make the system robust the *sp* silence model was added only after the models received some training. This was done because the pause model, which models optional short periods of silence, is very susceptible to picking up vectors belonging to neighboring phones. So in this particular case it is beneficial to explicitly define what type of data this model is supposed to represent. This was realized by creating a one state *sp* model, which has a direct transition from entry to exit node. It was then tied to the middle state of the silence model, so these two states now shared the same set of parameters. To the silence model *sil* transitions were added from the second state to the fourth state and back from the fourth state to the second state to make sure that the model could handle great variations in durations, from milliseconds up to a few seconds. This also had to be done after initial training to prevent that the silence model absorbed large parts of the utterances.

After re-estimation the models created so far were used to create new transcriptions using pronunciations that fitted the acoustic data embodied in the models. This was done by performing Viterbi alignment. These new transcriptions were then used in subsequent re-estimation cycles.

The performance of the single Gaussian monophone system that was created this way was tested by using the Viterbi decoding algorithm on the development test set. The transcriptions output by the Viterbi algorithm were compared to the original word level transcription files by a analysis tool, that uses a dynamic programming based string alignment procedure that is fully compatible with the one used in the standard US NIST scoring package. The percentage of word that was recognized correct was calculated, as was the so-called word accuracy. This measure compensates for the fact that the scoring algorithm may align incorrect inserted words with correct words by subtracting the number of insertion errors from the number of recognized words before calculating percentages.

In this test and in all other tests conducted during development that are described below, the first half of the development test set was used. This subset comprises 240 utterances spoken by about 100 different persons. The subset contained 1060 different words. A bigram language model containing these words was used in these tests. Table 1 shows the recognition results from the monophone system. To show the progress that has been made during training the results of various steps are included. Although the improvements made were considerable the overall results were very poor. This is due to the simple model topology and the fact that the system does not take into account linguistic effects like coarticulation and it does not make up for possible unbalances in the training data.

Table 1: Results Monophone System

System	Percentage of words recognized	Word accuracy percentage
Initial model set	17.15%	-77.17%
Fixed silence models	30.00%	-48.75
Monophone system	38.34%	-28.42%

To find more realistic density functions for the acoustic models Gaussian mixture densities were used. These were obtained from the simple Gaussian densities by iteratively incrementing the number of mixtures. For the distribution in a state the mixture with the largest weight was split until the required number of components was obtained. To prevent defunct mixtures with zero weight a floor mixture weight was defined, mixture weights were not allowed to fall below this floor. It turned out that during training the number of floored mixtures increased rapidly. To reduce this number and prevent overfitting the data the minimum number of examples necessary to allow for mixture incrementing of a model was also incremented in each step. The mixtures were incremented in seven steps until a 15-mixture system was obtained. The recognition results for these systems on the development test set are shown in table 2.

Table 2: Multiple Mixture Systems

Number of mixtures	Percentage of words recognized	Word accuracy percentage
2	40.11%	-25.42%
3	42.70%	-16.33%
5	45.74%	-7.57%
7	47.92%	-1.81%
10	51.54%	4.77%
12	54.50%	9.54%
15	56.81%	15.51%

To capture coarticulation effects context dependent models were introduced. In this project it was decided to use word internal triphones (which implies the use of biphones at word boundaries). These were created by cloning the monophone models as often as needed and creating triphone transcriptions to re-estimate the new model set. This way a set containing 8570 triphones was obtained. For many of these models there only was a single example in the training data, so the models could not reliably be estimated.

To find the right balance between the number of models and their modeling accuracy data-driven state clustering was used to obtain a smaller set of generalized triphones. A weighted Euclidean distance between the means of the Gaussian distribution functions was used to create clusters for each of the three model states. 15 different clusterings were tried using different minimum cluster sizes to find the right number of triphones. Out of these clusterings five different systems were build and re-estimated. For each system the transition matrices of the triphones that corresponded to the same monophone were tied. Models that ended up having their states in the same clusters and which had the same transition matrix were tied. Thus in this case one physical model represents several logical triphones.

The systems were tested using the development test set and the corresponding bigram. As is clear from table 3 the recognition results improved as more triphones were used.

Table 3: Triphone Systems

Number of triphones	Percentage of words recognized	Word accuracy percentage
101	40.44%	-25.46%
563	46.48%	-10.28%
1050	49.57%	-4.77%
2526	54.92%	5.68%
8570	62.32%	18.59%

Unfortunately this also means a larger model set. The system containing 2526 triphones booked fairly reasonable results. It had at least three models per cluster and in most cases more. Each cluster had at least four examples in the training data. This system seemed to provide a good balance between the number of parameters and the modeling accuracy. It contained less than one third of the original triphones, but was still large enough to model different contexts, therefore it was chosen to be further developed during subsequent steps.

However, before these steps could be performed one problem had to be solved. A limitation of the data-driven clustering technique is that it does not deal with triphones for which there are no examples in the training data. This may be avoided by careful design of the training database but a little research showed that this was not an option in this case. The training data contained 8570 triphones out of 10205 in the dictionary that was used. Redistributing the available data among the data set would not have solved the problem as the evaluation data set and the test data set together only contained 394 additional triphones. Furthermore there was no guarantee that our dictionary contained all possible triphones.

A well known solution to these problems is the use of decision tree based clustering (Woodland et al., 1994a; Jelinek, 1999), however, this is a knowledge based approach, using phonetic knowledge to classify triphones and we were interested in using a completely data driven approach. Therefore we introduced a technique that can be described as 'backed-off triphone approach'. In this approach the original monophone models augment the triphone model set. During recognition the word network is constructed by inserting the HMMs in the language model, triphone models are used whenever available, otherwise the corresponding monophone is used. This is implemented by tying all triphones that have no model of their own to the corresponding monophone. Essentially the monophones become generalized triphones. Of course they are less specialized than the other generalized triphones because they are trained on all corresponding triphones but the ones they represent, but being monophones they are general enough to cover the unseen triphones. The overall result is a robust recognizer that uses triphones most of the time and does not break down when an unknown triphone is encountered

The triphone models still had Gaussian distribution functions. As with the monophone system the mixtures were incremented in steps of two or three mixtures a time, with subsequent reestimation cycles. This process was stopped at 17 mixtures because the relative gain in performance introduced by additional mixtures got to small. A 17-mixture system performed almost as well as a 19-mixture system. The recognition results of these systems are shown in table 4.

Table 4: Multiple Mixture Triphone Systems

Number of mixtures	Percentage of words recognized	Word accuracy percentage
2	56.27%	8.02%
3	58.82%	15.63%
5	61.00%	22.21%
7	64.05%	26.82%
10	66.64%	30.93%
12	68.61%	34.92%
15	70.55%	38.46%
17	71.00%	39.57%

EVALUATION EXPERIMENTS

The final set of acoustic models consisted of the 17-mixture generalized triphone set and the 15-mixture monophone set. This system was tuned using the development test set. A grammar scale factor was used to regulate the relative influences of the language model and the acoustic model. The optimal system, which relied mainly on its acoustic models, had a word recognition percentage of 95.27%, a word accuracy of 89.59% and 38.43% of all sentences were recognized correctly.

These results were obtained on the test set that was used during the development process to tune several parameters therefore they are likely to be to optimistic. To test the robustness of the system and to see how it will perform on other data we did a number of evaluation tests.

First recognition was performed on part of the evaluation test set. This set contained 100 sentences, each of which was spoken by a different person. The bigram language model used contained 5017 different words. The result of this and subsequent tests is shown in table 5. The experiment showed that the recognizer generalizes very well to speakers it was neither trained nor tuned on even when a large word network is used. Although the evaluation data did not occur in the training or development test set it also came from the Polyphone database, so it was recorded under similar conditions as the other two sets. In particular it was recorded over a telephone line while our recognizer should be able to recognize speech recorded by a PC microphone and it should be easily adaptable to specific tasks and applications.

Table 5: Evaluation Results

Test	Percentage of words recognized	Word accuracy percentage	Percentage of sentences correct
Develop. Set	95.27%	89.59%	38.43%
Eval. Set	93.55%	88.76%	32.56%
Other data	87.30%	84.59%	36.36%
Adapted	96.76%	95.41%	60.61%
4 person adapted	91.80%	93.44%	47.62%
Adapted, Polyphone input	32.16%	-21.44%	36.36%
Grammar	75.30%	72.59%	68.00%

To test how well the recognizer generalized to other data and other environments the data set we recorded ourselves was used. The data set was split in a training set containing about 500 utterances, that is, about 100 per person and a test set containing 30 utterances.

First the system was tested without any further training (second row in table 5). Although still reasonable the performance clearly decreased in comparison to the performance of the systems described in the last two paragraphs. These results were to be expected, since the data set was recorded using a video camera in a quiet laboratory, while the Polyphone data on which the system was trained was recorded over a normal telephone line. So the ambient noise, which is also modeled in the acoustic HMMs, will be quite different, which means that the acoustic HMMs will have slightly different distribution functions. As a result the acoustic vectors 'produced' by the HMMs are still similar to the acoustic vectors in the data set, but there is some distortion, causing classification mistakes. To make up for this effect, the system was adapted to the new situation by re-estimating twice, using the training part from the recorded data set.

As can be seen from table 5 the performance considerably increased. Actually, it is better than any of the results obtained in earlier tests; especially the word accuracy and the percentage of correct sentences showed a large improvement. The explanation for these results lies in the fact that the system not only adapted to the background noise in this data but it also adapted to the voices of these five persons. In fact the system has become speaker dependent.

The voices it adapted to it recognized very well, but now recognition of other speakers might give some trouble. To show these effects, two more experiments were performed. In the first experiment the system was adapted using only 4 different persons. Recognition was performed using data from the fifth person. The results show that although the performance is not as good as in the previous test, the system is still better than the undapted system. So the system adapted to the new environment but is still capable of generalizing and performing speaker independent recognition.

In the second experiment recognition was performed once again on the Polyphone test set using the system that was adapted to four persons. The recognition results were dramatic, showing that the system no longer recognized the data it was developed with; it completely adapted to the new environment. That the performance is this low is due to the fact that the utterance in the Polyphone database contains much more noise than the utterances used here, since they are recorded over a telephone line. Recognizing the PC recorded data with an unadapted Polyphone trained system worked because from the systems point of view these were just very high quality recordings. But form the point of view of the adapted system the Polyphone data set contains very noisy recordings, indeed many sounds were classified as mouth noise.

Our data set also contained sentences that adhered to a telebanking application grammar. This application allowed people to manage their bank accounts and conduct financial transactions by telephone. A language model was created that implemented this grammar. The last row of table one shows the results that were obtained from tests with this language model and the corresponding sentences from the data set. The adapted acoustic models were used.

One would expect the percentage of correct words to be higher as only a small vocabulary is used and the syntax of the sentences is constrained. A sentence by sentence inspection of the results showed what was actually going on here. In most cases the system did recognize the right sentence, but when it made a mistake, it often recognized a completely wrong sentence in which only a few words were correct, because the grammar forced the Viterbi algorithm to take a specific path. So about 25 percent of all sentences are responsible for most word errors.

CONCLUSIONS

In this paper we described the development of a large vocabulary speech recognizer for the Dutch language. The system was build using an incremental approach that started with a simple single Gaussian monophone system, which was refined in a number of steps. The final system uses backed-off triphones that solve the problem of unseen triphones without the need for specific linguistic knowledge. The final system performs well, more than ninety percent of all words are recognized correctly. The errors that occur are usually small and typically involve wrong conjugations of a verb or hesitations by the speaker. A more powerful language model or a postprocessing module that checks the syntax of the sentences could reduce this kind of errors. The recognizer can cope with noises like smacking or loud breathing and the system is speaker independent. It has been tested with vocabularies of more than 5000 words, the performance decreased only slightly in this case. The system can be adapted to other environments and performance can be further improved, especially on sentence recognition, by making it person dependent.

REFERENCES

Boogaart, T.I., Bos L., Boves, L., *"Use of the Dutch Polyphone Corpus for Application Development"*, *Proceedings 2nd IEEE workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 145-148, 26-27 September 1994, Kyoto, Japan

Damhuis M., Boogaart T., in 't Veld, C., Versteijlen, M.,W. Schelvis, W., Bos, L., Boves L., "Creation and Analysis of the Dutch Polyphone Corpus", *Proceedings International conference on Spoken Language Processing*, (ICSLP) '94, pp. 1803-1806, 18-22 September 1994,Yokohama, Japan

Grocholewski, S. "Acoustic Modeling for Polish*"*, *International Workshop Speech and Computers*, SPECOM, 2000

Jelinek, F., "Statistical Methods for Speech Recognition*"* (Language, Speech, and Communication) MIT Press, January 1999

Rabiner, L.R., Juang, B. H., "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, N.J., 1993

Young, S., Kersaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P. C., "The HTK Book" (for HTK Version 3.0*)*, Cambridge University Engineering Department

Wiggers, P. "Hidden Markov Models for Automatic Speech Recognition and their multimodal applications*"*, Master Thesis, Delft University of Technology, 2001, The Netherlands

Wojdel, J. "The Audio-Visual Corpus for Multimodal Speech Recognition in Dutch Language*"*, Internal report, Delft University of Technology, 2001, The Netherlands

Woodland P.C., Odell, J., Young S.J., "Tree-Based Tying for High Accuracy Acoustic Modelling", *Proceedings ARAPA Human Language Technology Workshop*, 1994

Woodland, P.C., Odell, Young, S.J., "Large Vocabulary Continuous Speech Recognition Using HTK*"*, *Proceedings International Conference on Acoustics Speech and Signal Processing*, 1994

KNOWLEDGE BASED SPEECH INTERFACING IN THE SWAMP PROJECT

C.K.Yang, L.J.M.Rothkrantz
Data and Knowledge engineering,
Delft University of Technology,
Mekelweg 4, 2628 CD Delft, The Netherlands
Email: L.J.M.Rothkrantz@cs.tudelft.nl

KEYWORDS
Speech interface, dialog management

ABSTRACT

Speech technology is rapidly developing and has improved a lot over the last few years. Nevertheless, speech-enabled applications have not yet become mainstream software. Furthermore, there is a lack of proven design methods and methodologies specifically concerning speech applications. So far the application of speech technology has only been a limited success. This Paper describes a project done at CMG Trade Transport & Industry BV. It is called SWAMP and is an example of the application of speech technology in human-computer interaction. The reasoning model behind the speech interface is based on the Belief Desire Intention (BDI) model for rational agents. Other important tools that were used to build the speech user interface are the Microsoft Speech API 5 and CLIPS.

INTRODUCTION

Speech is the most common mode of communication between people. Although speech communication is not a perfect process, we are able to understand each other with a very high success rate. Research has shown that the use of speech enhances the quality of communication between humans, as reflected in shorter problem solving times and general user satisfaction (Chapanis, 1981). Furthermore, speaking to humans subjectively seems to be a relatively effortless task (Nusbaum et al, 1995). The benefits mentioned above are some reasons that have moved researchers to study speech interaction systems between humans and computers.

In September 1999 CMG Trade, Transport & Industry BV started the Wireless Automotive Messaging (WAM) project. Its purpose was to develop new wireless services in the field of traffic and transport. The WAM application is based on the Client-Server model. The server is stationary while the client travels with the user in his car. Because the clients are mobile, communication is based on wireless techniques.

This paper discusses the SWAMP (Speech Interfacing in the Wireless Automotive Messaging Pilot) project started in October of the following year. The purpose of the SWAMP project was to analyse if a speech interface is better suited for the WAM pilot. Therefore the WAM client is extended with a speech interface: the SWAMP client. This offers a way for the driver to interact with the system while his hands and eyes remain free, ideal for car driving situations.

THE SWAMP CLIENT

Speech interaction between the user and the SWAMP application is based on dialogues. Generally, the user starts a speech interaction by indicating (via speech) what his desires are. The system then leads the user through a dialogue in which it tries to retrieve information regarding these desires. If eventually all the necessary information is collected, the application takes the appropriate actions to realise the user's desires.

The general assumption behind the speech interface is that the user wants to accomplish something with his utterances, i.e. he has a certain goal in mind. The set of all services the SWAMP application has to offer is just a subset of all the goals the user can possibly have. Goals that don't correspond to a service, however are beyond the domain of the speech interface and are ignored.

The speech interface is divided into 3 components.

1 The speech recognition or ASR component:
Its function is to recognise the user's utterance and transform it into a format that can be processed.
2 The dialogue management component:
Its function is to process the input from the speech recognition component to figure out what the user wanted to accomplish and take the appropriate actions to realise the user's wishes.
This component is the main focus in this paper.
3 The speech synthesis or TTS component:
Its function is to generate speech output to the user.

Figure 1 gives a graphical overview of how the speech interface is implemented. The main application is the original WAM client modified in such a way that it can communicate with the dialogue manager.

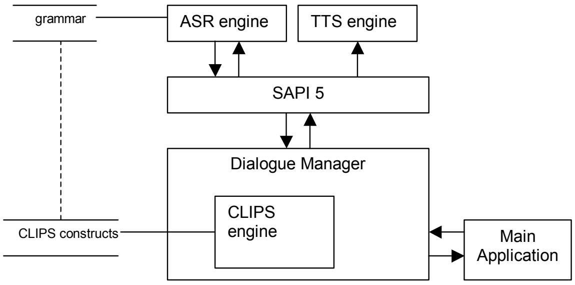


Figure 1: Overview of SWAMP Implementation

The SWAMP client is implemented in C++ (Microsoft Visual C++ 6.0 enterprise edition). Initially it was the intention to build the speech interface to run on Windows CE but due to limitations in software and hardware of handheld computers, Windows NT was ultimately chosen.

The Microsoft Speech Application Programming Interface 5.0 (SAPI5) is used as middle-ware between the engines and the SWAMP application. SAPI5 acts as a communication layer between the dialogue manager and the speech resources (ASR and TTS engine). It takes care of hardware specific issues such as audio device management and removes implementation details such as multi-threading. This reduces the amount of code overhead required for an application to use speech recognition and synthesis. Another advantage of using middle-ware is that the choice of the final ASR and TTS engine can be postponed till a later stadium (e.g. until there is more budget for better engines).

The CLIPS expert system tool is designed to facilitate the development of software to model human knowledge or expertise. CLIPS is embedded in the SWAMP client. It can be viewed as the knowledge processing and management unit of the dialogue manager.

DIALOGUE DESIGN

With each initial utterance from the user, the speech interface tries to find the corresponding service involved. Once the goals of the user are clear it tries to accomplish the service by checking whether all the information needed is available. If this is not the case, the speech interface must initiate a dialogue to retrieve the required information from the user until the task can be performed. All possible dialogues that the speech interface can be involved in must be designed beforehand. This includes speech prompts for each situation, and all possible user responses on those prompts. Furthermore design involves the definition of a grammar that captures the syntax of whole conversations into a few simple grammar rules.

Design approach

The goal of the speech interface is to give a user access to the SWAMP services by means of simple speech interaction. To achieve this, one can choose between two different approaches: 1) demand a longer learning time for the speech interface and require the user to adapt his speaking style or 2) make it easy for the user by allowing an extensive grammar and modelling more and more complex dialogues so that the user can speak to the system as with another human.

Speech User Interface (SUI) designers have learned that humans are extraordinarily flexible in their speech and readily adapt to the speaking style of their conversational partners. This is not a new finding: think about how easily we adjust our speech depending on whether we are speaking to children or other adults. This flexibility has useful implications for designing the speech interface: after extensive use of the speech interface (as the user gets acquainted with the grammar and has more experience) some dialogues become less and less common. Since the user will adapt his style of interacting and refrain to only those dialogues that were successful in the past. Because of this finding and the choice of our typical user ("he is familiar with current computer technology") the first approach was chosen: only model the most common utterances and let the user adapt to it.

Dialogue representation

Without a proper representation technique, the dialogues can quickly become very complex and unmanageable. In this project dialogues are represented by flow diagrams containing nodes representing start/begin points of a dialogue, boxes representing actions (e.g. an utterance from a user or an action from the system), diamonds representing decisions point and arcs to connect the nodes, boxes and diamonds. A dialogue always begins with a start node and ends with an end node. Within these nodes, the dialogue travels from box to box along the arcs and branching at the decision diamonds. A successful dialogue corresponds to a path in the flow diagram from the start node to the end node.

Speech dialogues are context sensitive. In our representation, the context is defined by the positions within the dialogue flow. Each box represents a certain state or context. The arcs branching from a box indicate the options available within that context and the branches leading to a box define how that context can be achieved.

The power of above dialogue representation technique lies in the fact that dialogues are represented in a generic way. E.g. the (user action) boxes define what the user can say at that moment in the dialogue, but not how it must be said (this is defined in the grammar). In this way, a single path in the dialogue flow diagram can represent whole categories of similar dialogues.

A well-modelled dialogue flow diagram is one where each possible dialogue flow can fit in. This implicates that common communication errors, such as misunderstandings, should be modelled as well as mechanisms for correcting and preventing these errors, such as requests for confirmation and roll back. Table 1 shows an example dialogue for the kilometre registration (KM registration) service. The flow of this dialogue fits into the flow diagram in Figure 2 (accentuated).

Table 1: Example Dialogue

U: Change trip type
S: Is it a business or a private trip?
U: It's a business trip?
S: OK, what's the project ID for this business trip?
U: Project ID is SWAMP
S: Do you want to set the project ID to SWAMP?
U: Yes
S: OK, trip type is set.

In practice the dialogues can become so complex and the dialogue flow diagrams so large that it is best to split them up into one main dialogue and several smaller sub dialogues. For each sub dialogue a separate dialogue flow diagram is designed and referred to in the main dialogue flow diagram (by means of sub dialogue nodes). Another use for the dialogue flow diagrams occurs during the testing phase. Since each path from the start node to the end node corresponds to a successful dialogue. The correctness of the implementation of the dialogues can easily be verified if all the paths in the dialogue flow diagrams can be traversed

Grammar

The SAPI5 design specification requires the grammar of an application must be a context-free grammar (CFG) written in a format specified in the SAPI5 grammar schema. This schema describes the SAPI 5.0 speech recognition grammar format and is based on the XML framework. The ASR engine uses the CFG to constrain the words contained in the user's utterance that it will recognise.

Basically the grammar file consists of a set of grammar rules in the grammar schema syntax. The complete specification of the schema can be found in the SAPI5 online help. Grammar rules can have an activation state, which can be set to active of inactive. SAPI5 recognises active rules and conversely does not recognise deactivated ones. The application may change the state of the rules during execution. So if a rule is no longer needed, it may be deactivated.

In order to indicate the functional parts of a sentence i.e. the parts that actually contain relevant information, the CFG can be extended with semantic information declared inside the grammar. This enables the ASR engine to associate certain.

Recognised word strings with name/value-meaning representations. The dialogue manager then applies these meaning representation associations to understand and control the dialogue with the user.

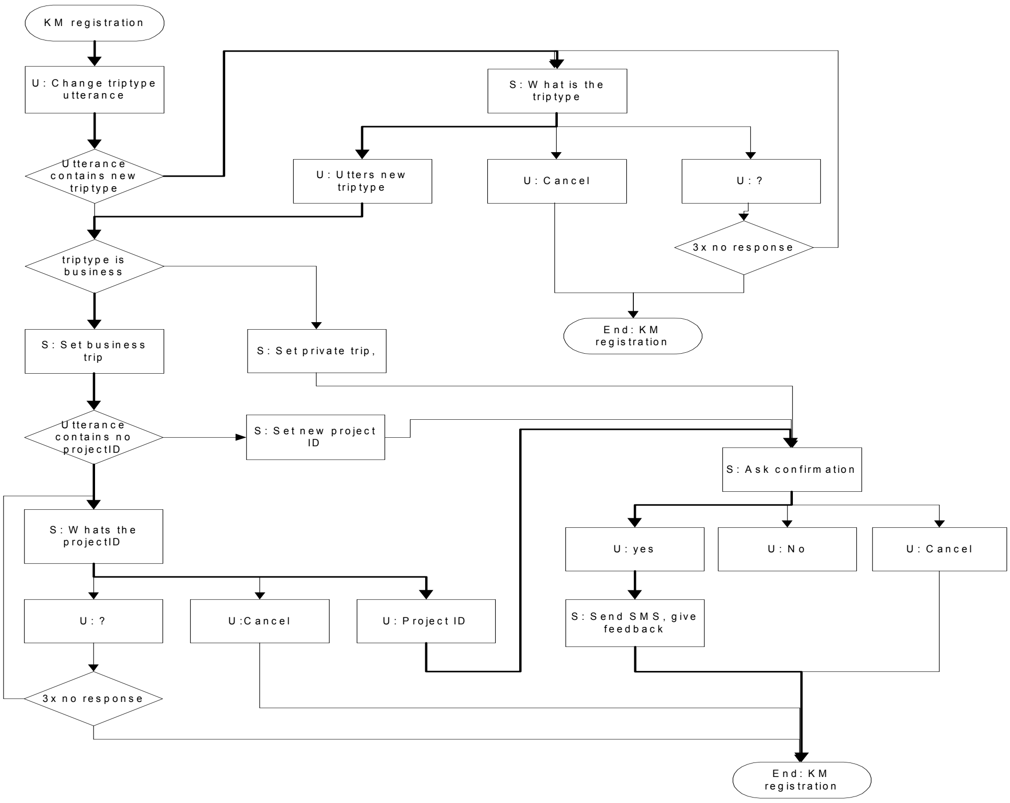


Figure 2: Dialogue Flow Diagram for the KM Registration Service

The grammar rules are derived from a corpus of utterances by hand. Crucial in this process is the determination where the relevant information is located within an utterance. Once this is accomplished, the derivation process is straightforward.

THE REASONING MODEL

In the search for a suitable reasoning model for the dialogue manager: one that is capable of adequately describing the reasoning behaviour of the dialogue manager, the Belief-Desire-Intention (BDI) model (Wooldridge, 2000) was chosen. In an implementation of a dialogue manager according to this model, the dialogue manager continuously executes a cycle of observing the world and updating its beliefs, deciding what intention to achieve next, determining a plan of some kind to achieve this intention, and then executing the plan.

There exist a correspondence between concrete CLIPS data structures and the attitudes in the BDI model. Beliefs in the BDI model are implemented as facts and rules in CLIPS. Facts are used to construct the dialogue manager's internal representation of the world. Facts can be seen as propositions and thus can only consist of a set of literals without disjunction or implication. Therefore special rules (belief rules) are used to complete the representation of beliefs. Belief rules represent the general relationship between facts (e.g. IF utterance=help THEN AlertLevel=high).

One way of modelling the behaviour of BDI reasoning (Rao and Georgeff, 1995) is with a branching tree structure, where each branch in the tree represents an alternative execution path. Each node in the structure represents a certain state of the world, and each transition a primitive action made by the system, a primitive event occurring in the environment or both. In this formal model, one can identify the desires of the system with particular paths through the tree structure. The above description of the branching tree structure is logically similar to the structure of the dialogue flow diagrams described in the previous section. In fact, both structures represent exactly the same: a path through the dialogue flow diagram is a successful dialogue, which is also a desire and therefore a path through the branching tree of the BDI reasoning model. As a result, the dialogue flows diagrams can be treated as the structures that describe the behaviour of the dialogue manager. They are directly implemented in CLIPS rules; each rule corresponds to a branch in the dialogue flow. Rules are both the means for achieving certain desires and the options available for the dialogue manager. Each rule has a body describing the primitive sub goals that have to be achieved for rule execution to be successful. The conditions under which a rule can be chosen as an option are specified by an invocation condition. The set of rules that make up a path through the dialogue flow, correspond to a desire.

The set of rules with satisfied invocation conditions at a time *T* (the set of instantiated rules) correspond to the intentions of the dialogue manager at time *T*. Obviously the intentions of the system are time dependent. The dialogue manager adopts a single-minded commitment strategy, which allows continuous changes to beliefs and drops its intentions accordingly. In other words the intentions of the system can be affected by the utterances of the user in contrast to blind commitment in which an intention is always executed no matter changes in beliefs.

AN EXAMPLE

In the previous section it was shown that the desires of the dialogue manager can be represented by dialogue flow diagrams. The flow diagrams are systematically translated into an executable system formulated in CLIPS rules. This section discusses the implementation of the desires. In particular the heuristics used for the translation from dialogue flow diagrams to CLIPS rules.

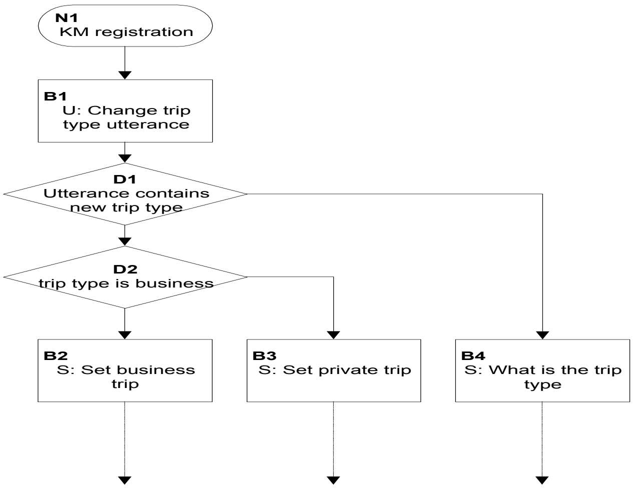


Figure 3: Part of the dialogue flow diagram for KM registration service

Suppose we must transform a dialogue flow diagram as in Figure 3. This dialogue is initialised when the user utters a phrase that matches the grammar for a change trip type utterance (box **B1**). Notice that box **B1** has 3 branches (to the boxes **B2**, **B3** and **B4**), furthermore we see that the action in **B1** is a speech action from the user. From this we conclude that the dialogue flow should be implemented using 3 speech rules. The invocation conditions for each rule are the evaluated values of the expressions in the decision diamonds **D1** and **D2**. The body of each rule contains the actions specified in the corresponding destination boxes. Furthermore, the body of the rules also contain actions to anticipate what follows after the action e.g. after box **B4** the user must supply the new trip type so the grammar rules for trip type utterances should be activated.

```
1    (defrule KM_Registration_Business
2       ?in<-(RECOGNISED 161 VID_
              KMREG_TRIPTYPE 50 TripType
              business)
3       ?pos<-(POSITION MAIN RUNNING)
4       =>
5       (printout t "<SAY>Do you want set
         the triptype to business?</SAY>
6           <ACT>VID_YESNO</ACT>
7           <DEACT>"?*Mainrules*"</DEACT
9           <REACT></REACT>" crlf)
10      (retract ?in)
11      (retract ?pos)
        (assert (POSITION MAIN KMREG))
        (assert (WANT CONFIRM))
        (assert (QUESTION KMREG business))
    )
```

Figure 4: CLIPS rule - part of the KM registration service

The CLIPS rule in figure 4 corresponds to the branch from box **B1** to **B2** (figure 3). The keyword RECOGNISED in line 2 indicates that a user's utterance is recognised. The grammar rule that matched the utterance is VID_KMREG_TRIPTYPE. Furthermore, the property name TripType with value business satisfies condition **D1** and **D2**. The actions taken satisfy **B2** (between the <SAY> tags in line 5) and anticipate future utterances of the user by activating the VID_YES_NO grammar rule and de-activating all other main grammar rules. Thereby limiting the user's input to only boolean values. The other actions in the rule body are used to update the internal representation of the world.

CONCLUSIONS

The model presented here allows for man-machine speech interaction. Indeed the speech interface of the SWAMP application implemented according to this model is capable of handling simple dialogues with the user. The dialogues are described and visualised as generic flow diagrams resembling branching tree structures (Rao and Georgeff, 1995). The chosen representation technique has also contributed greatly to the containment of the complexity in the dialogue models. Furthermore it allows an easy translation to executable CLIPS rules.

REFERENCES

Chapanis, A. "Interactive Human Communication: Some lessons learned from laboratory experiments", In: Shackel, B. (eds). "Man-Computer Interaction: Human Factor Aspects of Computers and people", Rockville, MD: Sijthoff and Noordhoff, page. 65, 1981.

Nusbaum, H., et al., "Using Speech recognition systems: Issues in cognitive Engineering", In: Syrdal A. et al. (eds), *Applied Speech Technology*, Boca Raton, CRC press, page. 127, 1995.

Rao, A. and Georgeff, M., "BDI Agents: From Theory to Practice", *in Proceedings of the First International Conference on Multi-Agent Systems* (ICMAS-95), 1995

Wooldridge, M., "Reasoning about Rational Agents", The MIT Press, Cambridge, Massachusetts, 2000..

MEDIA DISTRIBUTION AND ENCRYPTION

DISTRIBUTING MUSIC FROM IP NETWORKS TO UMTS TERMINALS: AN EXPERIMENTAL STUDY

Marco Roccetti Vittorio Ghini Paola Salomoni

Dipartimento di Scienze dell'Informazione
Università di Bologna
Mura Anteo Zamboni 7
I-40127 Bologna, Italy
E-mail:{roccetti,ghini,salomoni}@cs.unibo.it

KEYWORDS

Multimedia Services, Music-on-demand, Wireless Internet Applications, UMTS

ABSTRACT

We have implemented an Internet wireless application designed to support the distribution of Mp3-based songs to UMTS devices. In this paper, we report a large set of experimental measurements we have carried out to examine the performance of our developed system. The download time measurements we have obtained show that combining modern 3G mobile network technologies with an appropriate structuring of the Internet wireless applications may be very effective for the *fast* distribution of pre-recorded music to both fixed and mobile clients.

INTRODUCTION

The migration of part of the voice business to Internet service providers has been considered as a menace to the huge voice revenues coming from voice over circuits, even if it is expected that this will remain a large part of the operators' incomes until 2005 (Anquetil *et al*., 2000). Hence, it is considered that the best solution will be to offer new multimedia services mixing voice and other media on the Internet. Important revenues will come from emerging services that combine voice, video, data and music. Besides more *traditional* multimedia services, such as videotelephony, teleconferencing, instant multimedia messaging, teleworking, telemedicine, other important categories of multimedia services are emerging which include real time entertainment (e.g., movies, songs, …) interactive games (e.g., lottery, karaoke, distributed networked games, …) and infotainment (e.g., distributed learning, interactive consultations of multimedia information, …) (Anquetil *et al.* 2000). However, it is well known that, nowadays, a sound architecture for supporting multimedia services must be based on the concept of global mobility providing support for both fixed and mobile terminals. Due to the combination of the success of wireless networks and of the growth of the Internet, it is expected that there will be soon a greater demand for a mobile access to advanced Internet-based multimedia wireless services. Hence, it is easy to envisage that the success of these new communication technologies will depend on how much efficient the wireless radio access to those Internet-based enhanced application services will be (Kalden *et al*., 2000, Staehle *et al*., 2001). With third generation mobile systems using the *Universal Mobile Telecommunications System* (UMTS), a variable bit rate will be available of up to a few Mb/s, and this will permit connections supporting a wide range of existing and new Internet-based applications, including digital voice, video and data, as well as enhanced multimedia services (UMTS Forum). An important advantage of such 3G mobile network technologies is that packets originating from UMTS mobile devices can be directly transmitted to data networks based on the Internet Protocol (IP networks), and vice versa. This is due to the fact that UMTS networks support special *border nodes* (termed GSN) that use IP as the backbone protocol for transfer and routing of protocol data units.

However, in this communication scenario several important problems arise. The first problem is to decide if advanced TCP/IP based applications will be able to behave well over mobile radio communications protocols. With regard to this fact, it is important to notice that the Internet protocols TCP and IP have not been especially designed for wireless communications. Briefly, the standard Transmission Control Protocol (TCP) provides a sliding window based ARQ (Automatic

Repeat Request) mechanism that incorporates an adaptive timeout strategy for guaranteeing end-to-end reliable data transmissions between communicating peer nodes over wired connections. Since the ARQ mechanism of TCP essentially uses a "stop and resend" control mechanism for ensuring connection reliability, under question, here, is whether this TCP retransmission mechanism may trigger a TCP retransmission at the same time when the radio link level control mechanism is already retransmitting the same data. Secondly, a more significant problem of mobile wireless is that of temporary link outages. If a user, in fact, enter an area of no signal coverage, there is no way that the standard TCP protocol may be informed of this link-level outage (Huston, 2001). After having considered all the above challenges, a third and final problem is strictly related to the internal architecture of those advanced Internet-based applications that should be accessed through radio interfaces. Those applications, in fact, must exhibit a high rate of robustness and availability, since the mobile access to those applications should not be influenced by possible problems occurring at the Internet side.

In this context, we have designed, developed and experimentally evaluated an Internet-based wireless application that implements a *mobile music-on-demand service* to be enjoyed on UMTS devices. The developed application (whose functional architecture is depicted in Figure 1) permits to mobile users to download and to listen to Mp3 files through UMTS devices (Roccetti *et al.*, 2002). Specifically, our wireless application exploits the *background* traffic class of UMTS to provide support to the following categories of users:

- *Music Listeners*, i.e., single clients, equipped with a mobile UMTS device and connected to their UMTS cell, who may want to search for their favourite songs over the Internet, download them onto their UMTS devices, and playout them at their earliest convenience.
- *Music Producers,* i.e., single clients who may wish to exploit the system in order to distribute their own music songs to be listened to on UMTS devices. At the current state of the art of our system, this kind of users needs a regular wireline Internet connection in order to upload to the system their Mp3 music resources.
- *Musical Service Providers*, they may exploit the system to organise, build and maintain structured repositories of Mp3 resources over the Internet for use from UMTS devices.

The important experiences of *Peer-to-Peer* computing-based software systems, such as, for example, systems Napster, Gnutella and Freenet (Napster official site, Gnutella official site, Freenet Project Inc.), have inspired our work, but our system is essentially new, in the sense that it allows a reliable and distributed song sharing service over mobile UMTS terminals. In particular, in order to ensure both the availability and the responsiveness of our music-on-demand service, we have structured our system according to the special technology of the *replicated Web servers* (Conti *et al.*, 2001). In essence, according to this technology, a software redundancy is introduced at the Internet side, namely by replicating (some of the) the music songs composing the music-on-demand service across a certain number of Web servers which are geographically distributed over the Internet. In this context, a typical approach to guarantee service responsiveness consists of dynamically binding the service client to the available server replica with the least congested connection. An approach recently proposed to implement such one adaptive downloading strategy at the Internet side amounts to the use of a software mechanism, called the Client-Centred Load Distribution (C^2LD) mechanism (Ghini *et al.*, 2001). With this particular mechanism, each client's request of a given Web resource (e. g., a Mp3 file representing a certain song) is fragmented into a number of sub-requests for separate parts of the resource. Each of these sub-requests is issued concurrently to a different available replica server, which possesses that song. The mechanism periodically monitors the downloading performance of available replica servers and dynamically selects, at run-time, those replicas to which the client sub-requests can be sent, based on both the network congestion status and the replica servers workload.

As far as the protocol communication problems mentioned above are concerned, our wireless application has been structured based on the use of an *ALL-IP* approach (Huston, 2001), where the mobile UMTS device is simply considered as any other Internet-connected device. In essence, we have surmounted all the possible problems due to the time-varying characteristics, temporary outages, protocol interference, and high bit error rates of the radio link by resorting to a wireless session level we have developed on the top of the standard TCP protocol.

The aim of this paper is not to concentrate on the discussion of all the architectural and protocol design issues which are at the basis of the wireless application we have developed. The interested reader may find them in the already cited paper (Roccetti *et al.*, 2002). Instead, we wish to present here a complete set of experimental results that exhibit the performance of our system. With this in view, the reminder of the paper is organized as follows. In Section 2 we examines a large set of preliminary performance results we have gathered from real-world experiments by using our system. Section 3 provides some final conclusions we have drawn from our experience.

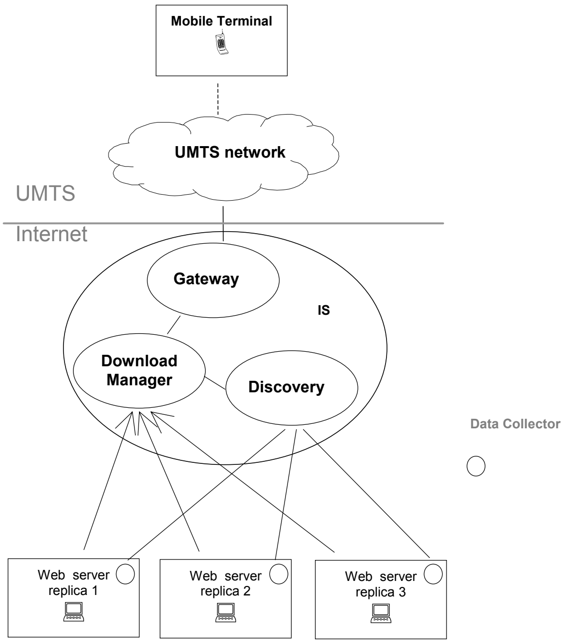


Figure 1: Music Distribution - Functional Architecture

EXPERIMENTAL ASSESSMENT

We present an experimental study we have developed in order to assess the effectiveness of our music-on-demand service. The intention behind our experimental study has been to investigate the quality of the Internet/UMTS download sessions carried out by our wireless application. In essence, we conducted around 4000 experiments consisting in the download of a set of different Mp3 files. Four different replicated Web servers were exploited, at the Internet side. Instead, the communications between the Gateway (located at the border between the Internet and the wireless UMTS link) and the mobile client was simulated by means of an UMTS simulator that is able to produce the transmission delay time of each frame at the radio link layer. Detailed information concerning the experimental models we adopted for our experiments are discussed in the following Subsections.

Application Level Model

We used four different Web servers, geographically distributed over the Internet, providing the same set of 40 different songs. The four different replica servers were respectively located in Finland, Japan, USA and New Zealand (Figure 2). Our designed Intermediate System (look at IS in Figure 1) was running over a Pentium 3 machine (667 MHz, 254 MB RAM) equipped with the Windows 2000 Server operating system, and was located in Italy (Bologna).

The UMTS network was simulated by means of an UMTS simulator provided by the "Fondazione Marconi" (a public Italian foundation for wireless computing). Finally, the UMTS device, on which the client of our application was running, was emulated by means of a Pentium 2 computer (266 MHz, 128 MB RAM) equipped with the Windows CE operating system. In order to provide the reader with an approximate knowledge of the transmission times experienced over the considered Internet links, it is worth mentioning that the Round Trip Times (RTTs), obtained with the ping routine, between the client and the four different servers (i.e., Finland, Japan, USA and New Zealand) measured, respectively 70, 393, 145 and 491 ms. As far as the downloading process is concerned, we have taken the two following basic assumptions:

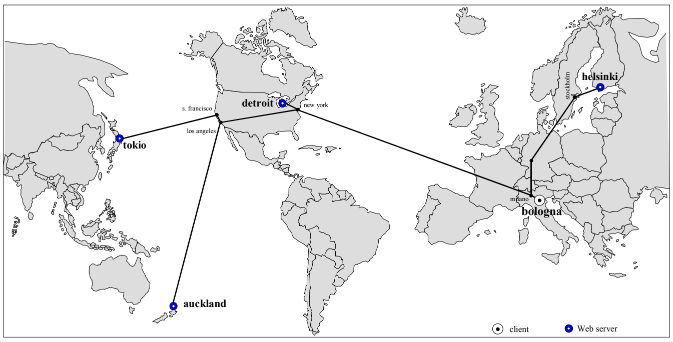


Figure 2. Web Server Replicas and Clients

1. *Mp3 file dimension:* in our experiments we used 40 different Mp3-based songs, whose correspondent file dimension ranged from 3 to 5 MB. The file dimension of 3-5 MB corresponds to the average file dimension of the songs maintained in the Napster system.
1. *Number of downloads activities*: our software application provides support to two different types of downloading services: the former consists of downloading a single song, the latter amounts to the downloading of a complete set of songs (compilation). To evaluate the performance of our system under both these circumstances, we conducted the following different experiments:
   - A set of independently replicated experiments consisting in the download of a single songs.
   - A set of independently replicated experiments, with each one consisting in the download of a set

of songs. The number of songs for each compilation ranged in the set of (3, 5, and 10). These three values were chosen based on the consideration that the average disk capacity of typical Mp3 players never exceeds 50 MB.

TCP/IP and UMTS Models

Since, currently, no real measurements of UMTS wireless data are available, in our experiments the communication between the developed Gateway system (at the Internet side) and the client application running over the UMTS device was carried out through a simulated UMTS network, by exploiting the *background* traffic class. It is well known that the UMTS protocol stack consists of: a PHY (Physical) layer, a MAC (Medium Access Control) layer, an RLC (Radio Link Control) layer that implements an ARQ mechanism for ensuring reliable data transmission, and finally, a PDCP (Packet Data Convergence Protocol) layer that provides data and header compression to improve channel efficiency. On the top of this UMTS stack we have the standard IP and TCP protocols.

The UMTS network simulator we adopted is able to return after simulations a complete *Wireless Network Transmission Time* (WNTT) value computed at the PDCP layer. Needless to say, these WNTT values depend on some operational parameters, such as the amount of traffic present in the simulated cell and the number of active clients and their speeds. WNTT measurements include also the time spent for possible retransmissions at the UMTS RLC level. In our experiments, different values of this WNTT measurements were taken based on the different possible sizes of the TCP segments coming form the Internet side (namely of 120/440/920 bytes). The unique problem that stems from this hybrid approach (i.e. both experimental and simulative) is that segment errors and resulting retransmissions at the TCP level are not taken into account. To circumvent this problem, our experiments have included the possible retransmission time delays incurred at the TCP level, by exploiting an external delay introduction mechanism that was designed to take into account the typical TCP error recovery mechanism based on received ACKs. Simply stated, this delay mechanism compares the WNTT values obtained through the UMTS simulation against the *timeout* values computed by TCP. If the simulated WNTT value is larger than the correspondent TCP timeout value, then we must conclude that a retransmission must occur at the TCP level. In such one case, the WNTT value of that given TCP segment is augmented by an additional value which is chosen as equal to the next WNTT value extracted from the set of the UMTS-based simulated values. Consequently, the TCP timeout value is updated as follows. If a retransmission at the TCP level has been detected according to the method mentioned above, then the subsequent TCP timeout value is calculated as double w.r.t. to the previously computed value. If no retransmission at the TCP level has been detected, then the traditional adaptive formula for the calculation of the TCP timeout value is followed:

$\text{Timeout} = \text{RTT} + 4 * D,$
$\text{RTT} = \alpha * \text{RTT} + (1-\alpha) * M,$
$D = \alpha * D + (1-\alpha) * |\text{RTT} - M|,$
$\alpha = 7/8,$
M = simulated value produced by the UMTS simulator,
RTT = Round Trip Time; D = Variation of RTT.

Experimental Results

This Section reports on a large set of preliminary results obtained during many experimental trials based on the above mentioned models. In particular, in the following Subsection, we present the measurements we obtained for our wireless application at the Internet side. In essence, we measured both the download time and the service availability at the Internet side (i.e., from the replicated Web servers to the Gateway). The second Subsection, instead, is devoted to present the time values obtained for the download of single songs on the wireless (UMTS) link. Finally, the last Subsection presents measurements of the downloading time values for entire set of songs (or compilations).

Downloading Time Values at the Internet side

The main objective of this Subsection is to report some preliminary results concerning the:

- Cumulative *WireLine Network Transmission Time* (WLNTT) values, that is the time spent over the wired Internet links to download a requested Mp3 file from the replicated Web servers towards our IS (see Figure 1) at the Internet side. These measurements have been compared with those that may be obtained by downloading the same Mp3 file with a standard HTTP GET method. The first row of Table 1 reports those results for Mp3 files whose size is 5 MB. The second row shows the average WLNTT percentage improvement obtained by our system that exploits the already mentioned C^2LD mechanism, w.r.t. the standard HTTP protocol. As shown in the Table, our system obtains an average percentage improvement over the fastest HTTP replica which is equal to 32 %;

- *Service Availability* (SA) values, i.e. the capability of carrying out a successful download of the requested songs (within a maximum time interval of 900s). As shown from Table 1, a full SA may be achieved with the use of the replication technology adopted by our C^{2}LD download mechanism. On the contrary, only a partial SA may be obtained by exploiting the standard HTTP protocol.

Table 1. Cumulative WLNTT and SA Results

	C^{2}LD (4 Servers)	HTTP			
		Finland	USA	Japan	New Zealand
Download time (seconds)	32.547	47.889	122.191	248.740	624.195
C^{2}LD improvement (percentage)	–	32%	73.4%	86.9%	94.7%
Successful download percentage	100%	98.5%	99.5%	95%	89%

Downloading Time Values for Single Songs on UMTS links

This Subsection is devoted to examine the Cumulative *Wireless Network Transmission Time* (WNTT) values that we have obtained, through UMTS simulations, to download single Mp3 files to the mobile devices on the wireless links. In particular, Figures 2 and 3 show the WNNT values (respectively, for 5 MB-sized songs and 3 MB-sized songs) depending on the two following traffic parameters:

- the speed at which users move throughout the cell (expressed in Km/h),
- the additional traffic in the cell (expressed via Erlang values).

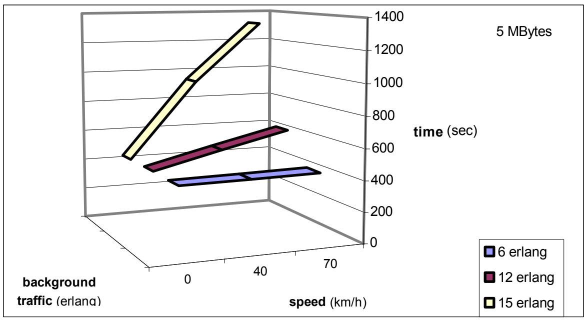


Figure 2: WNTT Values for 5 MB-sized Songs

Two main considerations about the results of Figures 2 and 3 are in order: i) the larger the traffic in the cell (and the users' speed), the larger the corresponding WNTT values, and ii) the best WNTT result may be obtained when the mobile device is completely still. (In such one case a data rate of about 15 KB/s may be obtained for a 3 MB-sized song.)

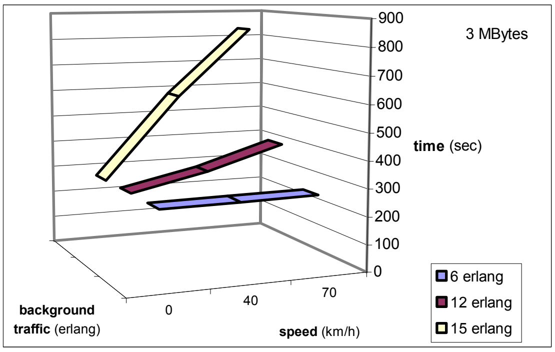


Figure 3: WNTT Values for 3 MB-sized Songs

Figure 4 summarizes the behavior of the WNNT values for some different traffic combinations, depending on the song sizes.

In particular from the top to the bottom of the graph represented in Figure 4, the curves for the following traffic parameters have been respectively plotted: 15 Erlang-70 km/h, 15 Erlang-40 Km/h, 12 erlang-70 Km/h, 12 erlang-40 Km/h, 6 Erlang-0 Km/h. It is easy to note that the more the song size increases (along with the amount of traffic in the cell) the more the WNNT values increase.

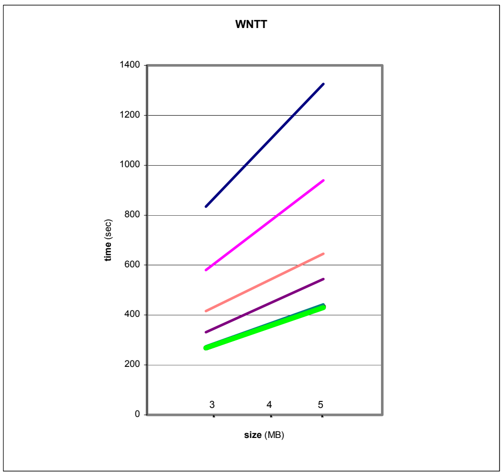


Figure 4: Summary of WNTT Values depending on Song Size

Figures 5 and 6 report on the average data rates that may be obtained on the wireless link for the download of songs of respectively 5 MB and 3 MB (yet again, depending on the user speed). As expected, the larger the user speed, the smaller the obtained data rate.

An important consideration is in order now, which is related to the impact that the download time values, obtained at the Internet side, have on the total time requested to download songs on UMTS terminals. The obtained average download delays at the Internet side (about 33 s) seem to be quite irrelevant if compared with the WNTT values which have been experienced on the wireless links (ranging from 250 to 1325 s, i.e. from about 4 to about 22 minutes). This optimal result at the Internet side is probably due to the use of the adopted Web replication technology along with the use of our distribution mechanism (C^2LD). Note, in fact, that if we try to download songs from a single Web server (such as the New Zealand Web server) with the standard HTTP, this can lead to an increase of the WLNTT value by about 600 s (10 minutes).

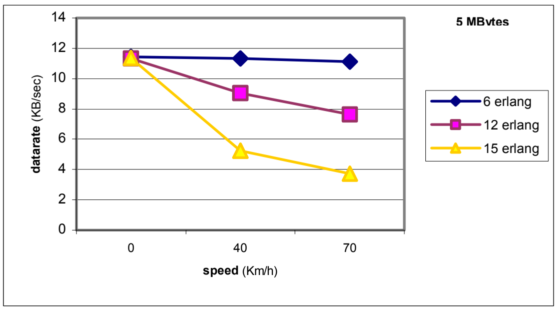


Figure 5: Data Rate for 5 MB-sized Songs

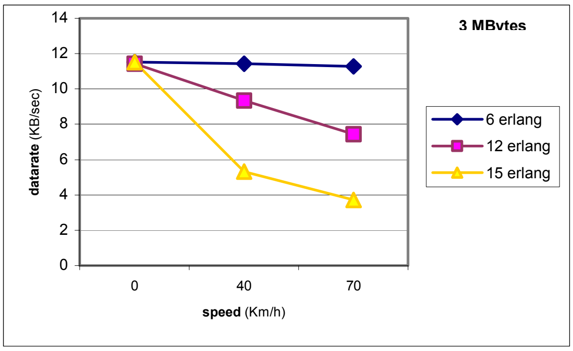


Figure 6: Data Rate for 3 MB-sized Songs

Downloading Time Values for sets of songs on UMTS links

It is well known that typically when a user wishes to listen to a set of songs, he would prefer to be able to play out each single song in sequence, without any interruption occurring in between different subsequent songs. Now, if we consider typical Mp3 songs with a 128 Kbit encoding (based on a 44100 Hz sampling rate) this may produce a needed data throughput of about 17 KB per second. Hence, we can be sure that the typical user's listening preference is satisfied only if our wireless application may guarantee a data rate data of 17 KB per second. Unfortunately, we already know from the results presented previously that, even in the best case (6 Erlang of traffic in the cell, user speed equal to 0 Km/h), the system data rate is not able to exceed the amount of 12/15 KB per second. A possible alternative solution to surmount this problem amounts to keep the user waiting for a certain initial time period before he can begin to listen to a consecutive playout of all the songs contained in the compilation. In Figures 7 and 8, we have sketched, in two particular cases, the theoretical computation of user's preliminary waiting time periods that may permit an uninterrupted playout of all the songs of a given compilation. The results shown in Figure 7 are different from those represented in Figure 8 since they depend on different traffic parameters.

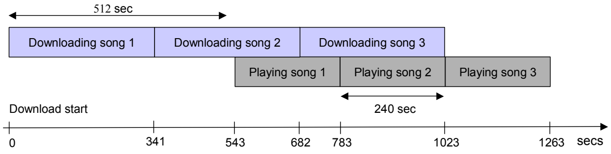


Figure 7: Waiting Time for a 3-song long Compilation (song size: 4 MB, traffic: 6 Erlang, user speed: 0 km/h)

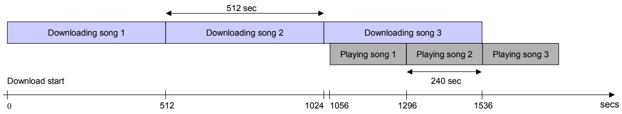


Figure 8: Waiting Time for a 3-song long Compilation (song size: 4 MB, traffic: 12 Erlang, user speed: 70 km/h)

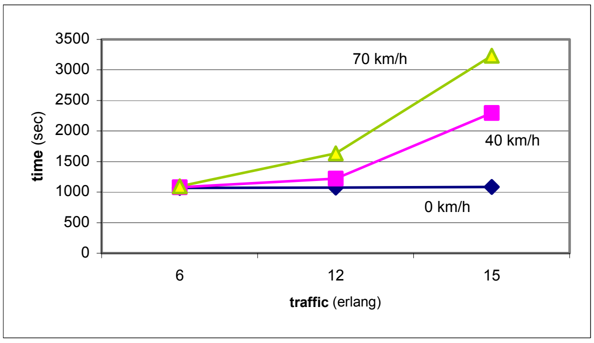


Figure 9: WNTT Values for a 3-song long Compilation (12 MB)

To validate the theoretical results presented in Figures 7 and 8, we carried out an additional set of experimental trials by downloading sequences of different subsequent songs. Figures 9 and 10 show, respectively, the WNTT and the data rate values for compilations composed by 3 songs, each one encoded with a 4 MB-sized Mp3 file. Finally, Figure 11 and 12 illustrate, respectively, the WNTT and the data rate values for compilation composed by 5 songs, each one encoded with a 4 MB Mp3 file.

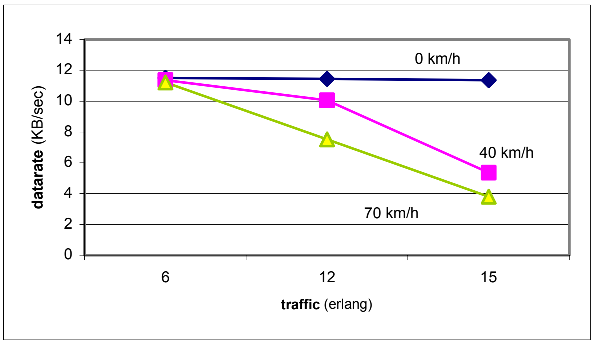


Figure 10: Data Rate Values for a 3-song long Compilation (12 MB)

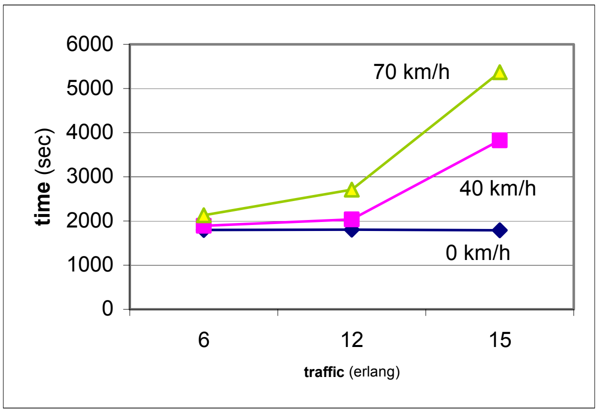


Figure 11: WNTT Values for a 5-song long Compilation (about 19 MB)

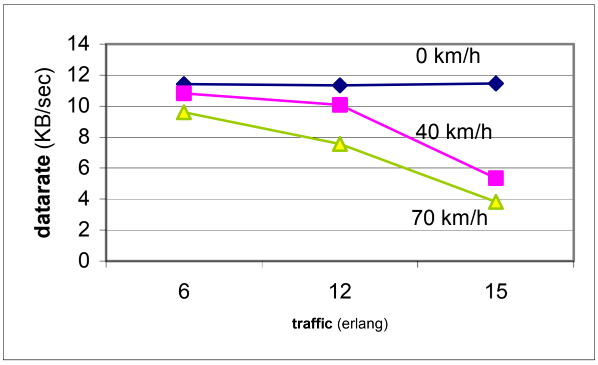
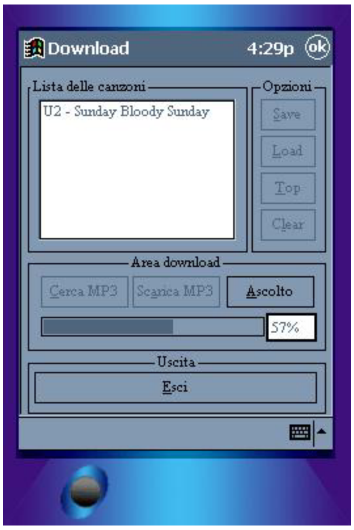


Figure 12: Data Rate Values for a 5-song long Compilation (about 19 MB)

By observing the different WNTT values represented in the curves of the presented plots, we may appreciate that the obtained experimental values are very close to the theoretical ones computed in Figures 7 and 8. Hence, we may conclude that, from an analysis of the curves of these plots, it could be possible to derive quite precise estimates of the extension of the waiting time periods a user should tolerate before beginning to listen to an uninterrupted playout of a compilation consisting of 3 (or 5) songs.

3. CONCLUSIONS AND FUTURE RESEARCH

We have developed an Internet wireless application designed to support music distribution to UMTS devices with the Internet as a backplane. Our application permits to mobile users to download and to listen to Mp3 files on UMTS devices, as shown in Figures 13 and 14. In this paper we reported on a large set of experimental results we obtained on the field by exploiting our wireless application. The download time measurements we have experimentally obtained show that combining 3G mobile network technologies with an appropriate structuring of the Internet wireless application may be very effective for the *fast* distribution of music to mobile clients. We conclude this paper by noticing that the WNTT values we have obtained from our experiments are in the range from 250 to 1325 s, for single songs, depending on the user's speed and on the external traffic in the cell. With different wireless technologies, and in the absence of other external traffic, we would have obtained theoretical WNTT values ranging from 1500 (GPRS technology at 28.8 Kb/s) to 3000 s (GSM technology at 14.4 Kb/s).



U2 - Sunday Bloody Sunday

Figure 13: Downloading Interface on the UMTS Device

In the near future, we expect to be able to assess the software architecture, which is at the basis of our wireless application, by using a real UMTS network infrastructure. In addition, future research efforts will be devoted to the activity of porting our Internet wireless application on different operating platforms.

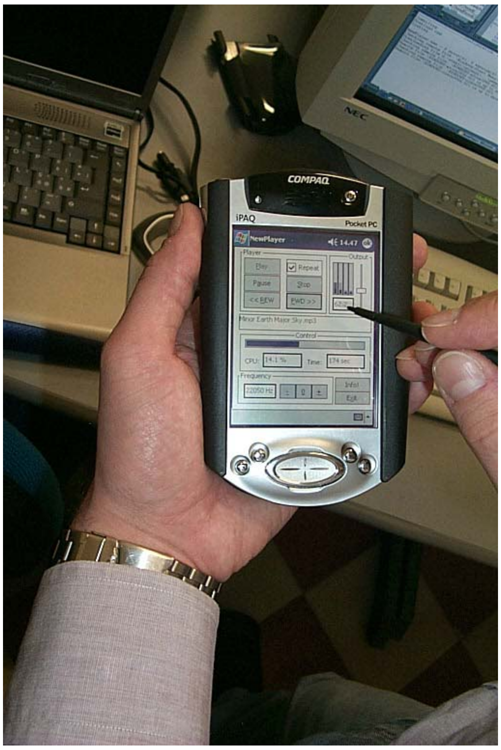


Figure 14: A Playout Session

Finally, we are planning to investigate further P2P-oriented design issues in order to provide support to a wider class of wireless multimedia applications including, for example, multimedia instant messaging applications.

ACKNOWLEDGEMENTS

We wish to thank the "Fondazione Marconi of Bologna, Italy", the Department of Computer Science of the University of Bologna, Italy, and Microsoft Research Europe (Cambridge, UK) for the partial financial support to this work under a grant. Finally, we are indebted to our colleagues Alessandro Gambetti, Davide Melandri, Orlando Orlandi, Mirko Piaggesi and Daniela Salsi for their helpful assistance during the implementation and testing of the software architecture discussed in this paper.

REFERENCES

L-P. Anquetil, G. Bonnet, M. Lapierre, W. Van Leekwijck, G. Willekens, "New Multimedia Services using a Server Architecture", Alcatel Telecommunication Review, 2nd Quarter 2000, 101 - 108.

R. Kalden, I. Meirick, M. Meyer, "Wireless Internet access based on GPRS", IEEE Personal Communications, Vol. 7 N. 2 , April 2000, 8 –18

D. Staehle, K. Leibnitz, K. Tsipotis, "QoS of Internet Access with GPRS", Proc. Fourth ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Rome, July 2001, 57-64

UMTS Forum, "What is UMTS?", http://www.umts-forum.org/what_is_umts.html

G. Huston, "TCP in a Wireless World", IEEE Internet Computing, March-April 2001, 82-84

M. Roccetti, V. Ghini, P. Salomoni, A. Gambetti, D. Melandri, M. Piaggesi, D. Salsi. "The Structuring of a Wireless Internet Application for a Music-on-Demand Service on UMTS Devices", Proc. ACM Symposium on Applied Computing, Madrid, March 2002, to appear

Napster official site, http://www.napster.com/

Gnutella official site, http://gnutella.wego.com/

Freenet Project Inc., *The Freenet Project*, http://freenet.sourceforge.net

M. Conti, E. Gregori, F. Panzieri, "QoS-based Architectures for Geographically Replicated Web Servers", Cluster Computing 4, 2001, 105-116

V. Ghini, F. Panzieri, M. Roccetti, "Client-Centered Load Distribution: A Mechanism for Constructing Responsive Web Services", Proc. 34th Hawaii International Conference on System Sciences, Maui, January 2001.

BIOGRAPHIES

Marco Roccetti earned an Italian Dott. Ing. Degree in Electronics Engineering from the Faculty of Engineering of the University of Bologna, Italy, in 1989. Since November 2000 he is a Professor of Computer Science at the Department of Computer Science of the University of Bologna. His research interests include: i) design, implementation and evaluation of multimedia computing and communication systems, and ii) performance evaluation and simulation of distributed and parallel computing systems.

Vittorio Ghini has an Italian Laurea Degree in Computer Science from the University of Bologna, Italy. He is currently a Ph.D. candidate at the Department of Computer Science of the University of Bologna. His research interests include issues of distributed multimedia systems and quality of service over IP-based networks

Paola Salomoni has an Italian Laurea Degree in Computer Science from the University of Bologna, Italy. She is currently an Associate Professor of Computer Science at the Department of Computer Science of the University of Bologna. Her research interests focus on the design and development of distributed multimedia systems, as well as on distance teaching/learning computer based environments.

A JAVA-BASED ADAPTIVE MEDIA STREAMING ON-DEMAND PLATFORM

Giancarlo Fortino, Angelo Furfaro, Wilma Russo
DEIS, Università della Calabria
Via P. Bucci s/n, 87036 Rende (CS), Italy
E-mail: {g.fortino, w.russo, a.furfaro}@unical.it

Juan C. Guerri, Ana Pajares, Carlos E. Palau
Universidad Politecnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
E-mail: {cpalau, jcguerri, apajares}@dcom.upv.es

KEYWORDS
VoD, Java Media Framework, RTP, RTSP, IP-multicast.

ABSTRACT

Nowadays, media streaming architectures and systems are gaining focus due to the widespread diffusion of IP-based, bandwidth-capable digital networks which can really support multimedia data-intensive on-demand services. In this paper, we present a Media on-Demand system which provides adaptive, multicast, and collaborative media streaming. Media streaming relies on the Real time Transport Protocol whereas streaming control is centered on the Real Time Streaming Protocol. A distinguishing feature of the proposed system is the ability of the media streamer component at the server side to adapt the media flow on the basis of the RTCP feedback of the client. The paper also describes the Java-based implementation of the system which uses the Java Media Framework library.

INTRODUCTION

Multimedia, in particular audio and video, is already being distributed in the Internet and in many other communications networks and systems (Steinmetz and Nahrstedt 1995). Thanks to many different products like RealNetwork's RealSystem (Realsystem 2001), Nullsoft's SHOUTcast (ShoutCast 2001), Windows Media (WinMedia 2001) or Apple's Quicktime (Quicktime 2001), the idea of broadcast has been brought to the web. Besides there are other multimedia applications based on the requests of the customers which are generally known as Video on-Demand systems (VoD). A VoD system allows customers to select movies from a large menu in order to view them (e.g., at home) at times of their choosing. Any VoD system is formed by three key structural blocks: the server, the communications network and the client. The main goal of a VoD system is to serve thousand of clients simultaneously. Each client connects to the communications network and gets in contact with the server to request a movie at a specific time (Wright 2001, Zink et al. 2001).
Today's information servers not only provide simple text-based data, but also a rich combination of text, still-images, animations, audio and videos. The main issues a VoD system has to deal with are storage spaces, bandwidth, synchronization, transport protocols, playback control protocols and scalability. There are several VoD systems developed using different techniques, architectures and philosophies, and each of them solve the above problems by means of alternative approaches.
There are several applications of VoD systems like broadband TV broadcast, education, medicine (Brett 1996, Guerri et al. 2000, Huang and Hu 2000). All these applications are usually based on the control of several multicast multimedia flows using a control protocol like RTSP (Shulzrinne et al. 1998).
Interesting example of VoD-like systems over Internet are: the KOM-player (Zink et al. 2001), Video Conference Recording On-demand (ViCRO) (Fortino and Nigro 2000), JMFVoD (Belda et al. 2002), multicast Multimedia-On-Demand (mMOD) (Parnes et al. 1997), and Universal Video-On-Demand (UVoD) (Lee 1999).
The main goal of this paper is to present the architecture and a prototype implementation of our Java-based media on-demand system providing multicast and adaptive streaming. Media on-Demand (MoD) is richer than Video on-Demand (VoD) in that MoD (Horn et al. 2001) provides for multimedia sessions which not only consist of synchronized audio and video streams, but also of other types of media such as text-tips, text-tickers, images, animations, slides, etc. To sum up, the playback can be a multimedia composite session.
The MoD system proposed centers on the following key technologies:

(i) *RTP as media transport protocol*. RTP (Schulzrinne et al. 1996), the Real-time Transport Protocol framework provides end-to-end delivery services for data with real-time characteristics. These services are suitable for various distributed applications that transmit real-time data, such as interactive audio and video. The companion protocol (RTCP) provides feedback to the RTP sources in the RTP session and to all the participants in the session. The same underlying transport protocol (i.e., UDP) is used for both, but different ports are used to differentiate packet streams;

(ii) *RTSP as playback control protocol*. The playback of the multimedia information is controlled by means of the Real Time Streaming Protocol (Schulzrinne et al. 1998). RTSP is an application-layer protocol that provides control over the delivery of real-time data. The protocol is typically applied for control over continuous time-synchronized streams of media. It usually acts as a remote control protocol for media servers. Our system makes the typical use of RTSP, not to deliver media by itself but to control streams delivered by RTP;

(iii) *Java Media Framework (JMF)*. JMF (Sun Jmf 2002) is an API defined jointly by Sun MicroSystems, Inc. and IBM Corporation. The main aim of the product is the provision of real-time delivery and control of multimedia data (video and audio) by means of streaming techniques. The API can be used in applets and in stand-alone applications developed in JAVA. A part from providing facilities for data processing (codecs, multiplexers and demultiplexers, renderers, etc) JMF allows transmission and reception of multimedia data using RTP in unicast and multicast modes, and the utilization of several capture devices like webcams. There are different versions of the API, the latest one is v. 2.1.1a, although new improvements are being included;

(iv) *Adaptive Server-to-Client streaming*. Adaptive streaming (Busse et al. 1996) is based on a media streaming rate control mechanism relying on RTCP feedback. The adaptation is performed in terms of bandwidth, losses and data encoding;

(v) *Friendly Applet-based client interface*. The choice to implement the client as a Java Applet allows easy system extendibility and dynamic software distribution.

The remainder of the paper is structured as follows. The second section introduces the main components of the media streaming on-demand architecture. The third section describes the RTSP protocol and the media control component based on it. The adaptive streaming component is introduced in the forth section. In the fifth section an extension of the architecture to support tightly coupled groups of distributed clients is discussed. Finally conclusions and directions of further and on-going work are reported.

THE BASIC ARCHITECTURE OF THE MEDIA STREAMING ON-DEMAND SYSTEM

As has been commented in the introduction section, there are several IP-based streaming systems developed in the literature (Gebhard and Lindner 2001, Huang and Hu 2000, Rowe 2001). At the same time new multimedia libraries and APIs are appearing or are being improved, like Java Media Framework (Sun Jmf 2002) in order to be used to develop such systems.

A main aim of our work is the validation of the viability of JMF libraries for the implementation of a media streaming on-demand system. The execution of the system will allow the real evaluation of the management mechanisms for multimedia sessions, RTP/RTCP protocols, the compression standards included in the libraries and the utilization of RTSP. The developed prototype will serve as basis for the future evaluation of new multimedia protocols that will substitute the traditional simulation studies in this field.

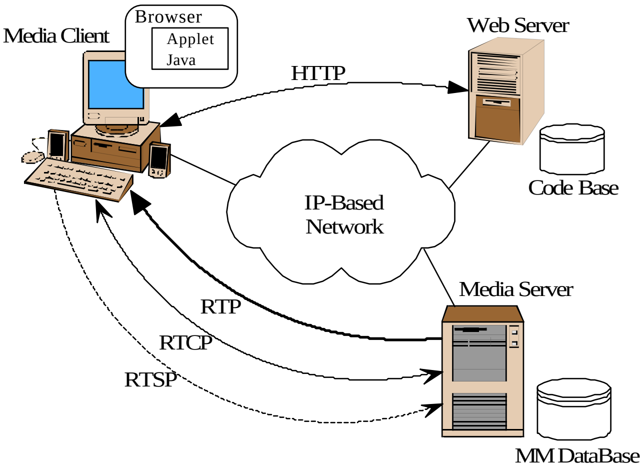


Figure 1: Architecture of the MoD system

With reference to a MoD system, four elements can basically be differentiated (Figure 1): client, web server, multimedia database and media server.

- *Client*: its development is based on a Java applet that is started each time the correspondent HTML page is loaded from the web server. Through the graphical user interface, the users can select from the remote multimedia database the multimedia session title they want to playback.
- *Web server*: it stores the start page and the Java applets. It also holds the authentication procedures and the database access management.
- *Database*: it keeps the information catalogue of the available multimedia sessions (e.g., movies).
- *Media Server*: it stores the media files and send them to the clients as soon as they request them.

Software structure

The Media on Demand (MoD) streaming system (Belda et al. 2002, Fortino and Nigro 2000) consists of three super blocks or applications (Figure 2) that communicate with each other through TCP and UDP –based connections, apart from the multimedia database containing the multimedia files published in the system.

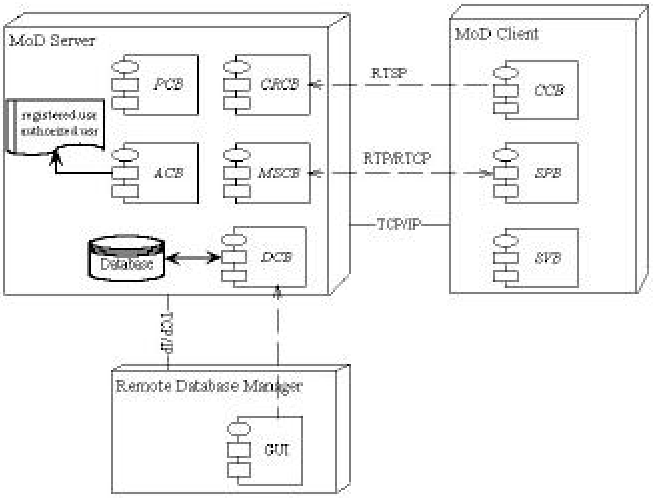


Figure 2: Components and Communications structure

- The *MoD streaming server*, continuously listens at a TCP port waiting for a client to make a connection request. The server exchanges messages defined according to the RTSP protocol with the clients through the opened sockets.
- The *MoD client*. There are two kinds of entities that can make a connection request: the client itself and the remote database manager. Both clients are applets and run in web navigators.
- The *Remote Database Manager*. The Database Manager is SQL based and provides access depending on the identity of the users. Some easy commands have been developed in order to communicate the server application with the database.

The server allows a fixed number of clients to simultaneously access to the system. It also includes a user authentication procedure along with an application that allows the list of users that can access the system to be managed.

Database structure

The published files database used by the application has been structured in three tables:

- *Main table*: contains the published multimedia files names and other related attributes.
- *Categories table*: contains the names of the categories to which a multimedia file can belong to.
- *Subcategories table*: contains the names of the subcategories to which a multimedia file can belong to.

Control protocol

The system control protocol consists of the exchanged messages and rules between the streaming MoD server and both the client and the database remote manager. The control protocol is based on RTSP and is described in the next section.

Server application

It is a multi-threaded unit whose main blocks are:

- The Connection Requests Control Block (CRCB) is created when the server is activated from the administrator's GUI. When it receives a connection request it creates a front-end handling the connection with the client.
- The Multimedia Streams Control Block (MSCB) provides the mechanisms to process multimedia files for creating and controlling the RTP sessions needed to send the multimedia streams to the client that has requested them.
- The Parameters Control Block (PCB) provides the mechanisms to read and write the configurable parameters of the MoD streaming system.
- The Database Control Block (DCB) accesses to, reads and modifies the database, by means of SQL instructions, using the JDBC-ODBC bridge driver provided by the library java.sql.
- The Access Control Block (ACB) manages two files: "registered.usr" and "authorized.usr". Both of them contain login names and passwords of the users along with their access rights.

Client Application

The client application is structured in the following blocks:

- The Control Connection Block (CCB) is committed for sending and receiving control messages exchanged between the client and the server.
- The Streams Player Block (SPB), it is responsible for streams synchronization and reproduction.
- The Statistics Viewer Block (SVB) gets the RTP sessions statistics from the SPB.

Remote database manager

The remote database manager is an applet embedded in a html page. When the applet is loaded, it generates a graphical user interface identical to the one generated by the local database manager. The database changes introduced by the user are sent through a TCP control connection to the server. This connection is established on the applet initialization. If there is not another user managing the database, the server will send the database content to the remote manager side through the TCP control connection block.

STREAMING CONTROL

RTSP: an overview

The Real-time Streaming Protocol (RTSP) (Schulzrinne et al. 1998) is a text-based application level protocol developed for on-demand delivery control of media streams both live and pre-recorded. It establishes and controls either a single or several time synchronized streams of continuous media such as audio or video. In other words, RTSP acts as a "network remote control" for multimedia servers.

RTSP extends HTTP 1.1 by introducing a stateful behavior and new methods. Thus, RTSP requires the notion of session. It defines a session identifier which is chosen by the server at the session establishment and used throughout the session lifetime. RTSP can be based on a single TCP connection activation, on multiple activations (each one for a request/reply interaction), on UDP.

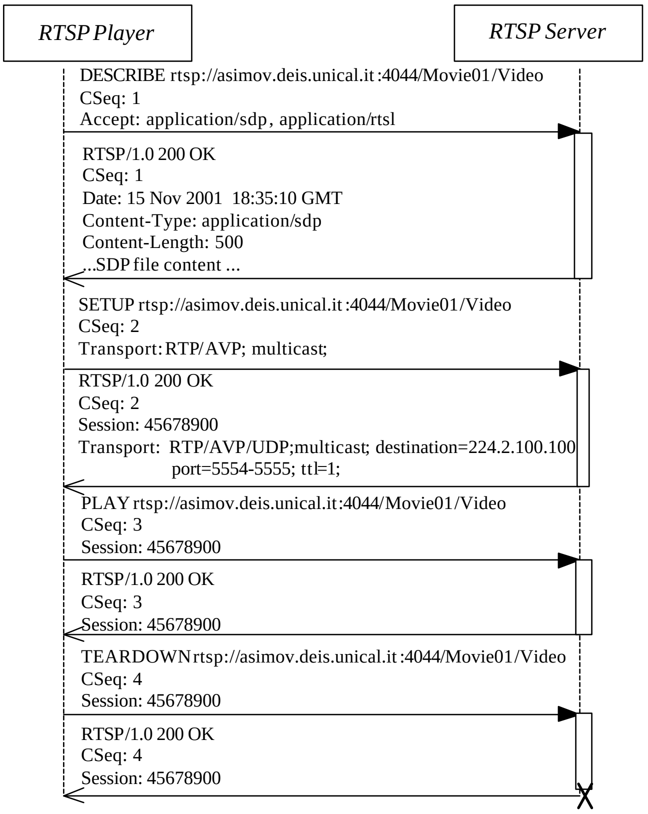
Multimedia presentations are identified by URLs, using a protocol scheme of "rtsp". The hostname is the server containing the presentation; whilst the port (the default port is the 554) indicates which port the RTSP control requests should be sent to. Presentations may consist of one or more separate streams. The presentation URL provides a means of identifying and controlling the whole presentation rather than coordinating the control of each individual stream. So, the *rtsp://asimov.deis.unical.it:4044/Movie01/videotrack* URL, identifies the video stream within the presentation Movie01, which can be controlled on its own. If the user would rather stop and start the whole presentation, including the video, then he/she would use the *rtsp:// asimov.deis.unical.it:4044/Movie01/* URL.

The standard RTSP methods are:

- The DESCRIBE method is sent from the client to the server and allows retrieving the description of a presentation or a single media object.
- The SETUP method issued by the client causes the server to generate the session identifier and allocate resources for the session. During the setup phase, client and server can negotiate transport parameters (addresses and ports) as well as media streaming parameters.
- The PLAY method signals the server to start streaming. PLAY can have a range header which contain a time parameter: start and stop instants within the multimedia session. Time can be encoded in NPT (Normal Play Time), SMPTE (relative time), and UTC (absolute time).
- The PAUSE method interrupts the media streaming delivery temporarily.
- The TEARDOWN method stops the streaming so that the server can free the resources allocated to the session.
- The SET_PARAMETER method allows setting a parameter associated to the current session.
- The GET_PARAMETER method retrieves a value of a parameter of a session.
- The OPTIONS method returns the supported methods of a server.
- The REDIRECT method informs a client that it must connect to the new server location contained in the Location header.
- The RECORD method allows starting the recording of an announced multimedia session.
- The ANNOUNCE method posts on the server the description of a multimedia presentation (e.g., to be recorded).

In addition, new methods, which extend the RTSP functionality, can be purposely introduced as reported in (Fortino and Nigro 2000, Fortino et al. 2001).

Figure 3 reports a typical control session of a video playback which consists of request/reply pairs according to the RTSP format.



```
DESCRIBE rtsp://asimov.deis.unical.it:4044/Movie01/Video
CSeq: 1
Accept: application/sdp, application/rtsl

RTSP/1.0 200 OK
CSeq: 1
Date: 15 Nov 2001  18:35:10 GMT
Content-Type: application/sdp
Content-Length: 500
...SDP file content ...

SETUP rtsp://asimov.deis.unical.it:4044/Movie01/Video
CSeq: 2
Transport: RTP/AVP; multicast;

RTSP/1.0 200 OK
CSeq: 2
Session: 45678900
Transport:  RTP/AVP/UDP;multicast; destination=224.2.100.100
            port=5554-5555; ttl=1;

PLAY rtsp://asimov.deis.unical.it:4044/Movie01/Video
CSeq: 3
Session: 45678900

RTSP/1.0 200 OK
CSeq: 3
Session: 45678900

TEARDOWN rtsp://asimov.deis.unical.it:4044/Movie01/Video
CSeq: 4
Session: 45678900

RTSP/1.0 200 OK
CSeq: 4
Session: 45678900
```

Figure 3: The UML sequence diagram of the RTSP-based control session of a multicast video stream

RTSP inside JMF

RTSP support has recently been added to the Java Media Framework library. This integration enables JMF based clients (both applets and applications) to communicate with RTSP enabled servers and request the streaming of specified mediafile. For instance, the Java Multimedia Studio application can be used as an RTSP client by opening an RTSP URL from the File menu. RTSP within JMF library adds more control mechanisms to multimedia information playback control.

JMF 2.1.1a allows to build RTP/RTSP players which interoperate with standard-based, third-party video streaming servers such as Sun StorEdge Media Central Streaming Server (Sun Storedge 2001) and Apple Quicktime Streaming Server (Apple Quicktime 2001).

Figure 4 depicts a simplified UML class diagram of the RTSP classes and interfaces which are inside JMF. The *RtspUtil* class can be employed by a programmer to implement an RTSP Player or an RTSP Server. It implements the *RtspListener* interface which defines two methods. In particular, the *rtspMessageIndication* method is invoked by the *RtpManager* when a new RTSP message arrives. The *RtpManager* class creates and manages the socket connections, and makes it available the *sendMessage* method to transmit a message over a connection identified by a specific ID. It's worth noting that the *RtspUtil* class is associated to one or more *RtpManagers.* If the *RtspUtil* is used as Player, the *RtpManager* instances are created after the SDP description of the requested media file has been retrieved and the SETUP message is being issued. In order to realize an RTSP Server we have first modified and then extended the *RtspUtil* class respectively by introducing a per-connection server state variable and by specializing the *processRtspRequest* method.

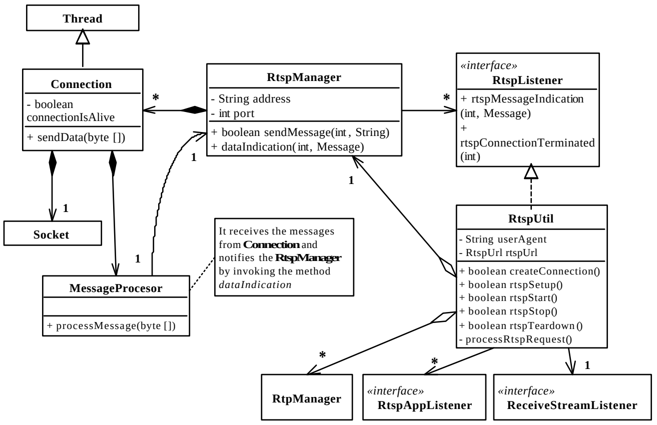


Figure 4: Class Diagram of the JMF implementation of RTSP

STREAMING ADAPTATION

A media streaming system (Dutta and Schulzrinne 2001) over Internet has to provide flexibility in coping with the bandwidth variations so as to deliver a certain degree of quality of service. In fact, in "traditional" design of integrated services networks, multimedia applications such as videoconferencing tools can negotiate a desired QoS during the connection setup phase and then the network guarantees such a quality. Conversely, Internet doesn't provide bandwidth guarantees, thus application level mechanisms are to be employed. In particular, we exploit a network-initiated in-call QoS adaptation control which bases the application target data rate on network feedback (Busse et al. 1996). Starting from a given bandwidth, low losses lead the application to slowly increase the media streaming bandwidth, while high packet losses lead to bandwidth decrease. The control algorithm is mainly based on the feedback information conveyed by the receiver reports of the RTCP companion protocol. This feedback information allows the source of the media streaming based on RTP to estimate the loss rates experienced by the receiver/s and to adjust its bandwidth accordingly.

RTCP: Receiver Reports and SDES reports

After receiving several data packets, the RTP-based client calculates some performance measurements (or statistics) which are then sent back to the media streaming component at the server site in the Receiver Report (RR) through the RTCP channel. The statistics are the following:

(i) *Total Packet Lost (TPL)*: the total number of packet lost, since the transmission beginning. It is computed from the number of packet actually received and the number of packets expected, which is detected on the basis of the sequence number of the last RTP data packet arrived.

(ii) *Fraction of Packet Loss (FPL)*: the fraction of the packet lost during the time interval between two consecutive transmissions of RR reports.

(iii) *Interarrival Jitter (IJ)*: the mean deviation (smoothed absolute value) of the difference in packet spacing at the receiver compared to the sender for a pair of packets (Schulzrinne et al. 1996).

Besides, RTCP messages contain an SDES (Source DEScription) packet which embeds information to identify the data sources such as the canonical name (CName), email address, telephone number, application specific info and alert messages.

End-to-End application control mechanism

The adaptive streaming control scheme (Busse et al. 1996, El-Marakby and Hutchinson 1997) is based on the following three phases:

(i) *RTCP analysis*. The RR of the receiver/s (in the case of unicast or multicast streaming) are analyzed.

(ii) *Network state estimation*. The actual network congestion state seen by the receiver/s can be classified as unloaded, loaded of congested (or simply as congested or uncongested).

(iii) *Bandwidth adjustment.* The bandwidth of the media stream is regulated according to the classification of the network state analysis.

Our adaptive streaming controller relies on the FPL statistics which is the indicator of the congestion. FPL can be used either raw or filtered. In the latter case, by applying a low-pass filter it is possible to smooth the FPL so reducing QoS oscillations. In the following two algorithms are described. The first is based on a basic algorithm proposed in (El-Marakby and Hutchinson 1997). The second is a variant of a more sophisticated algorithm proposed in (Busse et al. 1996). In figure 5 the basic algorithm Java-like is portrayed.

```
FPL=lastRR.getFPL();
if (FPL>lambda)
  currRate=Math.max(alpha*currRate, minRate);
else
  currRate=Math.min(beta*currRate, reqRate);
```

Figure 5: Basic rate control algorithm

```
FPL=lastRR.getFPL();
fpl_LAST=(1-a)*fpl_OLD+a*FPL;
if (fplLast>lambdaC)
  currRate=Math.max(mu*currRate, minRate);
else
if (fplLast<lambdaU)
  currRate=Math.min(currRate+chi, reqRate);
```

Figure 6: Rate control algorithm with filter and deadzone regulator

Lambda is the threshold parameter whose exceeding (keeping below) can cause a decrement (increment) of the media streaming rate. Conversely, [minRate, reqRate] is the admissible rate variation range. minRate is chosen by the streamer component according to the involved media whereas reqRate is the rate chosen by the client at media streaming setup time. Alpha is the fixed decrement fraction coefficient whereas beta is the fixed increment fraction coefficient. It is worth noting that: (i) no RTCP analysis is performed; (ii) the network state estimation (congested or not) is driven by lambda; (iii) bandwidth adjustment is done by using the max/min featured functions. In figure 6 the second algorithm Java-like is reported.

The algorithm first uses a smoothing (low-pass) filter, then a linear regulator with dead zone which classifies a receiver in congested, loaded and unloaded.

The introduced algorithms strongly depend on the choice of their characterizing controller parameters (*a, mu, chi, lambdaC, lambdaU*) in order to be really effective.

In case of unicast media streaming the presented algorithms can be directly applied. In case of point-to-multipoint media streaming, the decision to increase, decrease or hold the rate should be taken by considering all the receivers. A simple solution is to adapt the media flow to the slowest receiver so as to avoid congestion for all the receivers. The main drawback is that faster receivers obtains poor quality. A more interesting solution is to compute the percentages of the congested, unloaded and loaded receivers and to make a decision of rate decreasing, holding or increasing on the basis of these percentages (e.g, if the % of congested receivers is greater than 10, then the rate is decreased).

Another solution can be to adjust the media streaming by considering only the network conditions of the media playback owner, i.e., who requested the playback.

More powerful approaches to the streaming adaptation are the video gateways and the layered coding schemes supported by lightweight and tunable feedback protocols such as SCUBA (Scalable ConsensUs-based Bandwidth Allocation) (Amir et al. 1997).

GROUP SHARING OF A REMOTE MEDIA SESSION CONTROL

In traditional VoD systems, remote control of the media session can be only applied by the session owner. In case the requested media streaming is multicast, other clients can tune in and watch the multimedia session. However, they don't have session control and their view is affected by the control commands issued by the session owner client. In order to allow for sharing the session control among a group of clients, a shared media session controller is to be realized. In (Fortino et al. 2001) several schemes (unicast, hybrid and multicast) to implement such a virtual controller are proposed.

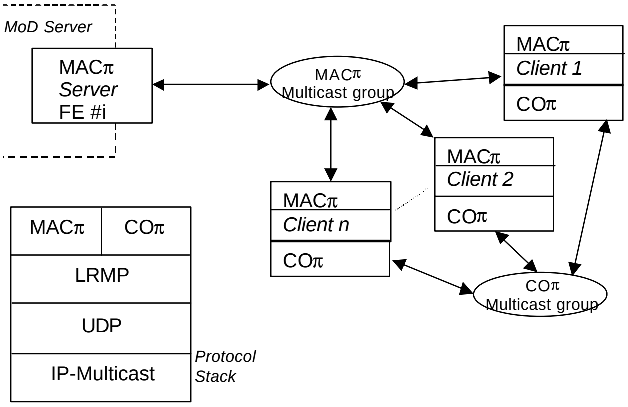


Figure 7: The VCC architecture and the protocol stack

Key issues concern with the media streaming control protocol and behavioral policy. The former has to provide rules to enable each client within a group to consistently send control commands. The latter is needed to regulate participants which exhibit behavior mining the media presentation view such as frequent pauses or seeks.

A first prototype of the virtual control component (VCC) relies on the MACπ protocol which is a multicast extension of the RTSP protocol and the COπ protocol supporting a voting-based coordination policy (Fortino et al. 2001).

Figure 7 portrays the distributed architecture of the VCC. For each requested media session to be shared, a Front-End (FE) at the MoD server site is spawn on a given multicast group. Each media client activates a COπ component on another multicast group. After the shared session is started, i.e., a media client issued the PLAY request and the FE replied to all, each command before being sent is constrained by the coordination policy adopted. The current available policy is based on a voting mechanism which is triggered by the seek (i.e., PLAY with range header) and the TERMINATE method (the PAUSE is not constrained). A seek is accepted if and only if the majority of the media clients agree whereas the TERMINATE needs the unanimity to be raised.

CONCLUSIONS AND FUTURE WORK

In this paper, we have described the basic architecture of a Java-based platform providing media on-demand services. The platform mainly relies on: (i) the multimedia internetworking protocols, namely RTP, RTSP, SAP and SDP; (ii) Java and the Java Media Framework.

Interesting feature of our platform is that it enables adaptive streaming based on receiver-based feedback algorithms and allows for sharing playback among multiple clients. The choice of Java as implementation language is strategic not only because it provides rich and powerful object-oriented libraries for distributed and multimedia computing but also since it allows to easily deploy software systems over heterogeneous environments like the Internet. The proposed MoD platform is the first prototype which will be used as base for the development of a cooperative and content-enriched media on-demand system over the Internet for distance learning and entertainment.

Our on-going and future work aims at: (i) refining the prototype and evaluating its performance; (ii) defining a comprehensive protocol suite for collaborative media on-demand which is mainly oriented to the learning and entertainment application domains.

ACKNOLEDGEMENTS

We would like to thank our colleagues at Università della Calabria and Universidad Politécnica de Valencia, particularly A. Belda, M. Esteve, L. Nigro, and F. Pupo for numerous helpful discussions and ideas. This research is supported in part by the Spanish and Italian Governments in the frame of a bilateral project.

REFERENCES

Amir E.; S. McCanne; and R. Katz. 1997. “Receiver-driven Bandwidth Adaptation for Light-weight Sessions.” In *Proceedings of ACM Multimedia*, Seattle, WA, (Nov).

Apple Quicktime. 2001. Apple Quicktime Streaming Server, available at http://www.apple.com.

Belda A.; A. Pajares; J. C. Guerri; C.E. Palau; and M. Esteve. 2002. “JMFVoD: Design and Implementation in Java of a Media Streaming on Demand System.” In *Proceedings of SCS Euromedia*, Modena, Italy, (Apr).

Brett P. 1996. “Using Multimedia: an Investigation of Learners' Attitudes.” *Computer Assisted Language Learning* 1 No. 2-3, 191-212.

Busse I.; B. Deffner; and H. Schulzrinne. 1996. “Dynamic QoS Control of Multimedia Applications based on RTP.” *Computer Communications* 19, No. 1, (Jan).

Dutta A. and H. Schulzrinne. 2001. “A Streaming Architecture for Next Generation Internet.” In *Proceedings of ACM Multimedia.*

El-Marakby R. and D. Hutchinson. 1997. “Delivery of Real-time Continuous Media over the Internet.” In Proceedings of the 2nd IEEE Symposium on Computers and Communications (ISCC'97), Alexandria, Egypt, (Jul).

Fortino G. and L. Nigro. 2000. “ViCRO: an interactive and cooperative videorecording on demand system over Internet Mbone.” *International Journal Informatica* 24, No. 1, 97-105.

Fortino G.; L. Nigro; and F. Pupo. 2001. “An MBone-based on-demand system for cooperative off-line learning.” *Proceedings of IEEE Euromicro*, Warsaw, Poland, (Sep).

Gebhard H. and L. Lindner. 2001. “Virtual Internet Broadcasting.” *IEEE Communications* 39, No 6 (Jun), 182-188.

Guerri J. C.; M. Esteve; C. Palau; M. Monfort; and M.A. Sartí, “A software tool to acquire, synchronise and playback multimedia data: an application to kinesiology.” *Computer Methods and Programs in Biomedicine* 62 (Jan), 51-58,.

G.B. Horn, P. Knudsgaard, S.B. Lassen, M. Luby and J.E. Rasmussen. 2001. “A Scalable and Reliable Paradigm for Media on Demand.” *IEEE Computer Magazine* 34, No. 9 (Sep), 40-45.

Huang S. and H. Hu. 2000. “Integrating Windows streaming media technologies into a virtual classroom environment.” In *Proceedings of the International Symposium on Multimedia Software Engineering*, 411–418.

Lee J. 1999. “UVoD: An Unified Architecture for Video-on-Demand Services.” *IEEE Communications Letters* 3 No.9 (Sep), 277-279.

Parnes P. et alter. 1997. “multicast Multimedia On-Demand (mMOD).” Available at http://mmod.cdt.luth.se.

Realsystem. 2001. RealNetworks's RealSystem, available at http://www.real.com

Rowe L. A. 2001. “Streaming Media Middleware is more than Streaming Media.” In *Proceedings of ACM Multimedia.*

Schulzrinne H.; S. Casner; R. Frederick; and V. Jacobson. 1996. “RTP: A Transport Protocol for Real-Time Applications.” *IETF RFC-1889*, (Jan).

Schulzrinne H.; A. Rao; and R. Lanphier. 1998. “Real Time Streaming Protocol (RTSP).” *IETF RFC 2326*, (Apr).

ShoutCast. 2001. NullSoft's ShoutCast, available at http://www.shoutcast.com.

Steinmetz R. and K. Nahrstedt. 1995. “Multimedia: Computing, Communications and Applications.” *Prentice Hall.*

Sun Jmf. 2002. Java Media Framework, available at http://java.sun.com/products/java-media/jmf/index.html

Sun StorEdge. 2001. Sun StorEdge Media Central Streaming Server, available at http://www.sun.com.

WinMedia. 2001. Microsoft Windows Media, available at http://www.microsoft.com.

Wright W. E. 2001. “An efficient video on-demand model.” *IEEE Computer*, (May), 64-70.

Zink M.; C. Griwodz; and R. Steinmetz. 2001. “KOM Player – A Platform for Experimental VoD Research.” In *Proceedings of ISCC'01*, Hammamet (Tunisia).

JMFVOD: DESIGN AND IMPLEMENTATION IN JAVA OF A MEDIA STREAMING ON-DEMAND SYSTEM

Angela Belda, Ana Pajares, Juan Carlos Guerri, Carlos Palau, Manuel Esteve
Universidad Politecnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain
E-mail: abelda@teleco.upv.es, apajares@dcom.upv.es,
jcguerri@dcom.uvp.es, cpalau@dcom.upv.es, mesteve@dcom.upv.es

KEYWORDS
Video-on-Demand, Java, JMF, RTP/RTCP

ABSTRACT

This article describes the main challenges of implementing a Java VoD system by using the RTP/RTCP protocols and the Java Media Framework library, JMF. This system, called JMFVoD, consists of a stand-alone VoD server application, a web integrated VoD client, a database which contains all the available multimedia files and a proprietary streams control protocol which provides the functionality of RTSP.

INTRODUCTION

Multimedia, in particular audio and video, is already distributed in the Internet and in many other communications networks and systems. For this purpose, many different products like RealNetwork's RealSystem, Nullsoft's SHOUTcast, Windows Media, Apple's Quicktime or Sun Microsystems IPTV have been developed. Multimedia applications based on the requests of the customers are generally known as Video-on-Demand systems.

VoD systems permit clients the selection of movies from a large menu in order to view them at times of their choosing. Any VoD system is formed by three key elements: the server, the communications network and the clients. Each client connects to the communications network and contacts the server to request a specific movie.

The main problems a VoD system has to deal with are storage spaces, bandwidth, synchronization, transport protocols, playback control protocols and scalability. There are several VoD systems developed using different techniques, architectures and philosophies, and each of them solve the above problems by means of alternative approaches (Lee, 1999).

Among VoD systems we can differentiate between near-VoD (NVoD) and true-VoD (TVoD). The first ones make use of multicast technologies in order to enable multiple user to share a single channel and reduce system cost. On the other hand, the second ones allocate a dedicated channel for every user to achieve short latency. The VoD system presented in this work is a true-VoD and its main features consist of:

- Use of the Java Media Framework (JMF) library.
- Utilisation of RTP as delivery protocol.
- Friendly browser-based interface.
- Proprietary playback control protocol.

Java Media Framework (JMF) is an API defined jointly by Sun MicroSystems, Inc. and IBM Corporation to develop multimedia applications. The main aim of the product is the provision of real-time delivery and control of multimedia data (video and audio) by means of streaming techniques. The API can be used in applets and in stand-alone applications developed in JAVA.

RTP (Schulzrinne *et al*, 1996), the real-time transport protocol framework provides end-to-end delivery services for data with real-time characteristics. These services are suitable for various distributed applications that transmit real-time data, such as interactive audio and video. The companion protocol, RTCP, provides feedback to the RTP sources in the RTP session and to all the participants in the session. The same underlying transport protocol is used for both (UDP), but different ports are used to differentiate packet streams.

The proprietary developed control protocol, is based on TCP connections and the exchange of commands between the client and the servers. The protocol keeps the interaction between the clients, VoD server and Database Server (Belda, 2001).

There are several applications of VoD systems like broadband tv broadcast, education, medicine (Brett, 1996) (Huang and Hu, 2000)(Erlandson, 2000) (Gebhard, 2001).

This paper is structured as follows: section II introduces the overview of the proposed VoD system called JMFVoD. Section III deals with the JMF libraries and RTCP/RTCP protocols. The implementation of the system is described in section IV and, finally, section V analyses the conclusion of this work.

OVERVIEW OF THE JMFVoD SYSTEM

As has been commented in the introduction section, there are several VoD systems developments in the literature and, at the same time, new multimedia libraries and API are appearing or are being improved like Java Media Framework.

The main aim of the VoD system we present (JMFVoD) is the validation of the viability of JMF libraries for the implementation of a broadcasting VoD system. This system will allow a real evaluation of the management mechanisms for multimedia sessions, RTP/RTCP protocols and compression standards included in the JMF libraries. The developed prototype will serve as basis for the future evaluation of new multimedia protocols that will substitute the traditional simulation studies in this field.

Figure 1 represents the architecture of the JMFVoD system and the communication flow between the different system elements.

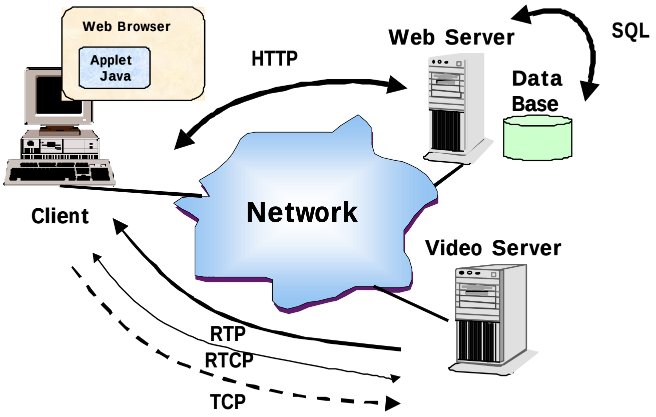


Figure 1: JMFVoD architecture

Basically four elements can be differentiated: client, web server, database and video server.

- *Client*: its development is based on a Java applet which is started each time the correspondent HTML page is loaded from the web server. Through the graphic interface, the users can select from the database the video title they want to playback.
- *Web server*: stores the start page and the Java applets. It also holds the authentication procedures and the database access management.
- *Database*: keeps the information catalogue of the available videos.
- *Video server*: stores the video files and send them to the clients as they request them.

With regard to the generic system operation, four phases can be considered:

1. The clients accesses the web server using an http connection in order to load the start page and the configuration applet. Without authentication the clients only will be able to visualize the available videos classified by title and category.
2. After selecting the video, the users run the authentication procedure using SHA-1 to cipher the information. The client personal information is validated by the web server. If the access is granted, then the client will be able to specify the features of the video formats provided by JMF (JPEG, H.263, MPEG, etc.).
3. At this point, the client starts the reception and playbacks the multimedia file (synchronized video and audio), using the RTP/RTCP implementation provided by JMF.
4. The client will be able to control the playback (stop, rewind, pause, etc.) through a TCP connection using an easy proprietary control protocol.
5. The last step corresponds to the disconnection from the system, closing RTP sessions and TCP control connections.

Additionally, it has been developed a management application of the database for modifying the available video file catalogue.

LIBRERIES AND PROTOCOLS: RTP/RTCP, CONTROL PROTOCOL AND JMF

RTP, Real-Time Transport Protocol, and its companion RTCP, Real-Time Transport Control Protocol, make it easier the transmission of audio and video flows over the Internet. RTP provides end-to-end delivery service for real-time applications, whereas RTCP provides feedback information about the session in progress. Both protocols work at the application level and are encapsulated in the underlying transport protocol, usually UDP, allowing unicast and multicast sessions. These protocols have been standardized by the IETF.

The motivation of using RTP/RTCP for audio and video transmission is significant. RTP provides load identification, sequence numbering and time stamp, whereas RTCP monitors the delivery packets and sends back information about packet lost and jitter variation. These services allow audio/video flow synchronization and the possibility of adapting the available bandwidth to current network state. These advantages have lead to many research groups to integrate RTP in its projects. This way, Kom Player (Zink *et al*, 2001), project has develop its own RTP library.

Among the different alternatives to implement our VoD system, we decided to use Java for its current importance in web-based multimedia applications. Java Media Framework (JMF) (Sun, 2001), developed by Sun MicroSystems, Inc. e IBM Corporation, is an application program interface (API) useful for including real-time data in Java applications and applets. This library supplies mechanisms to control the capturing, processing and rendering of multimedia flows. Moreover, the RTP API is completely integrated in the JMF architecture. This API consists of a set of Java interfaces that gather all the RTP/RTCP protocols tasks. This way, apart from providing facilities for data processing (codecs, multiplexers and demultiplexers, renders,…) JMF allows transmission and reception of multimedia data using RTP in unicast and multicast modes, and the utilization of several capture devices like webcams. There are different versions of the API, the most recent is v2.1.1.a, although new improvements are being included. From the JMF point of view, RTP is handled as any source or destination of multimedia flows. In our VoD system, the sender use the RTP API to deliver audio and video flows stored in a file and the client application displays the RTP packets received from the server .

The audio coders which are available in the last version of JMF are G.711, dvi, G.723 and GSM, whereas as video coders it is possible to use JPEG, H.261, H.263 and MPEG.

Another important point to take into account in a VoD system implementation is the protocol that controls the streams transmission. The current version of JMF supplies the RTSP (Real Time Streaming Protocol) client but does not include the RTSP server. For this reason, it was taken the decision of implementing a proprietary client-server control protocol that in a simple way was able to offer this functionality. This protocol is describer later in the next section.

JMFVoD SYSTEM IMPLEMENTATION

A. System definition

The VoD system (Figure 2) that has been developed consists of three super blocks or applications that communicate with each other and a database containing the multimedia files published in the system.

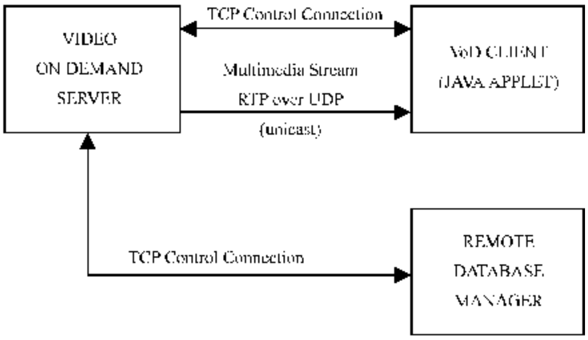


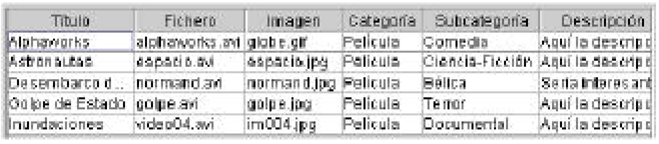
Figure 2: JMFVoD system

- The **VoD server** listens to a TCP port and waits for a client to make a connection request. There are two kinds of clients that can make a connection request: the VoD client and the remote database manager. Both clients are applets and run in web navigators. The server exchanges messages defined in a control protocol created ad-hoc with the clients through the opened sockets.
- The **VoD client** gets the data related to the published multimedia files, selects the file that wants the server to transmit and controls the received multimedia streams by sending messages to the server through the TCP control socket. Those streams are transmitted using the RTP/RTCP protocols over UDP in unicast sessions between the server and the client.
- The **remote database manager** shows the user the database contents and sends back the changes being made to them, so that the server can upgrade the database.

The server allows a fixed number of VoD clients to access the system simultaneously. This value and other parameters values can be configured from the server application.
Lastly, the VoD server includes a users authentication mechanism along with an application to manage the list of users that can access the system.

B. Database structure.

The published files database used by the application consists of three tables: the *main table*, the *categories table* and the *subcategories table*.
The main table contains the published multimedia files names and other related attributes. This table has six columns (Figure 3):



Titulo	Fichero	Imagen	Categoría	Subcategoría	Descripción
Alphaworks	alphaworks.avi	globe.gif	Película	Comedia	Aquí la descrip
Astronautas	espacio.avi	espacio.jpg	Película	Ciencia-Ficción	Aquí la descrip
Desembarco d...	normand.avi	normand.jpg	Película	Bélica	Sería interesant
Golpe de Estado	golpe.avi	golpe.jpg	Película	Terror	Aquí la descrip
Inundaciones	video04.avi	im004.jpg	Película	Documental	Aquí la descrip

Figure 3: Database structure

- *Title*: title assigned to the published files multimedia content.
- *File*: multimedia file name. It can be either the absolute path or a relative path referred to the multimedia directory specified in a configuration parameter.
- *Image*: image related to the multimedia content it is being published.
- *Category*: name of the category the file belongs to.
- *Subcategory*: name of the subcategory the file belongs to.
- *Description*: memo field with the file attributes or a description of the file content that contains a maximum of 65.535 characters.

The categories and subcategories tables allow to organize the multimedia files in a hierarchical way.

C. Control protocol

The system's control protocol consists of the exchanged messages between the VoD server and both the VoD client and the database remote manager. These messages are divided into four groups: *database management messages, authentication messages, selection messages and stream control messages*.

a) Database management messages

These messages allow the clients to connect to the database in two different modes: read-only and exclusive.
In the read-only mode the user cannot modify the database. This mode is only used to get the database content.
In the exclusive mode the user can modify the database but only one user at a time is allowed to access the database.
Some of these messages are followed by a SQL statement that has to be executed by the VoD server.
When a remote client requests the server to read the database, it sends the database contents in rows. The first row contains the headers of the columns of the table that is being read, next each row of the table is sent over the TCP control connection.
To send a row, the server sends the content of each column of that row in different lines. When reading, the client looks for the new line character to separate the different fields.
The last column data from the main table, *Description*, can contain many new line characters, so the server has to send a stamp to show that the description field has ended in order to recover the data properly at the client side.
The messages exchanged when reading and managing the database are:

- Exclusive mode reading request of the contents of a table from the database: *LIST EX*

SQL statement: *SELECT * FROM «table name» ORDER BY «column name»*

- Read-only mode reading request of the contents of a table from the database: *LIST*

  SQL statement: *SELECT * FROM «table name» ORDER BY «column name»*
- End of description stamp: *END OF DESCRIPTION.*
- Database upgrade: *UPDATE DB.*

  This message is used both to modify some field from the database and to add or eliminate rows. It is followed by the SQL instruction that specifies the desired change. The database manager only allows to modify the main table from the database.
- Database disconnection request: *DISCONNECT DB.*

  When accessing the database in exclusive mode, the client goes on being connected until it sends this message. When the server receives this message it releases the exclusive access so that another client can request it.

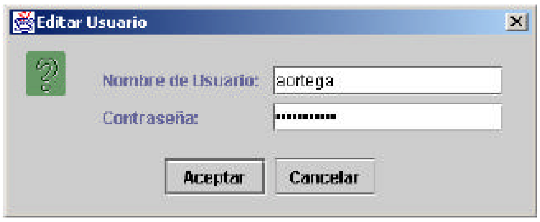
  In the read-only mode it is not necessary to send this message, because the client will disconnect automatically from the database when the query is completed.

b) Authentication messages

The messages exchanged during the authentication process are:

- Authentication request: *REGISTRATION* followed by the login name.
- Authentication passed: *ACCEPT*
- Authentication failed: *REJECT*
- Authentication information: time stamp, digest password.

When the user selects and requests a multimedia file for the first time, the client applet asks for a login name and a password (Figure 4). The authentication process begins when the user enters the requested data.



Nombre de Usuario: aortega

Contraseña: ••••••••••

Figure 4: Authentication interface

1. First the client sends the message *REGISTRATION* to the server, so it can also start the authentication process.
2. Next, the client sends the login name to the server.
3. The server looks for the received login name in the authorized users list and checks that there is not another opened connection with the same login name. If some of these verifications fail, the server sends the message *REJECT* to the client and the authentication process ends up.
4. If the previous verifications are correct, the server sends a time stamp to the client. This time stamp is taken from the server system clock.
5. Both the server and the client start calculating the digest password. In these calculations they have to append the time stamp to the password and then apply the **SHA-1** algorithm to the result. The client will use the password entered by the user whereas the server will use the password read from the authorized users file.
6. The client sends its calculated digest password to the server, that checks it with the one it has calculated. If both digest passwords are equal, the server sends the *ACCEPT* message allowing the transmission of the selected multimedia file. Otherwise the server sends the *REJECT* message.

c) Selection messages

This group consists of the following messages:

- *PLAY FILE*
- *BREAK*
- *SUPPORTED FORMATS*

Once authenticated, the client sends the selected multimedia file transmission request to the server. This is done by sending the message *PLAY FILE:* followed by the selected file name (in the same line).

The server tries to process the requested file. If an error happened (i.e. no file found) the server sends a *BREAK* message to the client showing that it cannot process the request. When the client receives a BREAK message it stops waiting for other messages or multimedia streams and turns back to the initial selection panel.

While processing the multimedia file for its transmission over RTP, the server requires that the client selects the transmission formats. The JMF library supports some multimedia streams transmission formats over RTP both for audio and video files. At the beginning of each list the server inserts a message about the content type of the list:

VIDEO FORMAT / *AUDIO FORMAT* / *OTHER FORMAT*

When the server has the information about the supported transmission formats it shall send them to the client so that the user can select the desired formats. To begin this data transference the server sends the message *SUPPORTED FORMATS*. Next the server sends the number of lists that are going to be transmitted (one per each stream the multimedia file contains). After this, the server starts transmitting the lists. For each list the server sends its length followed by every supported format in the list.

In the other side, the client waits for the *SUPPORTED FORMATS* message and, once received, show them to the user so that he can choose the appropriate transmission formats.

The selected formats are then sent to the server so that it can start the transmission of the multimedia file.

Finally, the server sends the number of RTP sessions that are going to be created and the full duration of the file specified in nanoseconds.

d) Stream control messages

The stream control messages gives local control of the remote transmission to the client. These messages are sent over the TCP control connection to the server so that it can start or stop the transmission and set the media time of the RTP streams. There are five stream control messages:

- *PLAY*: starts or restarts the stopped streams.
- *PAUSE*: pauses the streams reproduction at the current time.
- *REWIND*: stops the reproduction and sets it to the beginning of the media file.
- *POSITION*: changes the streams media time to the one specified by the parameter sent immediately after this message (in the same line) in nanoseconds.
- *EJECT*: ends up the streams reproduction and releases any exclusive resource used in the transmission and reproduction.

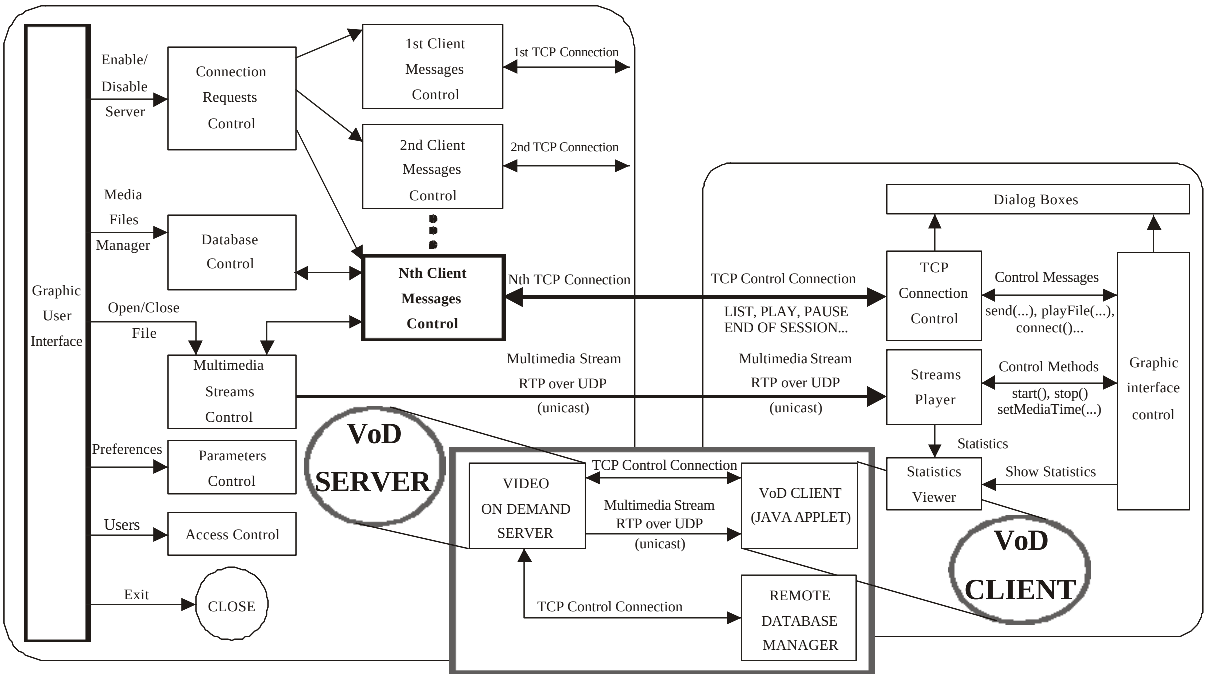


Figure 5: Block diagram

D. VoD Server application

The server application builds a simple graphic interface through which the user can access its different options (Figure 5).

The **Connection requests control** block is created when the server is activated from the user's interface. This block listens to a TCP port specified in a configurable parameter. When it receives a connection request it creates a new execution thread related to this connection. This thread will have to receive and process the clients requests and to send the appropriate responses while the applications keeps on waiting for other connection requests. The **Messages control for client i** blocks implement these execution threads.

The **multimedia streams control** block provides the mechanisms to process multimedia files both for playing them locally and for creating and controlling the RTP sessions needed to send the multimedia streams to the client. The *Messages control for client i* block have access to this block to process a multimedia file transmission request from the client it is related to.

In local reproduction, there is minimum stream processing and the server creates a new player frame separately from the server frame.

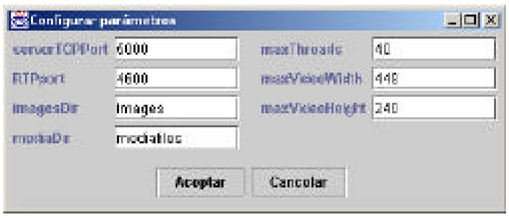
When the server receives a multimedia file transmission request, it calls the appropriate methods to process and configure the requested streams, creating as many RTP sessions as streams the requested file has. Once the transmission has started, the client can control its flow and temporal position by sending control messages to the *control messages* block assigned to its TCP control connection on the server side. This block will get into contact with the multimedia streams control block to make the requested changes on the streams.

The **parameters control** block provides the mechanisms to read and write the configurable parameters of the VoD system. This block dynamically creates a parameters presentation window according to the configurable parameters number found in the configuration file.

This block can be instantiated through the user interface of the VoD server. The *Preferences* option shows up a dialog frame with the configuration parameters and its current values. The names and values of the parameters are read from the HTML file that contains the VoD client applet. In this way, the parameters set up by the server are also read from the client so that it can configure itself.

The *Preferences* option is not enabled after starting the server to avoid the clients to be configured with different parameters.

The configuration parameters of the VoD system are (Figure 6):



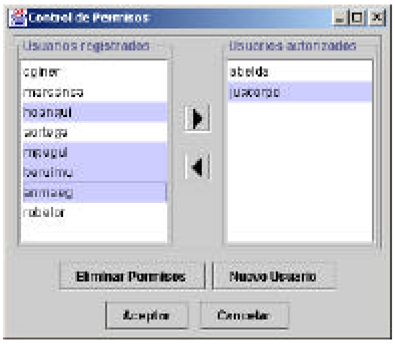
serverTCPPort 6000
maxThreads 40
RTPport 4600
maxVideoWidth 448
imagesDir images
maxVideoHeight 240
mediaDir mediafiles

Figure 6: Configuration interface

- *serverTCPPort:* the TPC port in which the server listens to the clients connection requests.
- *RTPPort:* the base port for the multimedia streams transmission. It must be an even positive integer less than 65535. Beginning by this port the successive multimedia streams will be transmitted over consecutive even ports.
- *imagesDir:* name of the directory where the images related to the published multimedia files are located. This directory must be located under the working directory path of the application.
- *mediaDir:* name of the directory where the published multimedia files are located. This directory must be located under the working directory path of the application. Any non-absolute path to a multimedia file in the database will be routed to this directory.
- *maxThreads:* maximum number of clients allowed to be connected to the server simultaneously. This limit depends on the capability of the hardware in which the VoD server is running.
- *maxVideoWidth:* the video maximum width in pixels. If a multimedia file has a video stream with a greater width, its size will be reduced maintaining the aspect ratio.
- *maxVideoHeight:* the video maximum height in pixels. If a multimedia file has a video stream with a greater height, its size will be reduced maintaining the aspect ratio.

The **Database control** block accesses to, reads and modifies the database, by means of SQL instructions, using the JDBC-ODBC bridge driver provided by the library *java.sql*.
The database management can be done locally, from the server itself, or remotely, from a database manager applet. In the first situation the database control block creates the graphic interface that allows the user to modify, add or remove database entries. This interface translates the changes made by the user to the appropriate SQL instructions. The database remote management is performed through a *messages control* block related to a *database remote manager* applet. That applet creates a graphic interface for the database management identical to the local interface, but in this case the SQL instructions are sent to the messages control block so that it executes them through the *database control* block.
In addition, any VoD client can request the database content to show it to the user. These readings are performed by the database control block through the messages control block related to each client. The system allows as many simultaneous database readings as simultaneous VoD clients are configured in the *maxThreads* parameter.

The **Access control** block manages two files (Figure 7): *registered.usr* and *authorized.usr*.



Usuarios registrados
cginer
marsanca
hoangui
aortega
mpagui
baruimu
anmaeg
robelor

Usuarios autorizados
abelda
juacorpo

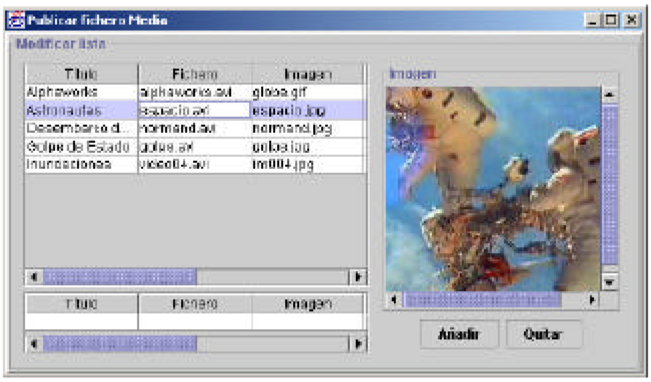
Figure 7: Control parameters interface

Both of them contain login names and passwords of the users. The users in the *registered.usr* file are registered in the VoD system but are not allowed to access to the VoD server. On the contrary, the users in the *authorized.usr* file are allowed to access to the server.
When a client requests the transmission of a multimedia file the access control block looks for its user login name and password in the *authorized.usr* file. With this information the *messages control* block tries to authenticate the user.
The registered and authorized users lists can be managed through the access control block user interface, adding or removing users and/or access permissions.

E. VoD Client Application

The VoD client application implementation is divided into different blocks. The main goal of these blocks is to act in response to user events. The **graphic interface control** block listens to user events and calls other blocks methods to take the necessary actions.
When opening the HTML page that contains the client applet the *graphic interface control* block is loaded. This block is also responsible for the VoD applet initialization. It follows the next steps:

1. The parameters from the HTML file are loaded.
2. It sets up a TCP control connection to the VoD server.
3. It sends a request to the server to read the database tables (Figure 8).
4. The values from the database are read and formatted appropriately.
5. The graphic interface images are loaded.
6. The graphic interface components are initialized.



Título	Fichero	Imagen
Alphaworks	alphaworks.avi	globe.gif
Astronautas	espacio.avi	espacio.jpg
Desembarco d...	normand.avi	normand.jpg
Golpe de Estado	golpe.avi	golpe.jpg
Inundaciones	video04.avi	im004.jpg

Título	Fichero	Imagen

Figure 8: Access Database interface

The **TCP control connection** block is committed for sending and receiving control messages exchanged between the client and the server. The graphic interface control block communicates with this block in order to send user requests to the server.

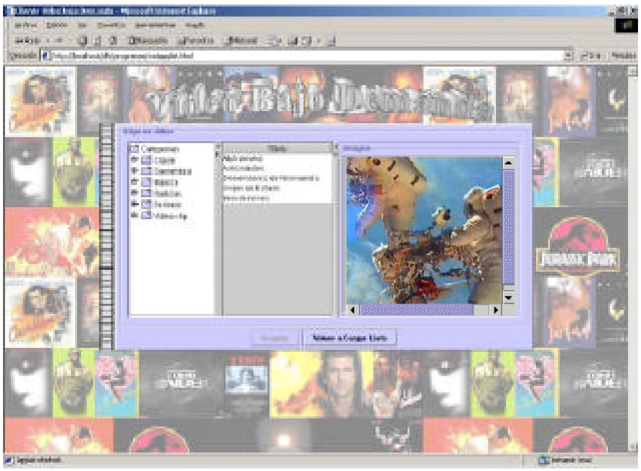


Figure 9: Client Playback interface

The **Streams player** block is initialized by the graphic interface control block when the user requests the transmission of a multimedia file. At that moment this block creates the necessary unicast RTP sessions to receive the multimedia streams the server is transmitting. This block is also responsible for streams synchronization and reproduction (Figure 9).

The **Statistics viewer** block gets the RTP sessions statistics from the *streams player* block. The user can bring up a dialog where the statistics from each stream are updated in real time within a second delay. These statistics display information of the number of duplicate, invalid, lost, miss ordered and processed RTP data packets. This block is created from the graphic interface control block.

F. Remote database manager

The remote database manager is an applet embedded in a web page. When the applet is loaded, it generates a graphic interface identical to the one generated by the local database manager. The database changes introduced by the user are sent through a TCP control connection to the VoD server. This connection is established on the applet initialization. If there is not another user managing the database, the server will send the database content to the TCP control connection block on the remote manager side.

The graphic interface has a table with the database contents. A user can modify the data of this table, as well as add or remove rows. This application generates the SQL statements required for each database update and sends them to the server so that it can execute them.

CONCLUSSIONS AND FUTURE WORK

This paper describes a VoD system, JMFVoD, based on the use of the JMF library and the RTP/RTCP protocols. The VoD server consists of a stand-alone application whereas the VoD client has been developed as an applet integrated in web. The implementation of this system shows the viability of the JMF library to develop multimedia applications. At the same time, this system provides a power tool for the evaluation of different audio and video formats transmission in a real network.

JMFVoD can be considered a true-VoD system which provides an easy interface to its users. The number of VoD clients which can connect to the system simultaneously depends on the capacity of the VoD server.

From the beginning it was though to use the RTSP protocol as the streams control protocol between the client and the server, but the actual version of JMF only supplies the client version and does not include the RTSP server. For this reason, it was necessary to implement a proprietary client-server control protocol which was able to commit this task. At the time the RTSP server version is available, it will be integrated as part of the JMFVoD system (Fortino *et al*, 2002).

Among the JMFVoD features, it can be pointed out its flexibility for application and protocols parameters configuration, the simple database local or remote management by using SQL statements and the security tools to access the system.

The audio and video formats that can be used in this system are all the coders available in the JMF/RTP library. As new coders are available in this library, they will be automatically integrated in the JMFVoD system.

As future work, new improvements will be included in the JMFVoD system like the integration of quality of service (QoS) mechanisms, both resource reservations and state network adaptation.

REFERENCES

Belda, A, 2001, "JAVA Client/Server Application for a Web-based VoD", Master Thesis, Technical University of Valencia.

Brett,P., 1996, "Using Multimedia: an Investigation of Learners' *Computer Assisted Language Learning* vol 1, n 2-3, pp. 191-212.

Erlandson, B.E., 2000, "Collaborative and multimedia medical *Information Technology Applications in Biomedicine*, pp: 153 –157

Fortino G., Palau C. E., Russo W., Guerri J.C., Furfaro A., Pajares A., 2002, "A Java-based adaptive media streaming on-demand platform", In *Proceedings of SCS Euromedia*, Modena, Italy, (Apr).

Gebhard, H. and Lindner, L., 2001, "Virtual Internet Broadcasting", IEEE Communications, vol. 39, no. 6, pp. 182-188 (June)

Huang; S. and Hu, H., 2000, "Integrating Windows streaming media technologies into a virtual classroom environment", *Proceedings of the International Symposium on Multimedia Software Engineering*, pp: 411 -418

Lee, J., 1999, "UvoD: An Unified Architecture for Video-on-Demand Services", *IEEE Communications Letters*, vol. 3, no. 9, pp.277-279 (Sept)

Schulzrinne H., S. Casner, R. Frederick and V. Jacobson. 1996., "RTP: A Transport Protocol for Real-Time Applications". RFC1889, (Jan).

Sun, 2001, Java Media Framework, available at http://java.sun.com/products/java-media/jmf/index.html

Wright, W., 2001, "An Efficient Video-on-Demand Model", *IEEE Computer*, vol. 34, no. 5 , pp. 64-70, (May)

Zink, M., Griwodz C. and Steinmetz, R., 2001. "KOM Player – A Platform for Experimental VoD Research", ISCC 2001, Hammamet (Tunisia)

JPEG2000 FULLY SCALABLE IMAGE ENCODER BY CONFIGURABLE PROCESSOR

Hiroshi Tsutsui, Takahiko Masuzaki, Masayuki Oyamatsu,
Tomonori Izumi, Takao Onoye, and Yukihiro Nakamura

Department of Communications and Computer Engineering, Graduate School of Informatics, Kyoto University
Yoshida-hon-machi, Sakyo, Kyoto, 606-8501 Japan
E-mail: {tsutsui, masuz, oyama}@easter.kuee.kyoto-u.ac.jp, {izumi, onoye, nakamura}@kuee.kyoto-u.ac.jp

KEYWORDS

JPEG2000, Scalable coding, Xtensa, Configurable Processor

ABSTRACT

In this paper, a JPEG2000 encoder for fully scalable image coding is described. To exploit different aspects of scalability inherent in JPEG2000, a set of novel mechanisms for pass termination, layering, and tile-part organization is devised. The proposed encoder is implemented through the use of Tensilica's configurable processor Xtensa optimized by user defined specific instructions.

INTRODUCTION

Recently, a variety of network systems is developed in terms of transmission media (wired or wireless) and bandwidth (Kbps to Gbps). In order to treat digital images in such a network system efficiently, a key to success is the concept of "*scalable*" coding, which can offer various image qualities at various bitrates by only one bit stream.

In January, 2001, *JPEG2000* (ISO/IEC JTC1/SC29/WG1 2000) was standardized by ISO/IEC JTC1/SC29 WG1 (commonly known as the JPEG), in which wavelet transformation is adopted to decorrelate images spatially to improve compression efficiency. With the use of this transformation, so-called *embedded stream* can be generated, in which code for low quality/bitrate image is included in that for high quality/bitrate image. Therefore, JPEG2000 can be regarded as the viable image coding scheme in the coming network era.

In order to attain the scalability, there are five progression orders possible in JPEG2000 standard. To explain briefly these progression orders, let us define *codestream*, *packet* and *layer*. In JPEG2000 coding scheme, a collection of bit streams and markers is called a codestream. Layer is a collection of compressed image data divided according to its significance. Packet is a collection of all compressed image data representing a specific position, layer, component, and resolution level, and it appears in a codestream as a contiguous segment.

In codestream, packets are interleaved to constitute a nest of loops in the following five progression orders; layer-resolution level-component-position (L-R-C-P), resolution level-layer-component-position (R-L-C-P), resolution level-position-component-layer (R-P-C-L), position-component-resolution level-layer (P-C-R-L), and component-position-resolution level-layer (C-P-R-L).

Among these five progression orders, R-P-C-L , P-C-R-L, and C-P-R-L can be easily implemented since less significance of layer, i.e. located in the most inner loop, enables the implementation which does not use efficient image layering.

On the contrary, in L-R-C-P and R-L-C-P, compressed image data must be separated into layers to achieve high scalability of the image. Motivated by this, the present paper describes JPEG2000 encoder for "*fully scalable*" image coding which supports all five progression orders.

The main procedure for JPEG2000 encoding is shown in Fig. 1. Details of each process are explained in Section 2. Here, let us show what must be the key processes for fully scalable image coding. The most significant part to realize scalable coding is the data ordering process. In this process, packets are ordered and codestream is organized. In addition, the scalability in terms of image quality can be achieved by the concept of layer in such a way that a packet organized by the data ordering represents a specific layer. To enable this, the data ordering needs some information for efficient layering such as possible boundaries of packets, which can be obtained from coefficient bit modeling and Arithmetic coding.

Thus, the present paper introduces new mechanisms for pass termination, layering, and tile-part organization.

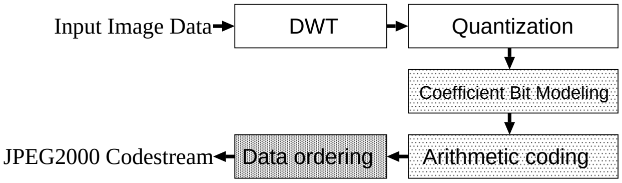


Fig. 1: Block diagram of JPEG2000 encoder

The proposed JPEG2000 encoder is constructed in C through the use of Tensilica's configurable processor Xtensa (Tensilica Inc.,2000) in conjunction with user defined instructions described by Tensilica Instruction Extensions (TIE) language (Tensilica Inc.,2000).

The rest of this paper is organized as follows. The JPEG2000 encoding algorithm is summarized in Section 2. In Section 3, pass termination mechanism, layering mechanism and tile-parts organization mechanism needed to realize scalable coding are described alongside of their simple but efficient implementation. The experimental results of our approach are shown in Section 4. The optimization of the proposed JPEG2000 encoder by TIE is also mentioned in Section 4. Finally, Section 5 concludes this paper.

JPEG2000 ENCODING ALGORITHM

In JPEG2000 coding scheme, first a target image is divided into square regions, called *tiles*. The concept of tile in JPEG2000 reduces computational complexity of image coding, since each tile can be coded independently each other.

Then 2-D discrete wavelet transformation (shortly DWT) decomposes a tile into LL, HL, LH and HH subbands. LL subband is a low resolution version of the original tile and again is to be decomposed into four subbands recursively. This decomposition is called *Mallat decomposition* (Mallat,1989). A subband is divided into code-blocks, each of which is coded individually by *coefficient bit modeling* described later. Fig. 2 depicts subbands, their *resolution levels*, and code-blocks.

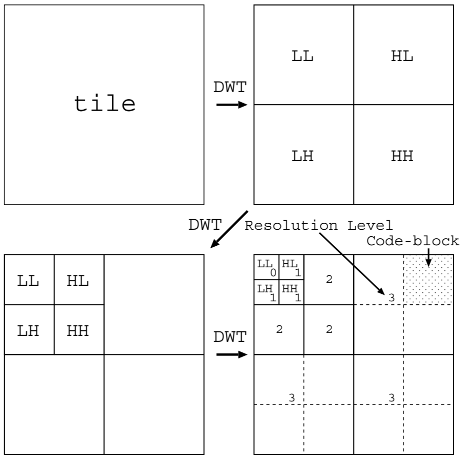


Fig. 2: Subbands and resolution levels resulted from DWT and code-blocks.

Wavelet coefficients in a code-block is separated to sign bits and absolute values, and so-called *bit-planes* are generated from the bits of absolute values such that each bit-plane refers to all the bits of the same magnitude in all coefficients of the subband. Then bit-planes are coded from the most significant one to the least significant one. Fig. 3 illustrates notion of bit-plane and how a bit stream is organized.

Coefficient bit modeling is a process to label bits of a bit-plane based on the statistical information through three different coding passes, which allows efficient compression by succeeding *MQ-coder*, a kind of arithmetic coder. MQ-coder generates a set of packets, each of which is a part of the codestream constructed by a header information and the compressed image data originated from a specific layer, position, resolution level, and tile.

Important processes and an essential concept of JPEG2000 encoding to discuss our approach, coefficient bit modeling, arithmetic coding, data-ordering, and the concept of tile-part, are described in the following.

Coefficient bit modeling

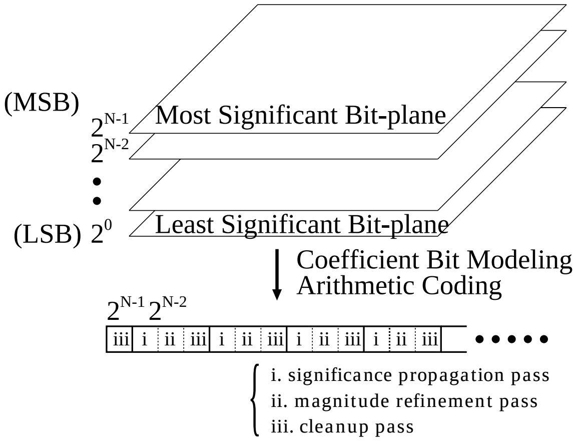

(MSB)
2^{N-1}
Most Significant Bit-plane
2^{N-2}
(LSB) 2^0
Least Significant Bit-plane
Coefficient Bit Modeling
Arithmetic Coding
2^{N-1} 2^{N-2}
iii i ii iii i ii iii i ii iii i ii iii
i. significance propagation pass
ii. magnitude refinement pass
iii. cleanup pass

Fig. 3: Bit-planes and coded passes.

Each coefficient bit of a bit-plane is encoded by one of the three passes according to *significance state* of eight nearest-neighbor coefficients. The three passes are *significance propagation pass*, *magnitude refinement pass* and *cleanup pass*.

Here, let us explain about these passes and significance state of a coefficient. Each coefficient in a bit-plane has an associated binary state variable called its *significance state*. First, significance state of each coefficient is initialized to 0 (insignificant) and the significance state changes itself from insignificant to significant at the bit-plane where the magnitude bit is found 1. The significance propagation pass only includes bits of coefficients that were insignificant (the significance state has yet to be set) and whose eight nearest-neighbor coefficients are not all insignificant. The magnitude refinement pass includes the bits from coefficients that are already significant (except those that have just become significant in the immediately preceding significance propagation pass). All the remaining coefficients are left as insignificant and included in the cleanup pass. Each sign bit is decoded in the significant propagation pass or cleanup pass if it has not been decoded yet.

Each bit-plane is scanned in order of pass and in each pass each coefficient is labeled by one of nineteen contexts. Then this context and a binary symbol are passed to successive MQ-coder.

Arithmetic coding

Binary symbols obtained by the coefficient bit modeling process are compressed by MQ-coder, which is a kind of binary arithmetic coder. MQ-coder compresses binary sequences by updating MPS (More Probable Symbol) and probability of LPS (Less Probable Symbol) of the context of each bit and returns an array of bytes. Original binary symbols are distributed in the compressed data unless *initialization* or *flush* of MQ-coder occurs.

Flush occurs at the end of each code-block or the end of each pass. Generally, flush only occurs at the end of each code-block so as to achieve high compression ratio. In this case, boundaries of passes in the compressed bit stream are not explicitly specified by JPEG2000 standard. However, it should be taken into account in determining a point of pass termination that all bytes needed to decode a pass must be included in its compressed bit stream.

Data ordering

The scalability of compressed image data, one of the novel features of JPEG2000 standard, is realized by the following data ordering.

A subband is divided into rectangle regions according to its original image position. These regions are called *precincts*. Here a *packet* is defined again as a collection of all compressed image data representing a specific precinct, layer, component and resolution level. In a codestream, packets are interleaved, as illustrated in Fig. 4, to constitute a nest of loops in the five progression orders.

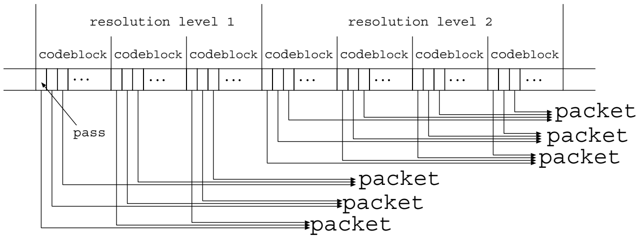


Fig. 4: Packet coordination

For instance, in the case of L-R-C-P progression order, layer must be processed the most outer loop in encoding. Packets included in a codestream are arranged according to the order of layer, packets included in one layer are arranged according to the order of resolution, packets included in one resolution are arranged according to the order of component, and packets included in one component are arranged according to the order of position.

Concept of tile-part

The concept of tile in JPEG2000 reduces computational complexity of image coding, since as mentioned before tiles can be coded independently each other. However, the scalability is limited inside of a tile and progressions over tiles of whole image can not be achieved. To facilitate scalability over tiles, so-called *tile-part* can be used in JPEG2000, which is a part or whole bit stream of a tile.

Given a set of packets for a tile, one or more adjacent packets are collected to constitute a tile-part accompanied with tile-part header. Generally, a group of tile-parts contains all packets in the tile, and these tile-parts can be interleaved as a unit in JPEG2000 code stream. In other word, a tile is divided into tile-parts and, by use of this concept of tile-parts, scalability over tiles are realized.

JPEG2000 ENCODER FOR FULLY SCALABLE IMAGE CODING

There are two approaches possible to realize progression orders. The one is to carry out coefficient bit modeling and arithmetic coding according to progression orders, and the other is to execute coefficient bit modeling and arithmetic coding according to the order of input data and then organize codestream to achieve progression. However, the former approach is resource consuming since nest structure of progression order usually differs from that of input order. Thus, for example in the case of L-R-C-P, the coefficient bit modeling and arithmetic coding need to stop and store their tables as temporal results into memory at every layer termination. On the contrary, in the latter approach, the coefficient bit modeling and arithmetic coding need not to stop and store their tables. The progression can be obtained after arithmetic coding by the data ordering process. In this paper, the latter approach is adopted aiming at efficient resource usage.

Since MQ-coder outputs a byte sequence in code-block order at the time of layering, the byte sequence of one layer is distributed. Thus, boundaries of layers of the output sequence of MQ-coder must be provided. It is specified in the standard that boundaries of layers must be boundaries of passes, and hence pass termination points in the compressed image must be determined.

From above discussion, the key faculties needed to realize scalable image coding can be resolved as

- pass termination in compressed image,
- layering, and
- tile-parts organization.

Henceforth simple but efficient mechanisms for these faculties are described.

It should be stated here that the concept of precinct in JPEG2000 coding scheme is not adopted here mainly because without this concept scalability based on the size of tile is obtained.

Pass termination mechanism

To realize layered coding, pass termination points in the compressed image must be determined since each of which must be identical with a boundary of passes. A trivial way to determine pass termination is to utilize the result of decoding compressed sequence. However, this approach is highly resource consuming one and impractical. Thus in this subsection, memory efficient simple pass termination mechanism is introduced.

MQ-coder has a register of four bytes in order to store the cumulative probability, from which the output bytes are generated. Therefore our proposed way is that when MQ-coder has encoded the last bit of a pass, the index of memory to specify the end of pass is increased by four so that, at the time of decoding, MQ-decoder can completely reconstruct the pass even succeeding passes are included in a different packet. If the pass is the end of a code-block and flush operation of MQ-coder occurs, the index is not increased and stored as it is.

The proposed pass termination mechanism is quite simple. By proposed pass termination mechanism, all indexes from which passes start and all indexes from which code-blocks start are stored in order to give flexibility for layering.

Layering

The ideal scheme for layering is to decode composed image data one after another and to generate layers by referring to PSNR (peak signal to noise ratio). However, it suffers from high costs in terms of processing power, time, and resources. Thus we employed simple layering scheme, in which each bit-plane is assigned to one layer.

Some bit-planes from most significant one whose all bits equals to zero are called *zero bit-planes*. Generally, more important information is included in lower resolution. Therefore each of zero bit-planes whose resolution level is 0 or 1 is not assigned to any layer and each bit-plane lower than zero bit-planes is assigned to one layer, while each bit-plane of the code-blocks whose resolution levels are larger than 1 including zero bit-planes is assigned to one layer. The layers are empty when there is no bit-plane assigned to layers.

To evaluate our layering approach, quality of sample grayscale images with varying the numbers of cumulative layers are investigated. Fig. 5 depicts PSNR attained by proposed layering scheme. The specification of the used software encoder is shown in the next section.

Table 1: Result of our layered coding scheme.

image name	SIDBA_TITLE		AERIAL_SCENE		GIRL		LENA	
image size	256×256		512×512		512×512		512×512	
	bpp	PSNR	bpp	PSNR	bpp	PSNR	bpp	PSNR
layer 0	0.01	10.6	0.02	15.5	0.02	17.4	0.02	17.9
layer 1	0.07	12.1	0.07	18.7	0.05	24.2	0.05	23.4
layer 2	0.30	15.1	0.14	20.8	0.11	29.0	0.09	27.2
layer 3	0.80	21.2	0.32	22.8	0.18	30.9	0.16	29.1
layer 4	1.36	26.8	0.69	26.2	0.26	32.4	0.27	31.4
layer 5	2.00	33.8	1.37	30.6	0.40	34.1	0.46	34.3
layer 6	2.57	39.8	2.32	36.0	0.85	36.5	0.86	37.4
all layers	3.99	∞	6.01	∞	4.61	∞	4.49	∞

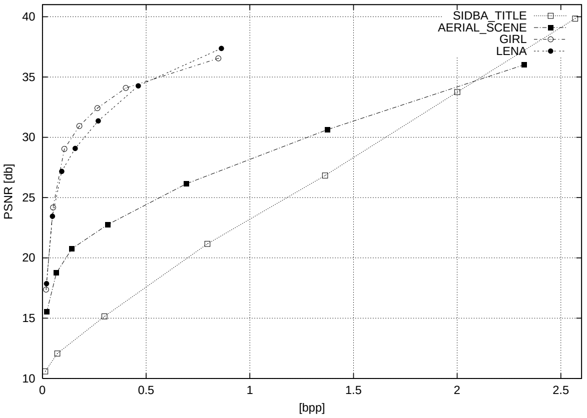


Fig. 5: Result of our layered coding scheme.

Tile-part organization

By gathering tile-parts of tiles in an image properly, scalability over tiles can be provided. In this paper, novel mechanism to set appropriate sizes of tile-parts for all progression order is devised.

As the size of a tile-part decreases, the overhead of tile-part headers to the codestream increases. Thus to maintain compression efficiency, the size of a tile-part must be set to maximum possible value. In our approach a tile-part whose size is maximum for each progression order is explored. For L-R-C-P and R-L-C-P, where position is located in the most inner loop, in order to realize scalability over tiles a tile-part representing a specific layer, resolution, and component is needed. Hence each tile-part corresponds to a packet. For R-P-C-L and C-P-R-L, a tile-part representing a specific resolution and a specific component is needed, respectively, in order to realize scalability over tiles. Thus a tile is divided into tile-parts whose numbers are same as the number of resolutions and components, respectively. For P-C-R-L, where position is located in the most outer loop, a tile-part representing a specific position is needed, so that a tile-part representing a specific position corresponds to a whole tile.

IMPLEMENTATION RESULTS

The proposed JPEG2000 encoder is implemented in C to evaluate our approach. The main specifications are summarized in Table 2.**Pass termination mechanism**

Table 2: Software JPEG2000 encoder specifications.

tile size	128×128
DWT	5-3 reversible filter
code-block size	32×32
decomposition level	3
use of precinct	no
progression mode	full support

To estimate the size of circuits and the delay time of MQ-coder and pass termination mechanism, these modules are designed in Verilog-HDL by using 0.15μm CMOS technology.

Through the use of Synopsys design compiler, MQ-coder and pass termination mechanism are synthesized to 5,900 gates and 1,300 gates, respectively. From this results, it is proved that our proposed pass termination mechanism can be implemented by about 20% additional hardware resource of MQ-coder.

Critical paths of MQ-coder and pass termination mechanism are 5.41 ns and 3.02 ns, respectively. The number of clock cycles needed by MQ-coder and pass termination mechanism to encode one binary symbol of sample images is 4.64 clock cycles on average, as summarized in Table 3.**Tile-part organization**

The tile-part organization mechanism is implemented by software, mainly taking account of flexibility of the layering.

The execution time of packet organization and our proposed tile-parts organization mechanism by Intel Pentium II (400 MHz, 128MB) processor. It can be seen from Table 3

Table 3: Experimental results for sample grayscale images.

image name	SIDBA_TITLE	AERIAL_SCENE	GIRL	LENA
image size	256×256	512×512	512×512	512×512
image type	artificial	airscape	portrait	portrait
symbols	457421	1788723	1331478	1332501
clock cycles	2036459	8425283	6296097	6270194
clock cycles/symbol	4.45205	4.71022	4.72865	4.70558
user time (sec)	0.09	0.49	0.38	0.37
compressed image size (bytes)	32698	197091	150971	147076
work memory (bytes)	8615	29057	21017	22380
work memory ratio (%)	20.85	12.85	12.22	13.21

that, in the case of images whose size are 512×512, our proposed packets and tile-parts organization mechanism can be done within 1 second.

As for the memory usage, in the progression orders except for P-C-R-L, the output order MQ-coder and the order of the JPEG2000 code stream are not identical. Therefore all image data compressed by MQ-coder must be stored once to memory.

The relation between the size of all image data compressed by MQ-coder and the size of work memory is summarized in Table 3. This work memory includes the memory needed to store indexes from which passes start, indexes from which code-block start, and some variables needed by software. Our proposed approach provides functionalities to support all progression orders with about from 10 $\sim$ 20% of additional memory on average.

Acceleration for Xtensa

Tensilica's configurable processor Xtensa has an ability to equip hardware resource described in TIE which can be execute machine codes directly. Hence performance improvement can be easily achieved by describing frequently executed functions by TIE.

Computational costs of all function by software JPEG2000 encoder in advance and concluded that coefficient bit modeling is the most computationally intensive one. Therefore coefficient bit modeling of software JPEG2000 encoder is optimized by TIE.

In this process, there is a function `element_of_plane` which checks whether the point (x, y) on a bit-plane is 0 or 1. The `element_of_plane` is a simple function, which costs 3 CPU cycles, but called many times. Thus this function is rewritten by TIE as a specialized instruction so that it costs only 1 CPU cycle. Function `reverse_yth_bit` reverses the bit of the point (x, y). Function `Sum_Vi` returns the arithmetic sum of $(x, y + 1)$ and $(x, y - 1)$. Function `four_elements_and` checks whether 4 successive bits are all 1 or not. Function `four_elements_or` checks whether 4 successive bits are all 0 or not. These four functions are also frequently called and can be implemented easily by TIE.

The numbers of cycles needed to encode sample images are evaluated on the Xtensa instruction set simulator as shown in Table 4 and Table 5.

Table 4: Profile of software JPEG2000 encoder.

function name	M cycles	cycles(%)	calls
`context_label`	335.79	20.02	1585948
`element_of_plane`	221.86	13.23	21088380
`coefficient_bit_modeling`	113.42	6.76	1628
`significance_pass`	88.64	5.28	1324
`cleanup_pass`	88.57	5.28	1628
total	1677.61		

Table 5: Profile of software JPEG2000 encoder with TIE acceleration.

function name	M cycles	cycles(%)	calls
`context_label`	135.58	13.67	1121078
`significance_pass`	83.89	8.46	1324
`soft_mq_tlc`	80.43	8.11	1
`cleanup_pass`	73.78	7.44	1628
`renorme`	55.76	5.62	933851
total	991.60		

As a result, about 40.9% performance improvement is achieved by TIE acceleration with small enough hardware overhead of 678 gates as summarized in Table 6.

Table 6: Sizes of functional modules.

function name	numbers of gates
`element_of_plane`	58
`four_elements_and`	52
`four_elements_or`	31
`reverse_yth_bit`	375
`sum_Vi`	162
total	678

CONCLUSION

The present paper has described an architecture of JPEG2000 encoder which supports fully scalable image coding. Novel effective mechanisms for pass termination, layering, and tile-part organization are introduced. In addition to its software implementation, acceleration of coefficient bit modeling by customized instruction, hardware implementation of MQ-coder and pass termination are described.

REFERENCES

ISO/IEC JTC1/SC29/WG1, WG1N1890R, "JPEG2000 Part I Final Draft International Standard Version 1.0," 2000.

Tensilica Inc., "Xtensa application specific microprocessor solutions — overview handbook," 2000.

Tensilica Inc., "Xtensa application specific microprocessor solutions — data sheet," 2000.

Tensilica Inc., "Tensilica instruction extension (TIE) language — user's guide," 2000.

Tensilica Inc., "Tensilica instruction extension (TIE) language — reference manual," 2000.

S. G. Mallat, "A theory for multiresolution signal decomposition : The wavelet representation," *IEEE Trans. PAMI*, Vol. 11, pp. 674–693, 1989.

MULTIMEDIA ON THE GO

AN XML-BASED FRAMEWORK FOR CONTENT ADAPTATION

Sam Lerouge, Boris Rogge, Dimitri Van De Ville,
Rik Van de Walle, and Jan Van Campenhout
Ghent University, Belgium
Department of Electronics and Information Systems - Multimedia Lab
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
Tel: +32 9 264 89 08, Fax: +32 9 264 35 94
E-mail: (slerouge,brogge,dvdevill,rvdw,jvc)@elis.rug.ac.be

KEYWORDS

XML, multimedia, content adaptation, terminals, negotiation protocol

ABSTRACT

The diversity of terminals that have access to multimedia resources over the internet is rapidly increasing. All of these terminals can have different capabilities and their users can have different preferences. Therefore, the need for adaptation of multimedia data to a specific terminal and/or its user has become indispensable. We need a framework that is able to adapt multimedia content to the terminal and the user that requested the data. This paper presents a content adaptation approach that is based on XML (eXtensible Markup Language). We also describe a basic implementation of our framework in which we incorporate existing XML technologies. Furthermore, requirements for a true content negotiation protocol are enumerated, in which the objective is to minimize computational overhead for the participating parties.

INTRODUCTION

In the recent past, the importance of multimedia applications has strongly increased, and multimedia information will become even more important in the future. It is thus obvious that the actual representation and the appropriate compression of the multimedia data are important research issues.

To be able to answer the demands of modern multimedia applications, new compression algorithms must support a large range of possible bit rates and high compression rates. In addition, they must be able to produce a strongly layered data stream to allow the adaptation of the compressed data to a variety of networks and terminals. Furthermore, these algorithms must allow scalability and guarantee a certain Quality of Service (QoS) for dynamic and heterogeneous environments.

Multimedia data is displayed to the user by means of a terminal. Recently, we have seen an increasing variety in the characteristics of different types of terminals. This means that resources may vary over these different types. Furthermore, the tasks of these terminals are not limited to the presentation of multimedia content. The application designer may also want to incorporate interaction capabilities of the user with the multimedia content. This interaction and the activity of other tasks executed by the terminal results in an unpredictable variation on the resources of the terminal. So, resources not only vary over different terminals, but they also vary in time for one specific terminal.

As a consequence, in addition to coding multimedia content in advance, we also need a terminal-dependent QoS negotiation mechanism between the provider of the data and the terminal. In this paper, we will make some suggestions for an XML-based adaptation of multimedia content to the terminal characteristics, or, more broadly, the *user environment*. This user environment consists of the following four aspects.

- The characteristics of the terminal (this captures both hardware and software).
- The network characteristics (such as average bandwidth and error rate).
- User preferences (such as preferred language).
- The natural environment of the user (such as location).

This paper is organized as follows. In the next section, we give an overview of existing tools, standards and frameworks that will play a role in our content adaptation framework. We also give some remarks on the transmission of XML in a network environment. Next, we describe the actual framework, first in a general way, and also by means of a description of the setup we use to test this framework. Then we explain how a true content negotiation framework could operate. The last section contains the conclusions of this paper and describes what our future work on content negotiation will consist of.

OVERVIEW OF EXISTING STANDARDS

Metadata

Metadata is commonly defined as *structured data about data*, and will be crucial in a content adaptation framework. Indeed, in order to adapt the content of a multimedia presentation to a certain user environment, we need to know some characteristics of the presentation that are not explicitly present in the multimedia data. Examples are: the color space and depth of an image or video, the language of a text, a textual summary of a video presentation.

The need for a metadata standard for multimedia data has been recognized by MPEG (Moving Picture Experts Group) with the development of the MPEG-7 *Multimedia Content Description Interface* (Martinez 2001). The main objective of this interface is to provide a language that enables complex search and retrieval operations in multimedia databases. However it is recognized by others (Smith 2000) that MPEG-7 metadata can be used in the context of content negotiation and adaptation.

The language for describing MPEG-7 metadata is defined by the DDL (Description Definition Language). Since this language is based on XML Schema, MPEG-7 metadata has an XML syntax. As a result, a number of tools are available for the processing

```
<Mpeg7 xmlns="http://www.mpeg7.org/2001/MPEG-7_Schema"
xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance">
  <ContentDescription xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="VideoType">
      <Video>
        <MediaInformation>
          <MediaProfile master="false">
            <MediaFormat>
              <Color>color</Color>
              <FrameWidth>320</FrameWidth>
              <FrameHeight>240</FrameHeight>
              <CompressionFormat>
                MPEG-1
              </CompressionFormat>
            </MediaFormat>
            <MediaInstance>
              <MediaLocator>
                 <MediaUri>
                   http://www.mysite.com/soccer.mpg
                 </MediaUri>
              </MediaLocator>
            </MediaInstance>
          </MediaProfile>
        </MediaInformation>
      </Video>
    </MultimediaContent>
  </ContentDescription>
</Mpeg7>
```

Figure 1. Example of an MPEG-7 document

of the metadata. Not only can we easily read and edit MPEG-7 data, we can also use the various existing techniques for navigating, searching, filtering and transforming XML data. An example of an MPEG-7 content description is given in figure 1.

Multimedia and XML

In the previous section, one example of the usefulness of XML within multimedia applications was shown. Not only the MPEG-7 standard, but also the most recent work of MPEG, called MPEG-21 *Multimedia Framework* (Bormans and Hill 2001), is based on XML. Nowadays, XML is being used in almost all aspects of multimedia. The following technologies are all standardized by the World Wide Web Consortium (W3C).

- Extended HTML (XHTML): a reformulation of HTML as an XML application.
- Synchronized Multimedia Integration Language (SMIL): for the synchronization of different individual audiovisual elements.
- Scalable Vector Graphics (SVG): a language for describing two-dimensional graphics.

The growing number of XML-based applications in the world of multimedia data shows the applicability of this language. Not only can it be used to markup a multimedia presentation, but it can also be used to describe the structure of non-XML multimedia data. An example of such an application can be found in (Devillers 2001), where scalable multimedia bitstreams, e.g., JPEG2000, are represented as XML in order to facilitate the adaptation of the bitstream to a specific user environment. In the remainder of this paper, we will suppose that the structure of the multimedia data is given in XML.

XML can be seen as a serialized form of a tree structure with one root node and child nodes. Each node has a name, can have an unlimited number of children and can have attributes with values. There are different ways to process XML data, but the following three options are most popular: DOM, SAX, and XSLT. We will elaborate these three options further in this section.

DOM, the Document Object Model, is a language-independent interface that allows programs to access and update the content, structure and style of any document that can be given a tree structure (typically HTML or XML). An advantage of working with DOM is that the processing application can freely access any node in the document tree. A disadvantage is that the whole tree is loaded in the memory before processing begins.

SAX (Simple API for XML) is an event-based API, in contradiction with DOM, which offers a tree-based API. This means that parsing events are reported to the processing application, where handlers are implemented to deal with such events. The major advantage of this approach is that for most applications the processing time is highly reduced while memory use remains very small. A disadvantage is that random access is not possible without building the whole tree in the memory, as DOM does.

The third way of processing XML data is by means of XSLT (Extensible Stylesheet Language - Transformations) (Clark 1999). This language originated from the need to translate XML documents to HTML, but it can capture any tree-to-tree transformation. An XSLT stylesheet consists of template rules, that produce a part of the output tree when they are matched with a node in the input tree. XSLT usually performs much less efficiently in speed and memory usage than a DOM or SAX approach. However, XSLT has one major advantage: it employs an XML syntax itself. This means we can use existing XML tools to edit XSLT stylesheets. Another consequence is that an XSLT stylesheet can be subject (input or output) to another XSLT transformation. We can use T-diagrams, as defined by Early and Sturgis (Early and Sturgis 1970), to show that the output of one transformation is used as stylesheet for another transformation. An example of the use of such T-diagrams can be found further in this paper (figure 6).

Shortcomings in the XML Domain

Nowadays, when XML data is to be transmitted over a network, it is sent as plain or compressed text, using a reliable protocol such as TCP. We believe more flexible solutions can be developed. In this section we give some suggestions that could lead to such solutions. We also touch the current problems of processing XML by means of XSLT in a network environment.

A first problem is the format in which XML data is transmitted in a network environment. Currently, XML is typically exchanged in a textual format. In our opinion, it is useful to consider other ways of transmitting XML data. The main reason is that when XML will have become a widely spread standard for exchanging data over the internet, multiple network nodes (like proxies and gateways) might interact with a stream of XML data. When we keep sending XML as text, every node will have to parse the data, process it, and then serialize the (possibly transformed) data again. The next node will have to start parsing from scratch, and does not take advantage of the parsing that has been done in the previous node. To avoid this, it might be appropriate to transmit parsing events (in the way SAX does) instead of textual data. To our knowledge, little or no research is done in this domain.

A second problem is the transmission of multimedia data in applications where data loss is acceptable (e.g., real-time streaming). To ensure a continuous stream, such applications often use

UDP instead of TCP, which can result in data loss. When we want to send XML over UDP, the data should be transmitted in such a manner that the loss of one packet does not corrupt the rest of the data. A solution would be to match every UDP packet with a subtree of the original XML tree, so the loss of one packet corresponds to the pruning of a subtree in the whole tree. This is of course a simplification of the problem, because we don't consider some obvious problems (e.g., cross-referencing can become invalid). We feel it is not yet clear whether it is possible to transmit XML data over UDP in specific applications.

A third topic is the evolution of the XSLT language and its implementations. In XSLT version 1.0, some useful operations were not allowed. Version 2 will make an end to some of these shortcomings. XSLT is constructed in a way that should allow a high level of optimizations, but the currently available XSLT processors make little or no use of these features. It should be possible for processing software to start sending output even before the complete input is received, but until now, no processors seem to be capable of doing so.

Another question is whether it is appropriate to add, remove or change template rules in an XSLT stylesheet during the processing of an XML data stream. To answer such questions, a clear formal model for XML and XSLT will be needed. Today, it appears that this kind of research is only happening in the context of XML databases (Vianu 2001).

Description of the User Environment

Recently, we have seen a few efforts in the domain of the description of terminal characteristics or even the complete user environment. The overview that we will give here is inspired by an input document for the MPEG meeting of January 2001.

The first work on content negotiation was done by the IETF (Internet Engineering Task Force), and resulted in the Transparent Content Negotiation (TCN) framework (Holtman and Mutz 1998). This is an extensible negotiation mechanism, layered on top of HTTP, for automatically selecting the best version of a certain resource when a URL is accessed. Conceptually, a request consists of two phases. In the first phase, the response consists of a list of variant descriptions, together with some metadata (MIME type, charset, language and features). In the second phase, the user agent selects the most appropriate version, and then requests this version. Figure 2 shows a possible response of the first phase. Proxy servers are allowed to cache the list of variants and the different versions. As such, they can reduce the overhead for the content server.

```
HTTP/1.1 300 Multiple Choices TCN: list
Alternates: {"doc.1" 0.9 {type text/html} {language en}},
            {"doc.2" 0.7 {type text/html} {language fr}},
            {"doc.3" 1.0 {type application/postscript}
                {language en}}
Vary: negotiate, accept, accept-language
```

Figure 2. A Response Header in the Transparent Content Negotiation framework

Within the W3C, a working group is developing a standard called CC/PP (Klyne et al. 2001), Composite Capability/Preference Profiles. The mission of this working group is to develop a framework for the management of device profile information. CC/PP is already in use in the UAProf (User Agent Profile) standard (Wireless Application Protocol Forum 2001), which is concerned with capturing capabilities of mobile devices. These classes include (but are not restricted to) the hardware and software characteristics of the device as well as information about the network to which the device is connected. Figure 3 shows a CC/PP description following the UAProf standard. Because CC/PP is RDF and thus XML-based, we will use this standard for the description of the user environment.

```
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:prf="http://www.wapforum.org/profiles/UAPROF/
ccppschema-20010430#">
  <Description ID="MyDeviceProfile">
    <prf:component>
      <Description ID="HardwarePlatform">
        <type resource="http://www.wapforum.org/profiles/
UAPROF/ccppschema-20010430#HardwarePlatform"/>
        <prf:ScreenSize>121x87</prf:ScreenSize>
        <prf:ScreenSizeChar>15x6</prf:ScreenSizeChar>
        <prf:ColorCapable>No</prf:ColorCapable>
      </Description>
    </prf:component>
    <prf:component>
      <Description ID="BrowserUA">
        <type resource="http://www.wapforum.org/profiles/
UAPROF/ccppschema-20010430#BrowserUA"/>
        <prf:CcppAccept>
          <Bag>
            <li>application/vnd.wap.wbxml</li>
            <li>application/vnd.wap.wmlscriptc</li>
            <li>image/vnd.wap.wbmp</li>
          </Bag>
        </prf:CcppAccept>
      </Description>
    </prf:component>
  </Description>
</RDF>
```

Figure 3. Example of a Terminal Description in UAProf

A BASIC FRAMEWORK FOR CONTENT ADAPTATION

General Framework

Figure 4 shows a general scheme for the adaptation of XML content to a specific user environment (UE). First, a request for content is sent to the web server. The server returns the XML data, which is adapted to the user environment by some engine, based on the metadata that can be found in (or is linked with) the XML data, and the description of the UE. The transformation engine can be any XML processing software that is aware of the target user environment. We will focus on the XSLT case, where the adaptation is defined by an XSLT stylesheet that corresponds to the UE. The transformation can be done at the server, at the client, or even at a proxy or gateway. All of these cases have their own advantages and disadvantages. Note that this framework fits within the HTTP request/response model, as long as the node where the transformation is executed knows the UE of the target terminal.

As we have said in the previous section, we want to use CC/PP for the description of the user environment. In order to incorporate this in the framework of figure 4, we will transform this description into an XSLT stylesheet that will take care of the actual content adaptation. Since these XSLT stylesheets are XML documents, they can be generated by another transformation, that might also be expressed as an XSLT stylesheet. Figure 5 shows the structure of such a content adaptation. Figure 6 uses

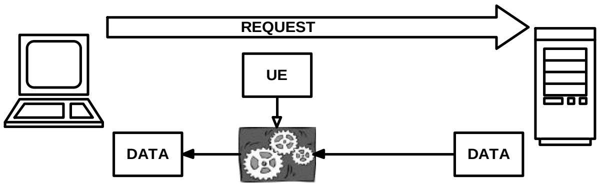


Figure 4. Content Adaptation for XML data

T-diagrams to show that the output of the first transformation is used as source for the second transformation.

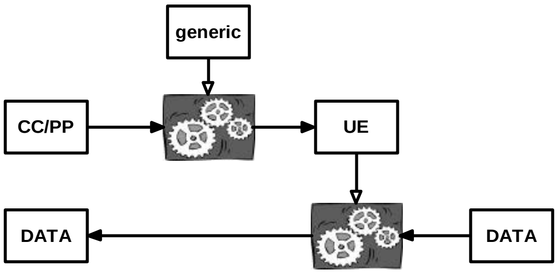


Figure 5. Using CC/PP with XSLT for the adaptation of XML content

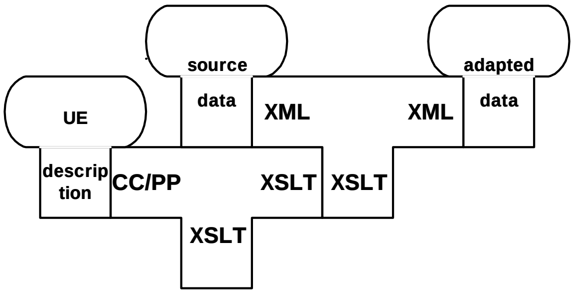


Figure 6. A T-diagram where the result of one Transformation is used as Source for another Transformation

The idea of using CC/PP for identifying a user environment and XSLT for the adaptation of multimedia content, has also been mentioned by the developers of the DELI software (Cowen 2002). DELI (Delivery Context Library), developed by HPLabs, is an open-source API to allow Java servlets to obtain a UE description from a HTTP request that may contain CC/PP data.

Experimental Setup

In our experiments, we try to incorporate existing technologies as much as possible. For the terminal simulation, we use X-Smiles, a Java-based XML browser capable of displaying documents written in various XML multimedia languages. X-Smiles was chosen for several reasons.

- It is possible to send CC/PP data within the HTTP request, as described in the W3C note “CC/PP exchange protocol based on HTTP Extension Framework” (Ohto and Hjelm 1999).
- Multiple XML multimedia formats are recognized (SMIL, SVG, XHTML, XSL-FO).
- It can simulate different types of terminals.
- The browser software is open source and written in Java.

At the server side, we chose Jigsaw, W3C’s experimental web server for similar reasons.

- Jigsaw has a basic implementation of CC/PP, based on the same note (Ohto and Hjelm 1999) X-Smiles uses.
- The server software is open source and written in Java.
- The code is very well structured and documented, making it easily extensible.

In our setup, the transformation is executed by the server, based on the UE information that is embedded in a HTTP request, and the metadata that is linked to the actual content. The format of such a request is shown in figure 7. The adaptation is executed by the server because we assume the server has more resources than the client. In addition, we want to minimize the amount of data that is sent over the network.

```
GET /menu.xml HTTP/1.1
Opt: "http://www.w3.org/1999/06/24-CCPPexchange" ; ns=19
19-Profile: "http://www.xsmiles.org/repo/desktop_en.rdf"
19-Profile-Diff-1: <?xml version="1.0"?><rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ccpp="http://www.w3.org/2000/07/04-ccpp#">
  <rdf:Description rdf:about="MyProfile"><ccpp:component>
  <rdf:Description rdf:about="TerminalBrowser"><rdf:type
  rdf:resource="BrowserUA"/><language>en</language>
  </rdf:Description></ccpp:component></rdf:Description>
  </rdf:RDF>
Accept: application/xml
```

Figure 7. A CC/PP Specification embedded in a HTTP Request

It is worth noting that the Jigsaw server can also be configured as a proxy server. As such, we can use our content adaptation software as a transcoding proxy. The advantage of such a solution is that our content adaptation framework remains transparent to existing software. The implementation of a transcoding proxy has been done before by others (Han and Bhagwat 1998), but always on one specific aspect of multimedia data (e.g., the adaptation of the resolution and color of images).

TOWARDS A NEGOTIATION PROTOCOL FOR CONTENT ADAPTATION

In the previous section, we described how we could perform an adaptation of multimedia content to a specific user environment. However, in this framework, there is no real negotiation. The involved parties do not negotiate the adaptation process. We feel that we can only talk of negotiation when the result is a true agreement between all participating parties.

In our opinion, the location of execution of the transformation should depend on a number of parameters. In the setup of the previous section, the transformation is executed in all cases by the server because we assume the server has more resources than the terminal, and network bandwidth is limited. However, in a real content negotiation framework, the location of the transformation should depend on different aspects of all nodes involved. In order to balance the usage of resources in a more rational way, we distinguish two phases in the transformation process.

1. Pruning of data in the original document that is not relevant for the target user environment.
2. Transcoding the remaining data to a format understood by the terminal.

The decision of where to perform each phase is based on the following metrics.

- The size of the original document.
- The estimated size of the pruned document (note that we want to know this before the actual pruning takes place).
- The bandwidth between successive nodes.
- Access to software that is capable of executing the required transformation.
- The availability of sufficient processing power.

Consider the scenario presented in figure 8. The terminal is a mobile phone accessing a web site. The device connects to the internet over the WAP protocol through a WAP gateway. Suppose the mobile phone is only capable of displaying WML pages (Wireless Markup Language, a HTML-like format for textual markup for wireless devices) and does not have sufficient resources to perform any transformation. At the web site, different versions of the content are embedded in the requested document: a high-quality video, a SMIL presentation with audio, video, text and still images, and an XHTML document that refers to some images. If other versions are needed, they will be generated on-the-fly, based on the version that is most suitable for transforming to the target data format.

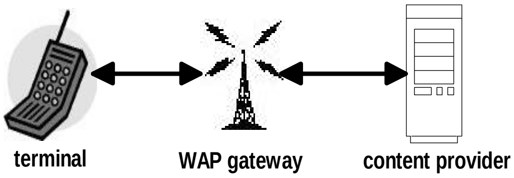


Figure 8. Content Negotiation in a Wireless Environment

Since the phone can only display text, the XHTML document will be selected to be transformed to WML. This selection is based on the metadata that is attached to each version of the content. The difference in file size between the original document and the data that will actually be used is significant, hence the pruning phase (extracting the XHTML document from the entire document) will take place at the server. When the processing time available at the server is insufficient, the transcoding phase (translating XHTML to WAP) could be done by the gateway, provided this gateway has the required knowledge and capabilities to do so.

CONCLUSIONS AND FUTURE WORK

In this paper, we showed how we can perform an adaptation of multimedia content to a user environment based on metadata. This approach is mainly based on XML: both the metadata and the description of the user environment are expressed by means of existing XML applications. Even the multimedia data can have an XML structure. We also described the environment we used to test this framework. What we proposed is however not yet a real content negotiation framework, because the different participating parties do not come to an agreement on how the adaptation should be executed.

Our future work will mainly focus on this content negotiation. Among other things, we will further elaborate the propositions of the previous section. This will consist of a further split up of phases of the transformation process (e.g., it might be appropriate to add a selection phase before the pruning phase), a clear definition of all relevant metrics for the negotiation process, and an algorithm that uses this information to resolve the negotiation question (i.e., the location of execution for each transformation phase).

In our opinion, it is not clear whether the HTTP protocol will be suitable for the negotiation framework we want to realize. HTTP is a stateless protocol that has no provisions to set up sessions. Therefore, we will consider the possibilities of other existing and emerging protocols, such as the XML Protocol. When we want to incorporate proxy servers in our framework in an efficient way, we will also have to consider caching issues.

A last point of interest is the relation between metadata and the description of a user environment. We will define a generic mapping mechanism between such a description of a user environment, expressed in a CC/PP vocabulary, and content description metadata, expressed in MPEG-7. This mechanism will be extensible, in order to allow other vocabularies to be incorporated in our content adaptation framework.

ACKNOWLEDGEMENTS

D. Van De Ville is a researcher funded by the Fund for Scientific Research-Flanders (FWO, Belgium). The research activities that were presented in this paper were funded by an OSTC-IUAP project, the FWO, and the Research Fund of the Ghent University (Belgium).

REFERENCES

Bormans, J. and Hill, K. 2001. MPEG-21 Overview. Technical Report ISO/IEC JTC1/SC29/WG11/N4511, MPEG.

Clark, J. 1999. XSL Transformations (XSLT) Version 1.0. Recommendation, W3C.

Cowen, A. 2002. Serving up device independence the extensible way. *M-Pulse, a cooltown magazine*. Available at http://www.cooltown.hp.com/mpulse/0202-developer.asp.

Devillers, S. 2001. XML and XSLT Modeling for Multimedia Bitstream Manipulation. In *Poster Proceedings of the Tenth International World Wide Web Conference*. IW3C2.

Early, J. and Sturgis, H. E. 1970. A formalism for translator interactions. *Communications of the ACM*, 13(10):607–617.

Han, R. and Bhagwat, P. 1998. Dynamic adaptation in an image transcoding proxy for mobile web browsing. *IEEE Personal Communications Magazine*, 5(6):8–17.

Holtman, K. and Mutz, A. H. 1998. Transparent Content Negotiation in HTTP. Technical Report RFC 2295, IETF.

Klyne, G., Reynolds, F., Woodrow, C., and Ohto, H. 2001. Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies. Working draft, W3C.

Martinez, J. M. 2001. Overview of the MPEG-7 Standard. Technical Report ISO/IEC JTC1/SC29/WG11/N4509, MPEG.

Ohto, H. and Hjelm, J. 1999. CC/PP exchange protocol based on HTTP Extension Framework. Note, W3C.

Smith, J. R. 2000. Interoperable Content-based Access of Multimedia in Digital Libraries. In *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries*. ERCIM.

Vianu, V. 2001. A Web Odyssey: From Codd to XML. In *Symposium on Principles of Database Systems*. ACM.

Wireless Application Protocol Forum 2001. User Agent Profile specification. Technical Report WAP-248-UAPROF-20011020-a, WAPForum.

ADAPTING MOBILE MULTIMEDIA APPLICATIONS TO CHANGING END-USER PREFERENCES

Robbie De Sutter, Boris Rogge, Dimitri Van De Ville, and Rik Van de Walle
Ghent University
Department of Electronics and Information Systems – Multimedia Lab
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
Tel: +32 9 264 89 08, Fax: +32 9 264 35 94
E-mail: {Robbie.DeSutter, Boris.Rogge, Dimitri.VanDeVille, Rik.VandeWalle}@rug.ac.be

KEYWORDS

Mobile Multimedia, Software Framework, Universal Multimedia Access, Metadata, MPEG.

ABSTRACT

At the moment there is an explosive expansion in the availability of mobile communication systems and mobile networks that support and enable the presentation of multimedia data. However, current design tools for mobile multimedia applications are unsatisfactory. A generic software framework is necessary to allow mobile multimedia system developers to create applications whereby the distinctive characteristics of mobile end-user terminals, the quality of the network, and the preference of the end-user can easily be incorporated. In this paper we present an application model whereby the content of a multimedia presentation is adapted to the preferences of the user and the possibilities of the end-user terminal amongst others.

INTRODUCTION

Mobile communication systems have recently known an explosive growth. This continuing expansion will incorporate new services in the near future, such as mobile multimedia applications. These applications are designed to deliver multimedia content – a combination of audio, still and/or moving images, speech and other kinds of data – to a mobile device. However, to ensure a breakthrough of such applications, there is still the need for further and extensive research and support to develop or to improve design tools. The huge variety of mobile devices (e.g., mobile phones, personal digital assistants, laptops …) results in a collection of very different device capabilities, device characteristics and bandwidth requirements.

Normally, a mobile device is – *at best* – optimized in handling one specific kind of multimedia data. For example, a mobile phone main purpose is to make telephone calls (audio data) and, to a lesser degree, send and receive text messages. Mobile phones are consequently optimized in handling audio and text data, but they are not the ideal devices for processing other multimedia data, such as video. Moreover a mobile phone is, like most other mobile devices, very limited in battery capacity, processing capabilities, display possibilities, and bandwidth. Other devices (e.g., laptops) have a better display or are optimized to process all kinds of multimedia data (e.g., tablet PC's).

In the remainder of this article we will first discuss the need for metadata in multimedia applications. Then we will show the differences between static metadata and dynamic changing metadata. The subsequent section explains the Universal Multimedia Access framework and discusses the possibilities of it as a negotiation framework between client and server. Finally, we will present an application model that makes it possible to adapt multimedia content to the changing parameters of a mobile device.

METADATA

Metadata is defined as *"data about the data"* (Ahmed et al. 2001). When applied to multimedia data, metadata can be seen as extra information associated with the multimedia data describing the content of the data. The MPEG-7 standard (Martinez 1999) (also known as *"the Multimedia Content Description Interface"*) is one of the latest and most extensive efforts to create a standard framework for using metadata in multimedia applications.

However the MPEG-7 metadata standard only describes *static* information linked to a multimedia presentation, such as the author, the date of creation, the encoding information and such. But some information is subject to change during execution of a multimedia application, e.g. the power supply of the device on which the application is running or the network load if the multimedia data is sent over the network. We call the metadata that is (more) likely to change, *dynamic metadata*.

This kind of information is very useful when a (mobile) multimedia application wants to guarantee continuously the most ideal representation of its multimedia data. For example, current internet streaming video applications could benefit by incorporating this information. If a user is watching a streaming video presentation on a laptop while the battery capacity is running low, the laptop usually reduces the clock frequency of the processor, resulting in a reduction of processing capabilities (AMD 2000). However, the video player does not react to this event and will continue to play the video as before. It would be better – in order to extend battery life and to cope with reduced processing capabilities – that the video player also reduces the playback quality of the video presentation. This can be done by playing the audio no longer in stereo, but in mono or reduce the color depth. Dynamic metadata can be useful to make this possible as we will demonstrate further in the application model.

UNIVERSAL MULTIMEDIA ACCESS

The application model will be compliant with the Universal Multimedia Access (UMA) framework (Christopoulos et al. 1999; Magalhães and Pereira 2001b). An application is compliant with the UMA framework if it has access to the complete description of every important component of a multimedia presentation. The application uses this information to deliver to the end-user terminal a perfectly suited presentation. The information needed to create such tailored presentation is:

- A description of the content of the presentation.
- A description of the end-user terminal.
- A description of the environment and circumstances in which the user and/or the terminal are residing.

The description of the content of a multimedia presentation is standardized in MPEG-7. The other two parts still need much more research and attention. At this time a (minimal) describing architecture that specifies the user environment consists of the following parts (Magalhães and Pereira 2001a):

- The network (e.g. bandwidth, bit-rate).
- User terminal (e.g. hardware, software, battery life).
- Personal preferences of the end-user (e.g. color depth, mono or stereo sound).
- Natural surroundings of the end-user (e.g. local time).

Note that this list isn't exhaustive. Future research must be conducted in order to have a complete describing architecture.
Each time a user, or a device on its behalf, requests a multimedia presentation through a UMA-compliant application, there has to be a mechanism that decides in what form the requested multimedia presentation must be sent. It is necessary that there is some kind of negotiation between the end-user terminal and the application on the server. There are three possible scenarios that result in a conclusion on how to send the presentation:

- Server decides: when the server receives a request of an end-user terminal for a multimedia presentation, the server gets the end-user terminal characteristics and capabilities. Then the server selects the most appropriate format which the end-user terminal can handle, and sends the presentation to the end-user terminal.

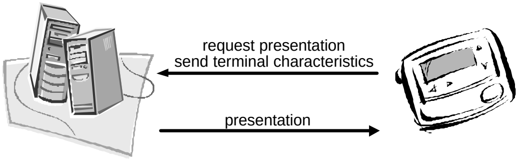


Figure 1: UMA: Server Decides

- End-user or end-user terminal decides: the server sends a list of all possible formats for the requested presentation to the end-user terminal. The end-user or the device selects one and receives the presentation in the selected format.

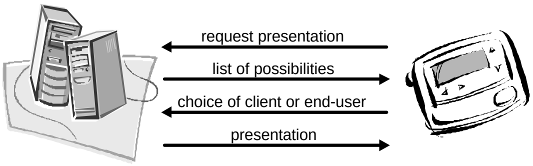


Figure 2: UMA: End-User Decides

- Application decides: The UMA-compliant application sends an application to the end-user terminal. This application will be executed local on the client and will take all decisions. If needed, the application will download additional libraries in order to process the multimedia presentation.

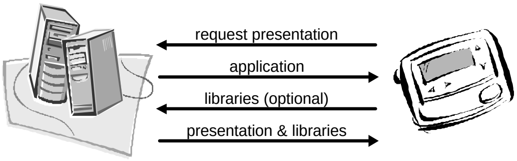


Figure 3: UMA: Application Decides

All three possibilities describe a negotiation process between a server (the multimedia application) and a client (the end-user terminal). The goal of this negotiation process is to ensure that the multimedia presentation send by the server can be handled, processed and presented by the client.

APPLICATION MODEL

The main goal of our model application is a proof-of-concept. We want to proof that it is possible to dynamically change the content of a multimedia presentation and this with the help of dynamic metadata.
The basic idea of the application is a slide show whereby the quality of the pictures sent to the client depends on end-user preferences. It is possible to extend this application so that it can send video or that the quality depends on other agents (such as battery life).
Our application consists of two distinct parts: the initial setup (just before the start of the presentation) and the actual playback (the slide show itself).

Initial Setup

The slide show consists of four distinct pictures. Every picture is saved several times, but in different quality, different size, different color depth, and different encoding formats. The client application can choose which kind of pictures it wants to receive.
During initial setup there will be a negotiation between the server (i.e. the slide show application) and the client (i.e. the end-user and the end-user terminal) in order to allow the end-user to select the quality. Because the goal was to create an application whereby the quality depends on the end-user preferences, the "*end-user decides*" UMA-scenario is the most appropriate.
The client initiates the application and receives the list of adaptable quality parameters from the server. The client also receives the allowed values of these parameters (e.g. the quality parameter "color preference" has the following allowed values: "black and white", "gray scale", "full color"). The end-user chooses the values of the quality parameters and can start the slide show.
If the user can also select the speed of the presentation (the tempo in which the images succeed each other), it is very important to remark that this parameter is not a quality parameter and will be processed separately. The quality of the content (the images) does not change if the user wants a new picture every three seconds instead of every five seconds. It is true that this parameter could have an influence on the bandwidth and the processing capabilities of the end-user device (for example if the pictures succeed each other too fast), but then the user should select lower quality parameter values. The speed parameter itself is not responsible for a

change in content; there is a supplementary step, namely the adjustment of the (real) quality parameters.

Presentation

During playback the multimedia application has to send a new picture to the end-user device after a certain amount of time. However, the end-user is allowed to change the initially chosen quality parameters. The application has to react by sending those pictures that meets the new parameters. This can be done without restarting the presentation.

There are several solutions for this problem, but we only mention a short list here:

- Every time the server needs to send a picture, it sends the picture in all possible quality formats. Then the client selects the picture that satisfies the quality parameters from the received batch. This is a very naive solution wasting a lot of bandwidth.
- The client requests a new picture and sends the quality parameters along, even if those parameters didn't change. Most slide-show applications currently available use this principle. This solution also wastes bandwidth, but not as much as the previous solution.
- The last solution notifies the server whenever the quality parameters changes. When a new picture must be sent, the server already knows the quality parameters. Concerning bandwidth, this is the most efficient of all options: no redundant data is sent over the network.

It is obvious that the last option is the preferred solution. While it is the best solution concerning bandwidth, it also breaks the usually tight coupling between the presentation and the parameters about the presentation. The end-user terminal sees the quality parameters simply as data. If a parameter alters, the terminal needs to transmit the changes to the server. A good implementation even makes it possible that the client isn't aware of the parameters and doesn't know about their existence. If we look at the first two solutions, it's natural that the client needs to know what the quality parameters are.

However this last solution needs a method to inform the server of changing parameters. The quality parameters are the metadata, i.e. data about the multimedia data such that it bears reference to additional information about the multimedia presentation. Furthermore this metadata isn't static, but is dynamic and time-dependent. In our application it describes end-user preferences, but it can easily describe end-user terminal characteristics (such as battery capacity). As such the quality parameters can be seen as dynamic metadata.

The problem of how to transmit the metadata to the server is solved by using XML. XML (eXtensible Markup Language) is the most important language in which metadata is written (Birbeck et al. 2001). Furthermore XML can easily be sent over the Internet thanks to SOAP (Simple Object Access Protocol). SOAP is described as *"a simple and lightweight mechanism for exchanging structured and typed information between peers in a decentralized, distributed environment using XML"* (Box et al. 2000).

Every adjustment of a quality parameter is sent to the server by using a SOAP message. The server accepts these SOAP messages and processes them. It checks which parameter has been changed and adjusts the quality parameters belonging to the client. When the server needs to send a subsequent picture to the client, it will lookup the quality parameters for that specific client and determines the quality of the next picture to be sent. Remark that the method to determine the next picture in the sequence of the slide show can't be deducted from the quality parameters. This is — analogously to the speed of the presentation — not a content related parameter.

The client will only receive pictures that meet the quality requirements selected by the end-user and this in real-time. The actions of the end-user have immediate impact on the presentation and this without any overhead regarding network access.

CONCLUSION

In this article an application model was presented. The main goal of this application is a proof-of-concept that it is possible to dynamically change a multimedia application if conditions of a mobile end-user terminal change. First, we have described the specific problems mobile devices have when used for multimedia applications. Next, we have shown the need for metadata and explained the difference in static and dynamic metadata. We explored the concept of a Universal Multimedia Access application as a framework that allows us to create a negotiation method between server and client. Finally we've presented an application model which proofs the usability of dynamical adapted multimedia presentations intended for mobile end-user terminals.

ACKNOWLEDGMENTS

D. Van De Ville is a researcher funded by the Fund for Scientific Research-Flanders (FWO, Belgium). The research activities that were presented in this paper were funded by an OSTC-IUAP project, the FWO, and the Research Fund of the Ghent University (Belgium).

REFERENCES

AMD. 2000. *AMD PowerNow! technology*. White Paper 24404A. AMD.

C. Christopoulos, T. Ebrahimi, V.V. Vinod, J.R. Smith, R. Mohan, and C. Lo. 1999. *MPEG-7 Application, Universal Multimedia Access Through Content Repurposing and Media Conversion*. Technical Report M4433. ISO/IEC JTC1/SC29/WG11.

D. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H.F. Nielsen, S. Thatte, and D. Winer. 2000. *SOAP v1.1 Specification*. Technical Report. W3C

J. M. Martinez. 1999. *Overview of the MPEG-7 Standard*. Technical Report N3158. ISO/IEC JTC1/SC29/WG11.

J. Magalhães and F. Pereira. 2001. *Describing User Environment for UMA applications*. Technical Report M7312. ISO/IEC JTC1/SC29/WG11.

J. Magalhães and F. Pereira. 2001. *User Environment Characterization in UMA applications*. Technical Report M6744. ISO/IEC JTC1/SC29/WG11.

K. Ahmed, D. Ayers, M. Birbeck, J. Cousins, and D. Dodds. 2001. *Professional XML Meta Data*. WROX Press.

M. Birbeck, J. Duckett, O.G. Gudmundsson, P. Kobak, and E. Lenz. 2001. *Professional XML, 2nd Edition*. WROX Press.

PALMTOP COMPUTERS FOR MANAGING INTERACTION WITH IMMERSIVE VIRTUAL HERITAGE

Luca Benini, Elisabetta Farella, Bruno Riccò
DEIS - Dept. of Electronics, Computer Science and Systems - University of Bologna.
Viale Risorgimento, 2 –40136 - Bologna - ITALY
E-mail: {lbenini|efarella|bricco}@deis.unibo.it

Maria Elena Bonfigli, Luigi Calori
°Vis.I.T. Lab - CINECA Supercomputing Centre
Via Magnanelli 6/3 – 40033 Casalecchio di Reno (BO) – ITALY
E-mail: {e.bonfigli|l.calori}@cineca.it

KEYWORDS

Realtime Systems, 3D Interfaces, Virtual Heritage, Palmtop Computer

ABSTRACT

The paper aims at exploring the potential of using portable devices in managing interaction in (semi)Immersive Virtual Reality (IVR) systems reconstructing cultural heritage objects and environments. The main advantage in this innovative mix of technologies is that portable devices are easily handled and enable a friendly interaction to the majority of users, with a reduction in learning time needed for accessing the contents. Moreover in a Virtual Heritage context, portable devices offer the opportunity to visualize, in a un-invasive way, sources – i.e. photos, drawings, plans, texts, etc.- necessary to validate virtual reconstructions and to explain to the user the real cultural value of the virtual world (s)he is visiting.

INTRODUCTION

Interaction is a fundamental part of any application designed for a Virtual Environment (VE). User interaction in IVR and Augmented Reality environments has traditionally been based on specialized interface devices (Brooks 1999) (gloves, 3-D joysticks, etc.). Even though these devices are specifically tailored to IVR interactions, they are subject to several limitations. First of all, most of them are connected to stationary appliances through wires (tethered), thereby limiting the freedom of movement in the VE. Moreover, these devices are designed and produced for a relatively small niche market, and therefore they cannot leverage the substantial industrial engineering effort, design optimisation, ergonomics tuning that are routinely performed for electronic appliances destined to the consumer market. As a result, Virtual Reality (VR) interfacing devices are either extremely expensive (as in the case of high-end military or industrial VEs), or they are plagued by substandard quality, performance, availability. Finally, the average user of computing and electronic equipment is not familiar with VR interfacing devices, because of their limited market penetration, at he/she might experience a steep learning curve when first entering a virtual reality environment.

Palmtop computers are rapidly becoming widespread: for instance, the recently released IPAQ PocketPC has reached the million-sold in less than a year, and an even deeper market penetration has been reached by lower-end PDAs, such as PALM products. Such a wide, and rapidly expanding, user basis implies two highly desirable consequences: first, a large number of people are familiarized with PDA interfaces; second the price of these devices is rapidly decreasing, while at the same time their quality, reliability and usability is improving. The most common PDAs are portable, lightweight and enhanced with fast processors, RAM and ROM capabilities, supporting complex tasks such as the visualization of colored graphical bi-dimensional and three-dimensional objects, multimedia audio and video applications, etc. Additionally, almost all palmtop computers of the last generation on the market today feature wireless networking capabilities: they are compatible with IEEE 802.11b wireless LAN cards, and more wireless interfaces (GPRS modems, Bluetooth cards) are becoming available.

Interaction through palmtop in immersive and semi-immersive VEs is not a new research direction. Fitzmaurice, Zhai and Chignell explored how virtual reality theories could be applied towards palmtop computers. Their project, "Chameleon", suggested that effective navigation and search could be supported by palmtop virtual reality, in rich and portable information spaces (Fitzmaurice et al. 1993). An interesting contribution was made by Watsen, Darken and Capps. They investigate the contention between 2D and 3D interaction metaphors and involved the use of a 3Com PalmPilot handheld computer as an interaction device to the VE, allowing the use of 2D widgets in a 3D context (Watsen et al. 1999). Another work by Wloka and Greenfield describes a user interface metaphor - the virtual tricoder - visually duplicating a 6-degrees of freedom input device in a VE (Wloka and Greenfield 1995). The JAIVE Project (Hill and Cruz-Neira 2000) extends previous effort integrating wireless networking, utilizing Java^{TM} and accommodating custom designed interfaces.

Also in Cultural Heritage field a lot of work has done to research innovative technologies for presenting cultural heritage sites with IVR and new interaction techniques using map based interaction metaphors. A research at the Fraunhofer Institute tried to apply some innovative metaphors in the immersive VR presentation of the cathedral of Siena (Behr et al. 2001).

The contribution of this work is the development of an hardware/software system for PDA-based interaction with IVEs realized in the field of Virtual Heritage (VH): Virtual Reality applied to Cultural Heritage. The software, that runs on PocketPC, is a Java based interface, which guarantees platform independence in appearance and operation, compatibility with other and new computing devices, to accommodate development and integration of new interaction types. In particular we introduce the use of maps for a 2D representation of the 3D world that help the user not to loose the perception of her/his position in the virtual world. Another contribution of our work is that PDA map-based interaction is performed through wireless LAN networking, exploiting the standard 802.11b, in existing IVEs realized by CINECA Vis.I.T lab with Performer/Vega in the field of Cultural Heritage within Nu.M.E., MU.VI and Pompeii "Casa del Centenario" projects (see http://www.cineca.it/visit/Researches/tecbec.html).

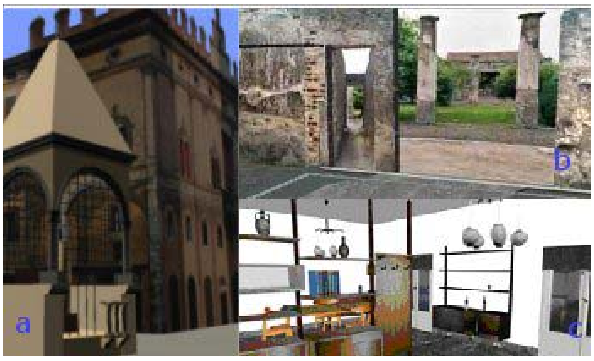


Figure 1: Pictures from Nu.M.E. (a), Pompeii Insula del Centenario (b), MU.VI. (c) VEs.

This paper is organized as follows. Section 2 presents interaction Techniques individuated as specific for Virtual Heritage. Section 3 shows the main features of the proposed system. Section 4 describes the architecture of the distributed system that manages the interaction between the portable terminals and VEs visualized in CINECA's Reality Center named *Virtual Theatre* (http://www.cineca.it/visit/virtualtheatre.html). Finally, section 5 concludes the paper and sketches the ongoing and future research directions.

INTERACTION TECHNIQUES FOR VIRTUAL HERITAGE

Virtual Heritage (Addison 2000) applications target a large and diverse user basis, ranging from scholars, which may gain better insight on the details and the overall structure of an archaeological site, to the general public of a virtual museum. For the vast majority of these users, we cannot expect any significant exposure to advanced VR environments and their specialized interfaces. The relatively long learning time of the "rules of interaction" may prevent active fruition by most users, and limit the VR experience to a passive virtual tour of an archaeological site, under the control of a single, well-trained operator. Even if rapid learning could be achieved (possibly thanks to effective real-life metaphors), the high cost and uncertain reliability of most VR interface devices may impose unacceptable operation cost for any mass-fruition facility, thereby strongly limiting the potential user basis of a virtual heritage environment. To address these issues, we propose the adoption of standard, consumer-market wearable interactive terminals, such as palmtop computers (Pocket PC and PDAs), as the preferred interfacing device with IVEs (Hill and Cruz-Neira 2000) in the field of Cultural Heritage. The use of portable terminals can in fact increase effectively the possibilities of interaction in IVEs, delegating the management of the principal interaction techniques to remote systems. This enables from a side to preserve the presence sense of the user (Hill and Cruz-Neira 2000), from the other to supply him with a tool for the visualization of cultural heritage data during the virtual experience.

Once defined the input devices - PocketPCs and PDAs - and the semi-immersive output device - Virtual Theater - the next step in designing the interface system is to identify which kind of interaction techniques are considered more useful in immersive virtual heritage applications.

In the field of 3D User Interfaces (3DUIs) the basic interaction tasks are: navigation, selection/manipulation and system controls (Kwon and Choy 2000). Navigation refers both to movement from place to place (*travel*) and decision making procedures to move in the VEs (*wayfinding*); Selection and Manipulation are respectively the specification of virtual objects and specification of virtual objects properties; finally System controls refers to the possibility to change the system state or the interaction mode. These basic interaction tasks can be combined to create more complex interaction tasks, some of which specific to an application domain. Among the most important Virtual Heritage (VH) specific interaction tasks: (i) spatial and time context defining, (ii) guided navigation, (iii) cultural information accessing. These tasks are outlined in the following subsections.

Spatial and Time Context Defining

To interpret Cultural Heritage data it is very important to understand their historical context both in terms of time and space. In particular in the field of Virtual Heritage the historical context can result determinant in underling originality and beautifulness of a building, a pictures, a sculptures, etc. and in recreating a connection with the real word that is always the main key for the communication process for virtual representations. Moreover, temporal context is important in order to understand how a artwork was modified over time or how it appeared different comparing it with similar pieces of the same time; while spatial context is important to understand how the real word around an artwork looks like, or, especially in the case of virtual reconstruction of buildings or of virtual archaeology, how different were the landscape and territory all around.

Interaction tools for helping users to understand the historical context can be the following:

- A “time bar” to select the historical period in which the users wish to take their virtual visit. Generally the user begins with the virtual reconstruction of the artwork as it is nowadays and travels backward in time (Bonfigli et al. 2000a).
- A “2D map” in which the users visualize their position and orientation in the virtual world (Bonfigli et al. 2000a), which regions can be reached by their current position (Behr et al. 2001), etc.

Both these tools are particularly suitable for the screen of a PDA.

Guided Navigation

Often virtual reconstruction of artworks have both the purpose to better understand their nature and to communicate a particular understanding that has already been reached. Moreover, due to the fact that there are physical limits to the scene complexity manageable by a VR system, there is no doubt that it is impossible to model all the 3D objects in a virtual scenario with the same precision or going into details in the same way.

Guided navigation enables the user interface designer to implicitly suggest fine views as well as details modeled better than others in order to give prominence to them or to underline a particular aspect of the virtual artwork.

Guided navigation of a VE can give users the possibility to select privileged point – e.g. on a 2D map (Behr et al. 2001) and to run automatic tours. This is a useful interaction task technique to provide the users with a quick overview of the virtual environment and of its objects, helping people to acquire spatial knowledge that is a fundamental requirement for obtaining wayfinding (Elvins et al. 1997).

Cultural Information Accessing

In order to avoid the risk of realizing misleading Virtual Heritage it is fundamental to connect VEs with all the background research material that lies under the virtual reconstruction activity. Thus, users will visualize models that are not only nice and realistic, but also historically accurate (Bonfigli and Guidazzoli 2000).

From the interaction point of view, when selecting a virtual object in the VE, the users should have the possibility of visualizing all the historical or archaeological sources (images, photos, written texts, etc.) that justify, validate and authenticate the reconstruction of that object in that particular way.

SYSTEM FEATURES

In CINECA’s Virtual Theatre, wearing stereoscopic glasses, a user can experience a “physical” immersion in VH environments: the area of Bologna City center from 13th century to nowadays (NuME VE), a typical middle-class italian house that changes during the 20th century (MUVI VE), an house in Pompeii before and after the eruption of the Vesuvio (“Insula del Centenario VE), etc. Moreover, according with the interaction techniques individuated in section 2, (s)he can interact with the User Interface (UI) of our system, visualized on the PDA. The UI features a two-dimensional map abstraction of the VR world, augmented with various navigation controls: eight arrows to more intuitively navigate in the environment, a time bar, and some buttons to enable an automatic tour and to change the VE draw status (see figure 2 below). Future works will include the capability to visualize in a browser multimedia data referring to the sources that justify the virtual reconstruction of single objects.

Interacting with the 2-D map reproducing the top view of the VE, the user is able to specify his/her movements and viewpoint changes. In particular the 2D map enables the user to visualize his/her position in the virtual world with a red rectangle and the direction of observation with a green rectangle (Bonfigli et al. 2000a) and includes the possibility to reach privileged viewpoint represented with colored dots. Moreover four arrows realizes the metaphor of moving forward, backward, left and right the avatar in the VE as if it is walking into it; and four arrows are used to turn around and to look up and down (Behr et al. 2001). This navigation mode is flexible towards curiosities and specific interests of the users.

The time bar (Bonfigli et al. 2000a) consists of a text field showing the year on display and a bar with a cursor showing the time line. Each time a year is entered in the text field or when the cursor is moved, the immersive VE is dynamically updated.

Finally, three buttons enable respectively the setting of a guided tour through the VE and the change of draw modes switching between textures and no textures and between solid and wire.

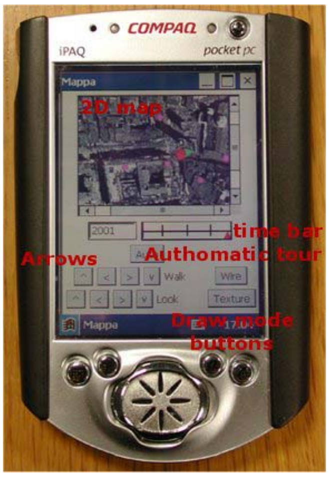


Figure 2: The User Interface Developed for Interacting via PDA with the Nu.M.E. VE

The choice of Java technology for the UI implementation guarantees cross platform design of the VR interaction application. There are several reasons for that. First, the palmtop computing market is rapidly evolving, and we need to guarantee rapid porting to new platforms as they become available on the market. Second, platform independence may enable end-users to download the application on their own PDAs as they enter the VR environment, thereby minimizing the learning curve and greatly enhancing the availability of the service.

SYSTEM ARCHITECTURE

The main purpose of this project is to deploy the hardware and to develop the software infrastructure needed to support PDA-based user interaction within a semi-IVR environment. The main components of the interactive system, shown in Figure 3 below, are: (i) the portable terminals, (ii) the wireless network, consisting of network interface cards (NICs) and base stations (that last one connected to the wired LAN of CINECA), (iii) the virtual theatre at CINECA and (iv) the client-server software that runs on the terminals (client component) and on the virtual reality engine (server component). That basic infrastructure implements the UI depicted in section 3.

The study has been performed using a PocketPC hardware platform, the Compaq iPAQ 3630, with a 320 x 240 resolution TFT color screen, 32Mb RAM and 16Mb ROM memory, running WindowsCE 3.0. The iPAQ was used with IEEE 802.11b High Rate standard wireless LAN interface working at 11 Mb/s. We tested our application with both wireless bridge for Ethernet networks and PC cards from Lucent and form Cisco.

The Wireless LAN card was setup using a given IP address but a test was made also using the Dynamic Host Configuration Protocol (DHCP), that means that the IP address is "leased" for predetermined periods of time.

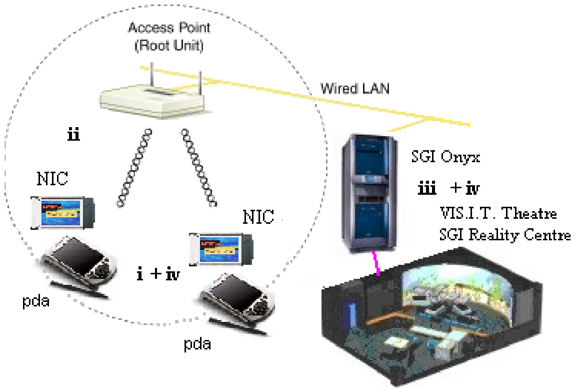


Figure 3: Components of the Interactive System: Portable Terminals(i), Wireless Network(ii), Virtual Theater (iii), Client-Server Software (iv).

Platform independence has been guaranteed by implementing the client-side software developing the code with Java^{TM} 2 Software Development Kit (SDK), v.1.3.1 and using Jeode *Runtime*, a Java virtual machine optimized for the PocketPC and WindowsCE platform. As of today all the Java Virtual Machine for WinCE and PocketPC support only AWT classes (not Swing classes for example), this was a limitation in building our GUI.

The server side of the application and the test of performances were realized in the VIS.I.T. Theatre, SGI Reality Center. The hardware platform consist of an Onyx2 with 3 graphic pipelines each with 64 MB texture memory and 8 processors MIPS R10000.

Client/Server Software Design

The current software implementation consists of a client component and a server component. Those two parts communicate through TCP/IP sockets. TCP guarantees reliable delivery of messages, while UDP makes no guarantees. TCP can be slower, and therefore less suitable for real-time communication than UDP. In the case of our application, however, the very simple kind of information transmitted, namely only numeric data representing coordinates, enables the choice of the more reliable, TCP protocol.

The Client-side block diagram is shown in figure 4 below. The higher level is a Java based GUI, realized with AWT classes, that runs on palmtop computer. This choice allows us to use familiar Windows, Icons, Menus and Pointer controls (WIMP). The UI on the client is designed in order to emphasize usability and short learning curve, especially for specifying movements on the third dimension on a two-dimensional representation of the virtual environment.

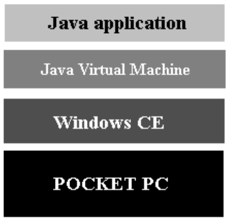


Figure 4: Client Block Diagramme

The reconstruction of the spatial context on the PDA is performed loading a simplified 2D map of the 3D reality and placing it in a canvas, able to capture events, in the higher part of the screen.
After receiving the request from the client to initialize the system, the server provides the client with correct data – i.e. image and virtual world coordinates - needed to set up the map itself. The resulting 2D map is different according to the VE projected in the Virtual Theater.

Interacting with the UI on the PDA, the user can create three kinds of change requests in the IVE projected in the Virtual Theater: (i) change system status (clicking the buttons named *wire* and *texture or* the button named *auto)*; (ii) change the time period (interacting with the time bar); (iii) change camera's position and orientation (interacting with the 2D map or with the arrows).
By touching the screen with the pen, a different event is generated and handled by a method of the Java application. Each method sends to the server, through a socket, a numeric identifier in order to specify the type of the request. It is worth noting that when requesting a change in time the

client has to append to the service identifier the information concerning the year, while requesting to change camera's position and orientation (s)he has to append to the service identifier a 6-dimensional vector that includes world coordinates *(x, y, z)* and the three angles: heading - the rotation about the Z axis - pitch - the rotation about the X axis - and roll - the rotation about the Y axis. In the case of interacting with the arrows this vector will be managed by the server in order to increment current camera's position and orientation, while, in the case of interacting with the 2D map, the vector includes directly new absolute camera's position and orientation.

The server component consists, in its highest level, of a C application in which the socket is opened and data received from a client are given as parameters to different specific Vega functions that cause the real-time rendering engine to provide the desired reaction to user request. The server-side C application that implements the socket connection and the real-time visualization of the IVE in the Virtual Theatre is built as a layer on top of Vega and OpenGL Performer: high-performance multi-process rendering graphic libraries that fully exploit the computing power of the graphic supercomputer Silicon Graphics Onyx2 to provide the necessary characteristics of realism and immersion. The system diagram (that include Irix Operative System and OpenGL graphic libraries) is shown in figure 5 below.

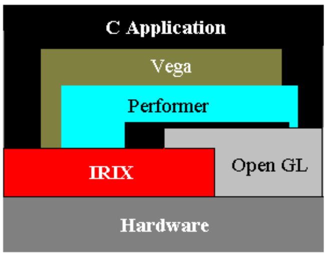


Figure 5: Server Block Diagramme

ONGOING AND FUTURE WORK

There are many directions for future research. Ongoing work focuses on adding more control movement capabilities on the client/server system to provide more freedom in navigating the IVE.

In future work we plan to expand features, adapting the application to multi-user presence. The navigation interface will be augmented with a set of controls that will allow the user to receive personalized information on various details of the scene. The availability of additional information (such as visualization of written commentary, pictures, audio captioning, etc.) will be made visible on the map by special icons. In this part of the project we will design several multi-user interaction modes.

AKNOWLEDGMENTS

Authors wish to thank Prof. Francesca Bocchi (Bologna University): scientific director of the NuME Project, Prof. Daniela Scagliarini Corlaita (Bologna University): scientific director of the Insula del Centenario Project, Prof. Donatella Vasetti, Dott. M.Chiara Liguori (Bologna University) scientific and technical directors of the MUVI project.

REFERENCES

Addison A., "Emerging trends in virtual heritage", *IEEE Multimedia*, April-June 2000.

Behr J., Fröhlich T., Knöpfle C., Kresse W., Lutz B., Reiners D., Schöffel F., "The Digital Cathedral of Siena - Innovative Concepts for Interactive and Immersive Presentation of Cultural Heritage Sites", ICHIM2001, Milan, Italy.

Benini L., Macii E., De Micheli G., "Designing low-power circuits: practical recipes," *IEEE Circuits and Systems Magazine* 1(1) , 2001.

Bonfigli M.E., Calori L. & Guidazzoli A., "Nu.M.E.: a WWW Virtual Historic Museum of the City of Bologna", Proceedings of SAC 2000 - ACM Symposium on Applied Computing (J. Carroll, E. Damiani, H.Haddad, D.Oppenheim Eds.) Villa Olmo, Como, Italy, March 2000, Vol. 2, pp. 956-961.

Bonfigli M.E., Guidazzoli A., "A WWW Virtual Museum for improving the knowledge of the history of a City", in *Virtual Reality in Archaeology*, (J.A. Barcelo, M.Forte, D.H. Sanders Eds.), ArcheoPress, May 2000.

Bonfigli M.E., Calori L ., Guidazzoli A., Mauri M.A., Melotti M.; "Tailored virtual tours in Cultural Heritage worlds", ACM SIGGRAPH2000, New Orleans, July 2000.

Brooks F., "What's real about virtual reality?", *IEEE Computer Graphics and Applications*, November-December 1999

Elvins T., Nadeau D., Kirsh D., "Wordlets – 3D thumbnails for Wayfinding in Virtual Environments", Proceedings of UIST'97, 1997, pp.21-30.

Fitzmaurice, G.W., Zhai, S., and Chignell, M. H., "Virtual Reality for Palmtop Computers," ACM Transactions on Information Systems, Vol. 11, No. 3, July 1993, pp 197-218, 1993.

Hill L., Cruz-Neira C., "Palmtop interaction methods for immersive projection technology systems," Fourth International Immersive Projection Technology Workshop (IPT 2000), 2000.

Watsen, K., Darken, D. P., Capps, W.V., "A Handheld Computer as an Interaction Device to a Virtual Environment," Third International Immersive Projection Technology Workshop (IPT 1999), Stuttgart, Germany, 1999.

Wloka, M. M., Greenfield, E., "The Virtual Tricorder: A Uniform Interface to Virtual Reality", UIST'95 Proceedings, 1995.

Kwon T., Choy Y., "A new navigation method in 3D VE (2D Map-based navigation)," International Conference on Virtual Systems and Multimedia, 2000.

APTEC

E-LEARNING AND ITS APPLICATIONS

Aspects of E-Learning in Europe

Katharina von Knop[1] and Harald Spiegl[2]

Summary

e-learning -- a word that was coined as a fashion term -- reflects the desire to introduce "new media" into the area of classic education. This development has only just begun.

Less than 15 years ago the development of home computers set off a revolution which gave rise to a new generation of young people. They are eager to learn, and their interest focuses on increasingly powerful technology. The elder generations had to accept the changes that occurred due to the global development towards a "media society". Meanwhile the internet, paving the way to global communication, has reached even the remote parts of the world.

And yet the teaching professionals are reluctant to adapt their methodics and to develop new educational concepts. The students' expectations are changing rapidly: most students would prefer to access study materials and supplements comfortably from their homes, using their computers and the internet.

Although the idea of purely "virtual" studies without any direct personal contact to professors and fellow students has often been discussed, it will certainly not be realized in the near future. Among other problems many legal aspects have to be resolved, for instance how to prove authorship of assignments and written exams.

Currently available software for the implementation of "virtual universities" is far from being standardized. Everyone may create and offer this software; recognized certifications do not exist. Students cannot make use of material offered by other universities because the curricula are not sufficiently coordinated, and also because achievements are not universally acknowledged

The software market is dominated by US-based companies. They tend to see their software as some sort of merchandise which is sold through marketing efforts, not primarily through its quality. Education in Europe, in contrast, is still largely influenced by the classic ideals; sales figures are not regarded as a measure of quality.

e-learning software requires powerful computer hardware. Thanks to the overproportional growth of computer performance in the recent past, suitable hardware is now available at reasonable prices. High cost for internet access from home and particularly for mobile connectivity, however, are another obstacle for the success of e-learning.

1. E-Learning in the society

Nowadays you cannot imagine universities, schools, authorities, companies or even private households without Internet connectivity. Many million people in almost any state on this planet are connected through the Internet.
While in the beginning the Internet was introduced by the army in order to ensure communication in the case of a nuclear war, today it serves mostly peaceful purposes, particularly the education and training of ordinary people.

The modern term of "e-Learning" combines the ideas of "further training in the spare time" and the "continuous presence of knowledge".

It comprises not only the substitution of conventional lectures by a choice of educational software but also the transfer of knowledge through "New Media". Many companies use this to their economic advantage and find cost-effective ways to educate their employees without sending them to expensive seminars. Libraries can offer their valuable resources to the public without the risk of losing their materials; schools and universities can make their services available

[1] Leopold-Franzens University, Innsbruck, Austria
[2] Heinrich-Heine-University, Computing Center, Duesseldorf, Germany

"around the clock", and they can pool their resources and found "virtual universities".

2. How does e-learning affect traditional learning?

As the process of learning through and about the new media requires a reorientation in the society, e-Learning breaks with the tradition of learning in classrooms and to meeting the teacher in person. The student will gain a new independance and organize his or her time for flexibly.

On the other hand e-Learning will probably never be able to replace conventional lessons completely. The personal contact with the teacher is of considerable emotional importance for many students. Oral and written exams will always require personal presence of students at certain times.

For the rather shy or introvert student e-Learning can be a way to take a more active part in the learning process due to the anonymity and the physical distance which the Internet provides, and personal appearance or ethnic characteristics become irrelevant.

E-Learning can be seen as a means of support for classical learning, albeit not as its replacement.

Today there is more demand for a broad personal knowledge than ever before. "Knowledge shows" dominate the television programmes. Recent surveys reveal that many spectators try to beat the candidates in finding the answers, often using the Internet in their quest for a solution. Knowledge databases (like www.wissen.de in Germany) report higher numbers of visitors than ever.

This trend is supported by continuously decreasing prices for Internet access and the availability of affordable flatrates for home users.

During the past years computing power has grown overproportionally and has become more affordable than ever, even if the prices have certainly not reached their minimum yet.

This country is still in the process of building up an all-embracing Internet infrastructure as we know it from the USA. New technologies will make the network ever so faster and safer.

E-Learning offers the individual chance to get prepared for the global markets of the future.

3. How can we market e-Learning?

When companies talk about "e-Learning" today, they use a fashionable term for one sole objective: Increased sales.

It is obvious the commercial enterprises do not develop software just for fun. Many well-known programs were and still are developed at universities, often as part of a thesis for a diploma or a doctor's degree. During the development process the authors publish intermediate results, communicate with fellow researchers over the Internet, offer their software for free and make it popular while it matures.

When the author leaves the university, there are only two possibilities: Either the software "freezes" or "dies", or it has shown to be marketable as a commercial product. In the latter case a software company may become interested and eventually buy the program. Further development is then funded by program sales, and often the users will then have to pay a considerable price.

Usually such software is only available as a licence product while all rights remain with the author. The more licences a customer buys, the cheaper he will get them; often the most reasonable way to use such a program is the purchase of a campus licence which grants the right to using the software to all members of a university. This looks like a good deal but is of no advantage if the number of users is small. The software company will proudly advertize the university as a reference customer -- but deceptively. Whoever is looking for "real" reference installations will only be able to sift the chaff from the wheat if he can find out about the numbers of real, i.e., not just licenced users.

Nevertheless it makes sense -- if the decision was made for commercial software -- not to rely on in-house development but to use the available off-the-shelf software.

The better a software is in terms of usability and marketability, the better are the chances that in the (not so rare) case of the original supplier going out of business another company will buy the product and continue its development.

In this country a university will only use legally acquired software. No staff member will dare to use pirated copies and thus run the risk of a lawsuit. The potential savings would not make up for the additional cost and the troubles which illegal software could cause.

Students may feel more inclined to use unauthorized copies of commercial software, often because of its -- from their point of view -- tremendous licence fees. Particularly for the leading software products, however, universities could reach agreements with the software companies which allow affordable access for students.

4. Which software is used (and is useful)

Nowadays there are quite a number of commercial software products which cover the field of e-Learning.

Specially designed packages for e-Learning are offered, for example, by the US-based companies WebCT (www.webct.com) and blaxxun (www.blaxxun.com). WebCT's products are, according to the company's sales figures, used at over 2,900 universities world-wide (most of them in the US, but many also in Switzerland).

Apart from these there are a number of so-called content management systems (for example by InterRed, www.interred.de) which are also suitable to present knowledge. Content management systems are a kind of database-oriented application software; they can be used to build well-structured and easily navigable presentations of large data sets but often lack a pedagogic background.

And then there is an immensurable supply of learning software, sold on CDs and designed to run statically on personal computer and which communicates in a way over the Internet, either downloading data from the manufacturer's server or offering communication with other users by the means of "chat rooms".

5. Mental differences in the learning environments

If we compare the teaching environments of the "classical" German universities with those typical for the US, we will find the lectures commonly held in Germany -- with only few exceptions -- to be "bone-dry".

The German student sits on his or her wooden flap seat in the same manner as generations did before him.

And as far as we could tell, the situation among our European neighbours is more or less the same.

American professors, however, take different approaches. In order to keep the students motivated, the fun factor is highly appreciated -- it is not uncommon that a professor in shorts and Hawaii shirt gives an open-air lecture or -- totally unbelievable for people from central Europe -- makes fun of himself.

From the American point of view, knowledge is a product that is to be sold, no matter how.

The mental environment is again, however, totally different if we look at the situation in China. It seems as if the fact that software cannot be "touched" physically implies for most of the users in China that it does not have to be paid for. The only thing you can touch when you deal with software is the CD -- the material has to be paid for, naturally, but the software that is "burnt" on the CD will not be licensed. The authorities seem to tolerate this mentality; the production of pirated copies of commercial software flourishes. Without proceeds from sales, Chinese software companies could not exist if labour were not so cheap; scientists and programmers working at minimum wages develop softwares that can be marketed in Europe and the US where people pay for their software licences.

6. What is the situation as of today?

6.1 General aspects

E-Learning experts demand a higher degree of interaction throughout the learning process; current projects are too static.

Lifelong learning should be and must be possible by means of mobile devices in the future. PDAs, cellphones, laptop computers etc. are already widely available and render the user independent of his office or his living-room.

Compared with other countries (particularly the US), Internet access for mobile devices like cellphones or PDAs is far too expensive.

If you are waiting in the departure lounge of an American airport you can watch other travellers with laptop computers who are online over their cellphone. Mobile connectivity at moderate flat-rate prices would make the market and the number of users explode.

While Virtual Universities in the US charge significant prices for their courses (approx. $100,000 for a full course of studies), and customers are prepared to pay these, this kind of market will only slowly gain a foothold here in Germany where the availability of learning materials at no cost is legally guaranteed.

Every student here is (still) entitled to a free course of studies; the majority of students could not afford to attend one of the "private universities" unless new models of financing studies through loans, scholarships etc. are developed.

This situation determines planning and calculation. Many professors ask themselves why they should take the trouble and prepare an online course instead of doing it the traditional way, and flatly reject any modernization of their teaching.

The age of the professors certainly plays an important role; many of them are beyond an age where they could easily adopt to new paradigms like e-Learning.

Furthermore a German particularity has to be taken into account which is a bureaucratic obstacle in the way of creative advances in the area of e-Learning. Most German scientists are employed by the universities as "public servants". The original idea of a public servant being that of a person who serves the state in a life-long position, not accepting money from anyone else, this scheme fits badly into the picture of today's dynamic scientists who work at the boundary between industry and scientific research. It probably is counter-productive for the deployment of new concepts in our universities.

6.2 Companies

Staff training courses cost working time and thus money.

Some small or medium-sized enterprises cannot afford to send their employees to seminars or other advanced training.

E-Learning offers them an alternative way to staff training.

This, however, opens up the question whether a staff member is able to accept a training offer. It does not make sense to pass him a CD and wish him good success; it cannot be assumed that he owns all the technical equipment necessary to run the learning software.

It is in the companies' own interest to improve their competitiveness through well-trained and motivated staff.

Furthermore it is a responsibility of the government to support the purchase and operation of e-Learning equipment; it should be possible, for instance, to get a tax deduction for a PC and other electronic devices that are necessary to take part in e-Learning.

Some large enterprises already make intensive use of e-Learning for staff training. Usually they install content management systems which offer, besides the presentation of knowledge, discussion groups, feedback functions and more features so that learning is not restricted to a one-way process.

6.3 Libraries

For a long time libraries have collected digital data about the books they own, be it cataloguing data or a digital transcription of the content.

When these data were made available over the network, the term "virtual library" was coined.

The users of the libraries' media can satisfy their thirst for knowledge in a much shorter time than before as they can browse the information more efficiently.

Comfortable search engines save enormous amounts of time.

In the not too distant future the "customers" of a library will be able to not only take the books home in printed form, but also, alternatively, in electronic form.

Although some users will insist on using "real" books as long as they cannot read their electronic books in bed, a virtual library is nevertheless well suited for a purposeful, supportive knowledge search.

6.4 Schools

The classical school with teachers and pupils will be supplemented by e-Learning, but never replaced.

Besides doing knowledge transfer, schools are expected to educate the young persons. This cannot be done via e-Learning.

Of course today's schools offer computer courses and are connected to the Internet, often thanks to the activities of volunteers.

Although it would be possible for a pupil to send his or her homework to the teacher via the Internet, this would open up to many opportunities for manipulation and therefore is not allowed.

In this country there are no uniform standards concerning the availability of electronic equipment (desktop and laptop computers, beamers) in schools.

Teachers have to buy their own hardware and software, which they need to prepare their lessons, and they have to arrange for their own computer training. Actually many pupils acquire a more advanced knowledge about computers than their teachers have.

6.5 Universities

Today it is fashionable to establish many "virtual universities" quickly. Concepts and regulations exist, and in Bavaria the first students of the"virtual university Bayern" are already registered.

It is a major goal to make "virtual teaching" more dynamic.

Some projects like VIROR (Virtual University "Upper Rhine" - www.viror.de) in the state of Baden-Wuerttemberg (in Germany) work only unidirectionally. The lectures are only transmitted into different local lecture-rooms. Thus there is no genuine interactivity between teachers and students. Students, however, should be have the opportunity to participate actively during the lessons. They should give talks or ask questions directly of their teachers.

Many students today possess mobile devices like laptop computers, PDAs or cellphones. Teamwork is far more important than to simply pursue a lecture, sitting in front of a computer screen or a projection canvas.

The recording of lectures for tele-presentations is quite simple. It only becomes technically difficult if the lecturer presents foils and writes comments on them. Presently there are no standardized procedures and programs for this functionality.

7. Special e-Learning-activities

7.1 Virtual University Bavaria (VHB)

The VHB is currently open only to regular students of one of the supporting universities. These are the Bavarian state universities and professional schools (but not the academies of art), the academy for television and film in Munich, the Bavarian school for public services, the catholic endowed professional school in Munich, the protestant academy in Nuremberg, the catholic university of Eichstaett and the university of the Federal Armed Forces in Munich. Other universities may join the VHB as supporters on request.

Before joining the VHB, a student has to register with one of these Bavarian universities as a regular or a guest student for the subjects he or she wants to study.

The "home" university, the classical university where a VHB student is enrolled, decides whether it credits the courses taken at the VHB. It is the student's responsibility to check the creditability of a course beforehand. The VHB collects each student's credits and visualizes them on a web page; the final decision on the validity of the credits, however, remains with the university's examination office.

The exams usually take place at one of the supporting universities where students are requested to appear in person.

7.2 E-Learning at the University of Heidelberg -- learning in net-based groups

For some years (since approx. 1996/7) the University of Heidelberg in cooperation with the neighbouring University of Mannheim has offered network-based lectures. Both universities are connected over an ATM network. A lecture can be held in one of the universities and students in Heidelberg as well as in Mannheim can attend the lecture at the same time. Video cameras and beamers are available in the lecture-halls for this purpose.

Furthermore some seminars are offered as a co-operation of the universities of Heidelberg, Freiburg and Karlsruhe as a video conference (using a shared whiteboard).

"Home Learning" (life participation with interaction) over ISDN or, preferably, DSL technology is still under construction.

The project of a "Virtual University Upper Rhine" (VIROR) operates as a regional activity. It has been supported by the local government for approx. 3 years. It aims at building a multimedia didactic program for different subjects, giving students the opportunity to study at four universities (of Heidelberg, Mannheim, Karlsruhe, and Freiburg) at the same time (www.viror.de). The data transfer is made over an MBONE network.

This project employs a media server based on WebCT software.

7.3 Virtual Campus Switzerland (VCS)

- Goals

The "Virtual Campus Switzerland (VCS) is a federal project founded by the Swiss University Conference (SUK). Funds are being granted to the institutions participating in this project for a duration of 3 years and usually amount to 1,000,000 SFR, of which 500,000 SFR are paid by the VCS and 500,000 SFR by the participant.

The VCS shall develop learning units for base and advanced learning programs that will be made available to students via the Internet. Particularly attractive courses will be preferably supported; low-level, repetition or preparatory courses will not be supported unless they complement another course for which support has already been granted.

Participation in this project is possible on request when the following preconditions are met:

There must be a partnership with at least two other partners at different universities.

A project sketch must be submitted, later followed by a detailed description of the project.

A network connection between the partners must exist.

The clarity of the educational concept must be recognizable from the project documentation.

The project must develop multimedia-assisted courses as a replacement for conventional courses.

A selection of suitable hard- and software must be specified, where it is regarded important not to promote unnecessary developments but to make use of existing commercial software.

The proposed courses must integrate seamlessly into existing curricula.

Multilinguality issues are important for Switzerland and therefore must be considered.

Cooperation with the industry is expected.

On the other hand the research and further training are not promoted.

Research and further education, however, are not supported.

As of October 2001, 50 projects were granted, 28 in a first and 22 in a second stage. The latter ones will start during the course of the year 2001.

- Mandates for the implementation of the project

1. Technical support, co-ordination and webserver

Following a report of the working group "remote studies at university level" of the HPK (university planning commission), the Swiss university conference appointed a team of experts: they will examine measures to promote the application of new information and communication technologies in the area of university education.

One of the measures suggested for the period of 1996-1999 was the installation of a national webserver, which covers the activities of Swiss universities in the field of remote studies. This is an essential prerequisite for any other action because researchers, teachers, students, developers, and decision-makers need access to the latest information about current activities.

In addition to this, discussions and evaluations are needed to extend the exchange of information in a systematical way.

Moreover this renovation of university pedagogics must be performed in close connection with the private training sector and based on current experiences in Switzerland and abroad.

The federal government hopes that this webserver becomes an efficient tool for information retrieval and exchange, enabling the universities to conduct cooperative pilot projects, which are of general usefulness.

The Federal Office for Education and Science and the Swiss University Conference assigned the implementation of this webserver to the university of Freiburg. Their NTE centre performs the practical realization of this project under the name "Edutech".

2. Educational/didactical support

At the beginning of the sponsoring programme this aspect did not receive enough attention. Only in the course of the time this support proved to be important. Although the specified preconditions for projects were respected, some courses found no acceptance with the students because of a lack of quality. These courses had innovative content but a dissatisfying appearance.

Hardware:

The used hardware is perfectly heterogeneous, although in Switzerland the share of Apple computers outweighs other manufacturers by far.

Software:

Standard software is being used only sporadically and not in uniform ways. The project demands the use of existing commercial tools instead of individually developed software. About 50% of the servers use WebCT software.

7.4 E-Learning in Austria

Approx. 33% of the households in Austria possess a PC and approx. a 30% of the households use the Internet.

Of a total of approx. 6,300 schools in Austria presently approx. 4,700 are reachable by email. For secondary schools this proportion is close to 90%. About 1400 schools have their own homepages.

In a first step the Austrian Federal Government has made available an amount of 80 million EUR for equipment and network connections for schools and universities. In addition to that a project called "new media for teaching" was initiated by the Federal Ministry and funded with 7 million EUR.

The pilot project started in 1996. There were no activities to be observed in 1997 as the project gathered only little interest among professors. By the end 2000 about 25-30 courses were offered, 20 of these as lectures. Lecturers criticized the time and cost necessary to install such courses.

In contrast to this approx. 80% of the working population regard such courses at least as useful, but only approx. 50% of the students share this opinion.

The year 2001 should introduce an uprise into the use of new media. WBT (Web Based Training) shall be promoted. An initiative by the FH Johanneum (Graz) presently supports 22 courses (with 14 instructors) and is under construction. Future goals are the development of the scenarios, the development of the use of different learning platforms and the strengthening of the community.

In Austria (much like Germany and probably Switzerland) students still prefer, according to recent opinion polls, to go into conventional lecture-halls rather than to sit at home in front of their PC and feel lonesome. Personal contact with fellow students still outweighs the advantages of virtual learning.

7.5 Other activities

If one searches the Internet for the terms "Virtual University" or "Virtual Campus", one finds lots of references. Certain activites of individual professors are announced by their universities as being part of a "Virtual University". These are usually limited, however, to the presentation of content and possibly a few chat rooms. The simple recording of lectures on video tape is sometimes regarded as a first step towards a virtual university. It is idle to speculate on the quality of these offers.

E-learning : New advanced learning models and the limits of our e-learning systems.

Joanna Schreurs, Rachel Moreau, Ivan Picart
Limburgs Universitair Centrum
Universitaire Campus
B-3590 Diepenbeek, Belgium
E-mail: jeanne.schreurs@luc.ac.be
rachel.moreau@luc.ac.be
ivan.picart@luc.ac.be

KEYWORDS

e-learning; learning portal; e-learning platform; webcourse

ABSTRACT

Students can have access to the electronic learning system of our university, from within electronic courses or web courses and other learning materials are available. By this way the students can learn on an anytime, anywhere base. Web courses are often used to deliver the basic learning content of a course. The learning process itself consists of several kinds of learning activities. All of them can be supported by e-learning, implemented as a sample of functions of the e-learning system. The definition of e-learning has been changed and will change as a consequence of the evolution in our learning models. The e-learning system has to follow this evolution too!

INTRODUCTION

Several years ago, we tried to automate the learning activity. In the late eighties and early nineties CBT became very popular. Several commercial authoring tools were soon available. Universities and other training institutes invested in those tools, in special hardware tools and in the development of computer based training modules. They believed in this self-paced learning model.
But the users, the students, didn't like reading the text on the computer screen.
As a consequence the return on investment was low and the enthusiasm disappeared.

Thanks to cheaper hardware, user-friendly software and especially thanks to the spread of the Internet , we see a new wave in online learning. A mass of web courses are published on the net. Many of them are free.
But research done on point of effectively of this way of online learning shows a bad result. Self paced learning doesn’t seem to be an effective way of learning. A student needs to learn in a learning environment in co-operation with a teacher and other students. The learning proces can follow several learning models. All of them are organized as a combination of classroom and online activities. The e-learning system, delivering the course content can also support all learning activities. The role of the teacher has been changed to a more supporting role.

The implementation of e-learning may not be based on technological opportunities, but must start from a new educational model

DEFINITION OF E-LEARNING

E-learning is the organisation of the learning process using ICT/Internet. The delivery of learning content (web courses and web documents) is one of the main goals. Even more important is the support of the other learning activities organized as classroom or as individual activities. The teacher is responsible to manage, to instruct, to control the activities and to evaluate the results

A HIGH QUALITY PEDAGOGICAL APPROACH IMPLEMENTED IN AN ADVANCED E-LEARNING SYSTEM : STATE OF THE ART.

A high quality pedagogical approach is ensured by the adaptation of the virtual class model combining *theory, examples and exercises* modules for training.

The way we are teaching the theory has been changed. First the classroom teaching duration time has been reduced. Only introduction in the topics and the relations between them are still presented on this way as a classroom activity. Our learning model has evolved to a mixed model of classroom teaching, self paced learning and other classroom or on distance learning activities. The teachers are supporting the learning process. Several kinds of classroom and on distance learning activities are organized.
In a classroom discussion session the team of students are interchanging their knowledge on point of the understanding of the theoretical content and of the applicability of it in real practice. The students are responsible to extend the delivered course content by searching for relevant information and to share it with the other students.
Examples of applications of the theory are partly included in the basic learning content. Students have to contact business enterprises to find a number of applications by themselves and to share them with other students.
The students have to make exercises on an individual or in a group wise way, as a classroom or as an on-distance activity. Papers on the topics have to be prepared and will be presented in the classroom.

On a more advanced level, a student team can be made responsible to solve a problem from business practice.

Delivering training based on the e-learning principle will result in an offering of a much higher flexibility than is currently possible through a typical classroom environment. Users will be able to follow advanced concepts using a simple learning scheme based on those three fundamental steps of a class: theory, examples and exercises. The theory mode provides access to image, text, audio, video, 3D animation and Internet supported explanations of the fundamental technical concepts in e-work practice. The example mode allows the user to follow the descriptive solutions to a number of pre-defined e-work problems incorporating computer simulation and multimedia graphical tools. Finally in the exercise mode students are asked to solve by themselves (with guidance from the tutor) various pre-defined cases of every day's information management practice.

A number of facilities and tools are available in an advanced e-learning system:

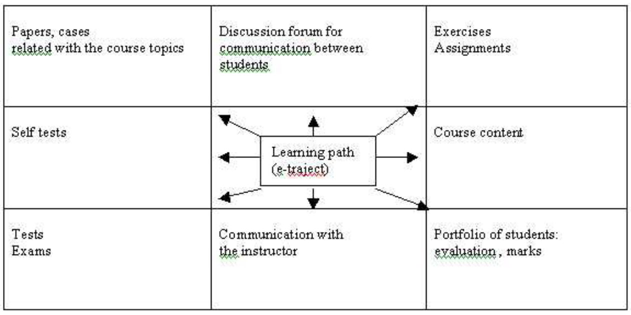
- Administration facilities for registration of students, general description of courses, follow-up of student curricula, input of new lectures and new courses, course management tools, ...
- Teaching facilities, including tools for creating and storing electronic documents, follow-up of student performance, solution of student questions, interactive communication with students.
- Library facilities, including storing, visualizing, publishing and downloading electronic documents on and from the Internet.
- General course information facilities, including possibility of storing relevant information related to the course content, relevant dates, registration details, answers sections, ...

An advanced e-learning system includes and is based on the following basic tools:

- a browser based interface for the teacher and for the student which allows for efficient document/info transfer .
- a powerful search facility
- a learning documents database
- flexible document/database connection allowing for easy adaptation to different learning applications.
- flexible database connection allowing linkages with the student and teacher data in the university central database
- (XML) authoring tools for preparing e-learning materials

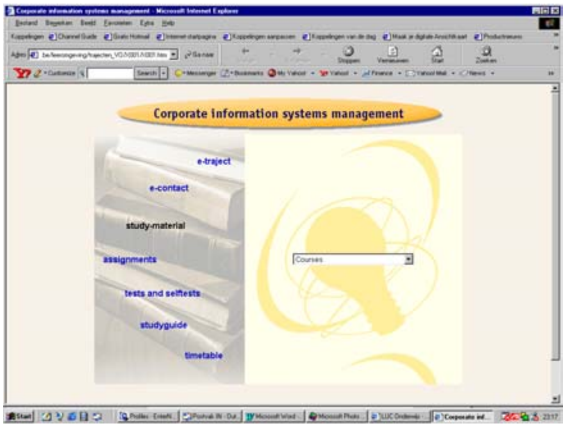
AN EXAMPLE OF A WEB-LEARNING PORTAL AND BUILT-IN E-LEARNING FUNCTIONS.

In the e-learning portal one can find different entries to different applications.
In the "e-traject" one can find the general information of the course. Information on point of content, learning model, planning and evaluation are delivered. All activities and the linked content documents are planned in a modular way.
The discussions will take place in the "e-contact". The student can find here the question/answer facility.
The "assignment" entry is the virtual classroom.
The entry "selftest" delivers the self testing facility.

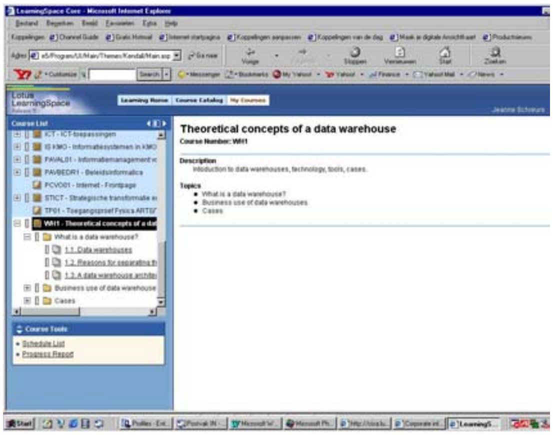


Papers, cases related with the course topics	Discussion forum for communication between students	Exercises Assignments
Self tests	Learning path (e-traject)	Course content
Tests Exams	Communication with the instructor	Portfolio of students: evaluation, marks

example of a learning portal



example of an "e-traject"


LearningSpace Core - Microsoft Internet Explorer
Lotus LearningSpace
Learning Home
Course Catalog
My Courses
Course List
ICT - ICT-toepassingen
IS KMO - Informatiesystemen in KMO
PAVAL01 - Informatiemanagement
PAVBEDR1 - Beleidsinformatica
PCVO01 - Internet - Frontpage
STICT - Strategische transformatie
TP01 - Toegangsproef Fysica
WH1 - Theoretical concepts of a dat
What is a data warehouse?
1.1. Data warehouses
1.2. Reasons for separating th
1.3. A data warehouse archite
Business use of data warehouse
Cases
Course Tools
Schedule List
Progress Report

Theoretical concepts of a data warehouse
Course Number: WH1

Description
Introduction to data warehouses, technology, tools, cases.

Topics
- What is a data warehouse?
- Business use of data warehouses
- Cases

THE EVOLUTION OF THE EDUCATIONAL MODEL AND THE REQUIRED EXTENSION OF THE E-LEARNING SOLUTION

.

A high degree of practical relevance in the learning process and the provision of up to date information on point of the theory and on point of best practices have become a high priority. The learning model has to be flexible to be able to adapt to the learning style of the student and to be able to deliver personalized education.

The system supporting the fundamental *theory-example-exercise* triangle is further enhanced by the opportunity of access to commercial computer application programs and, to various databases and to up to date information on the Internet.
The existing systems already have a real-time discussion- and chat function and soon the systems will be able to learn about the user's problems and preferences and as a consequence will deliver "just in case" information. By this way than the user will have available a personalized e-learning system.
In case of learning of business applications, the integration of benchmarking models will be also a key element of the learning/teaching modules, in the sense that, on the one hand, an "ideal" virtual enterprise is created and, on the other hand, "real" business processes are provided to enable students in their role of business managers, to derive problem solutions.

An advanced e-learning system will create an "interactive" environment for the students, which offer an immediate assessment of their learning actions and will provide a context for the exchange of information, which can be accessed locally or at a distance. Because it is a user-friendly and collaborative e-learning environment for personalized education, it will therefore stimulate the pleasure of learning and creativity and will result in a more effective way of learning.

Additional dynamic information collected by web-search agent will also be integrated into each learning module. The contents are complemented by constantly updated benchmarking solutions, and chat facilities.

The content delivery model makes a difference between the static information and the dynamic/recent information.

dynamic/recent information	static information
intelligent agents real-time expert discussion (mailing service)	learning general information learning content modules cases (and benchmarking cases)

Simulation and problem solving via benchmarking cases has to be implemented in a simulation tool and must be linked with the learning-portal.
The facility of real-time expert discussion session is a standard facility in some e-learning systems. The LearningSpace solution includes sharing of documents and voice and video support.
Intelligent agents are not included as a standard facility.

THE DISSEMINATION OF THE INTELLIGENT AGENT TECHNOLOGY IN THE E-LEARNING SOLUTION

Web-search agents will gather information on the relevant know-how topics and other areas of interest specified by individual users.
Intelligent agents implemented in the e-learning solution will provide:

- information retrieval from several (internet) sources.
- information filtering according to the personal characteristics of and delivered by the user.
- coaching of the student throughout the learning process.

By implementing the intelligent agent technology in the e-learning solution, a real flexible tool will result that allows individual and constant adjustments to user needs.
This facility has to be integrated with the basic e-learning functions of the used e-learning system into a web-portal.
This portal provides easy access to virtual learning modules anywhere and anytime by employing both online learning tools and agent technology for training purposes. The composition of the learning modules can be individually adapted based on the business characteristics of the student that will be specified in the participant profile upon registration to the course.

CONCLUSIONS

As a consequence the possibilities of e-learning will no longer be limited by the features of existing software products. Due to their flexible composition, virtual learning modules can be adjusted to changing educational/ informational needs of participating students. Some modules can be updated or even eliminated and subsequently replaced by new ones that prove to be of major relevance for users. Tracing agents and counting hints can measure information relevance.
Benchmarking models, best practice methods, high quality and personalized information inputs into the system as well as a virtual enterprise to derive problem solutions will boost the level of e-learning

REFERENCES

Cuppers L. and J. Schreurs. *The implementation of a virtual learning environment at the Limburgs Universitair Centrum*. Euromedia 18-20 april 2001.

Schreurs, J : *The learning portal of a collaboration learning system.* Euromedia 18-20 april 2001.

K.Kruse&J.keil : Technology-based training. The art and science of designing, development and delivery.
Jossey-Bass pfeiffer 2000.

M.J. Rosenberg : e-learning. Strategies for delivery knowledge in the digital age.
McGrawhill 2001

TO REACH CONFIDENCE IN MATHS "LEARNING BY DOING"

Ciro D'Apice
Rosanna Manzo
Department of Information Engineering
and Applied Mathematics
University of Salerno
Via Ponte Don Melillo
84084 Fisciano (SA), Italy
E-mail: {dapice, manzo@diima.unisa.it}

KEYWORDS
Computer Based Education, Learning by discovery and exploration, Interactive tools.

ABSTRACT

Blackboard and chalk are no longer the unique tool to teach mathematics. Students now require to learn using computer in a more interactive way. What do students expect from an interactive tool in maths teaching?

Collecting students suggestions we have realized friendly packages in Mathematica to motivate some mathematical concepts learning.

The goal of the packages is to enhance student insight, in the spirit of learning by experimentation. A package, able to motivate the learning of Laplace Transform and z-Transform, is presented.

INTRODUCTION

The principal goal of education is to create people who are creative, inventive, discoverers, who have minds which can be critical, can verify and not accept everything they are offered.

Computer based education (CBE) can help teachers in reaching this goal, offering the possibility to create various interaction tools and feedback forms, able to develop student critical mind and curiosity.

Even if current research on CBE is addressed on how to represent the learning content and tends to neglect the impact of the user-interface in the learning process, we have to take into account that the user-interface has a direct influence on knowledge acquisition. An integration of interface design with the representation of the learning content is thus fundamental for the success of the teaching-learning process [Shär et al.].

Today with the rapid growth of students' number and the grater diversity of their academic backgrounds, the task to teach mathematics in higher education is becoming much harder. Students appear to be less confident with mathematical subjects than before.

Interaction tools that increase the cognitive load on students when learning (e.g. by asking the user to type command in order to interact, to verify…) can be more effective than traditional learning tools.

TOWARDS INSTRUCTIVISM - CONSTRUCTIVISM PARADIGMS

According to Rieber (1992) and others (Duffy and Jonassen, 1992; Papert, 1993) teaching and learning can be realized using instructivism and constructivism paradigms.

In the instructivism approach, the learning process is focused on goals and objectives drawn from a domain of knowledge, e.g. algebra, or collected from behaviours' observations of experts in a given domain, e.g. surgeons.

Once goals and objectives have been individualized, a hierarchical scheme to reach them is planned, then direct instruction is designed to address each of the objectives in the hierarchy.

Little emphasis is given to the learner who is considered as a passive recipient of instruction. In CBE tools based on instructivist pedagogy, learner is treated as an empty vessel to be filled with knowledge.

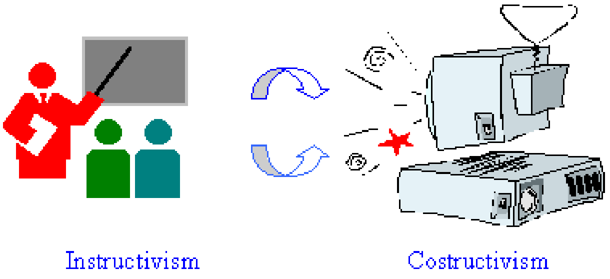
Instead constructivists stress the importance of learners' intentions, experience and the learning process is viewed as involving "individual constructions of knowledge" [Carrol]. They try to ensure that the learning environment is as rich as possible. Learner is not regarded as an empty vessel but as an individual with pre-existing knowledge, aptitudes, motivations, interests, styles, capabilities to which learning process has to be tailored.

Knowledge is not regarded as something that exists outside students which they have passively ingest, but as being socially and individually constructed on the basis of the experience.

In the constructivism paradigm often direct instruction is replaced with self-direct exploration and discovery learning.

It is obvious that some knowledge due to its nature can be efficiently and effectively acquired through direct instruction and other knowledge can be so creative or of a higher level that direct instruction is inappropriate.

The approach we have followed in mathematics teaching in Engineering Courses uses both instructivism and constructivism paradigms: we have combined traditional lectures with learning through discovery and exploration.



CBE ROLE IN EDUCATION

"Mathematics: sometimes it is so abstract, removed from real world experience". In traditional instruction information is presented in abstract lectures and texts in such way that students are not able to connect conditions (as a problem) and actions (as the use of knowledge to solve the problem). They "regurgitate" memorized information and often they don't have the ability to retrieve the necessary information in situation of problem solving. Instead if knowledge and skills are learned in a context of use, they will be used in that and similar contexts.

Carroll (1968) told us: "By far the largest amount of teaching activity in educational settings involves telling things to students…"[Rieber]. Today CBE allows us to say: "Where teacher exposition is an appropriate instructive strategy, CBE can be planned to support, reinforce and extend teacher presentations.

Applying information/computer to education is not just about presenting subjects via new media, it is about how to harvest the power of the technology to create a new way of teaching, learning and even thinking.

Students spend a lot of time assenting to what teachers assert mathematically. This does not imply that they have blindly accepted it: indeed they may have worked hard in order to agree with it. But agreement is different from understanding. To reach a real and deep understand they need to be able to assert things for themselves. The role of a teacher is to guide rather then to instruct, to suggest rather than to transmit. In this way, students can have ownership of the knowledge gained and mathematical knowledge can be constructed or discovered by the learners using interactive tools that are important in the process of discovery.

In realizing packages able to support mathematics learning we have applied the old maxim "experience is the best teacher", which reflects the belief that we learn much in life through trial and error.

Students can become confident with the subject of study "learning by doing". Experiential learning is highly valued because it provides opportunities for us "to learn from our mistakes. Obviously, reflection and criticism are the features that are involved in this learning experience.

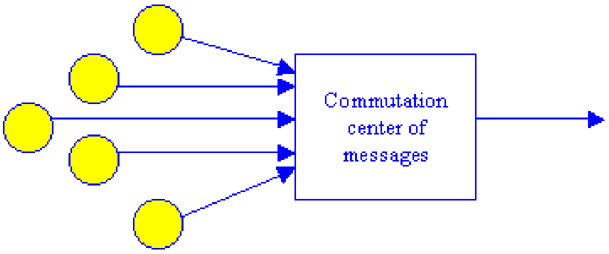
TEACHING LAPLACE TRANSFORM AND Z-TRANSFORM

The idea of the realised package was born from the consideration that there is an effective difficulty in the learning of mathematics. In fact, students have difficulties in understanding some basic mathematical concepts, and sometimes the utility of them, because they don't consider connections with the concrete world. From a pedagogical point of view, a task of the educators is to show these links and underline the importance of mathematics in solving problems of the real world.

The students of the third year of Electronic Engineering during the courseware of Mathematics IV have to deal with Laplace Transform and z-transform.

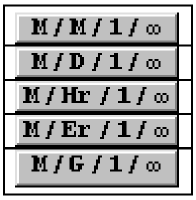
How can we motivate students in learning Laplace transform and z-transform? We have implemented a package through which the students can utilize abstract concepts in problem-solving situations. The package has been realized with the aim to satisfy students need to learn with practical applications and with more sympathy.

Let us consider a commutation center of messages with a big number of computers and one transmission line with high speed. Let us assume that the offered traffic is described by a Poisson flow with intensity equal to λ. The messages length is exponentially distributed with mean value μ and the transmission speed is equal to ν. Which is the waiting time of a generic message?



We can describe the behaviour of the system through a queuing system of type M/M/1∞ in which the customers are represented by the arriving messages and the service is executed by the transmission line. Using Laplace and z-transform we can obtain the stationary probability to have n messages in the system and the required waiting time.

The implemented package allows student to deal with some classes of problems that can be described through queueing systems with one server and waiting queue of infinity capacity. Using a palette the student can choose a class of queueing systems



where in the Kendall notation, M/G/1/∞, the first and second symbol denote, respectively, the distribution function of inter-arrival times and of service times.

Symbol	Distributiom
M	Exponential
D	Deterministic
H	Hyperexponential
E	Erlanghian
G	General

In the first experimental lecture the software Mathematica is introduced. This introduction includes what Mathematica is and can do, then students learn how to use the package.

The choice of Mathematica™ among the software systems suitable to develop didactic applications finds motivation in its features among which powerful numerical and symbolical calculus capabilities and high-level expressive programming language.

The developed package encourages independent thinking. Its features enable students to actively learn with discovery and exploration. Students can change the problem parameters and see the impact of these changes on the results.

We report an example of learning using the package.

We suppose that a student has chosen the class M/M/1/∞, then the package proposes him/her to insert the input parameters:

M/M/1/¥

Insert [illegible] intensity , Service Time

Insert [illegible]

The utilization coefficient is ugual to 0.666667. It is less than 1, so the system is ergodic and we can apply the Pollschek-Khintchin theorem to obtain the z-transform [illegible] tionary probabilities

The z-transform is given [illegible]

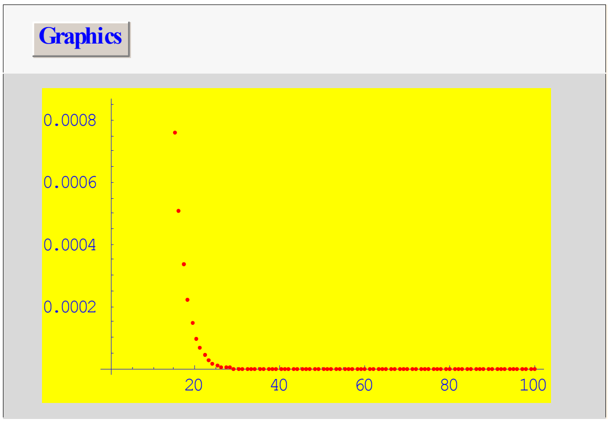
$$P\,[\text{illegible}]\,\frac{z}{3\,[\text{illegible}]\,\frac{2}{3}+z}\,[\text{illegible}]$$

Apply the z-inversetransform to obtain the probabilities to have n customers in the system:

$$p_n = 2^n 3^{-1-n}$$

Then the student can visualize the graphics of the stationary probabilities to have n customers in the system and the value of some useful performance indices.

The graphical visualisation encourages the students to explore mathematical properties, by arousing their curiosity.



Performance indices

Probability of free system = 0.333333
Mean number of customers in the system = 2
Mean waiting time = 0.666667
Mean sojourn time = 1

We report a session in which the learner has chosen the class M/H$_2$/1/∞. He/she has to insert the input parameters:

$M\ \ H_2$/1/¥

Insert [illegible] Intensity, Hyperexponential Order

Insert [illegible]

Insert [illegible] a_2, l_2 [illegible]

Insert [illegible] 3 4, 2 [illegible]

You have chosen the following hyperexponential distribution function :

$$\frac{1}{4}\,[\text{illegible}]\,-e\,[\text{illegible}]\,\frac{3}{4}\,[\text{illegible}]\,-e^{-2x}\,[\text{illegible}]$$

Which is the mean service time?

Insert [illegible] ice time [illegible]

Insert [illegible]

Try again!!!

Insert [illegible]

Now go on!!!Compute the utilization coefficient!!Is the system ergodic?

Answer [illegible]

Compute Laplace-transform of the distribution density function.

b $\frac{6}{2+z}$ 4 1 + z

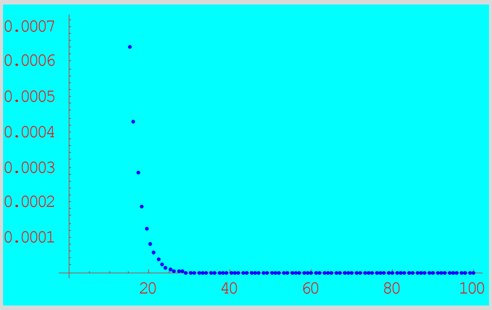
Apply the Pollschek-Khintchin formula to compute the z-transform of the stationary probabilities.
The z-transform is given by:

P + 5 z
8 - 1 z + 15 z^2

Apply the z-inversetransform to obtain the probabilities to have n customers in the system

$p_n = 2^{-5+n} 5^{-n} + 3^{2-n} 5^n$

Graphics



Performance indices

Probability of free system = 0.333333
Mean number of customers in the system = 1.79167
Mean waiting time = 1.16667
Mean sojourn time = 1.79167

Finally we analyze a M/E_2/1/∞ queueing system.

M/E_2/1/¥

Insert Intensity,
Hyperexponential Order, Mean Service Time

Insert 3

You have chosen the following hyperexponential distribution function :

1 - $\tilde{a}^{-6x}$ + 6

Now go on!!!Compute the utilization coefficient!!Is the system ergodic?

r = 2 3. The system is ergodic.

Compute Laplace-transform of the distribution density function.

b + z

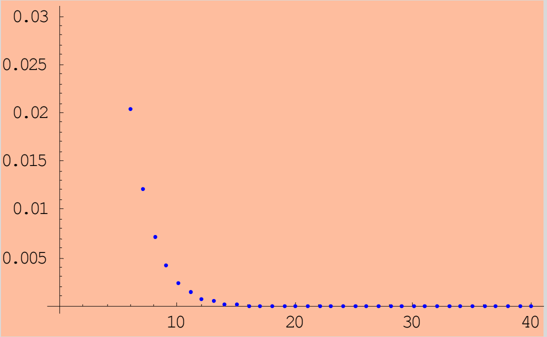
Apply the Pollschek-Khintchin formula to compute the z-transform of the stationary probabilities.
The z-transform is given by:

P 1 + z 7 + z

Apply the z-inversetransform to compute the probabilities to have n customers in the system

$p_n = \frac{18 \quad 7 + \quad 13 \quad 1 \quad {}^n + \quad 13}{6 \quad 13}$

Graphics



Performance indices

Probability of free system = 0.333333
Mean number of customers in the system = 1.66667
Mean waiting time = 0.5
Mean sojourn time = 0.833333

As we can deduce from the reported examples, the package meets three fundamental qualities of learning:

- active: learners have an active role in the learning process; in fact, they insert data, manipulate information and they are responsible for the results;
- contextual: learning tasks are situated in meaningful real world tasks or they are introduced through case-based or problem-based real life examples;
- reflective: learners articulate what they have learned and reflect on the observed results.

An evaluation questionnaire is given out at the end of the experimental lectures asking the students what they think of the package. In general students appreciate the package, showing great interest.

Differences and similarities were examined between the class with the package and without the package in view of the understanding of mathematics.

CONCLUSIONS

It is true that technologies are widespread almost everywhere, but the issue is how to use the available technology in more effective and efficient ways to help students better understand mathematics.

Great efforts has to be addressed towards the analysis of the pedagogical relevance of the use of CBE in the learning process evaluating the capabilities of CBE to realize powerful instructional interactions, monitor learner progress, empower effective teachers, accommodate individual differences, etc.

To improve mathematics teaching-learning process is a great challenge due to the fact that mathematics is present in so many other subjects and courses that it is possible that failure at maths is the biggest reason for the first year failure in higher education.

REFERENCES

G. Albano, C. D'Apice, R. Manzo: “An Interactive tool in teaching statistics”. 1999. Proceedings Sixth International Conference on Statistics, Combinatorics, and Related Area, Mobile, Alabama.

G. Albano, C. D'Apice, B. D'Auria, R. Manzo: “A new approach to teaching/learning mathematics”. 2000. Proceedings EAEEIE (Annual Conference on Innovations in Education for Electrical and Information Engineering), Ulm, Germania, 1-5.

Carrol. J. B. “On the learning from being told”. 1968. *Educational Psychologist*, 5, 4-10.

C. D'Apice, R. Manzo, E. Zappale: “Learning power series with computer tools”. 2000. Proceedings 4th International Derive-TI89/92 Conference, Liverpool.

Keller, J. M. “Strategies for stimulating the motivation to learn”. 1987. *Performance and Instruction*, 26(8), 1-7.

Rieber, L. P. “Computer-based microworlds: A bridge between constructivism and direct instruction”. 1992. *Educational Technology Research and Development*, 40(1), 93-106.

Shär S. G., S. Schuep, C. Schierz C., H. Krueger H. “Interaction for Computer-Aided Learning”. *IMEJ (Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*).

BIOGRAPHY

Ciro D'Apice was born in Castellamare di Stabia, Italy and obtained PH degrees in Mathematics in 1997. He is researcher at the Department of Information Engineering and Applied Mathematics in University of Salerno. His research interests include: computer aided learning, queueing theory, hybrid systems, homogenization problems, spatial behaviour for dynamic problems.

Manzo Rosanna was born in Polla, Italy. She studied Mathematics and obtained her degrees in 1996. From 1998 she works at the Department of Information Engineering and Applied Mathematics in University of Salerno. Her research area are queueing theory and computer aided learning.

VR in civil engineering education. Which is the student's expectancy?

Gerardo Silva Chandía

Department of Civil Engineering, The University of Santiago, Santiago, Chile

Abstract

Professional engineers and engineering students are commonly very practical human beings, so they could appreciate the potential of a computational tool or technique by working hands on in a sample project a few months. At the end of 2001 a survey was realised in a course of civil engineering students that worked near four months in a virtual engineering project named Island Project where they planned three virtual projects: airport, maritime port and a road in between. The results of the survey are given in this paper.

Keywords:

VR, VE, civil engineering, engineering education

1. Introduction

Professional engineers and engineering students are commonly very practical human beings, so they could appreciate the potential of a computational tool or technique by working hands on in a sample project a few months. What is the attitude from civil engineering students after the use of the Virtual Reality (VR) techniques in a case project by working four months in it, that is the motivation of our research conducted in the last year of studies of civil engineering education, at the U. of Santiago, Chile, since 1999.

In the Construction Projects course, a VR introduction was realised by the last three years primarily as a "natural" sequence to the 3D modelling of civil projects developed by using CAD tools since half '80s period. In this methodology, students construct 3D models of terrain and modify them by including esplanades and roads at a conceptual level.

In precedent years, students did construct basic 3D models of houses, buildings, bridges, earth dams, tunnels and industrial buildings, each one developed by using well known 3D CAD techniques. In the final stages of the project, the 3D CAD model was "placed" in the virtual terrain and was published in a web page to show the project both the teacher and the other students. In the Construction Projects course, the VR work is a complement to the principal task consistent in the development of construction drawings, quantities and cost estimations, planning and programming the construction tasks of selected civil projects.

In precedent years, at the end of the course a survey was conducted with the students, asking them their opinion about the use of the VR techniques in the civil engineering education. The author published papers with these results in 2000, and they did show a positive attitude from the students both to use VR techniques and web pages as support to their work.

At the end of 2001 the question included in the survey was different and assuming a positive attitude from the students, they were asked to propose VR applications they would like to count with to support their studies.

In the next section, results of the previous surveys are given. In Sections 3 and 4, this last experience is detailed and finally some conclusions are remarked.

2. Previous Work

At the end of 1999 and 2000, surveys were conducted with the students to appreciate their response to the use of VR techniques in the civil engineering education. The results are given in the next tables.

2.1 At the end of 1999

After work in basic 3D model of houses and buildings, the following were the principal results of the student's survey:

Question:

"Give us your opinion about the use of VR in civil engineering education"

An analysis of the 17 freely written answers gave the next conclusions:

1. VR is an attractive visualization tool in the development of civil engineering projects (65%)
2. VR gives an added value to the development of civil engineering projects (35%)
3. VR requires a previous engineering software training to be productive (29%)
4. Walkthrough or interact with virtual worlds or virtual environments do not replace the learning results of real visits or inspections of civil works, but in absence of these, VR works fine (24%)
5. VR is a very fine tool for a learning process because it captures the student attention (18%)
6. VR is more efficient in architectural teaching than civil engineering teaching (12%)
7. VR is an efficient tool to detect changes needed in a project in development (12%)

2.2 At the end of 2000

After work in basic 3D model of civil projects as bridges, earth dams, tunnels, etc., the principal results of the survey with the students are presented below.

Question:
"Give us your opinion about the use of VR in civil engineering education"
The contents of the 37 freely written answers given by the students are being summarized in the next tables.

Use of VR in civil engineering education is:	**%**
Useful or very useful	35,1
Important	18,9
Interesting, excellent (the same percent for each one)	8,1
A powerful tool, a fundamental tool (the same percent for each one)	5,4
Necessary, positive, a good innovation, indispensable (the same percent for each one)	2,7
Total positive attitude:	**91,9**
Not fundamental	10,8
Note: Percentage add don't summarize 100 because several students included more than one adjective in their answer.	

Reasons given:	**%**
Projects can be presented to non technical users in a clearly way	32,4
Projects can be visualized in a realistic way	24,3
VR abilities are potentially a plus at working search time	21,6
We can visualize a project in a short time	21,6
It is a new tool to support technical presentations	18,9
We can visualize complex situations derived from the construction project	18,9
Presentations with VR are more pedagogical	16,2
Presentations with VR are more entertained	10,8
Engineering projects are better understood by students without experience	10,8
Project's errors are discovered	8,1
Physical scaled representations are almost unneeded	8,1
We can visualize VR models from existing and not yet existing projects	5,4
We can interact with project's parts or elements	2,7

Comments included by the students:	**%**
VR must be used before the last year of studying	21,6
VR must be learned hands on	8,1
VR used as a support tool needs dedicated PC and projection equipment	8,1
VR use distort the course's objectives	8,1
More PCs are needed, the other courses are unattended (same percent)	2,7

3. The year 2001 experience

As in precedent years the students work begins by doing the terrain modelling by using some of the next software packages: Surfer from Golden Software or Civil Design from Autodesk (actually Landscape Development).

At this time each student received a set of fictitious topographical data corresponding to a "virtual island" located in the south of the country (Chile).

With the purpose of the island can be inhabited and natural resources can be exploited and exported, students must plan, locate and construct 3D models of three projects: airport, maritime port and roads in between.

The methodology included the next steps:
a) Construction of the Digital Terrain Model (DTM) of the island, to get the contour level curves and 3D visualisations from different viewpoints.
b) Convertion of the DTM of the island into a virtual environment (VE) by using the software named WalkThrough from AutoDesk. This VE was presented to the teacher by using a video in an AVI format.
c) After inspect the contour level plan of the island, the students did select alternative locations to place an airport and a maritime port. In the airport case, calculations were done to set the level of the esplanade, so the earthwork was of minimum volume and to compensate cut and fill volumes. In the case of the port, conditions as the bathimetry and others were considered.
d) Construction of the basic 3D model of the piers of the port.
e) Construction of the basic 3D model of the airport facilities.
f) Construction of the basic 3D model of the road between the port and the airport.
g) Convertion of the 3D model of the island with the civil projects into a VE, including animations, to present the principal aspects of the project.

In the 2001 experience, lectures included presentations of VR applications obtained from the WWW, the TV and other sources.

4. The 2001 survey

This time the question was:
"From your experience as a student of the course, which applications you would like to get in civil engineering?
The answers from 26 students are summarised below:

No.	Application:	Reasons:
14	Step by step building process models	To support the learning process of civil projects not shown in the regular studies
7	VR models of finished civil projects	Aids to the full building project understanding. It allows seeing alternatives, the context with the surround and possible errors. Gives support to public presentations of the project and to the planning process
3	VR models of details and finishes	To support the communication with clients and others engineers
3	VR models of earthquake structure interaction	To provide a better understanding of the structure movements
3	VR models of roads and highways	To support studies in the fields of roads security and roads design
1	Topography	To support land design

5. Bibliographic discussion

Use of VR tools in AEC industry was leaded in the last decade by the Architectural component and at a long distance followed by the Engineering and Construction areas.

The application spectrum in the last two areas is well reflected by the papers presented to the AVR II and CONVR 2001 conference at Chalmers, Gothenburg, Sweden in October 2001.

A very good synopsis of the state of the research is included in the introduction of the paper presented by Ms. Jennifer White in that conference.

In the engineering education arena, VR tools has been used by basic sciences researchers developing applications to support mathematics, physic and chemistry principally, but at the graduate level, there are few attempts to use these tools.

Dr. Nelson Baker at Georgia Tech., USA, has included in his EPITOME research program, VR applications to support some postgraduate courses.

At Chalmers, Sweden, the VR applications in education are developed in connection with Finite Element Method (FEM) applications.

In the USA, Dr. Martin Fisher and other researchers has worked in the development of the integration between VR tools and scheduling tools to see or track the building construction evolution in the time.

6. Conclusions

At the end of 2001, after three years of experience with last year civil engineering students, we can conclude that they recognize the importance of to have at least basic abilities in VR modeling and Internet web pages creation.

No one rejects the learning and use of these techniques in the engineering studies and practice, but all of them claim lack of time and too much to study in the other traditional courses.

They suggest beginning the learning process of VR and Internet technologies early in the engineering studies, to promote its use in the others courses of the curriculum and to accelerate the introduction of the engineering software in the last courses.

None in this experience suggested such as in previous years, that VR is only for architects or other professionals.

Students not having PCs at home or not having Internet connection expressed they felt themselves working in lower condition in respect the ones that do have them.

We believe that civil engineering students become appreciate these technological tools because they can use them in an active way by doing their own VR worlds and web pages.

At the end of 2001 they visualize several areas of civil engineering projects, where VR can give a support to the learning process or to the practice of the engineering.

Especially they point out the need to develop models that are visualized step by step to support the learning and the understanding of complex projects.

There are two ways to generate this kind of models: the first is that they were developed by the teacher or their assistants, the second way is that the models were developed by the same students, as part of the course, under the guide of the teacher.

We believe that the last way is more possible, and as a student wrote in the survey, the kind and size of the projects considered must be smaller and the time dedicated must be greater.

Only future time will give the answer to the question about these expected applications could become exist to support the students expectancy.

6. Further research

Working with a selected students group, we plan in 2002, to conduct a research about the student’s perception when immersive VR is used.

The methodology adds to the previously explained aspects in this paper, the use of the VFX1 head set and the MindRender software from ThemeKit, UK.

7. Some pictures

The kind of work developed by the student is better reflected by their animations, but an idea can be obtained from the next pictures.

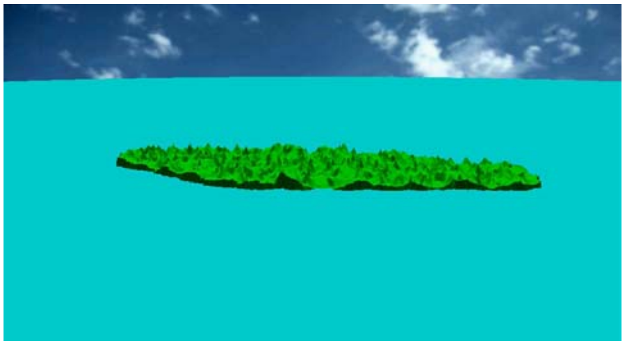


Figure 1: *South viewpoint of a fictitious island*

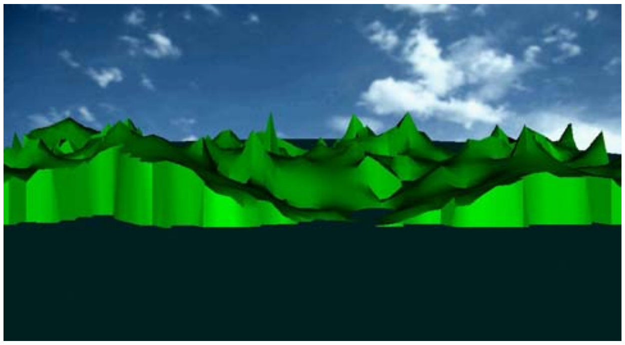


Figure 2: *Near viewpoint of the port location proposed*

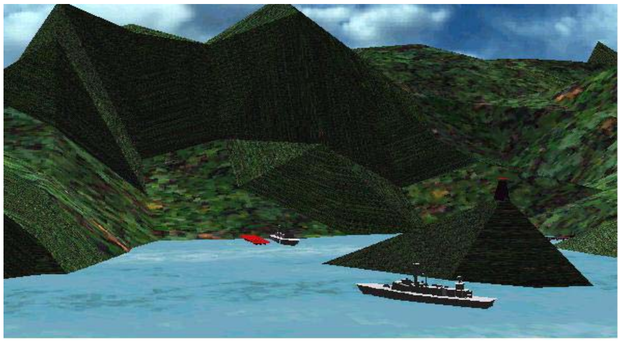


Figure 3: *Pier proposed by one student*

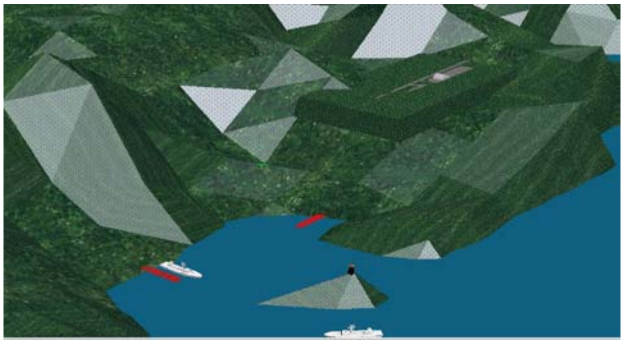


Figure 4: *Different viewpoint of the same pier*

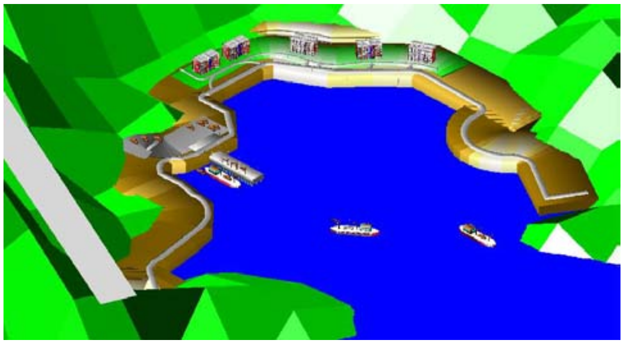


Figure 5: *Maritime port proposed by other student*

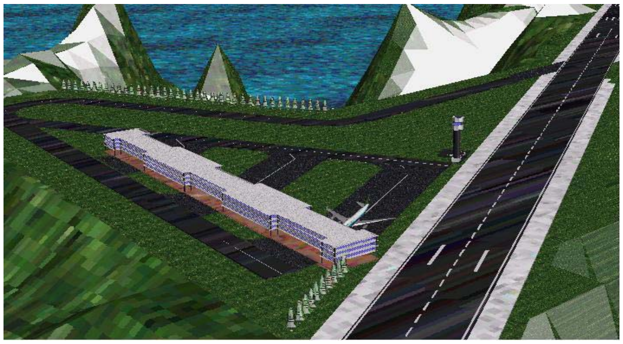


Figure 6: *Airport facilities proposed by one student*

References

1. J. Whyte (2001). Business Drivers for the use of Virtual Reality in the construction sector. *AVR II and CONVR 2001 Proc.*, 99-105
2. B. Koo and M. Fischer (2000). Feasibility study of 4D CAD in commercial construction. *Journal of Construction Engineering and Management,* Vol. 126, (4): 251-260.
3. A. Op den Bosch and N. Baker (1995). *Simulation of Construction Operations in virtual environments.* Proceedings of the Second ASCE Congress for Computing in Civil Engineering, Atlanta, USA.

INFORMATION MANAGEMENT

Learning content management in a University

Ivan Picart, Rachel Moreau, Jeanne Schreurs
Limburgs Universitair Centrum
Universitaire Campus
B-3590 Diepenbeek, Belgium
E-mail: ivan.picart@luc.ac.be
rachel.moreau@luc.ac.be
jeanne.schreurs@luc.ac.be

KEYWORDS

Knowledge pool; warehouse; collaborative learning; learning material; e-learning platform.

ABSTRACT

Because often a university wide access to learning materials is important, a central knowledge pool for our university has been built. The idea of warehousing has been copied from business practice. The system has already been implemented as a prototype in the business informatics department of our university. While we cannot fully guarantee the safety, we are analysing the security rules available in ASP.NET. In our solution we proposed some basic requirements such as reusability, interactivity, and personalization. Ongoing research points in the direction of standard XML file format to store the learning materials.

INTRODUCTION

Research and learning materials are often locally created and/or organised by the users (authors) themselves on different ways and on different systems. As a consequence the access is rather limited. Because often a university wide access is important, a central knowledge pool for our university has been built. As a consequence, we improved the efficiency of the information search activity. Staff members will find the information they want to have, even if it is owned by other users and created and stored initially on local systems of individual users. This idea of warehousing has been copied from business practice. The system has already been implemented as a prototype in the business informatics department of our university

THE SYSTEM DEVELOPMENT PROGRESS

Starting from the problem definition, the system architecture has been defined. Several commercial solutions have been evaluated. Initially, preference was given to a relational database approach. The documents would be organised as attributes in a database table. Later on, the solution of document management systems came in the picture.

In this phase we analysed the IBM database system DB2, Lotus, domino.doc system and Microsoft SharePoint/Webstorage Technologies

The advantages and disadvantages have been weighed thoroughly. None of those systems seemed to be suitable/affordable for our University.

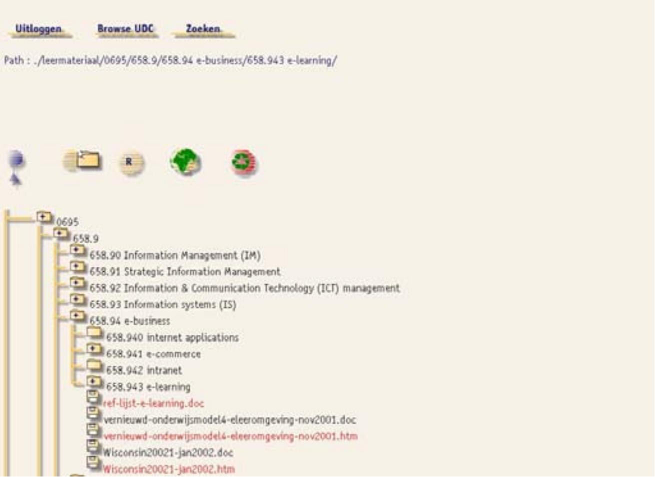
Finally we changed again and we opt now to split the physical storage of the document on one hand and the definition of the document on the other hand.

The metadata or the characteristics of the document are stored in a database-table.

The documents themselves are centrally organised in a directory/ subdirectory hierarchical and multilevel structure. The documents are belonging to a domain/discipline and have to be published in the corresponding discipline directory. The directory organisation is following the well-known library catalogue system UDC (Universal Decimal Classification) complemented with some home-defined UDC codes. The document is linked with its metadata in the database table.

USERS OF THE WAREHOUSE SYSTEM

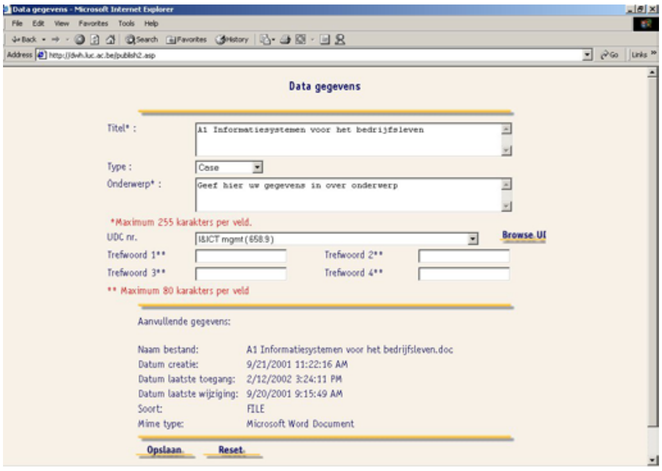
An author/user can decide to centralise his files/documents on the warehouse server and to publish them immediately or later on.



Path : ./leermateriaal/0695/658.9/658.94 e-business/658.943 e-learning/

0695
658.9
658.90 Information Management (IM)
658.91 Strategic Information Management
658.92 Information & Communication Technology (ICT) management
658.93 Information systems (IS)
658.94 e-business
658.940 internet applications
658.941 e-commerce
658.942 intranet
658.943 e-learning
ref-lijst-e-learning.doc
vernieuwd-onderwijsmodel4-eleeromgeving-nov2001.doc
vernieuwd-onderwijsmodel4-eleeromgeving-nov2001.htm
Wisconsin20021-jan2002.doc
Wisconsin20021-jan2002.htm

A file/document can be stored in a private personal directory as long as it is not yet ready to publish. The access to this directory is limited to this user. To publish the file/document, the metadata has to be defined.


Data gegevens - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dwh.luc.ac.be/publish2.asp
Go Links
Browse UI
Opslaan
Reset

Data gegevens

Titel* : A1 Informatiesystemen voor het bedrijfsleven

Type : Case

Onderwerp* : Geef hier uw gegevens in over onderwerp

*Maximum 255 karakters per veld.

UDC nr. I&ICT mgmt (658.9)

Trefwoord 1** Trefwoord 2**

Trefwoord 3** Trefwoord 4**

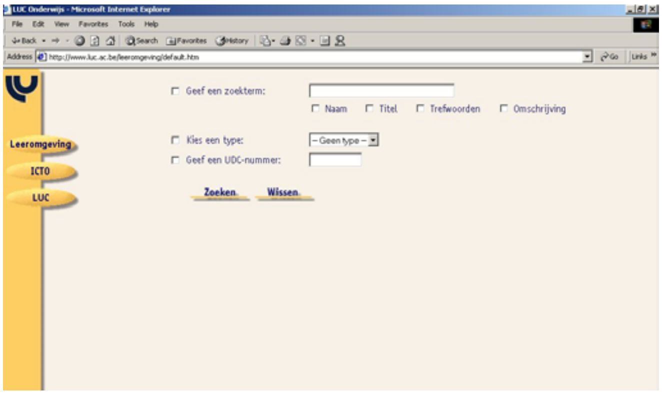
** Maximum 80 karakters per veld

Aanvullende gegevens:

Naam bestand:	A1 Informatiesystemen voor het bedrijfsleven.doc
Datum creatie:	9/21/2001 11:22:16 AM
Datum laatste toegang:	2/12/2002 3:24:11 PM
Datum laatste wijziging:	9/20/2001 9:15:49 AM
Soort:	FILE
Mime type:	Microsoft Word Document

The UDC code is one of the metadata items and will force the storage of the file/document in the corresponding UDC directory.

Once the file/document is published it will be accessible for other users through the Internet. A user will query the knowledge pool to become some selected information. A query interface in the browser (as an ASP based application) has been built for all staff members and students. The query is executed on the metadata table and a query report of the corresponding linked documents has been generated by this application.


LUC Onderwijs - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.luc.ac.be/leeromgeving/default.htm
Go Links
Leeromgeving
ICTO
LUC
Zoeken
Wissen

□ Geef een zoekterm:

□ Naam □ Titel □ Trefwoorden □ Omschrijving

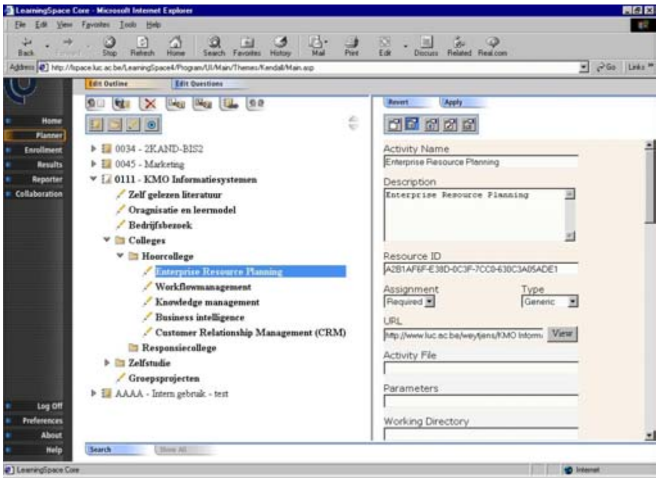
□ Kies een type: – Geen type –

□ Geef een UDC-nummer:

Another kind of use is the integration of the warehouse learning documents in the e-learning platform to be used in the courses. An important requirement is the independence of the learning materials from the learning system. Our warehouse guarantees this independency. The documents are stored as html-files in our warehouse and the most advanced learning systems can create linkages to those html-pages.

In our University we are using LearningSpace 5 (Lotus Notes). A main characteristic of this learning system is the independency of the learning content. In our learning process we are following a learning path organised as a series of modules. The content is preferably stored as html documents in the warehouse and is simply linked with its URL.

These learning documents can be created by a normal text processor or by special authoring tools. Animation can be added by using Macromedia Flash, Dreamweaver, Microsoft Frontpage, …


LearningSpace Core - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Related Real.com
Address http://lspace.luc.ac.be/LearningSpace4/Program/UI/Main/Themes/Kendall/Main.asp
Go Links
Edit Outline
Edit Questions
Home
Planner
Enrollment
Results
Reporter
Collaboration
Log Off
Preferences
About
Help
Revert
Apply
View
Search
Show All
LearningSpace Core
Internet

- 0034 - 2KAND-BIS2
- 0045 - Marketing
- 0111 - KMO Informatiesystemen
  - Zelf gelezen literatuur
  - Oragnisatie en leermodel
  - Bedrijfsbezoek
  - Colleges
    - Hoorcollege
      - Enterprise Resource Planning
      - Workflowmanagement
      - Knowledge management
      - Business intelligence
      - Customer Relationship Management (CRM)
    - Responsiecollege
  - Zelfstudie
  - Groepsprojecten
- AAAA - Intern gebruik - test

Activity Name: Enterprise Resource Planning

Description: Enterprise Resource Planning

Resource ID: A2B1AF6F-E38D-0C3F-7CC0-630C3A05ADE1

Assignment: Required

Type: Generic

URL: http://www.luc.ac.be/weytjens/KMO Inform

Activity File

Parameters

Working Directory

SECURITY ASPECTS AND FUTURE IMPROVEMENTS

The security aspects of our warehouse are based on the NTFS security settings of our Windows 2000 advanced server. A Windows 2000 user will be created for each author/user of the warehouse. With this account, the authors/users can save their files into their personal directory. The access is limited to themselves. To make the documents accessible through the Internet, they must be organised into the right UDC category in the UDC directory. We can't allow the users, even the authors/users to have write access in this collective, shared directory. Therefore, we created a 'dummy user'.

While we cannot fully guarantee the safety of this user now, we are analysing the security rules available in ASP.NET. We do believe that the ASP.NET Impersonation technology will increase the security aspects.

ONGOING RESEARCH ON POINT OF IMPLEMENTATION OF XML FILE FORMAT

We proposed some basic requirements such as reusability, interactivity, and personalization. They have to be fulfilled in the near future. Ongoing research points in the direction of standard XML file format to store the learning materials. XML provides a standard way to tag or mark up information documents to become easy accessible and exchangeable. As an example the National Library of Medicine prescribes XML to support the input, the storage, the process and output of the data from its MEDLINE database. (http://www.fcw.com/fcw/articles/2001/0108/cov-xmlbx4-01-08-01.asp)

Another succeeded example is NADA (National Archives and Records Administration) that is using XML technology in their electronic archives program to help ensure that all documents can be read without needing the software program that produced them (http://www.fcw.com/fcw/articles/2001/0108/cov-xmlbx5-01-08-01.asp

REFERENCES

A Lotus Development Corporation White Paper: Lotus and IBM knowledge management strategy. September 2000.

Schreurs, J. 19-20 October 2000. *Development of a Virtual Learning Environment and Implementation of Courses in ICT and in ICT Management*. Conference Proceedings, Bolton International Conference.

Schreurs, J. a.o. 9-10 May 2000. *A multimedia warehouse supporting on line learning via internet.* Euromedia 2000: Building a global business. Ed. F.Broeckx & L.Pauwels.

Cuppers L. and J. Schreurs. *The implementation of a virtual learning environment at the Limburgs Universitair Centrum*. Euromedia 18-20 april 2001.

Schreurs, J : *The learning portal of a collaboration learning system.* Euromedia 18-20 april 2001.

http://www.fcw.com/fcw/articles/2001/0122/tec-xml-01-22-01.asp

ON THE VALIDITY OF SIMULATION MODELS IN PROCESS ENGINEERING AND OPERATOR TRAINING

Prof., Dr. Victor M. Dozortsev [1], Dr. Dmitry V.Kneller [1], Dr. Mark J.Levit [2]

[1] Institute of Control Sciences RAS,
65, Profsoyuznaya St., 117997, Moscow, Russia
E-mail: {victor, dvk}@petrocom-jv.ru
Homepage: http://www.petrocom-jv.ru

[2] The Security Industries Automation Corp.,
New York, NY, 11201 USA
E-mail: marklevit@juno.com

KEYWORDS

First principles simulation model (FPSM), Validity, Operative control task, Decomposing.

ABSTRACT

The paper discusses a new approach to the validation of first principles-based simulation models applied in process engineering and personnel training. The approach uses both process design data and expert judgements and is based on decomposing the plant's model in accordance with the structure of operative control tasks accomplished by the user. A numerical example, which illustrates the application of the approach to process simulation, is adduced.

INTRODUCTION

Process modeling and simulation (**M&S**) have been just in our sight growing into a powerful and sometimes indispensable tool for solving key process engineering problems where precise representation of plant's interactive dynamic behavior in a wide range of environment changes, operation modes and control actions is necessary. Among such tasks, one should mention process decision-making support, operation mode fitting and adjustment (Thomas, 1999). Training the skills of safe and efficient process operation with the help of computer-based training systems (Dozortsev et al., 1999; Dozortsev and Kneller, 2000) is of especial interest here.

Successful solution of such tasks is a key issue of enterprise's competitive strategy and directly depends on the validity of the process models involved. At the same time, the problem of ensuring model validity subject to the contensive M&S features and the specificity of process engineering tasks is still waiting for being understood and solved. In view of the desired complexity and profundity of representing steady-state and dynamic behavior of real process plants, one could say about the so-called *first-principles simulation models* (**FPSM**), which as compared with the conventional ones used in control tasks demonstrate the following important features:

- They are derived against physical and chemical background of process technologies
- They are complete pertaining to control impacts and measurable process variables and operate in a wide range of variables changing
- They operate in the interactive (simulation) mode and allow sudden (either in time or in content) interventions in the process under simulation
- They represent a wide range of standard and special process situations.

These properties of an FPSM determine the specificity of the problem of providing the process model validity in engineering applications. First and foremost, the validity should be ensured over a wide set of steady-state and dynamic operation modes. (It must be noted here that just construction of this set for an FPSM is a nontrivial task as compared with traditional control tasks, where small model's dimension allows easy determination of a limited set of testable operation modes.) Second, the model's vast dimension and a great number of cross-connections predetermine inevitable recursiveness of the parameter adjustment procedure with backspaces to the modeling stage for structure correction (Dozortsev et al., 2001). Third, the nature of the information used for FPSM parameters adjustment drastically differs from the case of parameter estimation in control tasks, where the validity criterion to be minimized is some measure of the offset between model's and plant's outputs. In the FPSM case, actual data samples can be used only for some special (usually steady-state) operation modes, while the primary source of information about plant's behavior is the judgements of experts: process engineers, high-skilled operators, etc. Finally, the objectives of process engineering themselves define the focus on ensuring the validity under 'arbitrary' manipulations with the model rather than on accomplishing the hardly realizable highest accuracy of parameter estimation. Here, the validation accuracy required also becomes the result of expert judgements. So, in M&S practice for USA A-plants the 98%-accuracy is required for critical variables, while the 90%-accuracy is admitted for non-critical ones, both in steady-state modes. For transients, qualitative similarity of plant's and model's behavior is sufficient (American National Standard, 1981).

Against this background, the paper now states the problem of FPSM validation and proposes a formal procedure of its solution. The procedure is illustrated with a numerical example from operator training simulator development for oil refining industry.

1. PROBLRM SPECIFICITY

The lack of works in FPSM validity for sophisticated processes in author' opinion is accounted for apparent baffling complexity of the task. In fact, the study of dynamic behavior of a nonlinear differential and finite-difference equation system with the dimension of several thousands variables under arbitrarily changing input actions seems at first sight to be unrealistic (let alone the adjustment of its behavior to desired patterns). Nevertheless, the actual difficulty seems to consist just in the understanding of the model's 'validity', because only such understanding can imply the criteria of modeling accuracy and the ways of their accomplishing.

The key question can be formulated as follows: *what an FPSM should be adequate to*? In parameter estimation tasks for control purposes the model must approximate (subject to some measure) the plant's behavior in some identification experiment, the number of both the variables to be approximated and the modes to be validated being rather small. Such approach is inapplicable to FPSM, because the experiments for the most of the modes under simulation are impossible on principle, while the number of variables to be approximated is exceedingly big. Besides these technical difficulties there is also a problem of the repeatability of process' dynamic (as well as, of course, steady-state) operation modes, the proximity to which the FPSM should ensure. The problem can be formulated as follows: *is it possible to provide the repeatability of plant's behavior in real environment under identical control impacts and external disturbances*? (It should be noted here that even in case of expert judgements about plant's behavior, the problem of the repeatability of some operation mode 'imagined' by an expert still exists.)

Even a non-specialist in process control would come to a negative answer to this question. Here, while some speculations about the identity control impacts may somehow seem pertinent (only theoretically, because it is impossible to provide in practice the repeatability of control impacts implemented with real control devices), nothing can be said about the identity of uncontrollable (and often even unobservable) disturbances. That is why no operator would undertake providing the repeatability of a steady-state (let alone a dynamic) operation mode with acceptable accuracy and over a wide range of observable variables.

Besides the lack of the 'object' to be approximated the identification approach seems incorrect also from the problem setting viewpoint. The destination of FPSM is an interactive representation of process dynamics under wide range of control impacts and disturbances. Here the observable part of process output variables should comply with the user's (operator's, process or control engineer's, maybe also the expert's) current vision of the real process plant. According to the ideas of modern cognitive engineering about the decision-making process nature, a user having his/her own 'conceptual' process model (i.e. more or less comprehensive image of the process under control) while accomplishing each particular task uses some 'cut' of the conceptual model: a so-called *operative process model*. Its composition and properties depend on the task's features (Dozortsev et al., 2001).

This general statement can be reduced in the context of FPSM validation to the following fundamental features of user's process model perception in process engineering tasks:

- ➢ The user 'sees' the process through a family of operative models connected with some fragment of the whole process. This allows him/her while solving a specific problem to disengage from the whole sequence of cause-effect relations (i.e. plant's input-output dependencies) and to focus on local relations between the variables without extending the effect of current operative model's outputs on the behavior of other models. Therefore, FPSM must also ensure this separability property.
- ➢ Each operative model is described by relatively few critical operating variables, which are usually controlled automatically. Owing to this, steady-state and dynamic process modes can be simulated subject to these variables with sufficient accuracy, because the automatic control compensates the disturbances.
- ➢ Most of other variables are considered by an operator as uncritical operating variables. Their values are needed with much less accuracy than for the critical ones. The number of such variables in each operative model is also rather small.
- ➢ Finally, most of the signals are purely instrumental, their values being unimportant for the user within the current operative model.

These features of operative models, which are applied during decision-making in process control, engineering and training, can underlie the approach to FPSM validation.

2. FPSM VALIDATION AT THE MODELING STAGE

Usually, FPSM development starts with the design of a FPSM family, which, in general, can be defined as the following pair of mappings:

$$\boldsymbol{P}: (U, X, \Lambda^p) \to C(X) \quad \text{(set mapping)}$$
$$\boldsymbol{N}: X \to Y \quad \text{(observation mapping)}$$

Here:

- U is the space of controls functions at the interval $T = [0, \tau]$ with the values in some subspace of $\mathbf{R}^m$, whose dimension may be less than m (if not all controls are included in the FPSM)
- X is the admitted region of state variables (in the FPSM family it has a finite dimension q even if distributed parameter processes are simulated)
- Y is the admitted region of observable variables with the finite dimension n, which, in general, is less than the dimension of the vector of plant's observable variables and (this is especially important in the context of this paper) is considerably less than q
- Λ^p is the subset of feasible parameters of the FPSM family, p is its dimension
- $C(X)$ is the space of time-varying functions (without any smoothness requirements) with the values in the region X.

Given an initial condition x^0 and a control function $u(\cdot) \in U$, we introduce a set of mappings:

$$\{\boldsymbol{P}_\lambda \mid x(\cdot) = \boldsymbol{P}(u(\cdot), x^0, \lambda), \lambda \in \Lambda^p\}.$$

Thus, by definition we have:

$$x(t) = \boldsymbol{P}(u(\cdot), x^0, \lambda)(t) \text{ for any } t \in T, \qquad (1)$$
$$\boldsymbol{P}(u(\cdot), x^0, \lambda)(0) = x^0, \qquad (2)$$
$$y(t) = \boldsymbol{N}(x(t)). \qquad (3)$$

Here the choice of some parameter vector $\lambda_0 \in \Lambda^p$ chooses an FPSM from the family. And it is just the choice that is the matter of difficulties at the model tuning stage.

If we neglect time delays then the mapping $\boldsymbol{P}$ delivers the solution to the system of differential equations:

$$\frac{dx}{dt} = \psi(x, u, \lambda), \quad t \in T, \qquad (1')$$
$$x(0) = x^0, \qquad (2')$$
$$y(t) = N(x(t)). \qquad (3')$$

Usually, some of the observable variables y_i are at the same time the state variables x_i: $y_i(t) = x_i(t)$. For simplicity we assume that $i = 1, \ldots, n$.

The system (1′)-(3′) can be extended by differentiating (3′) subject to (1′) and adding the resultant equations to the system:

$$\frac{dy}{dt} = \frac{\partial N}{\partial x}\psi(x, u, \lambda),$$
$$\frac{d^2y}{dt^2} = \frac{\partial}{\partial x}\left(\frac{\partial N}{\partial x}\psi(x, u, \lambda)\right)\psi(x, u, \lambda),$$
$$\ldots \qquad (4)$$

Here $y_1, \ldots, y_n$ should be substituted in the second members of (4) instead of $x_1, \ldots, x_n$.

It is well known from the observability theory that such sequence of derivations can result that at some step k the second members of (4) would not contain state variables. Then the system (1′)-(3′) can be replaced with the equations:

$$\frac{d^k y}{dt^k} = \Theta_k(y, y', \ldots, y^{(k-1)}, \lambda), \quad t \in T, \qquad (5)$$
$$y(0) = y^0, \ \frac{dy}{dt}(0) = y'^0, \ldots, \frac{d^{k-1}y}{dt^{k-1}}(0) = y^{(k-1)^0}. \qquad (6)$$

If no combination of derivations allows excluding state variables from the system (1′)-(3′) it means that the model's state space is redundant and, generally, can be reduced.

This fact can be recognized as an argument additional to the aforementioned statement about the reduction of FPSM validity requirements to the conditions imposed on the dynamics of the 'narrow' set of observable critical variables. Moreover, contensive analysis of continuous process models indicates rather small order k of the system (5)-(6), because usually the second members of (1′) have rather simple algebraic structure while the nonlinearities arise owing to the boundedness of the feasible sets U, X, Y.

2.1 Model parameter estimation

The values of parameters λ, which single out a specific FPSM from the whole FPSM family can be derived by analyzing the data about the plant dynamics. Assume, for some operation mode $y(\cdot)$ (steady-state, for simplicity) there are some estimates y_j^0 of the observable variables and

$$\left|y_j^0 - y_j(t)\right| < \varepsilon_j, \qquad (7)$$

where ε_j are specified by an expert. By putting the first member of (1′) to zero we obtain the following system of the order q with parameters λ and q–n+p unknown nonobservable state variables:

$$\psi(x, u, \lambda) = 0, \quad t \in T, \qquad (8)$$
$$y(t) = N(x(t)). \qquad (9)$$

Each new piece of information about the processes $(u(\cdot), y(\cdot))$ adds new equations like (8) and (9). At some stage, the number of equations becomes sufficient for excluding unknown state variables and deriving with the help of (7) the following inequalities with respect to λ:

$$\Omega(\lambda,\varepsilon) \geq 0. \qquad (10)$$

The inequalities (10) specify the membership of λ in some set D whose boundary depends on the process $(u(\cdot), y(\cdot))$ and the estimates ε. There are some parameters, on which theses inequalities do not impose any constraints at all. In order to estimate such parameters dynamic behavior data is to be employed, such as the speeds of level or temperature changes, etc.
The information about plant's steady-state and dynamic operation modes can be withdrawn from the following sources:

A. Design information

This data allows direct calculation of some necessary parameters. Usually it includes the process equipment geometry data, material properties, etc.

B. Plant operation data (unit startup, normal operation, shutdown, etc.), heat and mass balances

This data provides some components of the vector y as the known constants y_j^0 or (more seldom) as the values of periodic functions $y_j(t_k)$ in some moments $y_{jk} = y_j^0(t_k)$.

C. Digitized sensor data $y_j^0(t_k)$ stored automatically in databases

This source provides the richest information on process dynamics. But here the above considerations about process repeatability are to be accounted for. The boundaries (7) of possible deviations should be provided by the experts for $t = t_k$.

D. Historic trends

The same as in ***C*** above with the only exception that graphical information itself is not precise, hence only interval estimate $y_j^{\min}(t_k) < y_j < y_j^{\max}(t_k)$.

E. Lab analyses data

This is valuable information about some state variables, which are not measured on-line. But the synchronization of this information with other process data is to be ensured.

F. Expert judgements

It is the final argument in FPSM development and validation, because the proceed directly from potential users of process engineering and training systems. But the expert requirements are not always formulated with enough clearness. The FPSM developer's task here is to take into account the aforementioned features of process operator activities and to formulate the questions to expert in the way that allows getting the necessary information. The second feature of this type of information is that an expert sometimes formulates his/her requirements and considerations not at once (at the early stage of FPSM development that is the most desirable) but sometimes with a considerable time lag, e.g. while studying a beta-version of the FPSM.

2.2 Model decomposing

The form of the inequalities (10) looks too general for making any conclusions about the approaches to model parameterization. But the structure of real process units and the features of control, engineering and training tasks allow splitting the units into several relatively independent functional blocks (process devices and groups of devices), which effect each other via process flows. As far as the unit's control system holds the variables (flow rates, temperatures, etc.) of the most of these flows as well as most of drum pressures and levels at setpoints, the unknown parameters of process device interaction are flow compositions. After these are identified (with the help of mass balances and other process information) the system disintegrates to several subsystems with nonoverlapping groups of unknown parameters. This facilitates considerably further parameter estimation.

2.3 The recurrence of parameter estimation process

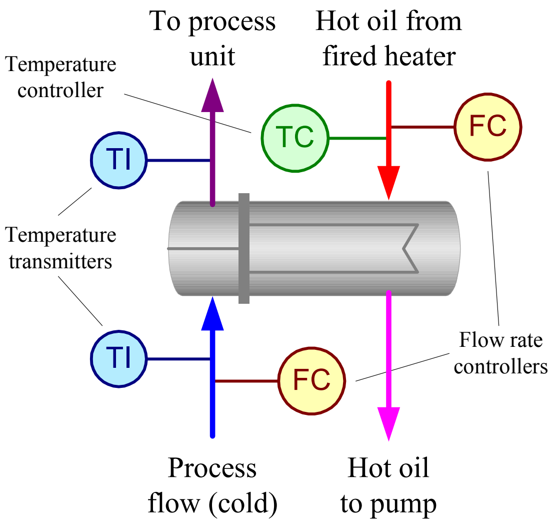
As it was said above, an expert little by little improves the requirements to FPSM. Each new requirement increases the number of the equalities (10). Thus we deal with a sequence of nested sets:

$$D_1 \supseteq D_2 \supseteq \ldots \supseteq D_n .$$

At the n-th step, an arbitrary parameter vector $\lambda_0 \in D_n$ matches for FPSM individualization. But any new additional requirement narrows the feasible set. Our hope that the sequence $D_1 \supseteq D_2 \supseteq \ldots \supseteq D_n \supseteq \ldots$ would have a nonempty intersection (i.e. that we can somehow satisfy all expert's reasonable requirements) is based just on the first-principles nature of the model development. But if at some i-th step the set D_i turns out to be empty, this means that either some expert's requirements are inconsistent, or some basic effects, which describe plant's behavior in some situations, were not accounted for at earlier modeling stages.
Taking into account such additional dependencies increases the dimension p of the parameter space Λ^p. At the same time, it requires the modeler to backspace for revising the set of basic concepts (laws) underlying the modeling strategy and constructing a new FPSM family. Therefore, the nature of the model adjustment stage is iterative. It should be noted that sufficient modeling profundity can be hardly predicted, because an expert may formulate some requirements later. Such situation presumes the need in special (more or less formalized) procedures of withdrawing such information from experts both at the early design stages and in process of FPSM development and tuning.

NUMERICAL EXAMPLE

A refinery hot oil belt is intended for preheating several process flows to different units. It includes a fired heater that supplies hot oil to several relatively small heat exchangers. The hot oil is pumped back to the heater and thus circulates in a closed loop.



A heat exchanger from refinery's hot oil belt

Consider one of such small exchangers (see Figure). Usually both process and hot oil flow rates as well as the temperature of the hot flow at the heater's outlet are maintained by controllers and, hence, known. The temperatures of the process flow at the exchanger's inlet and outlet are measured, while the hot oil temperature at the exchanger's outlet is not known precisely.
The simple model of an exchanger consists of the following 2 equations:

$$\frac{dT_{\mathrm{CO}}}{dt}=\alpha_1\left[\mu\left(\frac{T_{\mathrm{HI}}+T_{\mathrm{HO}}}{2}-\frac{T_{\mathrm{CI}}+T_{\mathrm{CO}}}{2}\right)-c_{\mathrm{C}}m_{\mathrm{C}}(T_{\mathrm{CO}}-T_{\mathrm{CI}})\right], \quad (11)$$

$$\frac{dT_{\mathrm{HO}}}{dt}=\alpha_2\left[c_{\mathrm{H}}m_{\mathrm{H}}(T_{\mathrm{HI}}-T_{\mathrm{HO}})-\mu\left(\frac{T_{\mathrm{HI}}+T_{\mathrm{HO}}}{2}-\frac{T_{\mathrm{CI}}+T_{\mathrm{CO}}}{2}\right)\right], \quad (12)$$

where
T_{CO}, T_{CI}, T_{HO}, T_{HI}, are exchanger outlet and inlet temperatures of cold process flow and hot oil respectively,
m_{C}, c_{C}, m_{H}, c_{H}, are flow rates and heat capacities of process and hot oil flows respectively,
μ, α_1, α_2 are the design parameters, which are not known precisely and are to be recovered from available data.
Assume for simplicity: $m_{\mathrm{C}} = c_{\mathrm{C}} = m_{\mathrm{H}} = c_{\mathrm{H}} = 1$,
$\alpha_1, = \alpha_2 = \alpha$, $T_{\mathrm{CI}} = 100$, $T_{\mathrm{HI}} = 400$.

Differentiation of (11) subject to t and substitution of (12) to (11) yields the following equation in the form (5):

$$T''_{\mathrm{CO}}=-\alpha(\mu+2)T'_{\mathrm{CO}}-\alpha^2[(\mu+1)T_{\mathrm{CO}}-400\mu-100], \quad (13)$$

For steady-state operation both derivatives in (13) are zero, hence $(\mu+1)T_{\mathrm{CO}}=400\mu+100$. The unknown parameter μ delivers a solution to this linear algebraic equation for some known steady-state value T_{CO}* of T_{CO}. For example, if T_{CO}* = 200, then μ = 2.
The parameter α responsible for dynamic behavior ('heating-up speed') of the exchanger can be recovered from dynamic operation data. For the known μ, the solution of (13) looks like:

$$T_{\mathrm{CO}}=c_1(\alpha)\exp[\lambda_1(\alpha)t]+c_2(\alpha)\exp[\lambda_2(\alpha)t],$$

where $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ are the eigenvalues, which depend on T_{CI} and all other measured and controlled variables. For calculating α, the initial conditions are needed as well as some dynamic operation data provided from a historic trend or by an expert, e.g. 'in case of step-wise T_{CI} increase from 100 to 150 T_{CO} increases from 200 to 240 in 2 minutes'. Together with 2 initial conditions (say $T_{\mathrm{CO}}(0) = 100$ and $T'_{\mathrm{CO}}(0) = 0$) this yields 3 transcendent equations, from which α, c_1 and c_2 can be recovered.

REFERENCES

Thomas, Ph. 1999. "Simulation of Industrial Processes for Control Engineers". Butterworth-Heinemann, 1999.
Dozortsev, V.M. et. al. 1999. "The Cost Effectiveness Criterion-Based Approach to Development, Implementation and Support of Computer-Based Training Simulators for Continuous Process Operators" In *Proceedings of the 4th Euromedia Conference "Euromedia'99"* (Munich, Germany, April 26-28, 1999), SCS, Delft, Netherlands: 197-201.
Dozortsev, V.M. and D.V.Kneller. 2000. "KTK-M: A Computer-Based Training Complex for Process Operators" In *Proceedings of the 5th Euromedia Conference "Euromedia'2000"* (Antwerp, Belgium, May 8-10, 2000), SCS, Delft, Netherlands: 171-174.
Dozortsev, V.M. et. al. 2001. "Developing a Process Simulation Model: A Scientific Challenge and a Business Process" In *Proceedings of the 6th Euromedia Conference "Euromedia'2001"* (Valencia, Spain, April 18-20, 2001), SCS, Delft, Netherlands: 269-273.
American National Standard 1981. "Nuclear Power Plant Simulators for Use in Operator Training." *ANSI/ANS – 35*:13.

FORMATION OF EFFECTIVE FAULT DIAGNOSIS SKILLS BY MEANS OF TECHNICAL SYSTEM'S SIMULATION

Prof., Dr. Victor M. Dozortsev
Institute of Control Sciences RAS,
65, Profsoyuznaya St., 117997, Moscow, Russia
E-mail contact: victor@petrocom-jv.ru

KEYWORDS
CBT, fault diagnosis skill, technical system simulation model, computer-based decision-making support system, active training of diagnosis strategies.

ABSTRACT

The paper describes a new motivation-oriented computer-based decision-making support system for formation and strengthening of technical system operators' effective skills of fault diagnosis. The system's information-functional structure is described. A new class of reference strategies of optimal information search in process of accomplishing diagnosis task is presented. Practical results of operator training and future developments of the approach proposed are discussed.

1. INTRODUCTION

Owing to practical importance of fault diagnosis in technical systems, cognitive mechanisms of operator's diagnostic decision-making attract the attention of scientific community for a rather long while (Rasmussen and Jensen, 1974; Marshall et al., 1981; Bainbridge, 1984; Hukki and Norros, 1993; Patrick, 1993). The research undertaken in this area has two basic objectives: first, to reveal optimal strategies of the failure cause search and thus to give a guideline for improving operator's diagnosis skills, and, second, to employ human search engines for automated diagnosis system design.

W.Rouse has segregated diagnosis strategies into *prescriptive* and *descriptive* (Rouse, 1983). The first type of strategy means that an operator acts according to some prescribed formal procedure of alternative search, as e.g. in a well-known 'Half-Split Strategy' (Goldbeck et al 1957; Kletsky, 1960) or in the strategies based on the probability of failure or on average operation time of system components (Bond and Rigney 1966). The second type of strategies is based on the search criteria, which are the most frequently chosen by human operators: effort minimization, risk minimization, etc. At the same time, Rasmussen and Jensen has established in their classical work (Rasmussen and Jensen, 1974) that the nature of mental procedures chosen by an operator is determined by the limitations of the short-term memory and reasoning volume, which dramatically reduce the opportunities of applying effective prescriptive strategies.

The attempts of using the models of diagnostical decision-making for improving operator's strategies are not numerous. The difficulties of full-scale training of effective diagnosis strategies are apparently connected with necessity of solving the following three problems:

- ➢ providing rich and available fault symptomatology, which on the one hand should not require the operation of real technical plants in potentially dangerous modes and on the other hand should go beyond the expert judgements, which are sometimes disputable and hardly augmentable
- ➢ guaranteeing positive transfer of the acquired skills both by ensuring the adequacy of the symptoms offered to operators and by representing skill formation mechanisms in the training process
- ➢ motivating operators to improve their search strategies.

An innovative tool for answering these challenges was given by for technical system (TS) simulation techniques[1]. At present, dynamic simulation models of the TSs become a source of practically unlimited and easily augmented fault symptomatology, while the simulation accuracy attained allows excluding the risk of acquiring 'false skills'. Such situation has to call into being a novel approach to forming and improving diagnosis skills (Dozortsev, 1999). In its context, a cognitive model for diagnosis skills formation was proposed and analyzed. The model was based on the mechanism of hypotheses generation and testing and included explicitly the functional elements responsible for search strategy selection (Dozortsev, 2001). Diagnosis skills training itself is implemented with the help of motivation-oriented computer game, where the operator's objective is the minimum of information

[1] The same situation takes place in automated diagnosis system design, where TS simulation model, which operates simultaneously with diagnosis system is used both for detecting process offsets and for checking the verity of the hypotheses about deviation causes (Jaimieson, 1991).

required and, hence, the minimum of time spent on troubleshooting.

This paper gives a development of the aforementioned studies and provides procedural and mathematical underpinnings of the computer-based decision-making support system (**CDSS**) including:

- information-functional structure of CDSS
- formal description and comparative analysis of selection criteria of information inquiry sequence
- the analysis of system application results and its further development.

2. INFORMATION-FUNCTIONAL STRUCTURE OF THE CDSS

The mechanism of diagnosis skills formation is described in (Dozortsev, 2001). It is a procedure of failure's cause detection as a goal-seeking inquiries of the symptoms observed and consecutive constriction of the feasible causes set down to a single true one. Here, the trainee makes each information inquiry based on both the available current-state hypothesis about the true cause of the process offset and the search strategy employed.

Formally, this procedure can be determined on a so-called incidence matrix (see Figure 1).

Symptoms *Causes*	Reflux flow increases	Reflux flow decreases	Bottom level increases	Bottom level decreases	…
Feed flow increases	*1*	*0*	*0*	*1*	
Feed temperature increases	*0*	*0*	*0*	*1*	
Feed pump failure	*0*	*1*	*1*	*0*	
…					

Figure 1. Incidence matrix

The rows of this matrix correspond to possible offset causes, while the columns – to different symptoms, i.e. to various deviations of TS's variables from their normal values. Several columns-symptoms can correspond to each variable. The cell in the intersection of *i*-th row and *j*-th column contains 1, if the *j*-th symptom ensues from the *i*-th cause and 0 if it does not. When some variable has now deviation in some cause, such situation is considered as an independent measurement result; therefore, a single (and only single) symptom for each cause equals 1 in each process variable.

One can speak about the following 3 sources of the incidence matrix (see Figure 2):

- **The TS itself**. This symptomatology is authentic; the opportunities of getting the symptoms are utterly limited: fortunately, not all TS malfunctions take place in practice and nobody will activate them just for accumulating the desired symptoms. At the same time, the opportunity of including such observations as a row in the incidence matrix seems a relevant tool for retaining the information about the malfunctions happened earlier in trainees' 'memory', even if those malfunctions had taken place long ago and maybe in some other TS.
- **An Expert.** This cost-efficient way of gaining information allows including malfunctions, which have never happened in reality. But it may be rather difficult for a human expert to create a diverse and non-trivial symptomatology for a wide circle of malfunctions.
- **A simulation model of TS.** If the model's validity is guaranteed, it is, certainly, the best source, because it allows augmenting the symptomatology practically with no limit as far as CDSS is utilized and developed. Moreover, the availability of the model enables turning from *troubleshooting by final symptoms* to *troubleshooting by situation development dynamics* and further to *troubleshooting with possible situation correction* (see Section 5).

In the ***DIAGNOST*** system developed by the author (Dozortsev, 2001), the formation of a certain diagnostical game is reduced to extracting from the incidence matrix a submatrix of causes and symptoms, proposing some cause i^* and making the 'initial throw-in', i.e. selecting some process variable j_0, whose value (symptom) $a(i^*, j_0)$ is reported to the trainee. The trainee's aim (see Figure 3) is to select the causes, which do not conflict with the symptoms accumulated (at the first step – with the initial throw-in), to form a hypothesis about possible failure cause (e.g., a hypothesis, which includes the causes i_1 and i_2) and to formulate the next (*j*-th) information inquiry to the system.

The CDSS offers the trainee the following 3 types of estimates:

- **An estimate of the hypothesis selected**. If a hypothesis contains only one cause the probability of its occurrence can be estimated. For complex hypotheses, context-based estimates are used (see Section 5).
- **An estimate of a current information inquiry.** The estimates of the trainee's move subject to various optimality criteria are used in ***DIAGNOST*** (see 3.1-3 in Section 3 below)
- **An overall estimate of the game.** The game time (overall cost of information inquiries and hypotheses testing) is calculated and compared

with the 'optimal' search time and with the maximum game time (when all variables are inquired, see 3.4).

3. OPERATOR'S DIAGNOSIS STRATEGIES AND THE SELECTION CRITERIA FOR THE INFORMATION INQUIRY SEQUENCE

In view of diagnosis skills optimization, an estimate of the trainee's strategy subject to possible optimal search criteria is a critical issue. Formal description of several such criteria is given below.

3.1 Minimax strategy

The 'Half-Split Strategy' mentioned in the Introduction presumes the following: At each step of the search, the operator selects an information inquiry (i.e. checks the value of some TS's measurand), the answer to which (at all possible events) would split the uncertainty related to the feasible set of causes into the most equal parts in view of his/her possible preferences (or subjective realization probabilities) of different causes.
Denote $\{H_k\}$, k=1, 2, …, N the set of causes, which do not conflict with current symptomatology, while q_k will stand for the probability of the cause H_k, Σq_k=1. For each i-th cause, each j-th process variable has the elements
$a(i, j_1)$, $a(i, j_2)$, …, $a(i, j_J)$ in the incidence matrix, where J is the total number of symptoms in each variable, and

$$\sum_{k=1}^{J} a(i, j_k) = 1 .$$

The information inquiry choice law subject to 'half-split' criterion runs as follows:

$$j* = \arg\min_j \max_{k\in 1,2,\ldots,J} \sum_{i=1}^{N} a(i, j_k)\, q_k. \qquad (1)$$

It should be noted here that this is a one-step strategy in the sense that the uncertainty is at most reduced at the current step of the search without recalculating the whole tree of future inquiries and corresponding subsets of causes. Some authors suppose erroneously that such 'one-step' search under arbitrary relation between possible causes and the symptoms generated by these causes guarantees that a single cause can be found in a finite number of steps, which equals $\log_J N$, where N is the total number of causes considered (of course, if the true cause is among them). It can be easily proved that it is not always like this (in particular, even for $N > J + 3$ the one-step strategy is not always globally optimal).

It is clear that the minimax strategy (1) is a considerable bondage for operator's memory and attention, because he/she should keep in mind: (i) the set of causes ordered by preferences, (ii) the set of symptoms, and (iii) prediction results. Furthermore, this strategy in view of its cognitive structure 'depersonalizes' different causes and therefore mismatches the operator's diagnostic decision-making paradigm, which is based on hypotheses generation and testing mechanism. A class of strategies directly based on such mechanism is introduced and analyzed below.

3.2 Checking the most probable cause

This strategy presumes that at each search step the operator: (i) selects some cause H_i from the set $\{H_k\}$, k = 1, 2, …, N, which do not conflict with available symptoms and (ii) determines the validity of H_i using information inquiries of TS variables (the case of several causes, i.e. making up compound hypotheses, will be considered in 3.3). It should be noted that operator's preference of a certain cause is rather subjective; in particular it can rest upon the aforementioned subjective probabilities.
We now introduce a normalized vector $\boldsymbol{r}$ of probability estimates for the occurrence of the residuary causes (i.e. the causes that were not selected by the operator at the current step):

$$\sum_{m\neq i} r_m = 1 .$$

At each search step, the operator makes some information inquiry, the result of which segregates the selected cause from the rest ones more effectively than in case of any other inquiry. For each j-th variable we define (on the whole of non-selected causes) the 'probability of not detecting' the same symptom as in the cause under test:

$$\mathbf{P}(i, j_k) = 1 - \sum_{m\neq i} a(m, j_k) r_m . \qquad (2)$$

As far as k is determined uniquely for the couple (i, j), it is then clear that

$$\mathbf{P}(i, j_k) = \mathbf{P}(i, j) . \qquad (3)$$

The Appendix shows that in view of the time T_j necessary for variables inquiry (testing), the best strategy subject to the criterion of minimum mean time of checking the validity of i-th cause consists in the maximization (with time correction) of the probability of not detecting the same symptom in the set of non-selected causes:

$$j* = \arg\max_j \frac{\mathbf{P}(i, j)}{T_j} . \qquad (4)$$

In particular, if all non-selected causes are equiprobable: $r_m = 1/(N–1)$ and, hence,

$$\mathbf{P}(i,j) = 1 - \frac{1}{N-1}\sum_{m \neq i} a(m, j_k),$$

and if all test times are also equal (T_i = const), then the best inquiry is determined due to the following rule:

$$j^* = \arg\max_j \mathbf{P}(i,j) = \arg\min_j \sum_{m \neq i} a(m, j_k). \quad (5)$$

Another extreme case takes place, when direct cause tests are included into the number of variables. Generally, the time spent on such tests exceeds considerably the time needed for measuring the status of conventional variables because of the necessity picking up the process data from remote instrumentation. In such situation the incidence matrix is supplemented (assume from the left) with an identity submatrix, which consists of cause-columns $a(l, l) = 1$, $a(l, j) = 0$, $l \neq j$, $j = 1, 2, \ldots, N$. For this submatrix: $\mathbf{P}(i, j) = 1$, i.e. the probability of not detecting another cause with the same symptom is maximal, though the test time is longer, and the parameter, which is less 'probable' but cheaper in testing, can win the inquiry selection.

3.3 Checking compound causes

Strictly speaking, the diagnosis mechanism presumes the an operator works with hypotheses that combine several causes rather than with separate causes. Here, in the simplest case the operator rejects (for the time being) splitting several causes combined in the hypothesis and discards the variables whose symptoms would not agree completely with all causes included into the hypothesis. After the field of variables is reduced like that, the task would not change significantly, and the aforementioned rules for information inquiry selection survive. If the hypothesis is denied then all its component causes are removed from the search process, otherwise the operator considers the hypothesis as a set of causes and continues the search.

At the same time, the strategies, which operate the hypotheses in the entirety of their symptomatology, are not in the least excluded. The task is now reduced to checking the validity of conjunctive-disjunctive statement, in which both separate symptoms for the variables with completely coincident symptomatology and the disjunctive groups of symptoms that can be met in the rest of variables appear conjunctively.

Assume the j-th variable appears in some hypothesis disjunctively and the set of corresponding symptoms met in this hypothesis $\{j_k\} = J^D \subset \{j_1, j_2, \ldots, j_J\}$. Then the probability of detecting none of the symptoms from the subset J^D subject to all causes that do not belong to the i-th hypothesis equals to

$$\mathbf{P}(i, J^D) = \mathbf{P}(i,j) = 1 - \sum_{m \neq i} \sum_{j_k \in J^D} a(m, j_k) r_m. \quad (6)$$

Each disjunctive component will appear in the optimal hypothesis testing rule (4) with the probability (6). The rule for optimal checking the validity of disjunctive elements is adduced in the Appendix as well as the mean time of the test.

The strategy (4) based on the selection of several causes by the operator describes the mechanism of hypotheses generation and testing quite sufficiently. Here the hypothesis selection itself is based on their matter, in contrast to the strategy (1). The following fact can be demonstrated: if on the minimax-optimal splitting of the cause set any its subset is assumed as the operator's hypothesis, then the strategy (4) yields the solution j^*, which coincides with (1). In other words, *the more successfully the trainee aggregates the causes into a hypothesis, the closer is the 'rational' solution (4) to the 'optimal' one (1).*

3.4 An overall estimate of the game

When the target cause is detected the quality of the game can be evaluated based on the total time t_{GAME} spent by the trainee on symptom inquiries and causes testing. The time t_{OPT} necessary for the game optimal subject to the criterion (1) can serve as a reference point here. On the other part, if all symptoms are inquired, the cause can be identified uniquely; the time spent on this will obviously be:

$$t_{\mathrm{MAX}} = \sum_{j=1}^{M} T_j,$$

where M is the total number of variables in the game (except the initial throw-in).

In general, the case of $t_{\mathrm{GAME}} < t_{\mathrm{OPT}}$ is possible if a one-step strategy is non-optimal and the trainee has 'guessed right' the best sequence of inquiries. The case $t_{\mathrm{GAME}} > t_{\mathrm{MAX}}$ is also possible if in addition to variable inquiries the trainee also undertook expensive tests of incorrect causes. But the bases case presumes the validity of the two-sided inequality:

$$t_{\mathrm{OPT}} \leq t_{\mathrm{GAME}} \leq t_{\mathrm{MAX}},$$

and it is just the position of the game time t_{GAME} on the range $[t_{\mathrm{OPT}}, t_{\mathrm{MAX}}]$ that allows evaluating the quality of the player's strategy.

4. MOTIVATIONAL ASPECT OF DIAGNOSIS SKILLS TRAINING

In the author's opinion, one cannot speak about positive transfer of diagnosis skills if the motivational structure of operator's activities during the training does not reflect his/her motivation in accomplishing actual search tasks. This motivation is determined by a specific set of stressful factors experienced by the operator in his/her practical activities. (It must be emphasized that only ‘internal’ motivation of decision-making process is discussed here, while the ‘external’ motivation, such as economical, social and other considerations lie beyond the scope of this study.)

The following 3 types of TS operator’s stress can be emphasized:

- **Hazardous operation stress**. The stress caused by the operation under dangerous values of some process variables cannot be simulated within a training session because it is first of all connected with emotional experience of danger and its sequences for operator’s life and health.
- **Loss expectation stress.** The feeling of responsibility for the consequences connected with TS malfunctions can be simulated by calculating the overall game cost (e.g. of summable values of t_{GAME}), which depends on the quality of accomplishing diagnosis tasks and is initially set as rather large in order the operator would have enough to lose.
- **Limited diagnosis time stress.** This is the most important issue of stirring up the operator’s activities aimed at just the improvement of diagnosis techniques rather than the motivation of the solution itself. Direct simulation of the real time of diagnosis task accomplishment seems hardly possible because the approach described in this paper presumes that the diagnosis procedure is ‘dismantled’ into separate actions of information inquiring, selection of consistent hypotheses, prediction of possible intervention effects, etc. However, motivational likeness seems to be attainable if *the lack of decision-making time will be replaced by the lack of available information*. The game directive focuses the operator on searching a sequence of information inquiries that provides the quickest solution.

5. APPLICATION EXPERIENCE AND FUTURE WORK

DIAGNOST was tested in laboratory environment. The developers of process simulation models (but not of the model used in the ***DIAGNOST***-based training!) and unfamiliar with CDSS design acted as trainees. Process operators of a vacuum distillation unit at one of the biggest Russian refineries were also trained with the help of ***DIAGNOST***. The trainees poll and the analysis of the training system logs evidences (see Figure 4) that as the number of game increases:

- the process unit becomes more ‘familiar’, the evidence of that is the decrease of the number of trainees’ addresses to process diagrams and textual description (curves 1 and 2 respectively)
- the diagnosis task accomplishment is mastered that can be seen from the decrease of game duration (curve 3)
- finally, and it is the most important, the solution strategy is changed. This becomes apparent from: (i) the number of erroneous cause tests comes practically to naught (curve 4) and (ii) the number of necessary symptom inquiries is reduced (curve 5) towards the minimax-optimal number (curve 7) as compared with the maximum number of possible inquiries (curve 6).

It should be noted that the shape of the learning curve (curve 8), which reflects the overall game time t_{GAME}, complies with the classical ideas of the learning theory and includes along with the initial period of low quality also a sharp improvement, a plateau and the final segment of the developed skill.

As it was expected (Dozortsev, 2001) most operators work with a single cause. The cases of using compound hypotheses, which consist of several causes, are reduced to grouping the causes subject to TS’s ‘topology’ (according to the plot plan of plant’s equipment) or subject to ‘phenomenology’ (malfunction type). These data can be used for tuning the decision-make support system in order to estimating the hypothesis chosen by the trainee, as it was noted in Section 2.

The most promising development lines are connected with:

- further study of real diagnosis, selection of effective strategies and including them into reference patterns used in decision-making support
- switching from a game with final fault symptoms to a game with time evolution of symptoms (Dozortsev, 2001)
- implementing the opportunity of combining fault diagnosis with partial compensation of consequences, i.e. including trainee’s impacts on the simulation model that change dynamic symptomatology of the system (Dozortsev, 2001).

6. CONCLUSION

High-fidelity process simulation models by themselves cannot guarantee the optimization of search strategies in fault diagnosis skills training. Moreover, they can even result in the perseveration of non-effective skills. The new proposed, implemented and tested paradigm of such training consists in simulation model-based representation of the whole variety of fault symptomatology, which is offered to a trainee for searching its original causes within the motivation-oriented game environment of CDSS. The results obtained give hope to non-expensive and non-time-consuming formation, strengthening and polishing of effective diagnosis skills.

REFERENCES

Bainbridge, L. 1984. "Diagnostic Skill in Process Operation." *Proc. Intern. Conf. on Occupational Ergonomics*, Vol. 2. Reviews. May 7-9, Toronto, Canada: 1-10.

Bond, N.A. and J.W. Rigney. 1966. "Bayesian Aspects of Troubleshooting Behavior." *Human Factors,* 3: 377-383.

Dozortsev, V.M. 1999. "Technique and didactics of computer-based training for industrial plant operators." In *Proceedings of the 4th Euromedia Conference "Euromedia'99"* (Munich, 26-28 April, 1999), SCS, Delft, Netherlands: 189-196.

Dozortsev, V.M. 2001. "*Diagnost*: A Software for Developing Efficient Decision-Making Strategies." In *Proceedings of the 6th Euromedia Conference "Euromedia'2001"* (Valencia, 18-20 April, 2001), SCS, Delft, Netherlands: 261-268.

Goldbeck, R.A. et. al. 1957. "Application of the Half-Split Technique to Problem Solving Tasks." *Journal of Experimental Psychological,* 2: 330-338.

Hukki, K. and L.Norros 1993. "Diagnostic Orientation in Control of Disturbance Situations." *Ergonomics,* no.35, 1317-1328.

Jaimieson, J.R. 1991. "Model-Based Reasoning for Industrial Control and Diagnosis" *Proc. 17 Annual Control Confer.,* Purdue Univ. W.Lafayette (IN), USA: 119-127.

Kletsky, E.J. 1960. "An Application of the Information Theory Approach to Failure Diagnosis." *IRE Transactions*, PRQC-9, No.3. 29-43.

Marshall, E.C. et. al. 1981. "Panel Diagnosis Training for Major Hazard Continuous Process Installations." *The Chemical Engineer,* 365: 66-69.

Patrick, J. 1993. "Cognitive Aspects of Fault-finding: Training and Transfer." *Le Travail Humain*. 56: 185-210.

Rasmussen, J. and A. Jensen. 1974. "Mental Procedures in Real-Life Tasks: a Case Study of Electronic Trouble Shooting." *Ergonomics,* no.3: 293-307.

Rouse, W.B. 1983. "Models of Human Problem Solving: Detection, Diagnosis, and Compensation for System Failures." *Automatica*, no.6: 613-625.

APPENDIX

The whole of symptoms of the *i*-th cause selected by the trainee is a conjunctive expression. The mean time of its validation under the sequence of the inquiries of the values of the variables $k = 1, 2, \ldots, M$ equals:

$$T_{\mathrm{con}} = \sum_{k=1}^{M} T_k \prod_{j=1}^{k-1} [1 - \mathbf{P}(i,j)], \qquad (7)$$

where $\mathbf{P}(i, j)$ corresponds to the weighted probability (2). In accordance with (7) interchanging the adjacent inquiries j and j+1 yields the following increment of mean time:

$$T_{\mathrm{con}}(j, j+1) - T_{\mathrm{con}}(j+1, j) = [1-\mathbf{P}(i,1)][1-\mathbf{P}(i,2)]...[1-\mathbf{P}(i, j-1)]$$

$$\times \{T_j + [1-\mathbf{P}(i,j)]T_{j+1} - T_{j+1} - [1-\mathbf{P}(i, j+1)]T_j\}.$$

It is easy to notice that this expression is non-positive if

$$\frac{\mathbf{P}(i,j)}{T_j} \geq \frac{\mathbf{P}(i,j+1)}{T_{j+1}}.$$

Since any inquiry sequence can be obtained by pairwise interchanges of elements and in view of the above inequality the minimum time for checking the conjunctive expression can be derived if at each step the inquiry is chosen that brings maximum to $\mathbf{P}(i,j)/T_j$. In the same way it can be shown that the minimization of the mean time for checking the disjunction, which equals

$$T_{\mathrm{dis}} = \sum_{k=1}^{M} T_k \prod_{j=1}^{k-1} \mathbf{P}(i,j),$$

it is necessary the inequality

$$\frac{1-\mathbf{P}(i,j)}{T_j} \geq \frac{1-\mathbf{P}(i,j+1)}{T_{j+1}}$$

to be true for any 2 adjacent elements of the inquiry sequence. This means that the variable that brings maximum to $[1 - \mathbf{P}(i,j)]/T_j$ should be chosen.

BIOGRAPHY

Victor M. Dozortsev was educated in Russia. He has received his Diploma degree in control science in 1976. In 1987 he got Ph.D. degree (Cand. Sc.) in control engineering. In 1999 he has received his doctorate (Dr. Sc.) after defending the thesis dedicated to design, development and application of operator training simulators. Since 1976 he has been working in the Institute of Control Sciences of Russian Academy of Sciences (Moscow, Russia), presently, as chief of a research group. His research interests include system analysis, mathematical modeling, cognitive engineering, and computer-based training and decision-making support. He also has professorship in several technical universities of Moscow.

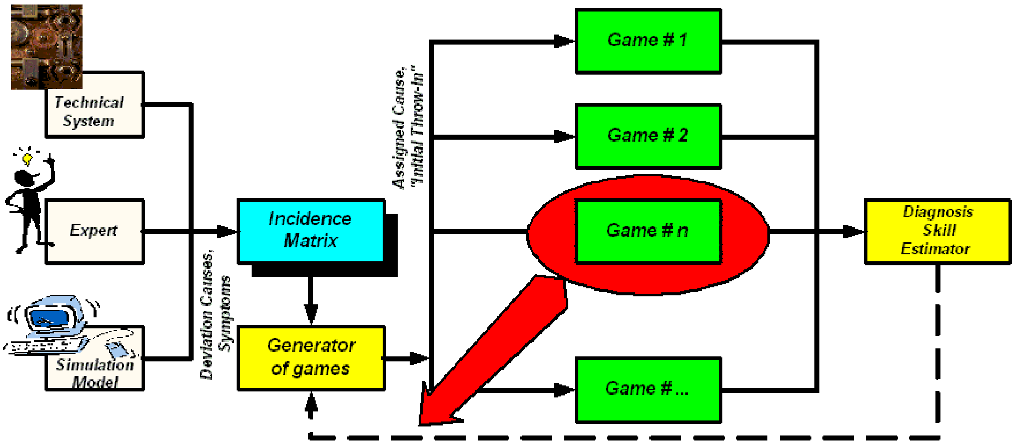


Figure 2. Information structure of ***DIAGNOST*** system

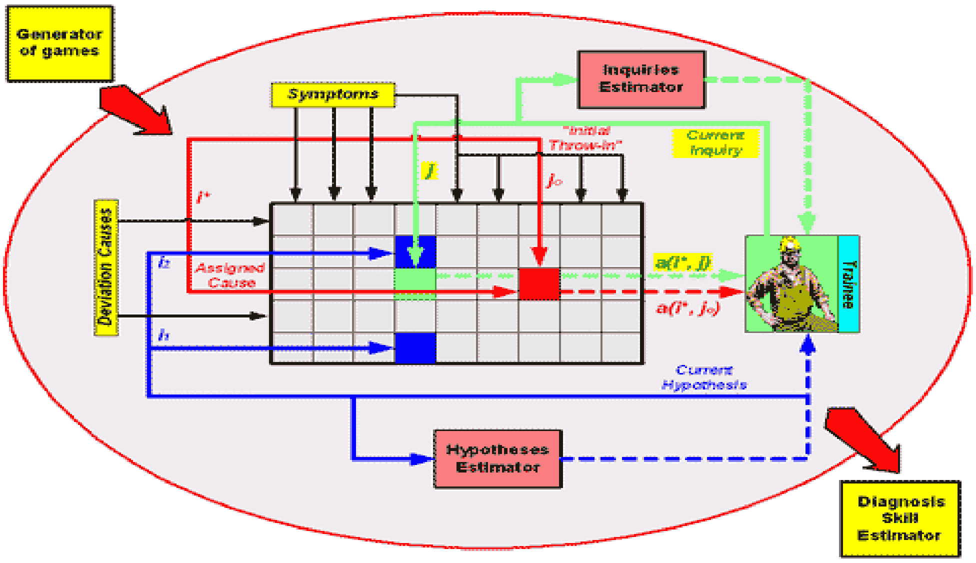


Figure 3. A diagnostical game structure

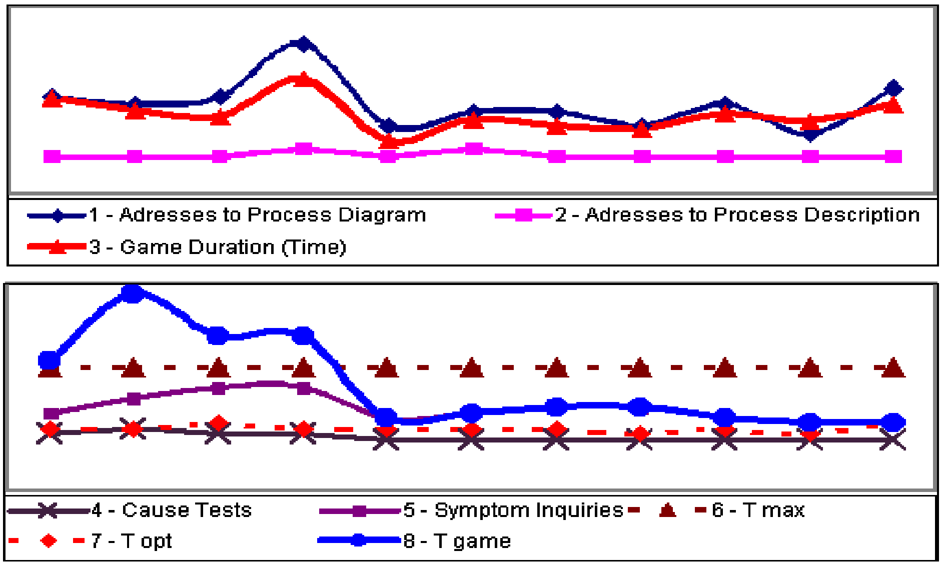


Figure 4. The results of training using ***DIAGNOST***

LATE
PAPERS

RTP-BASED FEEDBACK GLOBAL PROTOCOL INTEGRATED IN MBONE TOOLS

Fernando Boronat Seguí, Juan Carlos Guerri Cebollada, Manuel Esteve Domingo, J. María Murillo Safont
Area de Ingeniería Telemática, Departamento de Comunicaciones
Universidad Politécnica de Valencia - Escuela Politécnica Superior de Gandia
Ctra. Nazaret-Oliva S/N, 46730 Grao de Gandía (VALENCIA)
Telf: 96 284 93 41, Fax: 96 284 93 13
E-mail: fboronat@dcom.upv.es, jcguerri@dcom.upv.es, mesteve@dcom.upv.es

KEYWORDS

Multimedia Applications, Synchronisation, RTP, RTCP, multicast.

ABSTRACT

A modification of several Mbone tools to test a new method for multimedia streams synchronisation is presented in this paper. This method called RTP-based Feedback Global Protocol is based in the combination of RTP/RTCP (Real-time Transport Protocol/ Real-time Transport Control Protocol) and a called Feedback Global Synchronisation Protocol, based in Global Time (NTP) and feedback techniques. RTP is used for data distribution and RTCP for the exchange of control information, accordingly with Feedback Global Synchronisation Protocol.

The current implementation is based in RELATE integrated tool, grouping Mbone RTP-based tools such as *vic* and *rat* for video and audio, respectively, and a non RTP based tool, *nte*, for text. Consequently, we have implemented our method between *vic* and *rat*.

Finally, several results were obtained in verifying the behaviour of our tool and the correctness of our synchronisation protocol using our laboratory LAN as the test scenario.

1. INTRODUCTION

In the past, we used Internet only to communicate with other friends and colleagues all around the world. Nowadays, we find ourselves using it replacing the telephone, for teleconferencing and for delivering classes and seminars to remote students. We also think that this is the way that communication will be achieved, in general, for a wide variety of applications in the future thanks to the rapid evolution that telematic industry is suffering these years to high speed networks and systems

From early research experiments in the late 1970s and early 1980s, through to the deployment of the 1980s and 1990s, multimedia has grown as a presence in the Internet. Interactive multimedia, whether for media-on-demand from World Wide Web servers or between users in multimedia conferencing systems, is becoming quite omnipresent.

Every time, new requirements appear in more heterogeneous environments, such as remote teaching, video on demand, Telework or Telemedicine, all of them multimedia services or applications. The new exigencies set new problems out, that we necessarily have to tackle and solve, to render an acceptable service with a minimum quality of service despite the changeable network conditions (delay, error rate, jitter and throughput). These distributed multimedia applications are characterised by their real-time requirements. With word *multimedia* we mean the integration of different types of media (text, images, audio and video). These are real-time applications because all of them have different exigencies on transfer time requirements.

In this kind of applications, transmitted information becomes *time-dependent data*, and both the value and generation and playback instants are very important. Here, the concept of multimedia refers to the process of integration, at the playback instant, of different kinds of media, and, due to the temporal dependency between the different media, it is necessary to coordinate and order in time the different multimedia streams. This task is called *'Multimedia Synchronisation'*.

As network delays are neither deterministic nor constant and the playback speeds of all the devices involved are not perfectly adapted, the synchronisation process needs some protocols and mechanisms to be guaranteed. Moreover, random delays or jitter are added to the multimedia information stream because of the existence of congestion and queues in the network nodes. In (Boronat 2001) we presented an algorithm to correct this random delays and synchronise a group of receivers distributed in the network, based in RTP/RTCP (Schulzrinne et al. 1996), that we called RTP-based Feedback Global protocol, explained later in this paper.

Our main objective has been to implement the RTP-based Feedback Global Protocol (Boronat 2001) and test objectively the smooth running of this protocol in a closed environment, in our laboratory Local Area Network.

We can find lots of distributed multimedia applications in Internet to integrate our protocol in, for example, (Gerri et al. 2000), (Esteve et al. 1999), MBONE tools: vic (McCanne and Jacobson 1995), rat, vat, sdr, nte, MINT, NEVOT, etc., most of them also based in RTP/RTCP (Schulzrinne et al. 1996). We chose Mbone tools (vic, rat, nte) to implement our method in, because the source code is public and we have been able to modify it accordingly with our protocol to test its behaviour.

2. RTP-BASED FEEDBACK - GLOBAL PROTOCOL

This Protocol is based in a previous synchronisation protocol called *Feedback Global* or *Adaptative Multimedia Synchronisation Protocol, with flow control, based in a Global Time end feedback Techniques* (Guerri 2000), and uses NTP for acquiring global time, and RTP (*Real-Time Control Protocol*) and RTCP (*RTP Control Protocol*) (Schulzrinne et al. 1996) to its implementation. Obviously, RTP packets will be used for data distribution and RTCP packets for the exchange of control information.

2.1. Feedback Global Protocol

The Feedback-Global Protocol is described in (Guerri 2000), and its technical innovations are to include new synchronisation techniques for multimedia on demand services over integrated networks using Global Time. It also uses NTP (*Network time Protocol*) (Mills 1992) to establish the system global time. Feedback-Global can be considered as a hybrid synchronisation method between Global Time protocols and Feedback protocols.

Multimedia servers use short messages (*feedback units*), transmitted by receivers to the sender to guarantee fine synchronisation (i.e., lip-synchronisation) between media streams from distributed sources and between different receivers. These feedback messages contain the sequence number of the last played unit and the local time of the instant of its reproduction.

2.2. Network Time Protocol (NTP) (Mills, 1992)

In the implementation of the previous protocol NTP was already chosen as the global time protocol because it has been proposed as the standard time synchronisation protocol to use in the Internet and it provides precision of milliseconds.

RFC 1305 (Mills 1992) presents the formal specification of the NTP, version 3. It is used to synchronise clients and servers in a distributed environment. We can find in this document the architecture, algorithms, entities and protocols used. The current version of NTP is 4.

We use it to synchronise all the hosts in the multimedia communication system.

NTP works well with network and transport protocols, IP and TCP/UDP respectively, the most used in multimedia environments. It is very easy to implement and it can be used by a lot of operating systems and networks.

NTP uses short control messages to monitoring, so its influence is minimum on the system behaviour.

2.3. RTP/RTCP Protocol (Schulzrinne et al. 1996)

This is the Internet transport protocol for real time multimedia streams. Complete specification of it is in (Schulzrinne et al. 1996).

It provides a standard packet format, with a header containing media-specific timestamp data, as well as identifiers, payload type information, number of sequence, etc. (see figure 1). This way, Real-time Transport Protocol (RTP) provides end-to-end transport functionality suitable for Real-time applications, over multicast or unicast networks.

The main function of RTP is to carry the real-time information, i.e. audio and video, over an IP Network, normally carried using UDP, which provides checksumming and multiplexing.

RTP is supplemented by Real-time Transport Control Protocol (RTCP). Together with RTP packets sent to the receiver or multicast group, RTCP packets are sent in a unicast or multicast way too, but with a different port number. This control protocol provides the participants in a multicast group to be able to exchange control information (i.e., sender identifier, information about Quality of Service of the received data, looses, etc.).

The function of the control protocol RTCP we have mainly used is the inter-media synchronisation one and the possibility of defining new special packets (APP packets) for our application. RTP can be malleable to provide the information required by a particular application.

Different RTP sessions are set up for different data types in the same communication session (see figure 2). For example, voice (audio) and video of the same communication are sent in different RTP sessions (Schulzrinne et al. 1996). Sources are identified by the *SSRC* (*Synchronisation Source Identifier*) field in RTP packet headers, which is guaranteed to be unique in each session.

However, to make the synchronisation possible, a globally unique identifier is necessary, all sessions inclusive. This identifier is called *CNAME* (*Canonical Name*) and it is obtained from the user name and the host name. RTCP packets show the correspondence between a source *SSRC* and the *CNAME.* So, a receiver can group different RTP sources by their *CNAME* in only one logical entity, that represents a session of a particular participant.

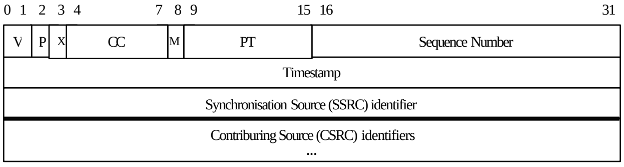

0 1 2 3 4 7 8 9 15 16 31
V
P
X
CC
M
PT
Sequence Number
Timestamp
Synchronisation Source (SSRC) identifier
Contributing Source (CSRC) identifiers
...

Figure 1: RTP Header

Timestamps contained in RTP header are useful to smooth out the effects of network jitter. The timestamp is media specific and different reference clocks are used for each media. For example, audio uses its own device interface as a clock since it is available and the format of the timestamps used for video stream depends on the type of compression used (Kouvelas et al. 1996).

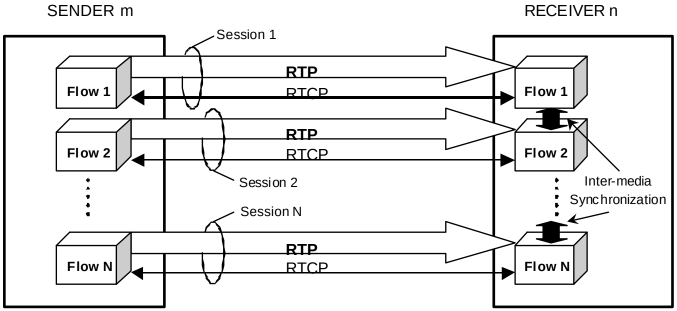


Figure 2: Multimedia transmission using RTP/RTCP

As RTP can be malleable to provide the information required by a particular application

2.4. RTP-Based Feedback Global Protocol

This protocol described in (Boronat 2001) is the result of mapping Feedback Global Protocol messages to RTP/RTCP packets, and the implementation of the resynchronisation algorithms developed for Feedback Global protocol. In that paper it is described without formulas (which can be obtained from (Esteve et al. 1999)) and without formal specification.

We have defined new RTP packet extensions to get our objective and map Feedback-Global control packets to RTCP packets. The new packet formats and the basics are described in (Boronat 2001).

3. VIDEOCONFERENCING TOOL

3.1. Application appearance

This application includes Mbone tools modified to implement our synchronisation method.

We can see the main window of the application in figure 3. As we can see, it includes modified vic (McCanne and Jacobson 1995), as the RTP-based application for video; modified rat (developed at UCL) as the RTP-based application for audio; and nte (non RTP-based) for text edition.

3.2. Architecture

The architecture in which we are going to use our tool is the one shown in figure 4, with one multimedia server or source and one or more receivers, all of them synchronised with a NTP Global Time reference.

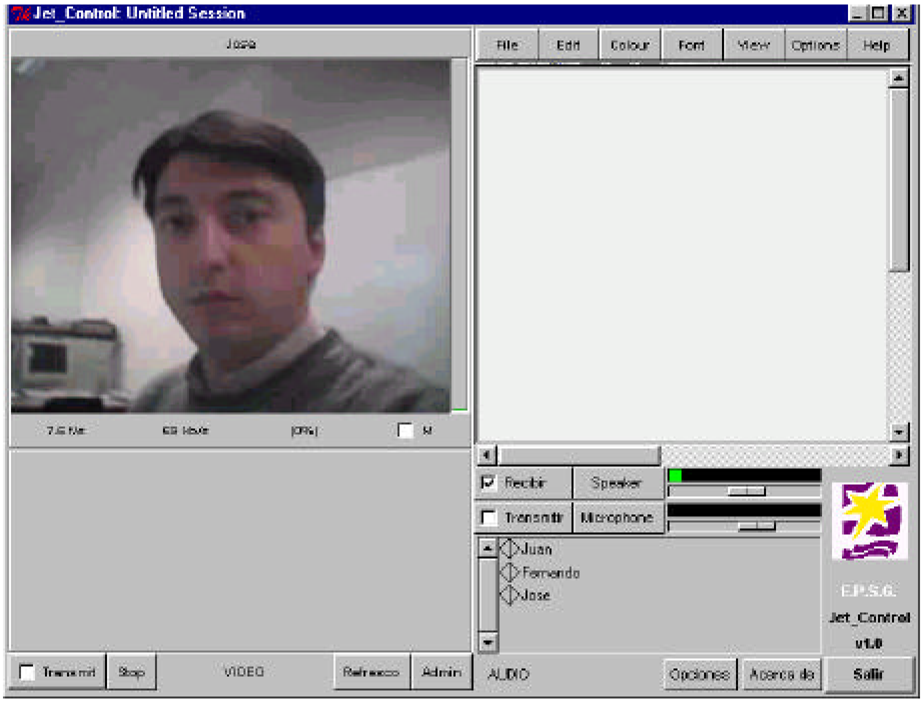


Figure 3: Videoconfencing tool

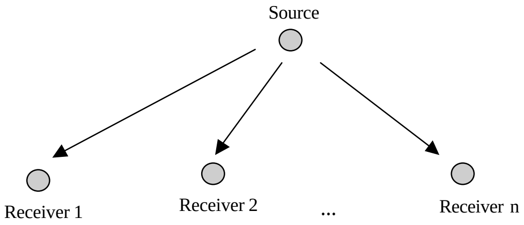


Figure 4: Architecture

The source captures and sends through a network (LAN or WAN) multimedia information streams in real-time, encapsulated in RTP packets, to the receivers, which decode them, reconstruct the streams and reproduce the multimedia information all synchronised.

To get synchronisation between receivers, RTP-based Feedback Global Protocol requires that source and receivers perform additional simple tasks that we already explained in (Boronat 2001) but summarise in the next paragraphs.

3.3. Initial Playback Time

In order to achieve that all the receivers begin the reproduction at the same time, prior to send data, the source sends a first control packet to tell them the precise moment to initiate the reproduction of each stage (period of time that starts when the source have any kind of multimedia information to send and ends when there is no information to send during a determined period of time). See figure 5.

As all the receivers are NTP synchronised, this moment will coincide in all of them, so the presentation will begin in all of them at the same global time.

To obtain this Initial Time the source has to estimate the delay (Del_i) between source and receivers (see figure 6). The first time, it can calculate it sending *IP echo packets* (RFC 791) to all the receivers. Next times this delay will be calculated by the mechanism defined in (Schulzrinne et al. 1996). When all the *Replies* are received the source can estimate the exact time (T_{ic}) in which all the receivers can start the reproduction synchronised. We have considered the worst case (maximum delay suffered by one receiver, in eq. (2).

$$T_{ic} = t_0 + Del_{max} \qquad (1)$$

, where

$$Del_{max} = \max \{ Del_i , \forall i \} \quad (2)$$

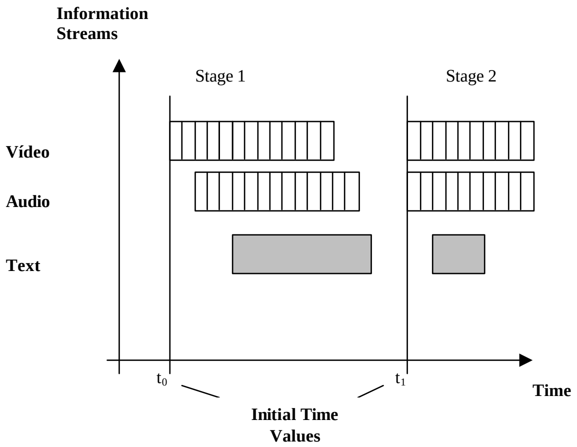


Figure 5: Initial Time Values

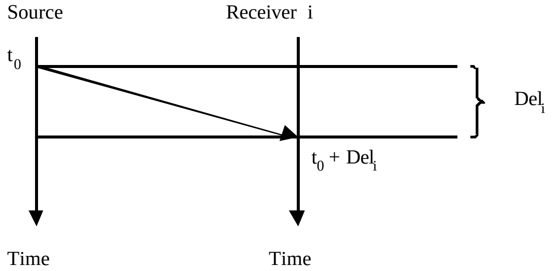


Figure 6: Receiver i Delay

To send this initial time to the receivers we need to use the RTCP APP packet, we called ***Tini APP Packet***, whose format is described in (Boronat 2001). It will include the initial time estimated, obviously, in NTP format. This packet will be sent at the beginning of each stage and then the source can send the RTP packets with the media information of that stage.

3.4. Fine Synchronisation between receivers

When receivers start to reproduce multimedia information, they have to go on being synchronised, so we need a mechanism to maintain that synchronisation and each receiver has to coincide with others in the playback instants. To get this, we make use of the existence of NTP Global Time, the simplicity of the feedback mechanism of RTP/RTCP and the timestamps of RTP packets.

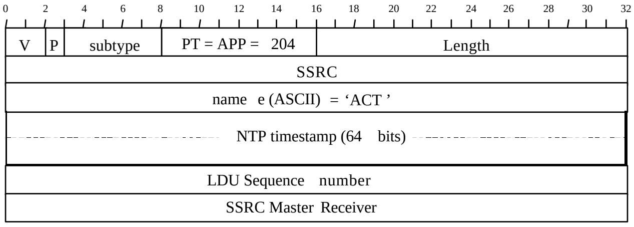
RTP-Based Feedback-Global Protocol implements a mechanism that consist of the receivers sending RTCP feedback reports (RR modified packets (Boronat 2001)) to the source including a timestamp and the number of sequence of the LDU (Logical Data Unit) that the receiver is playing at that time.

With that feedback information, the multimedia server can calculate how many LDUs each receiver has to skip or pause to be synchronised with the others receivers.

As we told before, we have used RR feedback reports that receivers send to the sources in RTP/RTCP but with a modified format, as described in (Boronat 2001), to include that feedback information that it doesn't contain.

When the source has received the RR packets from all the receivers, it can determine which receiver is the master (we have chosen the most delayed receiver). Then the source will build and send a new control packet (APP Act packet, defined in (Boronat 2001)) to make the other receivers to pause a determined number of LDUs to be synchronised with the master.

The packet used by the source to communicate a receiver to skip or pause actions is called ***Action APP Packet*** but it is different from the one described in (Boronat 2001). We have included in it the SSRC of the master receiver to avoid the master receiver tries to synchronise with itself. The new APP packet format is shown in figure 7.



V	P	subtype	PT = APP = 204	Length
SSRC				
name e (ASCII) = 'ACT '				
NTP timestamp (64 bits)				
LDU Sequence number				
SSRC Master Receiver				

Figure 7: New *Action APP packet* format

The fine synchronisation process is shown in figure 8.

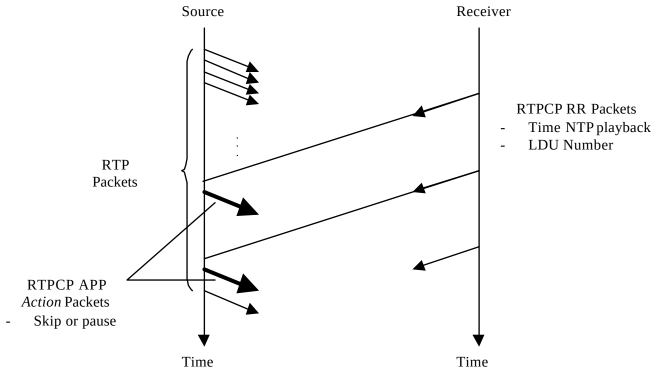
3.5. Inter-stream Synchronisation

An important feature of the digital audio system is its inherent transmission delay. The different processing to which video and audio signals are subjected basically produces different delays. For a videoconference close lip synchronisation (the reduction of the time lag between a remote user's audible words and the associated movements) is desirable. Any noticeable deviation can be extremely annoying to all participants. Therefore, the processing delays of the video and audio channels should be accurately equalised.

In our tool, as shown in figures 9 and 10, we include communication inter-stream via a local internal bus (mbus) to get inter-stream synchronisation. In (Kouvelas et al. 1996) there is an example of lip-synchronisation for use over the Internet.

The Mbus

The Mbus (Perkins 1998) is an interprocess communication channel designed around the needs of local coordination in multimedia conferencing systems. It is based in the Conference Control Channel Protocol (CCCP) (Handley et al. 1995), developed at UCL. It provides the mechanism for conference coordination, but policy and message contents are not defined.



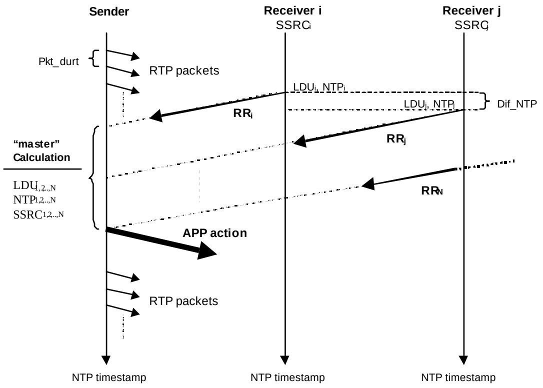


Figure 8: Fine Synchronisation Process

In our application, the inter-stream communication is necessary to interchange the playout delays between them. We have used the Mbus library for this proposal.

As playout delay negotiation between stream processes needs to be in a common format and with a common reference clock, we take advantage of RTCP messages providing individual mapping between the media and NTP timestamps.

In our application, only audio and video processes communicate to each other the required playout delay over the local bus and then both tools enforce the larger of the delays requested.

Each moment, we define a stream (audio or video) as the master flow and the other stream as the slave, according to the playout delay. The text tool is not synchronised in this preliminary version.

As shown in figure 10, we also consider two reconstruction buffers, each one for each media, to smooth out the effects of network jitter and to adapt to the playout delays. The video reconstruction buffering is also needed to guarantee that video frames are displayed at regular predictable intervals (Kouvelas et al. 1996).

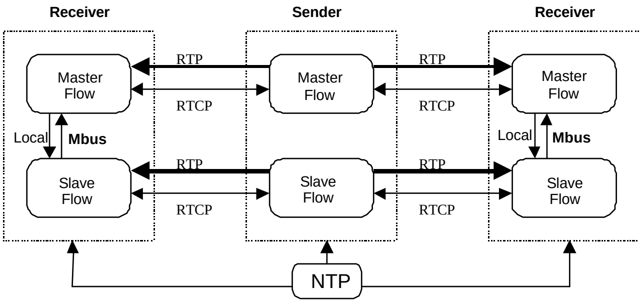


Figure 9: Local Mbus: Inter-stream Communication

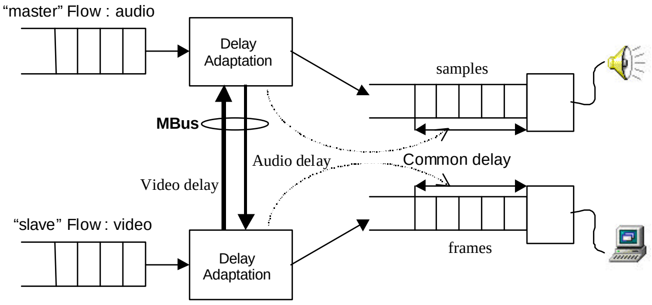


Figure 10: Playout Coordination

4. MEASUREMENT SCENARIO

We have tested our protocol using workstations running the videoconferencing tool in the LAN environment represented in figure 11. We use a Multimedia server with audio and video hardware devices to acquire and send in real-time multimedia data to a couple of different receivers. The characteristics of all of them are: Server is running in a Pentium III, 650 MHz and 128 MB RAM); Receiver 1 (better) is a client running in a Pentium III, 950 MHz and 128 MB RAM; and Receiver 2 (worse) is running in a Pentium MMX, 166 MHz and 32 MB RAM.

The results of playout delay for audio tool (rat), network jitter in each receiver and adjustments from others receivers are shown in figure 12 and 13. In both cases we have taken an initial delay of 2 seconds.

We can see in figure 12 the effects of the jitter in the playout delay in each receiver, and the adjustments in both receivers. Playout delay of Receiver 1 decreases with time because is faster than server, but it synchronises well with receiver 2, delaying determined amounts of time obtained from our algorithm (*dif_playouts* parameters in the figures). In figure 13, we can see what happens when there is no adjustment between receivers and how playout delay of receiver 1 decreases constantly.

In both figures we can see the effects of the jitter in the calculation of the playout delay.

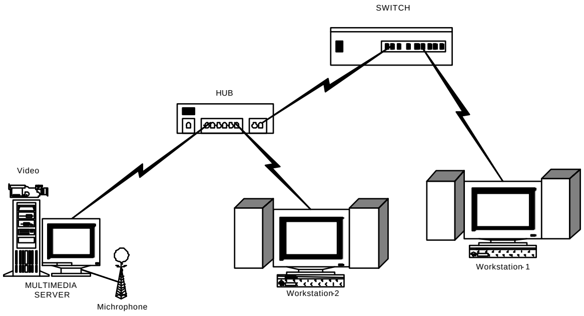


Figure 11: Measurement Scenario

5. CONCLUSIONS

We have got the same benefits than with RTP-Based Feedback-Global protocol, improving the results obtained with Feedback protocol:

- We have reduced the interval of reproduction to only one point in the temporal axis.
- We synchronise several receivers to reproduce multimedia streams all at the same time.
- We make the synchronisation precision independent of the network jitter.
- We maintain the simplicity of the Feedback protocol implementation but using RTP/RTCP, the protocol for multimedia transmission over Internet.

Other researchers have centred their research in synchronisation between streams in a local machine. We have also obtained synchronisation between receivers.

As it is shown at the graphs, it works quite well in a LAN environment but we have not made measurements for WAN environment.

Of course, we have to bear in mind the increase on the number of packets interchanged between source and receiver due to the global time network protocol but it is minimal.

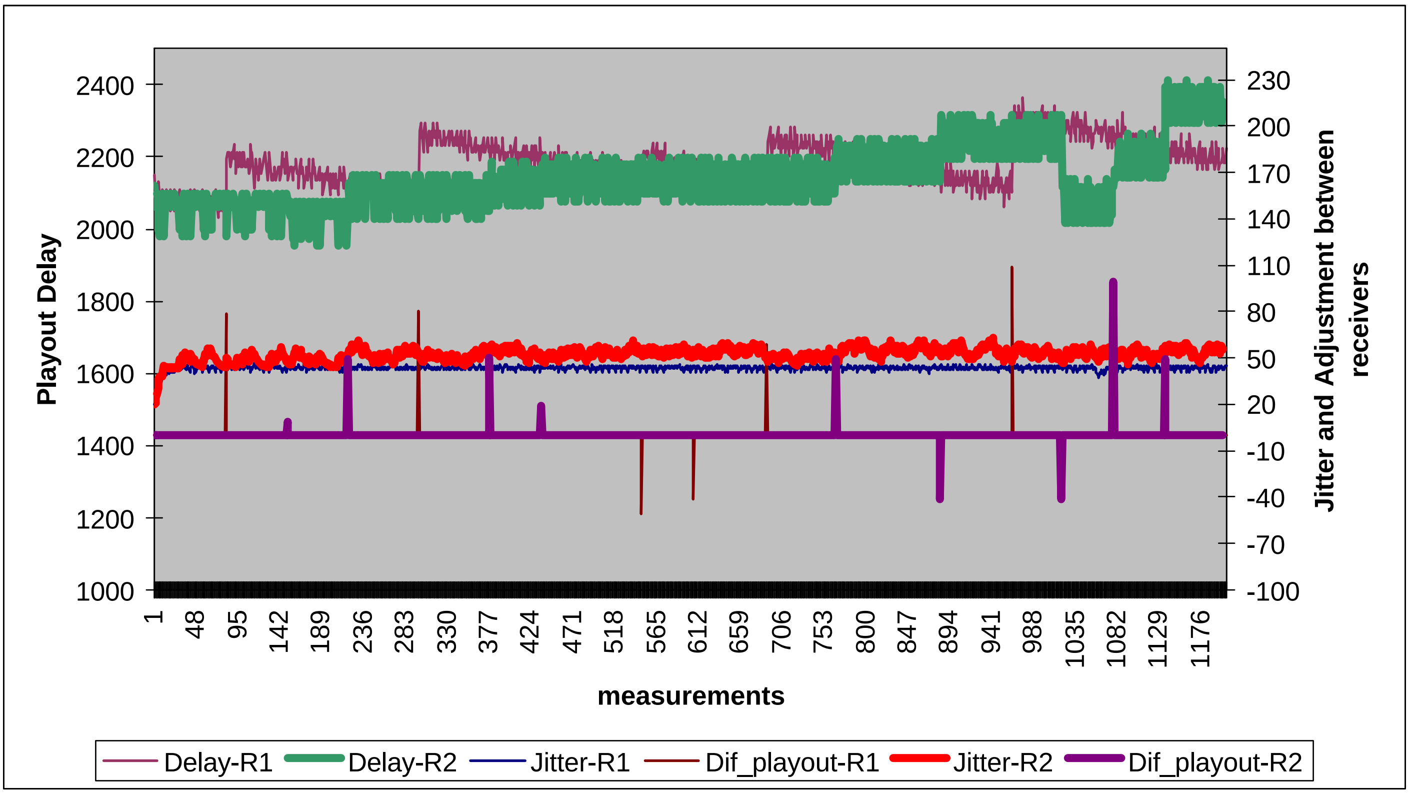


Figure 12: Measurements with adjustments between receivers

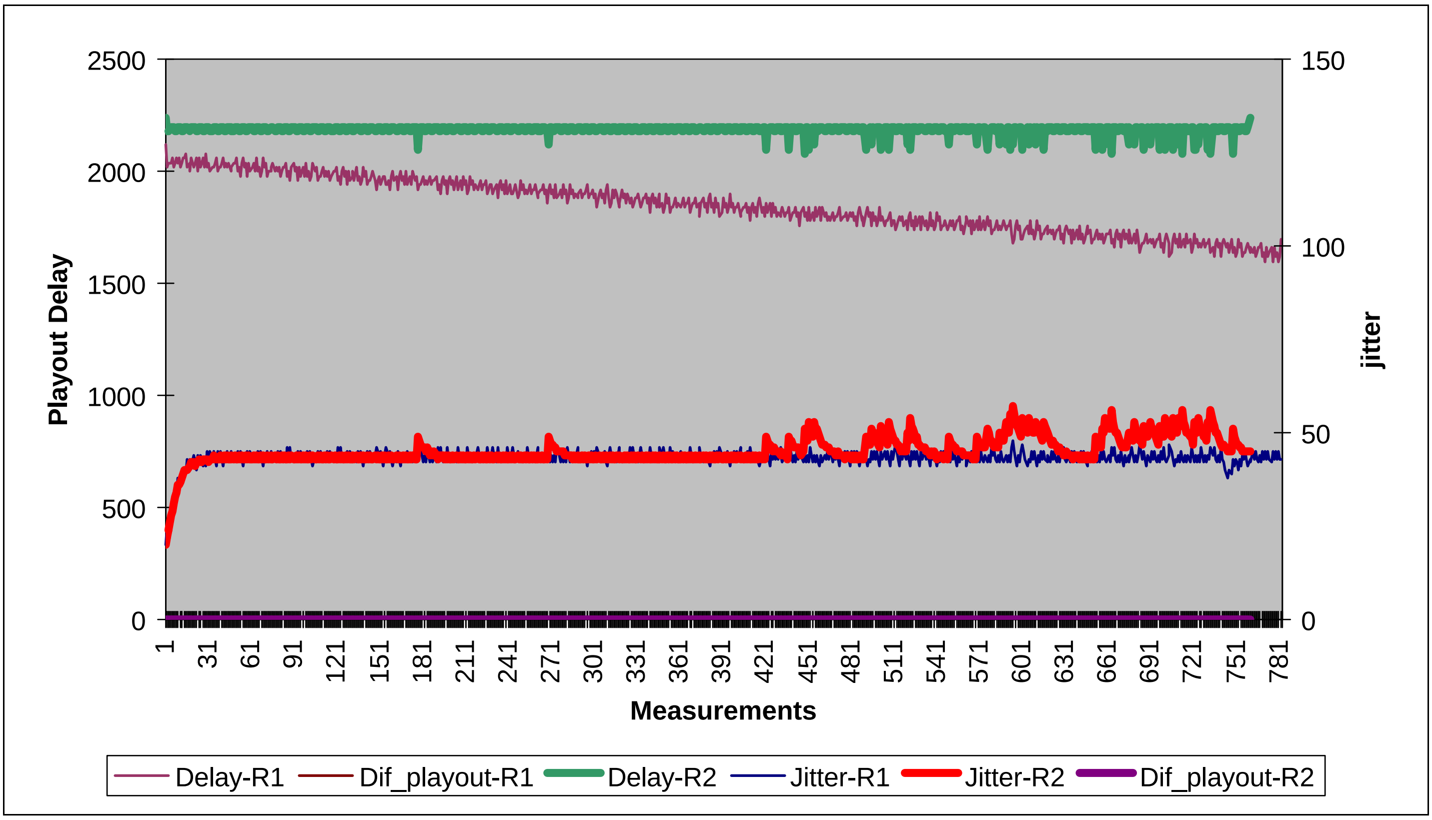


Figure 13: Measurements without adjustments between receivers

6. FUTURE WORK

First, we have to find a method for the server to find the initial playout delay, from the packets received initially from all the receivers.

Then we will evaluate objectively the protocol performance over a WAN environment (Mbone if it is possible).

Next steps are to evaluate subjectively the synchronisation protocol performance (taking into account human perception of a multimedia presentation in several receivers at the same time, like the one presented in (Steinmetz 1996)) and to add a RTP based tool for text transmission (We are developing a new tool for text transmission accordingly with RFC 2793).

Finally, we would like to implement that the source, if there are problems with bandwidth and accordingly with the QoS feedback information, may modify dynamically the transmission parameters (transmission rate, digital coding systems, etc.) to suit the current state of the network.

7. ACKNOWLEDGMENTS

This work has been financed by Comisión Interministerial de Ciencias y Tecnología (Spain) (CICYT) under project numbers CICYT (TIC-1999-0769).

8. REFERENCES

Boronat, F "Multimedia Streams Synchronisation Protocol based in RTP/RTCP (Real-Time Transport Protocol): RTP-based Feedback Global Protocol" *EUROMEDIA 2001*, Valencia (Spain), April.

Esteve, M.; Palau, C.; Guerri, J.C.; León, A. and Boronat, F. "A Distributed Medical Software Tool". *Simposio Español de Informática Distribuida* , Santiago de Compostela (España)

Guerri, J.C.; Boronat, F; Palau, C; Esteve, M; Berenguer, R and Pajares, A. "An Adaptive Multimedia Application for Remote Control and Synchronized Video Distribution" *EUROMEDIA 2000'*, Amberes (Belgium), May.

Handley, M.; Wakeman, I. and Crowcroft, J. "The Conference Control Channel Protocol (CCCP): A scalable base for building conference control applications" *SIGCOMM*. Pp. 275-287, Cambridge, Massachusetts, September 1995

Helstrom, G. "RTP Payload for Text Conversation·", RFC 2793, may 2000.

Kouvelas Y., Hardman V. and Watson A. 1996. "Lip Synchronization for use over the Internet: Analysis and Implementation". Technical Report, Department of Computer Science, University Collegue London. Analysis and Implementation, *Proceedings of IEEE Globecom'96*, November 1996, London UK.

McCanne, S. and Jacobson, V. "vic: A Flexible Framework for Packet Video". *ACM Multimedia*, November 1995, San Francisco, CA, pp. 511-522

Mills, D. L. "Internet Time Synchronisation: The Network Protocol", *IEEE Transactions on Communications*, vol 39, nº 10, pp. 1482-1493, October 1991.

Mills, D. L. "On the cronollogy and metrollogy of computer network timescales and their application to the Network Time Protocol", *ACM Computer Communications Review* 21, pp. 8-17, October 1991.

Mills, D. L. "Network Time Protocol", RFC 1305, 1992.

Ott, J; Perkins, C.; Kutscher, D. "Requeriments for Local Conference Control", draft-ietf-mmusic-mbus-req-00.txt, December 1999.

Perkins, C. "Using the Mbus Library", 1998

Ramanathan, S and Rangan. V. "Adaptative Feedback Techniques for Synchronized Multimedia Retrieval over Integrated Networks". *IEEE/ACM Transactions on Networking*, Abril 1993.

Rangan, V; Ramanathan, S and. Vin. H.M. "Designing an On-Demand Multimedia Service". *IEEE Communications Magazine*, Julio 1992.

Schulzrinne, H.; Casner, S.; Frederick, R. and Jacobson, V. "RTP: A Transport Protocol for Real-Time Applications". RFC 1889. January 1996.

Steinmetz R. 1996. "Human Perception of Jitter and Media Synchronization." *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, (Jan.): 61-72.

Xiao, X. And Ni, L.M. "Internet QoS: A Big Picture", *IEEE Network Magazine* March/April 1999, pp. 8-18.

E-BUSINESS REQUIRES
THE SIMULATION OF LOGISTICS PROCESSES

Wilfried Sihn, Tom-David Graupner, Jörg Mandel
Fraunhofer-Institute for
Manufacturing Engineering and Automation (IPA)
Nobelstr. 12
D-70569 Stuttgart, Germany
{whs,tdg,jrm}@ipa.fhg.de

KEYWORDS

E-Business, e-Services, Web-Services, Simulation, production, logistic management

ABSTRACT

E-business offers the chance to open new markets for selling and buying. E-business networks include both in-plant and cross-company processes. Therefore they are marked by a high complexity and strong demands regarding the management of the network. The Fraunhofer-Institute for Manufacturing Engineering and Automation developed several high-performance tools for designing and managing e-business networks. Three of these tools namely E-Simulation, E-Optimisation and the Internet-based Logistic Analysis. are described in this paper. E-Simulation allows users to simulate their own scenarios via Internet and benefit from the advantages of a dynamic simulation. If required, models can be adapted to changing conditions of the production environment. E-Optimisation is an Internet service for the determination of optimised production schedules based on flexible optimisation algorithms. The third tool described in the paper, the Internet-based logistics analysis is an electronic service that identifies rationalisation potentials in production logistics in a fast and cost-effective way.

INTRODUCTION

E-business opens up new opportunities to businesses while at the same time raising the demands on the structure and operation of new, more complex supply chains and networks. The spread of e-business and the related production networks have, without doubt, had a considerable impact on the organisation of order management processes throughout the supply chain. They include both in-plant and cross-company processes. In order to realise the benefits gained from e-business, a well functioning logistics chain is an absolute necessity. To enable order management in production networks it is no longer sufficient to consider the production capacities only. Existing or expected restrictions in supply, storage, and transport must also be included in planning. Therefore, when it comes to production and distribution, it is a must to use a high-performance tool to handle both E-business and Supply Chain Management. At our institute, several high-performance tools for designing and managing e-business networks have been developed. Figure 1 shows a selection of these tools, which will be described in the following (Sihn et al. 2001), (Sihn and Graupner 2001).

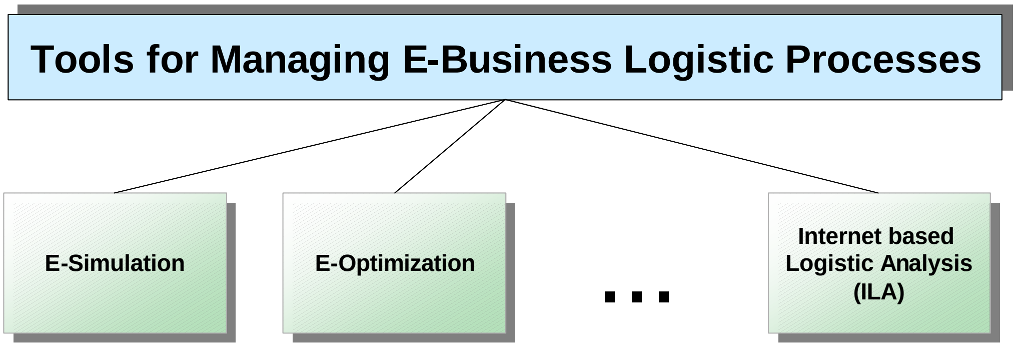

Tools for Managing E-Business Logistic Processes
E-Simulation
E-Optimization
...
Internet based Logistic Analysis (ILA)

Figure 1: Tools for Managing E-Business Processes.

E-SIMULATION

Meaning of e-Simulation

Today, simulation systems are deployed in the planning of production and logistics systems. By visualising processes and evaluating simulation results, production planners are in a position to better understand complex systems. Difficult decisions are made more easily when a solid basis for decision-making is provided.

Simulation projects create customised simulation models for production and logistics systems. These models perfectly match organisational and technical processes with the required resources. Simulation models can thus be referred to as information stores. After completion of the simulation project, these information stores are usually no longer available to the user, unless the simulation systems are maintained and advanced by the enterprise. This requires that, apart from license fees and high-performance hardware, the company also has to provide skilled personnel (Kulis and Paul 2000), (E.H. Page 1998).

The aim of the e-Simulation service is to use and enlarge the information store for as long as the production and logistics system exists (see Figure 2). Therefore, it allows clients to use the generated models even after a project has been completed. By coupling simulation models to the Internet, users get constant access to the simulation tool. To make use of e-Simulation only an Internet connection is required.

Benefits of e-Simulation

E-Simulation allows users to be actively involved in the modelling of processes during the early stages of the project. Thus, they are in a position to influence the model design at an early stage. They can simulate their own scenarios and benefit from the advantages of a dynamic simulation.

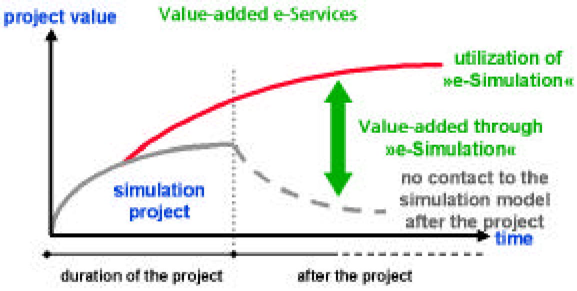


FIGURE 2. E-Simulation continues to increase the project benefit after the simulation project.

If required, models can be adapted to changing conditions of the production environment. This kind of customised support ensures the safeguarding of a company's investment. Adjustments can be made in a time and cost-saving manner to create the highest possible benefit for the entire life cycle of a production or logistics systems.

Functioning

When creating a simulation model, the required data are specified in close coordination with the user. These data include cycle times, set-up times or capacities needed for the simulation process. They are stored externally, e.g. in MS-Excel or MS-Access tables. The simulation models themselves are filed at the Fraunhofer Institute.

The user can now employ these data to create individual scenarios and upload or send them to the e-Simulation service via e-mail. On arrival, the transferred data are automatically checked for inconsistencies and transmitted to the simulation model for processing. Then, the simulation is automatically started. After the conclusion and evaluation of the simulation, the user receives an email with the results in tabular and graphic form. These evaluations consider the previously analysed company-specific parameters.

Practical example

E-Simulation has been successfully used for production optimisation in an automotive supply chain. During the project, simulation was employed to determine the necessary logistical areas and interpret the number and size of the cage boxes. Moreover, the future machine utilisation and throughput had to be established. Within a few days, a highly detailed simulation model for the production process was generated. The simulation model mapped 15 different machines and 40 different articles with specific cycle and set-up times.

Both the process and the control parameters of the simulation models were embedded in a MS-Excel table, ensuring the simple manipulation of production data. By applying the e-Simulation service, users at different sites could access the constantly updated models. Once the MS-Excel table was sent to a specified e-mail address, it took only 30 minutes until the user received the evaluation of the simulated scenarios, edited as graphics and nominal parameters. The simulation helped to carry out detailed analyses, identify bottlenecks and weaknesses before serial production started, and find appropriate solutions.

E-OPTIMISATION

Meaning of e-optimization

As mentioned above, e-business significantly increases the need for co-ordination in the production and logistics processes. Often, the problems are so complex that modern planning tools are required to facilitate decision-making in order to create and implement proposals in a reasonable period of time.

E-Optimisation is an Internet service for the determination of optimised production schedules based on flexible optimisation algorithms. E-Optimisation solves varied problems. The optimisation considers all the relevant constraints, such as stock costs, delivery performance, lead times and workload, and provides an optimal solution regarding the desired criteria.

Benefit of e-Optimization

By applying e-Optimisation, the user saves time and money. Time is saved because a quick and uncomplicated utilisation of the planning tool is possible via Internet. The tool is based on flexible algorithms and runs on the service provider's own high-performance computers. Using the planning tool via Internet also helps to save costs. The only thing required to get access to e-Optimisation is an Internet connection. As the applications and hardware belong to the service provider, they do not have to be purchased. Thus, the user keeps both investment and operating costs for the maintenance of infrastructure at a low level. Only the service provider's performance has to be paid for.

Functioning

Using e-Optimisation is very simple: Either the user is directly connected to the e-Optimisation service via data link, or an encoded channel is used to transfer the production data at regular intervals.

The scheduler himself sets the operational planning frameworks through a specific graphic user interface, so that he can easily adjust them. Special attention has been paid to the design of the graphic user interface, as the man-machine interface is decisive for the correct and efficient use of a system.

E-Optimisation performs the optimisation tasks with the help of several evolutionary algorithms running in parallel on a computer network. The system is learning with every optimisation run and autonomously changing the parameterisation of existing algorithms. Thus, the system adapts to changing tasks, reducing maintenance costs and improving the system's performance.
After the optimisation, the results are shown on the desktop as loading plans or directly imported into an ERP system (see Figure 3).

Practical Example

E-Optimisation serves a wide range of applications mainly in the fields of automotive and steel industry, but also in the aircraft industry.

Taking an example from the steel industry, e-Optimisation was used to work out optimal loading plans for a cold rolling mill. A cold rolling mill produces extremely precise steel strips to be used, for example, in the automotive industry. It is the task of the optimisation planner to align the processing sequence of the steel strips in such a way as to achieve maximum throughput while letting the steel strips pass through the mill without delay. The system includes a number of restrictions and dependencies calling for the global optimisation of the overall process.

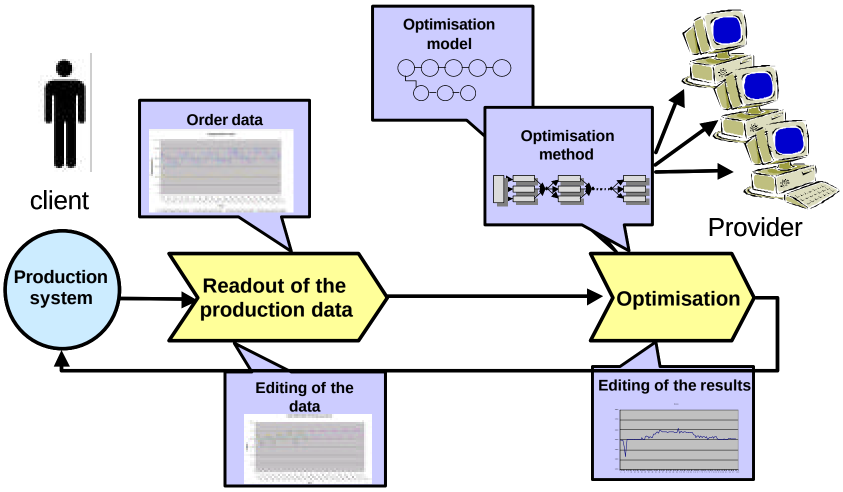


FIGURE 3. The E-Optimisation process.

INTERNET-BASED LOGISTIC ANALYSIS

Meaning of Internet based Logistic Analysis

With the aid of logistics analysis it is possible to rapidly and inexpensively identify and evaluate rationalisation and improvement measures in production logistics. Thus, logistics analysis becomes a suitable tool for controlling production planning and management (PPM); but it is also an appropriate means for dimensioning and structuring assignments within factory planning.

Internet-based logistics analysis is an electronic service that identifies rationalisation potentials in production logistics in a fast and cost-effective way. This service is based on the Hanover Funnel Model, automating the processing of existing production data. The latter are compressed to obtain relevant logistical parameters and be appropriately represented. The user gets information about the logistical behaviour of his production system sorted according to resources, i.e. machines and facilities, as well as production orders.

Benefit of the internet-based Logistics Analysis

This service is available to every production plant with single-item production. Being Internet-based, the service avoids costly consulting visits and is available independent of time and location. For easy usage, the Internet-based logistics analysis provides a brief manual to interpret the obtained analysis results. The clarity of presentation and the tight construction of the underlying Funnel Model contribute decisively to the comprehensibility of the service.

The origin of the logistics analysis is the Funnel Model of the Institute of Production Systems (University of Hanover, IFA Institut für Fabrikanlagen). It links the production logistical parameters of stock, performance and turnover time, offering a valuable descriptive model for production logistics. The defined parameters of stock, performance, order structure, lead times, utilisation and schedule deviations help to rapidly identify those machines and facilities that significantly influence the logistical performance. This allows taking appropriate actions focusing on (Nyhuis and Wiendahl 1999).

- lead time reduction,
- inventory reduction and
- improved adherence to schedules.

Most companies possess the necessary database by way of production data collection (PDC). Existing systems, however, create primarily parameters relating to monetary terms, from which production logistics improvements cannot be deduced directly.

The Internet-based logistics analysis is distinguished by an excellent cost-benefit ratio. By avoiding costly external consulting and giving simple access via World Wide Web, by exploiting available corporate data and making the modification of existing information systems superfluous, the potential benefits can be unlocked for production logistics.

Functioning

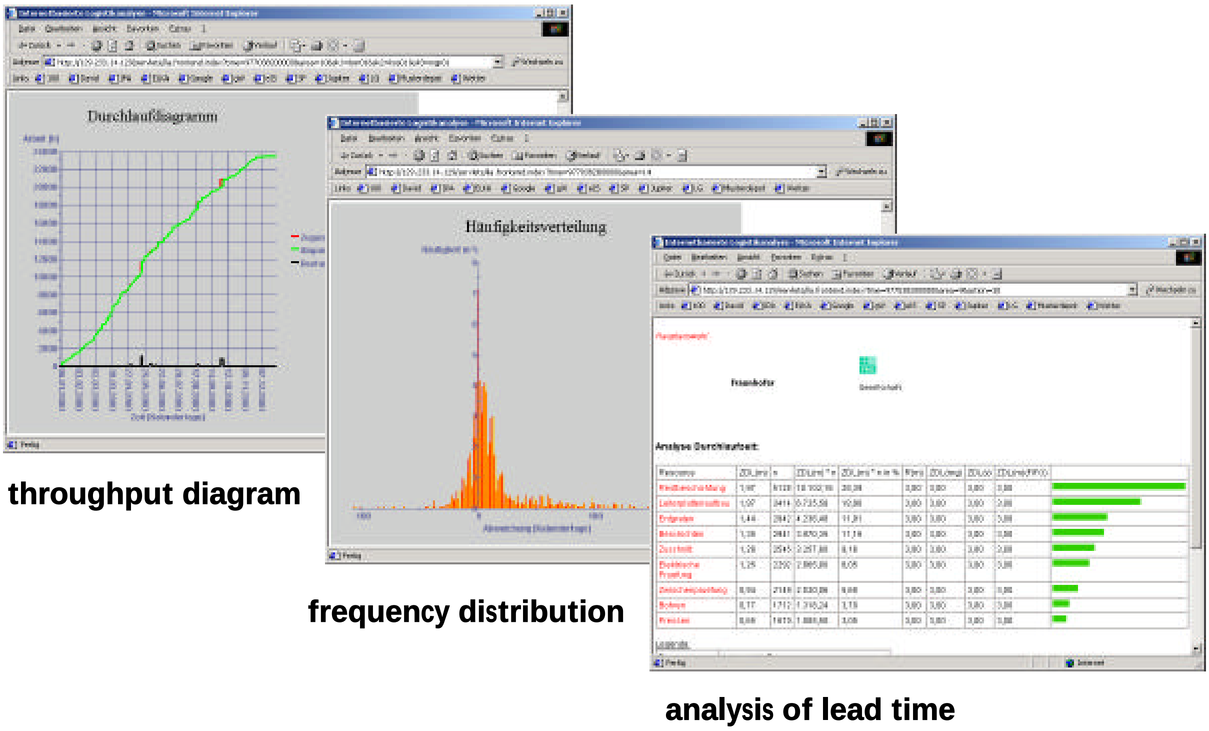
The Internet-based logistics analysis is divided into two parts, i.e. user registration and analysis. During registration the user enters his password at the Internet service’s homepage. This initiates the creation of a "personal area" on the web pages from which the user can carry out the logistics analyses.

When performing the logistics analyses, the first step is to fill in a text file according to a pre-designed format. To this end, a template is available to be downloaded from the homepage (Figure 4). Next, the file including the company data is uploaded to the Internet service, which automatically carries out the analysis. The results are edited as web pages and stored in the personal area of the user, who is informed by e-mail. Thus, accessing results is possible independent of time and location.

The user may perform an unlimited number of analyses. The results remain available, but may be deleted by the user.

Practical example

The Internet-based logistics analysis is a brand-new tool. The development of the tool was finished a short time ago. Therefore a practical example is not available up to now.



throughput diagram

frequency distribution

analysis of lead time

FIGURE 4. Analysis Results (Example).

CONCLUSION

The paper describes a selection of tools to support the production and logistical management processes in e-business networks. Our service range is constantly expanded thanks to several research and development projects being carried out in this field. Presently, we intend to develop new modules for an easier set-up of e-business simulation models.

REFERENCES

Sihn, W., Halmosi, H., Mandel, J., (2001), Komplexe, simulationsgestützte Internetdienstleistungen – die Beispiele Produktionssimulation und Logistikanalyse. In: Schulze, Th., Schlechtweg, S., Hinz, V. (Hrsg.), Proceedings zur Tagung „Simulation und Visualisierung 2001“ am Institut für Simulation und Graphik der Otto-on-Guericke-Universität Magdeburg, March 22nd / 23rd, 2001.

Sihn, W.; Graupner, T.-D., (2001), e-Industrial Services: Value-Added Services for the Producing Sector. In: Chryssolouris, George (Hrsg.); University of Patras / Dept. of Mechanical Engineering and Aeronautics / Lab. for Manufacturing Systems and Automation: Technology and Challenges for the 21st Century : CIRP 34th International Seminar on Manufacturing Systems, 16-18 May, 2001, Athens, Greece. Athens, Greece, 2001, S. 413-424

J. Kulis & R.P. Paul. (2000) A review of web based simulation: whither we wander? In: In: Proc. 2000 Winter Simulation Conference (Eds.: Joines, J.A., Barton, R., Kang, K., Fishwick, P.), 1872-1881.

E.H. Page. (1998) The rise of web-based simulation: Implications for the high-level architecture. Proc. 1998 Winter Simulation Conference, (Hrsg.: Medeiros, D.J., Watson, E.F., Carson, J.S., Manivannan, M.S.), 1663-1668.

Sihn, Wilfried; Graupner, Tom-David: (2001) E-Industrial Services Value-Added Services for the Producing Sector Using the Example of Simulation. In: Jiang, Pingyu (Hrsg.) u.a.: ICECE 2001 : Abstracts for Proceedings of Internaional Conference on e-Commerce Engineering: New Challenges for Global Manufacturing in the 21st Century, Sept. 16-18, 2001, Xi'an, China. Beijing, China : China Machine Press, 2001, S. 40

Nyhuis, P.; Wiendahl, H.-P.:Logistische Kennlinien.Berlin et al: Springer, 1999

TOWARDS A NEXT STEP FOR WEB-BASED ASSEMBLY MANUAL

Teruaki Ito
Department of Mechanical Engineering
University of Tokushima
Tokushima 770-8506, Japan
E-mail: ito@me.tokushima-u.ac.jp
Phone: ++81 88 656 2150

KEYWORDS
Web-based learning, assembly manual, VR, evaluation

ABSTRACT

Virtual product development paradigm shifts a large amount of information into the CAD models, which enables web-based applications using 3D models in various areas including virtual reality, simulation, and distance learning. Technologies to increase the reality in virtual environment are important, however, human factors are also getting more and more important. In order to clarify the human factors, we have conducted comparative experiment using different kinds of assembly manuals, with which participants carried out assembly operation to complete a fishing reel. This paper describes the overview of the experiment and its results, and discusses the human factors to be considered for assembly manuals, which can be a hint towards a next step for a web-based assembly manual.

INTRODUCTION

Virtual reality (VR) technologies enabled virtual product development in various areas. One established method in this categories is Digital Mock Up for automotive industries, where the concurrently designed parts and assemblies were checked for collision, clash and clearance [Schiemenz 2001]. When this technology is used in naval engineering, a virtual ship can be accurately modeled and its operations are realistically simulated on computers, where its behavior and interaction of the simulated components are modeled in accordance with laws of physics and principles of virtual ocean and maneuvers and operations conducted in various sea states and under selected environmental conditions [Zini 2000]. VR as an innovative technology has also been applied in the latest manufacturing simulation areas. Flight simulation is a pioneering application that has been used to train pilots for many year in airline industries. An automotive assembly simulator has been used to evaluate process installation feasibility. A virtual window tunnel was a flow field simulator that allows the user to explore fluid dynamics phenomena. VR-based simulations also include major manufacturing applications such as product design and prototyping, facility layout design and visualization, assembly process planning and simulations, operation training, remote operation of equipment, etc. VR-based system also demonstrates advantages on maritime and harbour training to save the height cost of traditional methods, and to simulate real life crisis situations and actual work environment, which are not available in existing training methods. [Broas 2000]

Meanwhile, more and more training is being delivered via the Internet or the Intranet. Web-based learning is the delivery of interactive training or education over the Internet or Intranet. It is the structured transfer of skill or knowledge that takes place using the World Wide Web as the distribution channel. For this purpose, interactive systems [Newman 1995] are designed based on multimedia technologies [Agnew 1996] including VR technologies and simulation technologies mentioned before. Those interactive learning systems are designed and implemented, which have changed the way in learning, from self-instructional to instructor led and collaborative, from synchronous learning to asynchronous learning, from built in authoring to third party authoring, from rich media to lean media, from low interaction to high interaction, or from course delivery to course management.

Considering the current and future trends that everything be available on the Web, we assume that web-based manual will also be getting popular not only for computer related fields such as software installation, but also various fields including product assembly using 3 D manipulation or simulation. VR-related technologies mentioned above will also be available for these web-based manuals. A Web-based maintenance system for fishing reel products using assembly/disassembly manual is under development by the authors, considering its potential for a web-based learning system. As for the assembly/disassembly manual module in the system, the objective of our research is not to replace paper manuals but to figure out a next step for web-based assembly manual. To increase the reality in web-based manual is important, however, human factors should be more considered. In order to clarify the human factors, we have conducted comparative experiment using different kinds of assembly manuals, with which participants carried out assembly operation to complete a fishing reel. This paper describes the overview of the experiment and its results, and discusses the human factors to be considered for assembly manuals, which can be a hint towards a next step for a web-based assembly manual.

FISHING REEL MAINTENANCE SYSTEM

Fishing attracts people all over the world and many people enjoy fishing as a hobby. Among various fishing gears, fishing reels (Figure 1) play a very important role. They are mostly delicate products composed of various kinds of tiny parts, however, environments of their usage are mostly under very severe conditions, such as windy and rainy weather, with low temperature, sometimes with salt water from the sea, etc. For the best usage, its maintenance is very important but it is not always easy for general users to do so by themselves. Maintenance can be taken care of by technical people from the shop, however, it takes time and money. User maintenance is better to save time and money, and also it can increase the pleasure of fishing.

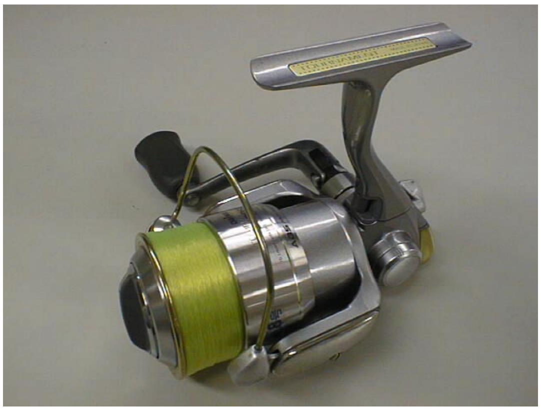


Figure 1: A fishing reel

We are developing a web-based fishing reel maintenance system for general users who enjoy fishing. The system is composed of corrective maintenance and preventive maintenance modules, both of which allow users to run web-based assembly/disassembly manuals including animation. Since the maintenance system is out of the scope of this paper, its description will be given somewhere else. This paper only focuses on its assembly/disassembly manual module.

ASSEMBLY OPERATION MANUAL

The assembly operation manual is one of the important modules in the reel maintenance system. The operation manual shows how to disassemble/assemble the fishing reel using pictures and figures, with description of operation procedures. In addition to that, each step provides 3D animation, which help the user to understand the structure of echo component, and to carry out the operation easily. Figure 2 shows a snapshot of 3D animation. In this paper, this manual is called 3D-manual.

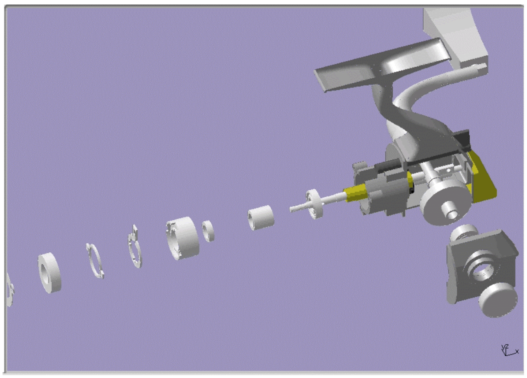
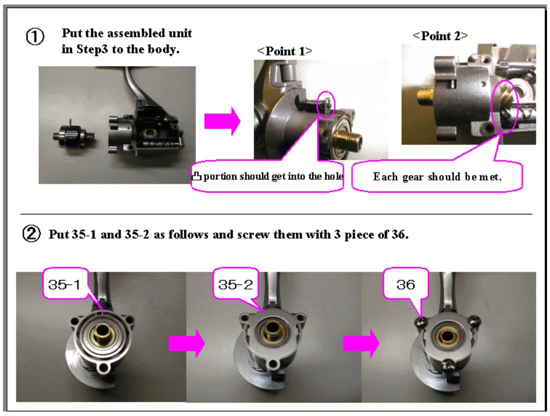


Figure 2: Animation of 3D manual

PAPER-BASED AND WEB-BASED ASSEMBLY MANUALS

A paper-based manual has been used for a log time as the standard media, and its effectiveness has been recognized by most of the people in various areas. Without using an existing manual for the fishing reel, we have prepared a paper-based manual from 3D-manual so that the contents of information the user obtains should be the same. Figure 3 shows a snapshot of this manual.



① Put the assembled unit in Step3 to the body.

<Point 1>

<Point 2>

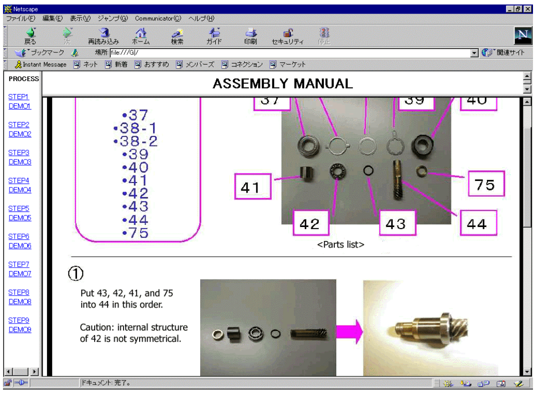
portion should get into the hole

Each gear should be met.

② Put 35-1 and 35-2 as follows and screw them with 3 piece of 36.

Figure 3: A snapshot of paper manual

We have also prepared a web-based manual from the paper manual. Figure 4 shows a snapshot of the manual.



Netscape
ファイル(F) 編集(E) 表示(V) ジャンプ(G) Communicator(C) ヘルプ(H)
戻る 次 再読み込み ホーム 検索 ガイド 印刷 セキュリティ 停止
ブックマーク
場所: file:///G|/
関連サイト
Instant Message ネット 新着 おすすめ メンバーズ コネクション マーケット
PROCESS
STEP1 DEMO1
STEP2 DEMO2
STEP3 DEMO3
STEP4 DEMO4
STEP5 DEMO5
STEP6 DEMO6
STEP7 DEMO7
STEP8 DEMO8
STEP9 DEMO9
37
39
40
41
75
42
43
44
ドキュメント: 完了。

ASSEMBLY MANUAL

- •37
- •38-1
- •38-2
- •39
- •40
- •41
- •42
- •43
- •44
- •75

<Parts list>

①

Put 43, 42, 41, and 75 into 44 in this order.

Caution: internal structure of 42 is not symmetrical.

Figure 4: A snapshot of web-based manual

EXPERIMENTS OVERVIEW

A fishing reel composed of over 60 parts was used in the experiment. Each part was numbered and prepared in the numbered box as shown in Figure 5. The manuals refereed to these numbers so that participants can easily access to each part. All of the participants in the experiment had no prior knowledge about this fishing reel. Facing with the assembly parts for the first time at the time of experiment, they were asked to assemble the parts to complete the fishing reel, referring to one of the three manuals. The allotted time to each participant was 30 min, which means that experiment was terminated after 30 min. whether the operation was finished or not.

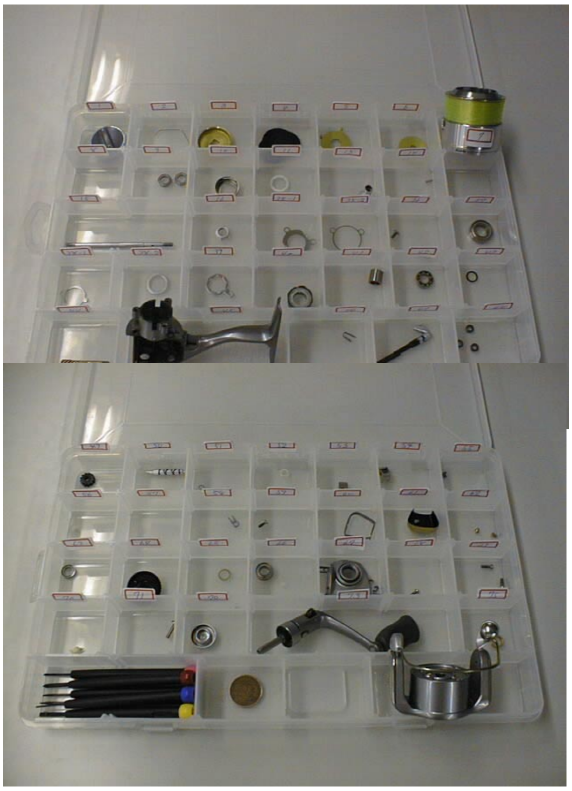


Figure 5: Assembly parts

The assembly procedure is composed of 9 steps, of which consumed time was recorded during the experiment. After the experiment, each participant was asked to fill out a questionnaire including their comments.

EXPERIMENTS RESULTS

Although required time to finish operation was upto each individual, 3D-manual and web-based manual users took more time for the operation. This is because those users needed some time to get used to the software operation in the computer. In overall results, however, there was no significant difference in which manual was used for the operation. Paper-based manual users gave us some comments, saying that they want to try 3D-manual or web-based manual. There were no comment from 3D-manual or web-based manual users that they wanted to use paper-based manual.

There are some operations which requires some skills, but they can be carried out if they refereed to any of the manual carefully. If the operation included more difficult steps, however, we assume that better results could be obtained from 3D-manual users. For a general ideas, a paper manual may be suitable to do this kind of operation, but no difficulty was observed from 3D-manual or web-based manual users.

Although the allotted time for assembly operation was 30 min., during which time some of the participants did not complete the assembly operation, most of them asked us to continue their assembly operation because they wanted to finish it. We believe that this kind of constructive and positive attitude was much more important factor even if they use any of the three manuals. More details of the experiments will be given at the presentation.

CONCLUDING REMARKS AND FUTURE WORKS

The paper describes the current trends regarding web-based applications using VR technologies and simulation, and their potential to web-based learning. The author is developing a web-based maintenance tool for fishing reel for web-based learning. To clarify the human factors in using assembly operation, we have conducted the comparative experiments using 3 kinds of manuals, or 3D-manual, paper-manual, and web-based manual. Before the experiment, we had expected a result that a paper manual would be better than the other two manuals. We found out that it was a good mistake. Since the target product of the experiment was not so much complicated, we did not find significant difference among users. For future work, we should pick up more complicated products, which are composed of many more parts, or which includes very complicated assembly operation, and compare the results.

REFERENCES

L. Jin, I. A. Oraifige, F.R. Hall and P.M. Lister: Distributed VR-based simulation for manufacturing, European Simulation Symposium: Simulation in Industry, Oct.18-21, 2001, Marseille, France, pp.289-293.

K. Schiemenz, Configured DMU - Generating and managing structure variants using tables, European Concurrent Engineering Conference: Concurrent Engineering: the path to electric business, Apr.18-20, 2001, Valencia, Spain, pp.191-194.

F. Boegershausen and K. Schiemenz, DMU in extended enterprise – cooperative product development with suppliers, European Concurrent Engineering Conference: Concurrent engineering in the framework of IT convergence, Apr.17-19, 2000, Leicester, UK, pp.255-259.

P. Broas, G. Granholm, and J. Hartimo, Collaborative virtual environments for maritime & harbour training, the international workshop on harbour, maritime & multimodal logistics modeling and simulation, Oct.5-7, 2000, protofinoe, italy, pp.106-111.

A. Zini, A. Rocca, M. Raffa, and R. Costa, Building a virtual ship, the international workshop on harbour, maritime & multimodal logistics modeling and simulation, Oct.5-7, 2000, protofinoe, italy, pp.35-40.

P. W. Agnew and A. S. Kellerman, Distributed multimedia: technologies, applications, and opportunities in the digital information industry, Addison-Wesley, 1996.

W. M. Newman and M. G. Lamming, Interactive system design, Addison-Wesley, 1995.

AUTHOR LISTING

AUTHOR LISTING

AUTHOR LISTING