

12<sup>TH</sup> ANNUAL EUROMEDIA CONFERENCE  
2006

ATHENS, GREECE

APRIL 17-19, 2006

Organized and Sponsored by:

EUROSIS

Co-Sponsored by:

EU-DG INFSO

BELGACOM

GHENT UNIVERSITY

ICCS

HOSTED BY

ATHENS IMPERIAL HOTEL



# **EUROMEDIA'2006**

FEATURING

TWELTH ANNUAL SCIENTIFIC CONFERENCE  
ON WEB TECHNOLOGY, NEW MEDIA  
COMMUNICATIONS AND TELEMATICS THEORY  
METHODS, TOOLS AND APPLICATIONS

Elpida Tzafestas

APRIL 17-19, 2006  
ATHENS, GREECE

A Publication of EUROSIS

Printed in Ghent, Belgium

## **EXECUTIVE EDITOR**

**PHILIPPE GERIL  
(BELGIUM)**

**Editor**

**Prof. Elpida Tzafestas  
National Technical University of Athens  
Inst. of Communications and Computer Science  
Zografou Campus, Athens, Greece**

## **Programme Committee**

**Prof. Marwan Al-Akaidi  
De Montfort University  
Leicester, United Kingdom**

**Fernando Boronat Seguíf  
UPV  
Valencia, Spain**

**Prof. Dr.J.Broeckhove  
RUCA-UA  
Antwerp, Belgium**

**Dr. Juan Carlos Guerri  
Cebollada  
UPV  
Valencia, Spain**

**Dr. Nathan Clarke  
University of Plymouth  
Plymouth, United Kingdom**

**Dr. Paul Dowland  
University of Plymouth  
Plymouth, United Kingdom**

**Dr. Steven Furnell  
University of Plymouth  
Plymouth, United Kingdom**

**Prof. Chris Guy  
University of Reading  
Reading, United Kingdom**

**Prof. Dr. Jan Knop  
University of Düsseldorf  
Düsseldorf, Germany**

**Ass. Prof. Qingping Lin  
Nanyang Technological  
University  
Singapore**

**Rachel Moreau  
LUC  
Diepenbeek, Belgium**

**Lorenzo Motta  
Ansaldo Trasporti s.p.a.  
Genoa, Italy**

**Dr. H. Joachim Nern  
Aspasia Knowledge Systems  
Dusseldorf, Germany**

**Dr.ir.Johan Opsommer  
Belgacom  
Brussels, Belgium**

**Dr. Carlos Enrique Palau Salvador  
UPV  
Valencia, Spain**

**Dr. Ana Pajares  
UPV  
Valencia, Spain**

**Maria Papadaki  
Symantec, UK  
8Southampton, United Kingdom**

**Prof. Jehan Francois Paris  
University of Houston  
Houston, USA**

**Prof. Marco Roccetti  
University of Bologna  
Bologna, Italy**

**Dr. Leon Rothkrantz  
Delft University of Technology  
Delft, The Netherlands**

**Prof. Paola Salomoni  
University of Bologna  
Bologna, Italy**

**Prof. Jeanne Schreurs  
LUC  
Diepenbeek, Belgium**

**Ph. D. Oryal Tanir  
Bell Canada  
Montreal QC, Canada**

**Ass. Prof. Vassilis Triantafillou  
Techn. Education Institute  
Greece**

**Prof. Rik van de Walle  
Ghent University  
Ghent, Belgium**

**Dr.Charles van der Mast  
Delft University of Technology  
Delft, The Netherlands**

**Prof. Matthew Warren  
Deakin University Geelong  
Victoria, Australia**



# **EUROMEDIA 2006**

© 2006 EUROSIS-ETI

Responsibility for the accuracy of all statements in each paper rests solely with the author(s). Statements are not necessarily representative of nor endorsed by the European Simulation Society. Permission is granted to photocopy portions of the publication for personal use and for the use of students providing credit is given to the conference and publication. Permission does not extend to other types of reproduction nor to copying for incorporation into commercial advertising nor for any other profit-making purpose. Other publications are encouraged to include 300- to 500-word abstracts or excerpts from any paper contained in this book, provided credits are given to the author and the conference.

All author contact information provided in this Proceedings falls under the European Privacy Law and may not be used in any form, written or electronic, without the explicit written permission of the author and/or the publisher.

All the articles published in these Proceedings have been peer reviewed.

EUROSIS-ETI publications are ISI-Thomson and INSPEC referenced

For permission to publish a complete paper write EUROSIS-ETI, Executive Director, Ghent University, Faculty of Engineering, Dept. of Industrial Management, Technologiepark 903, Campus Ardoyen, B-9052 Ghent-Zwijnaarde, Belgium.

EUROSIS is a Division of ETI Bvba, The European Technology Institute

**EUROSIS Publication**

**ISBN: 90-77381-25-2**

## **PREFACE**

EUROMEDIA'2006 follows on in the footsteps of past events in bringing together several strands of computer media research to converge into a single event, where media is the binding factor in all the applications, some of which are pushing back the boundaries of what is scientifically possible to achieve. This year's event is no different, and we are sure that all will appreciate the depth and expanse of this year's presentations, for which we should thank the authors.

As with any conference, EUROMEDIA also would not be possible without the help and support of a number of people, and we would like to begin by thanking all of the reviewers for their efforts, which have resulted in a truly interesting and varied conference programme. We are also most grateful to Fanuel Dewever of IBM Business Consulting Services for accepting to be our keynote for 2006, whose talk will cover the Open and Collaborative Innovation in a European Network of Living Labs, and of course to all of the paper authors and presenters for their hard work, and their willingness to share their results in the other sessions. Thanks also to the session chairs and other delegates who we are sure will guarantee us a lively and thought-provoking conference. Finally, special thanks are due to Philippe Geril, whose continued dedication and hard work as the conference organiser has enabled us to maintain the standard expected of EUROMEDIA events.

We sincerely hope that all of the delegates enjoy the conference, and that other readers of these proceedings will be encouraged to participate in EUROMEDIA events in the future. On behalf of all of EUROSIS, the International Programme Committee, we welcome you to this event and look forward to a successful conference.

Prof. Elpida Tzafestas  
General Conference Chair  
EUROMEDIA'2006



<b>Preface .....</b>	<b>VII</b>
<b>Scientific Programme .....</b>	<b>1</b>
<b>Author Listing .....</b>	<b>143</b>

## DATA SEARCH

<b>Searching Semi-Structured Data Using Landmarks</b> Andrew Davison.....	<b>5</b>
<b>A Model for Evolutionary Multimedia Objects</b> Georges Chalhoub, Richard Chbeir and Kokou Yetongnon.....	<b>10</b>
<b>Contour-Based Center of Gravity Evaluation of Characters</b> Akio Kotani, Yoshitaka Tanemura, Yukio Mituyama, Yoshimi Asai, Yasuhisa Nakamura and Takao Onoye .....	<b>15</b>

## CHARACTER EMOTION DETECTION

<b>Text-to-Emotion Analysis Engines -Theory and Practice</b> David John, Anthony Boucouvalas and Zhe Xu.....	<b>23</b>
<b>A Text-Based Synthetic Face with Emotions</b> Siska Fitrianie and Leon J.M. Rothkrantz .....	<b>28</b>
<b>Using a sparse learning Relevance Vector Machine in Facial Expression Recognition</b> W.S. Wong, W. Chan, D. Datcu and L.J.M. Rothkrantz .....	<b>33</b>

## NETWORK MANAGEMENT OPTIMIZATION

<b>Variability of Internet Traffic in Multiple Time Scales and its Relevance for Quality of Service</b> Gerhard Haßlinger .....	<b>41</b>
<b>Mill: Scalable Area Management for P2P Network based on Geographical Location</b> Matsuura Satoshi, Fujikawa Kazutoshi and Sunahara Hideki.....	<b>46</b>

## CONTENTS

### **Effect of QOS Algorithms on ATM Switches**

John D. Garofalakis, Dimitrios S. Goulas and Vassilis D. Triantafillou .....53

### **Hardware/Software Co-Design for H.264/AVC Intra Frame Encoding**

Jan De Cock, Stijn Notebaert, Peter Lambert and Rik Van de Walle .....56

## **MEDIA APPLICATIONS IN EDUCATION**

### **Augmented Instructions for Learning Molecular Structures**

Kikuo Asai, Tomotsugu Kondo, Hideaki Kobayashi and Norio Takase .....63

### **The Application of Streaming Media Technology in Educational Software**

Sid Satbhai and Charles A.P.G. van der Mast .....69

### **Design and Evaluation of a VRET System for Agoraphobia**

Charles A.P.G. van der Mast and Frans S. Hooplot .....77

## **MEDIA APPLICATIONS IN BUSINESS**

### **A Workflow Model for Supporting Legislation Changes**

Elias A. Hadzilas .....87

### **Enhancing Knowledge Management with Business Intelligence –**

#### **A Case Study**

Kay Grebenstein and Stephan Kassel .....92

### **A Formal Method for Modeling and Evaluation of Protocols of Electronic Documents Transfer and their Security on the Web**

Gilles Eberhardt, Ahmed Nait-Sidi-Moh and Maxime Wack .....98

### **Mobile Agent Based Large Scale Collaborative Virtual Environment System**

Qingping Lin, Liang Zhang, Irma Kusuma and Norman Neo .....105

## **SEMANTIC WEB**

### **On the Requirements to the Methods for Web Service Composition**

Hristina Daskalova and Tatiana Atanasova .....115

### **An Approach for Semantic Web Service Composition**

Tatiana Atanasova and Hristina Daskalova .....122

## CONTENTS

### **Using Case Based Reasoning for Creating Semantic Web Services: An Infrawebs Approach**

Gennady Agre.....130

### **Combination of Semantic Web Services by the Contrivance of the current WSMO Specification**

Vladislava Grigorova.....138





# **SCIENTIFIC PROGRAMME**



# **DATA SEARCH**



# SEARCHING SEMI-STRUCTURED DATA USING LANDMARKS

Andrew Davison  
Dept. of Computer Engineering  
Prince of Songkla University  
Hat Yai, Songkhla 90112, Thailand  
E-mail: ad@fivedots.coe.psu.ac.th

## KEYWORDS

data extraction, pattern matching, semi-structured data

## ABSTRACT

This paper introduces landmark search operators for extracting data from poorly formatted Web pages, plain text files, and XML/SGML documents lacking grammars. The emphasis is on ease of use, and a fast, simple implementation, which can be readily ported to a wide variety of host languages. There are two main operators: one using unique textual landmarks to divide text regions into smaller regions suitable for further search, and an operator that searches for XML/SGML tag pairs, and returns the matches as regions.

## INTRODUCTION

Our aim is to create a simple, efficient set of methods for document search and information extraction based on finding *landmarks* in semi-structured data. Semi-structured data includes: Web pages marked up with (often incorrect) HTML, ASCII files, and XML documents lacking schema.

Although grammars exist for HTML (e.g. XHTML (XHTML 2005)), the sad reality is that most Web pages would fail a parsing test. Nevertheless, pages returned from the same service often exhibit similar formatting, (e.g. book pages from Amazon, sale item pages from eBay, departmental home pages). This is either because the pages are generated dynamically by server-side scripts using common templates, or because the host organization requires certain information to appear in certain places on a page. In other words, although the pages may not be grammatically correct, they do contain unique formatting information, such as titles, section heading, and indenting.

An ASCII file is often treated as a stream of characters (or bytes), as typified by UNIX tools. However, text files also contain formatting elements, such as titles, headings, and indenting. In common with Web pages, ASCII files developed for a single site (e.g. a library's book catalog, a school's class timetables) utilize common layout rules.

The general point is that semi-structured data which fails grammar or scheme validation, or even lacks a grammar, can still be searched by using the data's formatting elements. These elements may be recurring strings (e.g. "Section" at the start of each section), or patterns of white

space (e.g. two newlines at the end of each 'paragraph'). We call these elements *landmarks*.

Although landmark search is aimed at data extraction from Web pages and text files, it's also useful for XML/SGML files. The markup tags can be treated as landmarks, allowing landmark search to be employed instead of grammar-based techniques.

In the next section, we introduce the basic landmark operations. The third section contains examples showing how landmark search can be applied to Web pages, text files, and XML documents. The fourth section compares our work with region algebras and regular expressions. The final section draws some conclusions.

## LANDMARK SEARCH OPERATIONS

The two basic search operations are `match3()` and `tagMatch4()`. They both treat a document (be it a Web page, text file, or XML document) as a region. `match3()` utilizes landmarks patterns to search through the region, extracting smaller regions delimited by the matching landmarks. `tagMatch4()` performs a similar kind of search but using XML-style tag pairs. A support class, called `RegionIterator` (a Java iterator), can employ `match3()` or `tagMatch4()` to search repeatedly through a region for matches.

### The `match3()` Operation

`match3()` carries out a linear search from the start of a region to find a landmark that matches the operation's start landmark pattern. The search then switches over to looking for a landmark that matches the operation's end landmark pattern. The result is a *region* separated into three parts: the left, matching, and rest regions. The matching region is usually of most interest since it lies between the start and end landmarks.

The search result is shown in Figure 1.

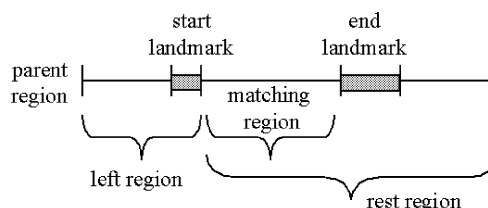


Figure 1: A Landmark Search using `match3()`

Landmark searches can be applied to the component regions, to 'zoom in' on areas of interest. When the information has been located, the region can be converted into a string, and manipulated using operations available in the host language.

`match3()` can be defined more formally:

`pr.match3(slp, elp)` returns  $\{lr, mr, rr\}$

where

`pr` is  $\{a_0..a_i, sl_0..sl_k, b_0..b_l, el_0..el_m, c_0..c_n\}$ ,  
the parent region,

`slp` matches  $sl_0..sl_k$ , the start landmark,  
and does not match anything in  $a_0..a_i$ ,  
`slp` is  $slp_0..slp_k$ , the start landmark pattern,

`elp` matches  $el_0..el_m$ , the end landmark,  
and does not match anything in  $b_0..b_l$ ,  
`elp` is  $elp_0..elp_m$ , the end landmark pattern,

`lr` is  $\{a_0..a_i, sl_0..sl_k\}$ , the left region,  
`mr` is  $\{b_0..b_l\}$ , the matching region,  
`rr` is  $\{b_0..b_l, el_0..el_m, c_0..c_n\}$ , the rest region

Otherwise `pr.match3(slp, elp)` returns null.

A region is treated like a character sequence when being searched. The `".."` symbol denotes a subsequence of characters.

The meaning of the "matches" operation will vary depending on the implementation; our Java prototype uses string equality: `x..y` matches `s..t` if the two subsequences contain the same text.

As the definition suggests, `match3()` can use string operations to perform a linear search over the parent region. The algorithmic complexity depends on the length of the sequence, and the cost of the "matches" operation. In general, if the landmark patterns are sufficiently small, then the algorithmic cost is on average  $O(n)$ , where  $n$  is the length of the parent region.

### The tagMatch4() Operation

`tagMatch4()` searches for tag pairs (e.g. `<P>` and `</P>`). On the face of it, this operation seems unnecessary since the tags could be represented by landmarks, and so found using `match3()`. However, there are two reasons for employing a separate method.

The first is that a start tag (e.g. `<title>`) may contain attributes, and we want to record them in an attribute region. The other reason is that tags may be nested (e.g. a `<ul>` list may appear as an item inside another `<ul>` list). The search ignores nesting, and finds the end tag which is at the same 'level' as the start tag.

`tagMatch4()` carries out a linear search from the start of a region to find a start tag that matches a supplied tag pattern. If the start tag contains attributes, these are stored in an attribute region. Then the search switches over to looking for an end tag that matches the tag pattern, with the proviso that the end tag must occur at the same level as the start tag.

The result is a parent region separated into four smaller regions: the left, matching, rest, and (potentially empty) attribute region.

Figure 2 shows the search graphically.

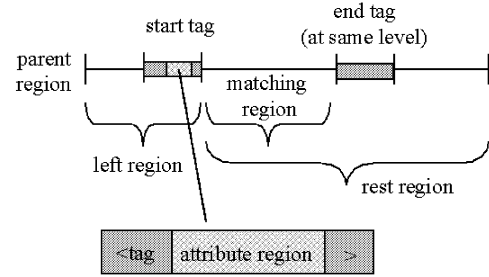


Figure 2: A Tag Search using `tagMatch4()`

Any of the four regions can be searched further by applying `match3()` or `tagMatch4()` to it.

`tagMatch4()` can be defined more formally:

`pr.tagMatch4(tag)` returns  $\{lr, mr, ar, rr\}$

where

`pr` is  $\{a_0..a_i, st_0..st_k, b_0..b_l, et_0..et_m, c_0..c_n\}$ ,  
the parent region,

( `"<"+tag+">"` matches  $st_0..st_k$ , the start tag,  
and `ar`, the attribute region, is null ) or

( `"<"+tag+" "+ ar_0..ar_p + ">"` matches  $st_0..st_k$ ,  
and `ar` is  $\{ar_0..ar_p\}$  ),

`tag` is  $t_0..t_r$ , the tag pattern,

the tag pattern does not match anything in  $a_0..a_i$ ,  
counter = 0,

`startTag` is `"<"+tag+">"` or `"<"+tag+" "` and  
`endTag` is `"</"+tag+">"`,

counter += count of `startTag` matches in  $b_0..b_l$  –  
count of `endTag` matches in  $b_0..b_l$ ,

`endTag` matches  $et_0..et_m$ , the end tag,  
and counter == 0

`lr` is  $\{a_0..a_i, sl_0..sl_k\}$ , the left region,

`mr` is  $\{b_0..b_l\}$ , the matching region,

`rr` is  $\{b_0..b_l, el_0..el_m, c_0..c_n\}$ , the rest region

Otherwise `pr.tagMatch3(tag)` returns null.

As in `match3()`, "matches" uses simple character comparisons between the tag pattern and the start and end tags.

The notion of level is captured with a counter which records the number of start and end tags encountered between the matching start tag and its corresponding end tag. The counter is incremented when another matching start tag is identified, and decremented when an end tag is encountered. The matching end tag is at the same level as the original start tag when the counter returns to 0.

The complexity of `tagMatch4()` depends on the length of the sequence, and the cost of the "matches" operation. The tag patterns are sufficiently small that the algorithmic cost is on average  $O(n)$ , where  $n$  is the length of the parent region.

## The RegionIterator Class

The landmarks operations are not intended to be a complete notation or language for text extraction. The host language is expected to have the usual control structures for looping, switching, and recursion, and (rudimentary) support for strings.

For example, repeated search through a parent region, looking for every matching region, can be coded using a while-loop and `match3()` (or `tagMatch4()`). A fragment of pseudo-code illustrates the idea:

```
Region r = /* region to be searched */
String slp, elp =
    /* start and end landmark patterns */

while (r.match3(slp, elp) ==
    {left, matching, attribute, rest}) {
    // use matching region...
    r = rest; // examine rest of region
}
```

However, searching for all the regions that match landmark (or tag) patterns is such a common task, that we have packaged it up inside a `RegionIterator` class. Several examples of its use appear in the next section.

## SEARCH EXAMPLES

This section contains three Java examples using the landmark operations: details are displayed about a specified Amazon.com book, price information is extracted from a text file of airline fares, and statistics are collected from an XML version of *Macbeth*.

### Extracting Amazon Book Details

Given a book's ISBN number, details about its title, prices, reviews, and sales ranking are extracted from the Amazon.com page for the book, and printed to standard output. Typical output is:

```
Retrieving Amazon's page for 0596007302
Accessing URL:http://www.amazon.com/
exec/obidos/tg/detail/-/0596007302
Title: Amazon.com: Books: Killer Game Programming
in Java
List Price: $44.95; Amazon Price: $29.67
```

```
Star Rating: 4-5; No. of Reviews: 3
Sales Rank: 6,236
```

The top-level region for the book's Web page is created, then searched in various ways:

```
String AMAZON_URL =
    "http://www.amazon.com/exec/obidos/tg/detail/-/";

System.out.println("Retrieving
    Amazon's page for " + isbn);
Region topR = new Region(AMAZON_URL + isbn);

System.out.println("Title: " +
    topR.tagMatch("title"));
showPrices(topR);
showReviewInfo(topR); // see below
showRank(topR);
```

The title is obtained by searching for the tag pattern "title". This was determined by looking at the source code for several Amazon book pages, and noting that their titles were always wrapped in a "title" tag pair.

The version of `tagMatch()` used here only returns the matching region, which is cast to a string in `System.out.println()` by `Region's toString()` method.

Amazon summarizes review details with a star rating (out of 5), and the number of reviews. This part of a book's Web page has the format:

```
 based on 3
reviews.
```

The long URL always ends with "common/customer-reviews/" and a GIF file for the stars image. The surrounding text contains many tables, links, comments, fragments of JavaScript, and white space. The search can ignore all of this by focussing only on the unique landmarks in the source fragment:

```
private void showReviewInfo(Region r)
{
    Region reviewsRegion =
        r.match("common/customer-reviews/", "review");
    Region starsRegion =
        reviewsRegion.match("stars-", ".gif");
    Region numReviewsRegion =
        reviewsRegion.match("based on ", " ");

    System.out.println("Star Rating: " +
        starsRegion + "; No. of Reviews:" +
        numReviewsRegion);
}
```

`reviewsRegion` is created with `match()`, a simpler version of `match3()`, which only returns the matching region between the two landmark patterns. For the example above, `reviewsRegion` will be:

```
{stars-4-5.gif"
height="12" border="0" width="64" /> based on 3 }
```

There is a space after the '3' at the end of the region.

starsRegion gets {4,5} from reviewsRegion, while numReviewsRegion extracts {3}.

The approach used by showReviewInfo() is quite common: first get *fairly* close to the required information with a region that cuts away *most* of the irrelevant data. This region should utilize landmarks which are unique across the entire document. The data is obtained in the second stage, using local landmarks next to the information, which only have to be unique within the region (e.g. in reviewsRegion).

## Looking for Airfares

We want to retrieve the cost of a roundtrip flight between two cities from "airfares.txt". The information for a city is formatted like the following example:

```
Roundtrip Fares Departing From BOSTON, MA To
-----
      $209    INDIANAPOLIS, IN
      $189    PITTSBURGH, PA
```

The collection of roundtrip fares for a city start with the "Roundtrip Fares" heading, a dotted line, then multiple price lines. The lines end with two newlines before the next city collection.

The first step is to iterate through the city information until we find the desired 'from' city:

```
Region topR = new Region("airfares.txt");
showTripPrice(topR, "PHILADELPHIA", "PITTSBURGH");

private void showTripPrice(Region topR,
                           String from, String to)
// show price of flight from-->to
{
    RegionIterator tripRegions =
        new RegionIterator(topR, "Roundtrip", "\n\n");
    // iterate through the trip regions
    while (tripRegions.hasNext()) {
        Region tripRegion = (Region)tripRegions.next();
        Region fromRegion =
            tripRegion.match("From ", " ", "");
        if (fromRegion.contains(from)) {
            // this trip is about <from>
            showToPrice(tripRegion, from, to);
            return;
        }
    }
    System.out.println("No fares from " + from);
} // end of showTripPrice()
```

The RegionIterator repeatedly searches for the landmarks "Roundtrip" and "\n\n" which delimit the collection of roundtrip fares for a city. Each call to next() returns the next collection, storing it in tripRegion. The 'from' city is extracted by pulling the region between "From " and " ", from tripRegion. This corresponds to {BOSTON} in the example above. If this is the desired city then showToPrice() looks at each price line to find the city we are interested in. This requires another RegionIterator:

```
RegionIterator priceRegions =
    new RegionIterator(tripRegion, "$", " ", "");
```

This iterator extracts the price and 'to' city information from each price line. For the fares table above, toRegions will deliver {209 INDIANAPOLIS} and {189 PITTSBURGH}. This example shows the utility of the RegionIterator for repeatedly applying a landmark pattern.

## How Worried is Macbeth?

We want to examine the play *Macbeth* by Shakespeare to discover just how many times Macbeth talks about "Birnam" and "Dunsinane" before his well-deserved end. This is admittedly rather silly, but it illustrates the ease of searching over a large XML file with complicated formatting, *without* employing a grammar/schema.

Landmark search means that most of the XML formatting can be ignored. The relevant parts for this task are the SPEECH tag pairs which wrap up speeches. A SPEECH block starts with a SPEAKER tag pair and one or more LINE tag pairs. For example:

```
<SPEECH>
<SPEAKER>MACBETH</SPEAKER>
<LINE>That will never be</LINE>
<LINE>Who can impress the forest,
        bid the tree</LINE>
:
<LINE>Of Birnam rise, and our
        high-placed Macbeth</LINE>
:
<LINE>Reign in this kingdom?</LINE>
</SPEECH>
```

The code iterates through each speech block, and, if the speaker is Macbeth, records the number of occurrences of the words "Birnam" and "Dunsinane":

```
Region topR = new Region("macbeth.xml");
Region speechRegion, speakerRegion;
String speechStr;
int numWords = 0;

RegionIterator speechIter =
    new RegionIterator(topR, "SPEECH");
while (speechIter.hasNext()) {
    // iterate through the SPEECH blocks
    speechRegion = (Region)speechIter.next();
    speakerRegion = speechRegion.tagMatch("SPEAKER");
    if (speakerRegion.contains("MACBETH")) {
        // is the speaker Macbeth?
        speechStr = speechRegion.toString();
        numWords += countString(speechStr, "Birnam") +
            countString(speechStr, "Dunsinane");
    }
}
System.out.println("No. words: " + numWords);
```

The RegionIterator uses a "SPEECH" tag pattern to iterate through the speeches. The "SPEAKER" text is pulled from the speech and if it contains "MACBETH", then the number of times that "Birnam" and "Dunsinane" appear in the rest of the speech are counted. Incidentally, the count for the play is 10.

countString() is a simple method (written by us) that uses String.indexOf() to search over the supplied string and count the number of times a given substring is found.



## COMPARISONS WITH OTHER APPROACHES

In this section we compare landmark search with region algebras and regular expressions.

### Region Algebras

Region algebras include PAT expressions (Salminen and Tompa 1992), overlapped lists (Clarke et al. 1995), and nested region algebras (Jaakkola and Kilpeläinen 1996). They treat a region as a contiguous portion of text, delimited by landmarks (also called anchors and match points). The algebras typically allow relationships to be expressed between regions, including 'precedes', 'follows', and 'contains', and support operations for creating region sets using union, intersection, and exclusion.

Landmark search employs a similar underlying model, but without sets and most of the region operators; this simplifies the model considerably. Also, tag-based landmarks are singled out for extra support, due to their importance.

WebL is a Web page manipulation language, with region algebras underpinning its text search capabilities (Kistler and Marais 1998). Its tag-based matching is similar to the version of tagMatch() that returns only a matching region. WebL employs regular expressions for matching against unstructured text.

### Regular Expressions

Regular expressions have problems searching over structured text, the foremost being their default use of leftmost longest match (Clarke and Cormack 1997). That search mechanism is a good choice when the text is being tokenized into numbers or words, but consumes too much data when applied to repeating text patterns. Regular expressions are also unable to count, making it impossible for them to deal with arbitrarily nested elements such as tags.

Regex libraries, as found in Perl 5 and java.util.regex, have introduced lazy quantifiers (Friedl 2002), which can simulate simple forms of landmark search. However, the formulation also needs to employ other extensions such as a multiline mode, back references, word boundaries for repeated search, and numerical quantifiers. Even with these additions, regexs are still unable to correctly handle searches involving nested tags.

## CONCLUSIONS

Landmark search consists of two basic operations, match3() and tagMatch4(). match3() utilizes landmark patterns to identify regions within a parent region, while tagMatch4() uses tag pairs to find regions. Repeated application of these operations can be carried out using a RegionIterator class.

Although landmark search only employs a few operators, it is capable of extracting information from poorly formatted Web pages, plain text files, and XML documents lacking grammars or schemas.

The underlying aim of this work was to develop a small set of operations, that could be easily understood, readily added to a host language, and efficiently implemented. This contrasts with feature-rich alternatives, such as region algebras and regex packages.

A prototype Java implementation of the landmark operators (together with examples) can be downloaded from <http://fivedots.coe.pau.ac.th/~ad/landmarks>.

## REFERENCES

- Clarke, C.L.A. and Cormack, G.V. 1997. "On the Use of Regular Expressions for Searching Text", *ACM Transactions on Programming Languages and Systems*, 19, 3, 413–426, May.
- Clarke, C.L.A., Cormack, G.V., Burkowski, F.J. 1995. "An Algebra for Structured Text Search and a Framework for its Implementation", *The Computer Journal*, 38, 1, 43–56.
- Friedl, J.E.F. 2002. *Mastering Regular Expressions: Powerful Techniques for Perl and Other Tools*, O'Reilly and Associates, 2nd ed.
- Jaakkola, J. and Kilpeläinen, P. 1996. "Using sgrep for Querying Structured Text", Department of Computer Science, University of Helsinki, Report C-1996-83, November.
- Kistler, T. and Marais, H. 1998. "WebL - a Programming Language for the Web", In *Computer Networks and ISDN Systems, Proceedings of the WWW7 Conference*, 30, 259–270, April.
- Salminen, A. and Tompa, F.W. 1992. "PAT Expressions: an Algebra for Text Search", *Acta Linguistica Hungarica*, 41, 1–4, 277–306.
- XHTML 1.0. 2005. "The Extensible HyperText Markup Language", 2nd ed., W3C, <http://www.w3.org/TR/xhtml1/>, August.

# A MODEL FOR EVOLUTIONARY MULTIMEDIA OBJECTS

Georges Chalhoub, Richard Chbeir, Kokou Yetongnon  
Computer Science Department, LE2I – Bourgogne University  
BP 47870 21078 Dijon - France  
gchalhoub@bdl.gov.lb, {Richard.chbeir, Kokou.yetongnon}@u-bourgogne.fr

## ABSTRACT

In recent years, an increasing number of applications, such as GPS, time series, vehicles, humans, orbital objects and medicine field, deal with multimedia objects and moving objects. Depending on the applications, we distinguish between continuously moving objects, where the database is continuously updated, and discretely moving objects, where a limited number of states are recorded. A multimedia object moves and evolves in time. Several multimedia object types evolution can be distinguished such as physical, semantic, relational, etc. In this paper we present a multimedia data model  $M^2$  and its corresponding query model  $M^2Q$ . The extended model is able to represent, on the one hand, the several features of multimedia objects and, on the other hand, consider the continuous, and the discrete, object evolution and movement in time.

## KEYWORDS

Multimedia temporal model, Multimedia temporal query

## INTRODUCTION

Multimedia data have become available at an increasing rate, especially in digital format. There has been a tremendous need for the ability to store, query and process non-traditional data in a wide variety of applications. Multimedia data content are described through several feature layers: Low level or physical features such as color and texture, semantic or metadata features, and high level or relational features. In recent years, an increasing number of applications such as GPS, time series, vehicles, humans, orbital objects, video surveillance and medicine field have two major focuses: Multimedia data and moving objects.

Moving objects are objects whose positions changes and states evolve in time. This evolution is widely studied in several works defining moving objects trajectories. (Abdessalem et al. 2000). Two types of moving objects can be distinguished: Continuously moving objects and discretely moving objects (Bohlen and Gutting 2000). The former type has been used in several video database management systems (VDBMS) (Ulusoy and Donderler 2004) providing an integrated support for spatio temporal queries. In the second type, databases cannot be continuously updated, so a limited number of states are recorded. Real states between two reported ones are not exact but they can be *estimated*. For this reason, several uncertainty models have been proposed and can be applied to each point of a moving object.

Moving objects queries contains spatio-temporal relations, numerical values expressing the velocity, the acceleration and other additional values.

Due to their multi-features representation, several types of multimedia object evolution can be distinguished:

- Physical evolution: Is a state modification of one or more physical feature in time. For instance, a multimedia object may undergo a shape deformation during a certain period of time. For example human cells deformation is observed to study a tumor situation.
- Semantic evolution: Is a state modification of one or more semantic features in time. The role of a multimedia object may changes and evolves in time. For example an assistant manager may become manager in two years.
- Relational evolution: Is a modification of the relations between objects. In a video scene, relations between moving regions evolve on time. For example in a formula 1 Grand Prix, two different situations may occur at two different times  $t_0$  and  $t_1$ : *Renault* BEHIND *Ferrari* and *Ferrari* BEHIND *Renault*.
- Position evolution: A multimedia object displaces and changes position, velocity and direction continuously or discretely.

The existing approaches does not take into considerations all of the above evolution types. We do believe that there is a real need of a multimedia data model and multimedia query model, able to represent both, the several features of multimedia objects, and the evolutionary nature of objects in time.

In (Chalhoub et al. 2004), the authors presented a multimedia data model and a multimedia query model considering the several types of features. However, this model does not take into consideration the time evolution of objects. In this paper, we extend the  $M^2$  data model to represent the time evolution of a multimedia object by integrating the time component. We also extend the  $M^2Q$  query model of (Chalhoub et al. 2004) in order to express moving objects queries and features evolution. These queries contain spatio-temporal relations, numerical values expressing the velocity, the acceleration and other additional values. The extended models of the multimedia data model  $M^2$  and the query model  $M^2Q$  are able to support *continuously* and *discretely* moving multimedia objects.


The paper is organized as follows: In the next section, we give the three motivation examples, then we present some interesting works related to moving objects, and in the sections that follow, we explain our proposal with validation examples.

## MOTIVATION

To motivate our proposal, we give the following three examples of multimedia evolutionary objects:

### Example 1

A police-monitoring camera takes continuously video shots of the circulation on main roads. The captured videos are stored in a database. The police can query the database to extract useful information about suspect circulations. For example:

*Q1: Find all videos containing a car like  going 3 miles toward the east with a velocity of 100 miles/h between 7:00 AM and 7:15 AM and then 2 miles toward the northeast between 7:16 AM and 7:30 AM*

### Example 2

In cancer research field (BC cancer research), cytotechnologists are interested about the localization and the states of the cells in time. They study the tissue growth process where each cell is defined by some internal states which include: its capacity to divide, its position in the tissue, its age, and its displacement capacity (speed at which a cell may migrate from the basal lamina to the lumen). In addition, they study the collision that may occur with its neighbors while moving from a location to another in time.

Several queries can be formulated to obtain statistics about cells having similar behavior inside the tissues in time. These statistics may help physicians to expect the evolution of the metastasis of the cancer. For example:

*Q2: Find all cancers cells images having the following trajectory inside the tissues between a start time  $t_0$  and  $t_0 + 10$ :*

- At time  $t_0=1$ , the cancer in position (3;5) with a velocity 12
- At time  $t_1=2$  the cancer in position (3.3; 5.9) with a velocity 13
- ...

### Example 3

To understand the effect of a nuclear weapon during an attack, one measures the width and height of the nuclear cloud and studies its development in time after the burst. (global security). In fact, the development of nuclear clouds is divided into three stages: fireball, burst cloud, and stabilized cloud. The duration of each stage depends on the yield of the weapon. The development of the clouds in time may give an idea about the type and the effect of the weapon.

Based on archive information stored in a multimedia database, the scientists may ask several questions about the cloud behavior in time. For example:

*Q3: Find all nuclear weapon type inducing a mushroom cloud development like the following images:*

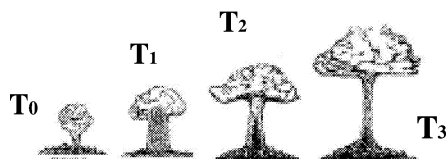


Figure 1: Nuclear cloud evolution in time.

In the first example, the movement of the red car is continuously updated in the database by the monitoring camera. The query contains physical, semantic and temporal features in addition to spatial relations. In the

second example, the cancer evolution is updated discretely and the query looks for images containing cancers evolving in time, with similar given trajectory. The third example describes the physical evolution of the object in time.

## RELATED WORK

Moving objects and trajectories applications are widely studied in the literature. Depending on the application domain, models for continuously or discretely moving objects have been provided. In this section we give an overview of each of the two moving object types modeling.

### Continuously Moving Objects

These objects are stored in a continuously updated database and has been used in several video database management systems (VDBMS)

Several works and applications on continuously moving objects are presented in the literature. In (Mokhtar and Ibarra 2002), the authors present the object location  $X$  in the space  $R^n$  using the linear function  $X = At + B$  where,  $A$  and  $B$  are two vectors in the space  $R^n$ . A trajectory is expressed by disjunction of linear functions. For example, an aircraft movement is represented by the disjunction of the two following functions:

$$X = (2, -1, 0)t + (-40, 22, 30) \quad 0 \leq t \leq 21$$

$$X = (0, -1, -5)t + (2, 23, 35) \quad 21 \leq t \leq 22$$

This disjunction represents the aircraft movement to the south and then changes direction at a time 21. This model is able to answer queries concerning the past, the present and the future aircraft movement in a given location.

In (kubica et al. 2004), the authors study the problem of intersection between a trajectory and a region. They define the trajectory (track) according to temporal space of  $D$  dimensions. Two types of trajectories are define: quadratic trajectory ( $g(t) = at^2 + bt + c$ ) and linear trajectory ( $g(t) = at + b$ ).

An interesting application on trajectories, concerning traffic calculation, is presented in (Nirvana 2002). The essential goal of this application is to transform the flux of raw data in a trajectory and then to make an aggregation of similar trajectories to form one of it. They define the spatial unit the minimal zone of interest in trajectory. Based on this definition, they propose a method for trajectory aggregation.

The techniques of the differential geometry and multi-dimensional vectors are used to represent the moving objects in (Xu et al. 2001). The authors define a moving point as a function of vectors  $P(t) = (p_1(t), p_2(t), \dots, p_n(t))$  where each  $p_i(t)$  is a continuous function. The velocity of  $P$   $vel(P)$  is defined as the first derivative of  $P$  and the acceleration  $acc(p)$  is the derivative of  $vel(p)$ .

Query processing for video databases is widely studied in the literature. In (Ulusoy and Donderler 2004) the authors define queries to retrieve moving objects segments within a video clip. The moving object trajectory is modeled as  $tr(v, \phi, \psi, \kappa)$  where:

- $v$ : Is the object identifier
- $\phi$ : Is the list of directions  $[\phi_1, \phi_2, \dots, \phi_n]$
- $\psi$ : Is the list of displacements of the object  $\psi = [\psi_1, \dots, \psi_n]$  where  $\psi_i \geq 0, i=1..n$
- $\kappa$ : Is the list of time intervals displacement.

In this model, the movement object displacements are recorded in the database with their corresponding directions and time intervals. The object trajectory query, as described

in this work, takes into consideration the displacements and the direction lists. In addition, the authors define a trajectory matching measure based on displacement and directional similarities. One more work on video modeling is presented in (Szafrom et al. 1996) where the Euclidian displacements and the directions of moving objects constitute the trajectory elements.

### Discretely Moving Objects

Due to several factors, like bandwidth limitation and sensor sensitivity, a limited number of states of discretely moving objects are recorded in the database. Several discrete models have been proposed in the literature.

Individuals' geo-spatial movement is studied in (Hornsby and Egenhofer 2002) where an individual's movement is presented by a *lifeline* including the different location visited by the individual with corresponding times. The basic element of this line is a triplet <Id, Location, Time>. This triplet expresses an individual's location at a certain time. In his displacement, from one point P1 ( $x_0, y_0, t_0$ ) to another point P2 ( $x_1, y_1, t_1$ ), an individual can visit several locations. This movement is modeled by two half cones Bead lifeline. The first half cone represents the progress of the individual going from location P1. The second represents the individual's displacement toward P2. This model is able to answer several queries. For example:

- Is it possible that an individual A was in the position ( $X_A, Y_A$ )?
- For how long B would have remained in ( $X_A, Y_A$ )?
- In which location the individual C can be at time t?

Another trajectory modeling is presented in (Nirvana 2005). In this work the authors design the trajectory as a sequence of positions <time, location>. They define the similarity between two trajectories according to time and location. They identify three classes of similarity of the trajectories: Temporal, spatial and spatio-temporal similarity.

Uncertainty model has been the focus of several studies on discrete moving objects. In (Kalashnikov et al. 2003), the authors define an uncertainty model and classify probabilistic queries. In (Bailey et al. 2004) one presents update policies in uncertainty model. In addition the authors define the linear interpolation between reported locations. Similarly, future locations are extrapolated by extending the most recently reported velocity. Another interesting study on moving objects and trajectories is presented in (Byunggu 2005). This work deals with CCDO (continuously changing data objects) due to the variety of spatio-temporal application such as vehicles, humans, economic indicators, etc. The authors define the concepts of CCDO database as follows:

- CDDO: consists of several temporal properties and sequences of trajectories
- Trajectory: consists of dynamics and a function  $f: \text{time} \rightarrow \text{snapshot}$ .
- Snapshot: represents a probability of every state in the data space at a specific point in time
- State: is a couple <P, O> where P is a position and O is an optional property list like velocity, acceleration, etc.
- Dynamics: represents the upper and lower bounds of the optional properties of all states in the trajectory

The representation of moving objects in GIS is addressed in (Abdessalem et al. 2000), where the discrete nature of instruments, the storage system limitation and bandwidth lead to representation problems of continuous movements. In this work, the authors formally present the coverage of all possible trajectories between two consecutive discrete observations. They define the movement of the object as a finite subset of the Cartesian product  $T \times G$  where  $T = \{t_1, t_2, \dots\}$  a set of time instants and  $G = \{g_1, g_2, \dots\}$  a set of geographical constants for points, lines and polygons. Using one more set  $N = \{n_1, n_2, \dots\}$  of integers representing speed and acceleration, the authors define several movement operations like projection, restriction and topological-temporal operations. These operations are useful in spatio-temporal queries.

### Conclusion

All the works in this section describe moving objects models and queries. Some of these works deal with continuously moving objects while others deal with discretely moving objects and corresponding queries. Moving multimedia objects may belong to both types. Therefore, there is a real need of a multimedia data and query models taking into consideration the evolution of semantic, physical and spatial features in time.

### PROPOSAL

In this section, we extend the  $M^2$  multimedia data and query models presented in (Chalhoub et al. 2004), by integrating the time component to express the multi criteria evolution of a multimedia object in time. We give query examples to show how the extended model is able to support continuously and discretely moving multimedia objects.

### $M^2$ Multimedia Data Model Extension

In (Chalhoub et al. 2004), the authors presented a multimedia data model  $M^2(\text{id}, O, F, A, R)$  able to improve multimedia management systems by providing a modeling framework to express the properties of data items and the meta-data at different levels. The key feature of the model is that it captures in a single modeling the low-level features, the structural and semantic properties, and the relationship descriptions.

As we mentioned above, multimedia object features evolve in time. To express this evolution, we integrate the time component T. This component represents a time interval expressed as:  $T = [t_s, t_e]$  where  $t_s$  is the start time and  $t_e$  is the end time. T can be reduced to one point when  $t_s = t_e$ .

To represent the status of the object at each time instant, we extend the model  $M^2(\text{id}, O, F, A, R)$  to the model  $M^2(\text{id}, O(t), F(t), A(t), R(t))$  Where:

- Id is a unique identifier associated to an object whatever the time. It is used to differentiate an object from any other object. Id never evolves with time
- $O(t): (O, T)$  is a reference to the raw data of the object at time interval T. The raw data of the object may evolve with times. The object O may undergoes a shape or a color evolution
- $F(T): (F, T) = ((\text{Descriptor}, \text{Model}, \text{Value}), T)$  is a feature vector representation of the object O at time interval T. In this component, we represent the physical evolution of the object. For instance, a colour histogram may have two different values in two different time intervals. We integrate also in this

component motion values expressing the displacement value and several numerical values expressing the velocity, the acceleration, and etc. These values are defined and integrated as follows:

- Dep: is the displacement normalized value or distance during the time interval T. Dep equal zero in the case of one point time interval T
- Pos: Express the position of the object in the space at the time ts of the interval T
- Op: Express numerical optional values like velocity and acceleration of the object
- A (T): (A, T)= ((ES, Sem\_F), T) contains meta-data at time T. In this component, we can also represent the semantic evolution of the object. For instance, the geographical location of the same moving objects may change between two time intervals without any physical evolution. For example the objects moves from a city to another.
- R (T): (R, T)= (: (S1 = {id<sub>i</sub>, i = 1..n}, S2 = {id<sub>j</sub>, j = 1..m}, Re = {Rel<sub>k</sub>, k = 1..p}), T) This component represents zero or more relationships between objects at time T. The sets S1 and S2 can be empty when they represent the object itself and to express relations between the object itself in two different time intervals. We integrate in this component the value Rel. This value represents the relation (directional or topological) between the object during [tsi, tei] and the same object during the previous interval [tsi-1, tei-1].

#### Examples

Let us study the evolution of the object image I presented in Figure 1 of Example 3. The nuclear cloud evolves in time as: At time T<sub>0</sub> the object represent the *fireball*, at times T<sub>1</sub> and T<sub>2</sub>, it represents the *burst cloud* with height evolution between T<sub>2</sub> and T<sub>3</sub>. And at time T<sub>4</sub> the object represents the *stabilized cloud*. We represent the object I evolution as follows:

- I.F(T) component at time T=T<sub>0</sub>:
  - Descriptor: *pos*
    - Model: *GPS*
    - Value: 100 ; 33;14
  - Descriptor: *Height*
    - Value: 15
- T: T<sub>0</sub>
- I.A(T) component at time T=T<sub>0</sub>:
  - Sem\_F.Type : *Model*
  - Sem\_F.Desc: *Fireball*
- T: T<sub>0</sub>
- I.F (T) component at time T=T<sub>1</sub>:
  - Descriptor: *pos*
    - Model: *GPS*
    - Value: 100 ; 33;14 (same)
  - Descriptor: *Height*
    - Value: 30
- T: T<sub>1</sub>
- I.A(T) component at time T=T<sub>1</sub>:
  - Sem\_F.Type : *Model*
  - Sem\_F.Desc: *burst cloud*
- T: T<sub>1</sub>
- I.R (T) component at time T=T<sub>1</sub>
  - ({}, {}, {"Double Height"})

○ T: T<sub>1</sub>

In this component, we express the height evolution of the object I between the two times instant T<sub>0</sub> and T<sub>1</sub>. The *Double Height* relation express the evolution of the nuclear cloud form a stage to another

In the above example, we represent a discrete moving object where the database contains the status of the object at discrete times: T<sub>0</sub>, T<sub>1</sub>, T<sub>2</sub> and T<sub>3</sub>. In the following we show how our model is also able to represent continuously moving of the object M (*red car*) of the example 1.

- M.F(T) at time T=[7:00 , 7:15]
  - Descriptor: *Color*
    - Model: *RGB*
    - Value: [255, 1;10]
  - Descriptor: *OP*
    - Model: *Velocity mile/H*
    - Value: 100
- T=[7:00, 7:15]
- M.A(T) component at time T=[7:00 , 7:15]
  - Sem\_F.Type : *Dep in miles*
  - Sem\_F.Desc: *3l*
- T=[7:00, 7:15]
- M.R(T) component at time T=[7:00 , 7:15]
  - ({}, {}, {"East"})
  - T=[7:00, 7:15]

The value of T can be different in two different components. Indeed, an object can evolve physically at a time T without evolving semantically.

The extended model is able to support and to represent the two types continuously and discretely moving objects. Indeed, in the first type, the movement is represented at each time interval [ts, te] with the displacement, relations and other movement related information. While in the second type, the time interval is reduced to one point (ts = te). The position of the object at time is recorded and the displacement value *Dep* is null.

In the next section we show how we extend the multimedia query model M<sup>2</sup>Q. The extended model will be able to represent multimedia queries able to deal with multimedia moving and evolving objects.

#### M<sup>2</sup>Q Multimedia Query Model Extension

In (Chalhoub et al. 2004), the authors define a query model M<sup>2</sup>Q (idq,Oq,Fq,Aq,Rq) of the data model M<sup>2</sup>. They consider several types of query: metadata-based, content-based and multicriteria-based query. In this section, we extend the M<sup>2</sup>Q to represent also multimedia queries able to deal with multimedia moving and evolving objects. We describe the extended query model as follows:

M<sup>2</sup>Q(idq,O<sub>q</sub>(T), F<sub>q</sub>(T),A<sub>q</sub>(T),R<sub>q</sub>(T)) where:

- T, integrated in each component, is a set of N consecutive time intervals. T={T<sub>si</sub>, T<sub>ei</sub>} I=1..N
- In the component Fq(T), the motion values are represented as follows:
  - Dep: is a set of N consecutive displacement values. Each displacement corresponds to a time interval of the set T

- $Pos$ : is a of  $N$  positions in the space. Each position corresponds to the time  $T_{si}$  of each time interval in the set  $T$ .
- $Op$ : is a set of  $N$  numerical vectors expressing  $S$  numerical optional values like velocity and acceleration of the object. Each numerical vector corresponds to a time interval of  $T$
- In the  $R_q(T)$  component,  $Rel$  is a set of  $N$  consecutive relations. Each relation corresponds to a time interval of  $T$  and to a displacement of  $Dep$ .

The time intervals and motion values express the movement trajectory query. In continuously moving objects, a trajectory is expressed by a series of displacement and a series of relations over a series of time intervals.

In discretely moving objects a trajectory is expressed by a set of triplet  $(pos, t, Op)$  where  $t$  represents the time interval reduced to a point ( $T_{si} = T_{ei}$ ). It expresses the recorded time in the database where the object was in position  $pos$  with the numerical options  $Op$ .

*Examples:*

To better illustrate our model, let us study the following query examples:

To represent a query of continuously moving multimedia object, we consider the query Q1 of the Example1. We represent the query as follows:

Q1:  $(id_q, \text{video}, F_q.* = ((\{3, 2\}, \{\}, \{100, 100\}), \{[7:00, 7:15], [7:16, 7:30]\}), A_q.ES = \text{'video'}, R_q.* = (\{\}, \{\}, \{\text{east, north}\}, \{[7:00, 7:15], [7:16, 7:30]\}))$  where:

- In  $F_q(T)$  component:
  - $\{3, 2\}$  is displacement set
  - $\{\}$  Positions set
  - $\{100, 100\}$  is the velocity set
  - $\{[7:00, 7:15], [7:16, 7:30]\}$  is the consecutive time interval set.
- In  $R_q(T)$  component:
  - $\{\text{East, north}\}$  is the set of consecutive relations
  - $\{[7:00, 7:15], [7:16, 7:30]\}$  The time interval set corresponding to the consecutive relations set.

The query Q1 can be expressed in SQL as follows:

$SELECTT * \text{From } M^2(t) \text{ Where } Q1.O_q \text{ Similar to } M^2.O \text{ and } M^2.F \text{ contains( } (Dep= 3 \wedge OP=100 \wedge T=[7:00, 7:15]) \wedge (Dep= 5 \wedge OP=100 \wedge T=[7:16, 7:30]) \text{ and } M^2 A_q.ES= \text{'video'} \text{ and } M^2.R \text{ contains}(((\{\}, \{\}, \{\text{east}\}) \wedge T=[7:00, 7:15]) \wedge (\{\}, \{\}, \{\text{north}\}) \wedge T=[7:16, 7:30]))$

To represent a query of discretely moving multimedia object, we consider the query Q2 of the Example2. We represent the query as follows:

Q2:  $(id_q, , F_q.* = (\{\}, \{Pos_i \ i=1..N\}, \{vel_i, \ i=1..N\}, \{T_i, \ i=1..N\}), A_q.ES = \text{'X-Ray'}, \{\})$  Where :

- In  $F_q(T)$  component:
  - $\{\}$ : Is the displacement set
  - $\{Pos_i \ i=1..N\}$ : Is the consecutive positions set
  - $\{vel_i, \ i=1..N\}$ : Is the consecutive velocities set
  - $\{T_i, \ i=1..N\}$ : Is the corresponding point times set
- $R_q(T)$ : null

The query Q2 can be expressed in SQL as follows:

$SELECTT * \text{From } M^2(t) \text{ Where } M^2.F \text{ contains( } (pos= pos_i \wedge OP=vel_i \wedge T=T_i, \ i=1..N)) \text{ and } M^2. A_q.ES= \text{'X\_ray'}$ .

## CONCLUSION

In this paper we presented a multimedia data model and multimedia query model, able to represent both, the several features of multimedia objects and consider the evolutionary nature of objects. The defined model is able to present and query different type of multimedia object time evolution. In addition, our model is able to deal with continuously and discretely moving objects.

Due the several features types of multimedia object and the evolutionary nature of each feature, cooperative query may be useful to improve the users requests and result. Query rewriting and constraint relaxation is one of the cooperative queries techniques. In our future work, we propose a query rewriting method in the framework of multimedia moving objects and trajectories

## REFERENCES

- Abdessaïem T. Moreira J. and Ribeiro C. 2000. "Query operations for moving objects database systems". *The 8th ACM international symposium on Advances in geographic information systems*. ACM Press .103-122
- Bailey T. B. Yu, S.H. Kim and R. Gamboa. 2004. "Curve-based representation of moving object trajectories". *IEEE DEAS International Database Engineering and Application symposium*, 419-425.
- Bc cancer Research [www.bccrc.ca/ci/tm01\\_modelling.html](http://www.bccrc.ca/ci/tm01_modelling.html) (last visited :10/01/2006)
- Bohlen M. and Gutting R. 2000 "A foundation for representing and querying moving objects". *ACM Transactions on Database Systems*, No 25(1) 1-42.
- Byunggu Yu 2005. "A framework of processing spatio-temporal database queries with uncertainty". To appear 2006.
- Chalhoub G. Chbeir R. and K. Yetongnon 2004. "Towards Fully Functional Distributed MultiMedia DBMS". *Journal of digital information management, JDIM*, No 2 (3) 116-121.
- Global security, [www.globalsecurity.org](http://www.globalsecurity.org) (last visited: 10/01/2006).
- Hornsby K. and Egenhofer M. 2002. "Modeling Moving Objects over Multiple Granularities", *Annals of Mathematics and Artificial Intelligence* No 36 177-194
- Kalashnikov D. R. Cheng and S. Prabhakar. 2003. "Evaluating probabilistic queries over imprecise data". *International Conference on Management of Data, ACM SIGMOD* No 3, 551-561.
- Kubica J, A. Moore, and A.J. Connolly 2004. "Spatial Data Structures for Efficient Trajectory-Based Queries". *Tech. Report Carnegie Mellon University Robotics Institute*
- Mokhtar Su. and O. Ibarra 2002. "On moving object queries". *Symposium on Principles of Database Systems (PODS) SIGMOD-SIGART*188-198.
- Nirvana M. 2005. "Towards Database support for moving object data". *PhD thesis, Center for Telematics and Information Technology (CTIT)*, The Netherlands. 14 – 16
- Nirvana. 2002. "Aggregation and comparison of trajectories". *The 10th ACM International Symposium on Advances in Geographic Information Systems*, McLean, ACM-GIS. 6 p.
- Szafron D. Li J. and Ozsu T. 1996. "Modeling of moving objects in a video base management system". *Technical report, University of Alberta, Alberta Canada*.
- Ulusoy O. and Donderler M. 2004. "Rule-based spatio-temporal query processing for video databases". *VLDB Digital Object Identifier (DOI)*, No.13, 86-103
- Xu H., J. Su and O. Ibarra. 2001. "Moving objects: Logical relationships and queries". *The Seventh International Symposium on Spatial and Temporal Databases*, Redondo Beach, USA, SSTD, 3-19.

# CONTOUR-BASED GRAVITY CENTER EVALUATION OF CHARACTERS

Akio Kotani,<sup>†,††</sup> Yoshitaka Tanemura,<sup>†</sup> Yukio Mituyama,<sup>†</sup>  
Yoshimi Asai,<sup>††</sup> Yasuhisa Nakamura,<sup>††</sup> and Takao Onoye<sup>†</sup>

<sup>†</sup> Department of Information Systems Engineering, Osaka University, 1-5 Yamada-Oka, Suita, Osaka, 565-0871 Japan

<sup>†</sup> {kotani.akio, asai.yoshimi, nakamura.yasuhisa}@sharp.co.jp

<sup>††</sup> Platform Technology Development Center, SHARP Corporation, 2613-1 Ichinmoto-Cho, Tenri, Nara, 632-8567 Japan

<sup>††</sup> {kotani, tanemura.yoshitaka, mituyama, onoye}@ist.osaka-u.ac.jp

**ABSTRACT** In the ubiquitous society, text information is one of the most important communication methods and a variety of text information is always displayed on cellular phone, navigation system, DTV, etc. Therefore, it is strongly demanded to improve legibility of text on displays. However, typical design process of fonts for specific devices requires a long development period, whereas print type has been gradually developed in long history. In this paper, considering gravity center on characters, which makes a large impact to legibility, a contour-based evaluation method of gravity center has been proposed. Based on the proposed evaluation method, the correlation between the legibility and the dispersion of gravity center on characters has been illustrated.

## 1. INTRODUCTION

The technical progress of electronic display devices drastically improves a resolution of displays. These electronic displays are expected to offer higher legibility and higher quality of fonts, which is one of the most important factors of contents. However, since fonts have been designed manually for each display and for each size of characters, in case of embedded system equipments, the quality of fonts depends on designer's sensitivity and the enormous design costs are required. Furthermore, due to the subjective evaluation of quality, it is very difficult to maintain the quality of fonts, and to transfer the know-how of fonts design.

In order to reduce the design costs of fonts and to improve the quality of fonts, the quantification of design processes, which depends on human sense, must be performed. However, it is very difficult to quantify the beauty, which is one of the main quality factors, since it depends primarily on one's sensuousness or preference. On the other hand, the legibility of fonts is less influenced by the individual sense, and thus can be quantified as an indicator of quality.

Motivated by this tendency, in the process of designing fonts, we have considered gravity center on characters (S. Okada et al. 2002, A. Kotani et al 2003), which has a large impact to the legibility of fonts. Although font size and serif may also have some relation on the legibility of fonts (M. L. Bernard et al. 2003, A. Arditì and J. Cho 2005), gravity center plays more important role in case of Japanese and Chinese character, where many of strokes are drawn crowdedly.

However, conventional methods (M. Yoshimura and T. Iijima 1970) can not obtain tolerable accuracy of gravity center on characters (A. Kotani et al. 2003). Thus, we have proposed the

contour-based evaluation method (A. Kotani et al. 2004, 2005) of gravity center on characters. This method can greatly reduce the design cost of fonts as well as contributing to the stabilization of quality of fonts, mainly because gravity center on characters, which is originally evaluated by a subjective method, can be quantified. In order to speed-up the evaluation of gravity center, it is preferable to accomplish a series of processes in the proposed method automatically.

In this paper, we propose the modified contour-based evaluation method with convex hull algorithm. The proposed method can perform completely-automatic and high-accuracy evaluation of gravity center on characters, and can be embedded to font designing tools. Moreover, we apply the proposed method to legibility evaluation of fonts.

The paper is organized as follows. In Sect. 2, we describe subjective gravity center on characters. In Sect. 3, the contour-based evaluation method of gravity center on characters is proposed. Section 4 assesses the correlation between the legibility and the gravity center on characters. Finally this paper is concluded in Section 5.

## 2. GRAVITY CENTER ON CHARACTERS

As shown in Figure 1, gravity center of a character is the psychological balanced point of the character, which is usually to nonidentical center of body frame.

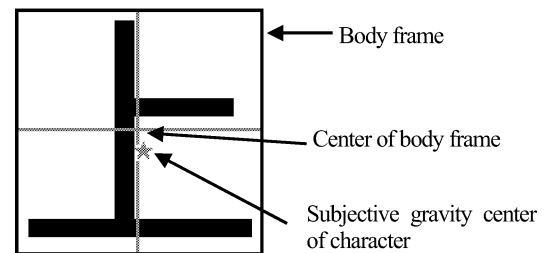


Figure 1: Subjective gravity center of character and center of character frame.

Figure 2 shows two typefaces on vertical and horizontal typesetting; typeface A is designed for both vertical and horizontal typesetting, while typeface B is designed for vertical typesetting. In case of vertical typesetting, the legibility of font depends on the horizontal dispersion of vertical lines passing through gravity center of characters. On the other hand, in case of horizontal typesetting, it depends on the vertical dispersion of horizontal lines passing through gravity center of characters. As

shown in Figure 2(a), in each typeface, the vertical lines, which are passing through gravity center of characters, almost form a straight line. This demonstrates that each typeface have high legibility in vertical typesetting. On the other hand, as shown in Figure 2(b), horizontal lines passing through gravity center of characters disperse in case of typeface B, which results in poor legibility.

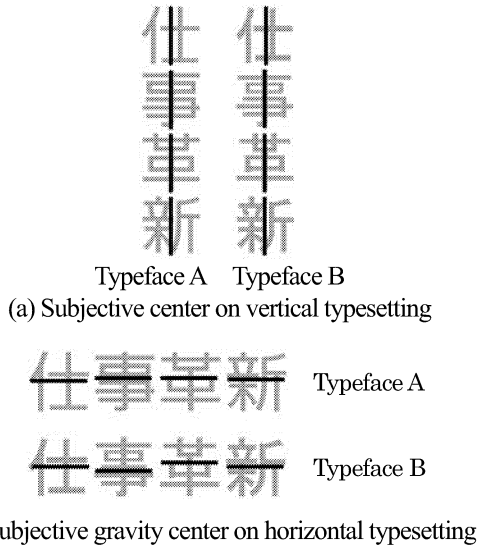


Figure 2: Subjective gravity center on character.

### 3. EVALUATION METHOD OF GRAVITY CENTER ON CHARACTERS

In this section, the circular frame-based subjective evaluation method of gravity center on characters (A. Kotani 2005) and the convex hull contour-based evaluation method of gravity center are to be described.

#### 3.1. SUBJECTIVE EVALUATION METHOD OF GRAVITY CENTER ON CHARACTERS BY USING CIRCULAR FRAME

Conventional subjective evaluation method of gravity center on characters, in order to achieve high accuracy, has to be performed by an evaluator who has adequate experience of font design. However, such an evaluator is very scarce, and even enormous evaluation cost, e.g. about 40 minutes per character, has been required for font design.

In order to reduce the evaluation cost for subjective gravity center on characters, we conduct hearings at the company with high degree of expertise and lot of know-how for lettering in the commercial printing. Herewith, we propose the subjective evaluation method of gravity center on characters by using circular frame. The procedure of proposed subjective evaluation method is as follows (see Figure 2).

- 1) A black circular frame with white background is displayed on the high-resolution LCD of a laptop PC. In this case, the circular frame has 450-dot diameter and 1-dot thickness, and the resolution of LCD is 178 dots per inch.
- 2) The target black character, which has 150-dot height, is

placed in the center of a circular frame. The position of the character is adjusted by using cursor control keys.

- 3) In the coordination system as shown in Figure 3, the center of circular frame is determined as gravity center on the character.

Hereinafter, in this paper, evaluations of gravity center on characters are performed on the coordination system as shown in Figure 3.

In order to maintain the high precision of evaluations, we make a subjective evaluation of gravity center on characters according the following rules.

- 1) Evaluation must be performed from various angles, as shown in Figure 4.
- 2) Evaluation must be performed by a pair of evaluators alternately.
- 3) Evaluators must be picked up from five evaluators, and the average of ten evaluations is adopted.

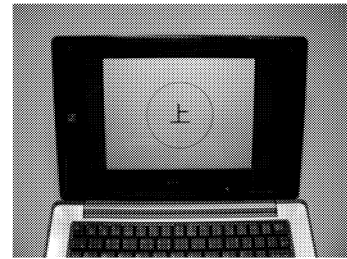


Figure 2: Environment of subjective evaluation.

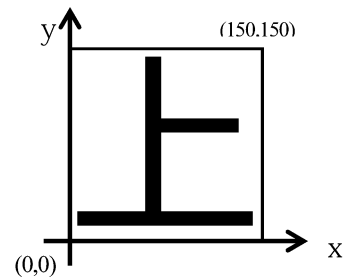


Figure 3: Coordinate system.

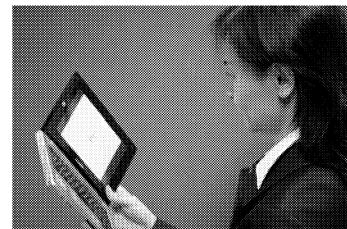


Figure 4: Evaluation trial.

The proposed subjective evaluation method is so simple that the evaluation can be performed not by designers but by testers, who have some experience to check the quality of characters. Therefore it is easy to assemble the human resources. It should be stated here that the proposed evaluation method also reduces the evaluation costs down to 24 minutes per character.



### 3.2 CONTOUR-BASED EVALUATION METHOD OF GRAVITY CENTER ON CHARACTERS

When designers perform evaluation of gravity center, designers generally does not measure from strokes of character, but a contour of character. Based on this fact, the contour-based evaluation method of gravity center on characters has been proposed.

In the contour-based evaluation method, as shown in Figure 5, pixels inside of contour of character are regarded as mass points of 1's. At the same time, pixels outside of contour of character are regarded as mass points of 0's. When a horizontal line is put on the center of mass of contour, the line is defined as contour-based gravity center on the character.

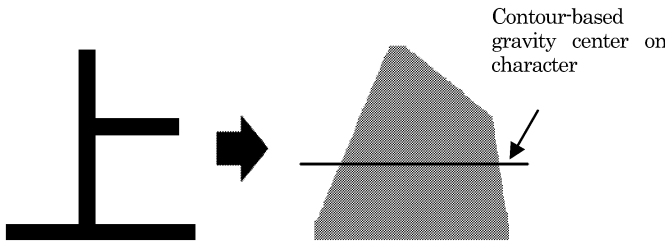


Figure 5: Contour of character and gravity center on contour of character.

Specifically, the contour-based gravity center on character  $G_c (G_{c_x}, G_{c_y})$  is defined as follows:

$$w = \sum_{x=0}^{X_{width}-1} \sum_{y=0}^{Y_{height}-1} p[x][y] \quad (1)$$

$$\begin{pmatrix} G_{c_x} \\ G_{c_y} \end{pmatrix} = \frac{1}{w} \sum_{x=0}^{X_{width}-1} \sum_{y=0}^{Y_{height}-1} p[x][y] \begin{pmatrix} x \\ y \end{pmatrix},$$

where  $X_{width}$  and  $Y_{height}$  are the width and height of a character respectively, and  $p[x][y]$  is the value of pixel of  $(x, y)$ .

However, considerable design costs are still required, since manual contour acquisition process of a character takes about 3 minutes, and thus enormous cost is required to design a font, which consists of about 7,000 characters in Japanese or about 27,000 characters in Chinese.

### 3.3 CONVEX HULL CONTOUR-BASED EVALUATION METHOD FOR GRAVITY CENTER ON CHARACTERS

In order to reduce design cost as described in section 3.2, the contour-based evaluation method is proposed. In this method, the contour of a character can be obtained automatically by using a convex hull algorithm, which seeks the smallest polygon containing the point group on a plane surface. In our evaluation method, the wrapping algorithm, which is the most basic algorithm for convex hull acquisition is employed.

The wrapping algorithm is defined as follow:

1. Initialize variables of  $\theta$  and  $i$  as 0.
2. Find the point where y-coordinate is smallest in the pixels on strokes of a character. When several points are

found, the pixel which has the smallest x-coordinate must be selected. As shown in Figure 6, this point is defined as  $P_0$ .

3. Calculate angle  $\theta$  for each point on strokes of a character with horizontal line. Find the point whose angle with horizontal line is the smallest and larger than  $\theta$ , and define the point as  $P_{i+1}$ . When multiple points which have the same angle are found, the farthest point must be selected.
4. When  $P_{i+1}$  is equal to  $P_0$ , the process is finished. In other case, Step 3 is repeated with incrementing  $i$ .

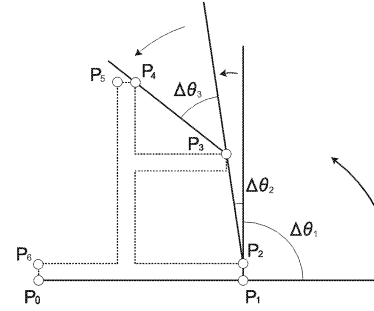


Figure 6: Wrapping method.

In this way, convex hull can be obtained as the convex polygon which is surrounded by the point set of  $P_0, P_1, \dots, P_i$ . In the convex hull contour-based evaluation method, gravity center on characters is calculated by Equation (1).

### 3.4 EVALUATION OF GRAVITY CENTER ON CHARACTERS

As shown in Figure 7, 94 characters of 16th group in JIS level-1 kanji set (JIS X0208-1990) are used in our evaluation. Table 1 and Figure 8 show the results of subjective evaluation by using a circular frame and convex hull contour-based evaluation. From these evaluation results, we can see that the maximum difference between subjective gravity center on characters and calculated value is 2.8 dots. The average absolute value and the standard deviation is 0.8 dot and 0.7 dot, respectively. Figure 8 shows that the characters, whose error is smaller than 1 dot, account for 68% of samples.

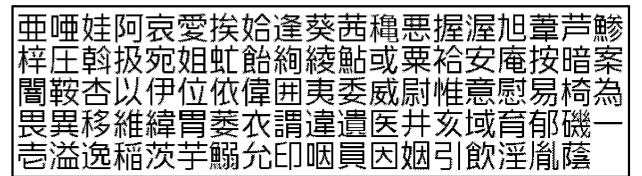


Figure 7: 94 characters of 16th group in JIS level-1 kanji set.

On various mobile terminals, the standard size of character is between 20 dots and 24 dots. The ratio between 1 dot and 24 dots is equal to 6.3 dots of 150 dots. Consequently, the difference which is less than 3.1 dots can not be recognized on a display and there is no problem on practical use.

From the evaluation results, we consider the proposed convex hull contour-based evaluation method has sufficient practicability.

Table 1: Subjective and calculated gravity center (y-coordinate) of 94 characters.

Character	Subjective center of gravity	Calculated center of gravity	Character	Subjective center of gravity	Calculated center of gravity	Character	Subjective center of gravity	Calculated center of gravity	Character	Subjective center of gravity	Calculated center of gravity
垂	73.9	75.0	姐	72.9	73.0	委	75.0	73.4	育	76.1	75.8
唾	76.1	75.7	虻	73.9	73.9	威	71.8	71.9	郁	77.1	76.9
娃	75.0	73.9	飴	72.9	72.3	尉	75.0	74.6	磯	76.1	76.2
阿	77.1	74.8	絢	73.9	73.9	惟	77.1	75.4	一	76.1	76.0
哀	73.9	72.9	綾	73.9	73.8	意	76.1	74.5	膏	77.1	77.0
愛	73.9	73.0	鮎	73.9	74.9	慰	78.2	76.3	溢	72.9	75.2
挨	72.9	73.2	或	73.9	74.5	易	75.0	74.0	逸	73.9	73.0
始	73.9	73.9	栗	76.1	75.9	椅	76.1	75.2	稻	76.1	75.8
逢	75.0	74.3	袷	73.9	75.5	為	71.8	71.5	茨	76.1	74.0
葵	75.0	74.6	安	73.9	71.9	畏	72.9	72.0	芋	82.4	81.2
茜	77.1	76.6	庵	75.0	73.4	異	73.9	72.2	鰯	75.0	74.9
穉	75.0	73.4	按	73.9	73.7	移	79.2	78.6	允	68.7	67.1
惡	78.2	75.6	暗	76.1	75.9	維	75.0	74.8	印	78.2	77.6
握	75.0	74.3	案	76.1	75.1	緯	75.0	74.9	咽	76.1	75.7
渥	75.0	74.9	閭	76.1	74.6	胃	76.1	75.8	昌	71.8	71.3
旭	73.9	72.9	鞍	75.0	75.0	萎	76.1	75.1	因	76.1	75.5
葦	77.1	77.1	杏	76.1	75.6	衣	72.9	72.2	姻	75.0	75.1
芦	84.5	83.1	以	71.8	73.1	謂	75.0	74.8	引	76.1	76.8
鰻	78.2	76.6	伊	82.4	79.6	違	73.9	74.8	飲	73.9	72.5
梓	77.1	76.1	位	75.0	73.9	遺	75.0	74.9	淫	76.1	76.4
庠	75.0	73.6	依	75.0	73.5	医	76.1	75.3	胤	72.9	72.8
幹	76.1	75.3	偉	75.0	74.3	井	76.1	74.6	薩	72.9	75.0
扱	73.9	73.0	圀	76.1	75.5	亥	73.9	73.3			
宛	75.0	72.7	夷	75.0	72.7	域	76.1	75.5			

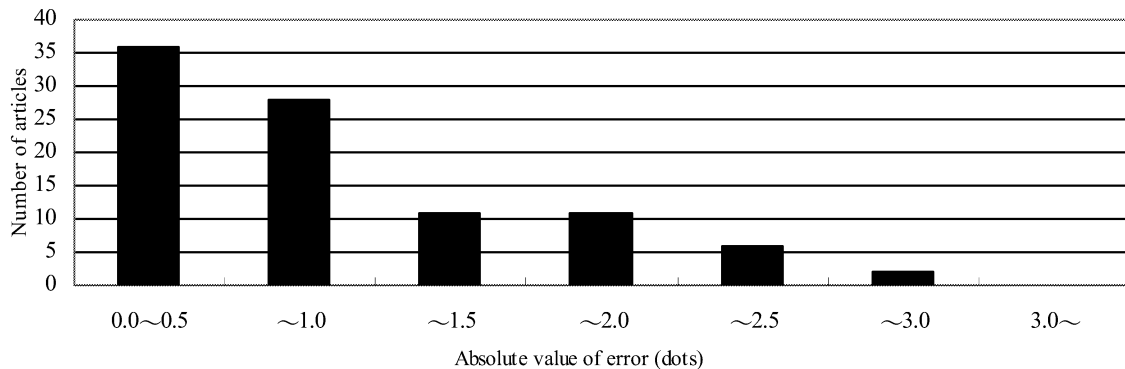


Figure 8: Absolute value of error between subjective and calculated gravity center of 94 characters.

#### 4. CORRELATION BETWEEN LEGIBILITY AND GRAVITY CENTER ON CHARACTERS

In this section, we describe the correlation between legibility of fonts and contour-based gravity center on characters.

##### 4.1 SUBJECTIVE EVALUATION OF LEGIBILITY

The impact of gravity center on legibility of fonts has been recognized from long ago. However, there is no experiment to reveal its correlation.

In order to investigate the impact of gravity center on legibility of fonts, as shown in Table 2 and Figure 9, we performed the subjective evaluation for three typefaces; Gothic, Mincho, and Kaisho. In this evaluation, four fonts of three typefaces are employed. Each font is a bitmapped font which is widely used. Considering the influence of stroke width upon subjective evaluation results, three typefaces of each font have the same stroke width. Figure 10 shows the four fonts of Gothic typeface for subjective evaluation.

Table 2: Subjective evaluation method for legibility of fonts.

Examinee	32 persons
Evaluation objects	Four fonts of Gothic typeface, Mincho typeface, and Kaisho typeface
Evaluation method	The images, which have 20 dots bitmapped font on the size of QVGA, are displayed on the 178ppi LCD of laptop PC. (Figure 9) In the method of pair comparison, the font which has higher legibility is preferred.



Figure 9: Condition of evaluation.

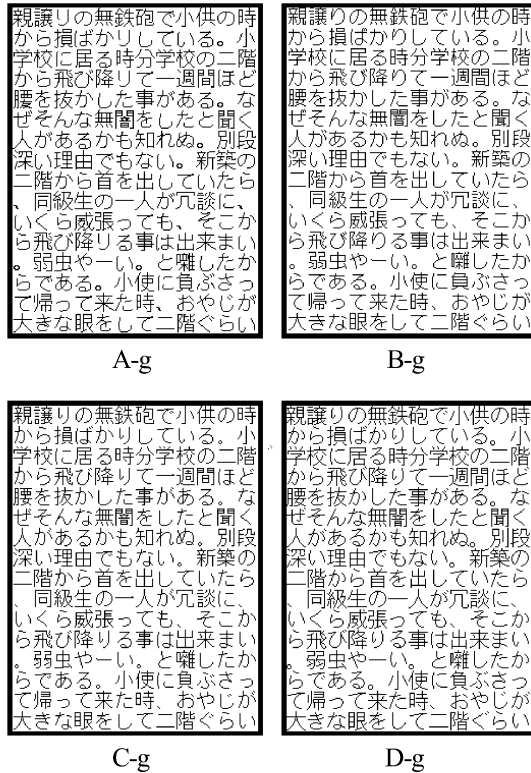


Figure 10: Fonts of Gothic typeface for subjective evaluation.

As shown in Figure 11, the size and resolution of display and the size of character have been determined with referring to these of cellular phones.

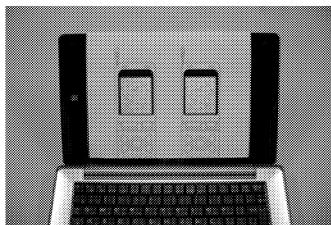


Figure 11: Environment of subjective evaluation.

Table 3, 4 and 5 show the result of pair comparison and the legibility rank of four fonts in Gothic typeface, Mincho typeface, and Kaisho typeface. Here, in each typeface, the legibility rank is determined based-on a normalized rank approach.

As shown in Table 3, the subjective legibility of fonts in Gothic typeface is B-g, C-g, D-g, and A-g, in that order. In the same manner, Table 4 demonstrates that the legibility of fonts in Mincho Typeface is in order corresponding to C-m, B-m, D-m, and A-m. As for Kaisho typeface, Table 5 shows that the legibility of fonts is in order corresponding to C-k, D-k, A-k, and B-k.

Table 3: Result of pare comparison of Gothic typeface.

Rank	Fonts			
	A-g	B-g	C-g	D-g
1	1	20	8	5
2	2	9	13	11
3	1	2	10	16
4	28	1	1	0

(Unit : person)

Table 4: Result of pare comparison of Mincho typeface.

Rank	Fonts			
	A-m	B-m	C-m	D-m
1	3	6	19	6
2	8	15	7	11
3	9	10	2	5
4	12	1	4	10

(Unit : person)

Table 5: Result of pare comparison of Kaisho typeface.

Rank	Fonts			
	A-k	B-k	C-k	D-k
1	2	0	21	9
2	5	6	9	14
3	16	6	1	8
4	9	20	1	1

(Unit : person)

## 4.2 DISPERSION OF GRAVITY CENTER ON CHARACTERS

Dispersion of gravity center on characters can be calculated by referring to the result of convex hull contour-based evaluation. Table 6, 7 and 8 show the results of calculated dispersion of gravity center on characters in Gothic typeface, Mincho typeface, and Kaisho typeface, respectively.

As shown in Table 6, the rank of fonts, which has smaller dispersion of gravity center on characters, is B-g, C-g, D-g, and A-g, in that order. In the same way, Table 7 shows the rank of fonts as C-m, B-m, D-m, and A-m. Table 8 shows the rank of fonts as C-k, D-k, A-k, and B-k.

Table 6: Dispersion of gravity center on characters in Gothic typeface.

Fonts	Dispersion of gravity center on character
A-g	0.70
B-g	0.45
C-g	0.58
D-g	0.64

Table 7: Dispersion of gravity center on characters in Mincho typeface.

Fonts	Dispersion of gravity center on character
A-m	0.49
B-m	0.38
C-m	0.36
D-m	0.43

Table 8: Dispersion of gravity center on characters in Kaisho typeface.

Fonts	Dispersion of gravity center on character
A-k	0.53
B-k	0.73
C-k	0.34
D-k	0.44

## 4.3. CORRELATION BETWEEN LEGIBILITY AND GRAVITY CENTER ON CHARACTERS

From the results of subjective evaluations of legibility in

section 4.1 and the dispersion of gravity center on characters in section 4.2, we can confirm that there is a correlation between legibility and gravity center on characters. Low dispersion of gravity center on characters leads to high legibility of fonts.

In the evaluation of legibility of fonts, the method of pair comparison can achieve high accuracy. However, in order to achieve high reliability of the evaluation results, the evaluation has to be performed by many evaluators. On the other hand, the proposed method can reduce the considerable cost for legibility evaluation of fonts, and enables to perform quantitative evaluation.

## 5. CONCLUSION

Gravity center on characters makes a large impact to legibility. In order to design fonts efficiently, which have high legibility and high quality, the establishment of an evaluation method of gravity center on characters is strongly required.

In this paper, we proposed the convex hull contour-based evaluation method of gravity center on characters. The proposed method can evaluate gravity center more efficiently than the subjective method or the conventional evaluation method. Furthermore, the proposed method can be employed to evaluate the legibility of fonts.

Further developments is to construct font designing tools based on the proposed method, with which designers can design fonts referring to an example from gravity center on a character in real-time, and can reduce the designing cost.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr. Toru Chiba, group general manager of corporate research and development group, SHARP Corporation. The authors would also like to thank to Prof. Isao Shirakawa from Hyogo University and the members of Platform Technology Development Center, SHARP Corporation for helpful advices and encouragements.

## REFERENCES

- S. Okada, N. Koyama, Y. Asai, and A. Kotani: "Resolution Enhanced, Smooth FONT for Color LCD (LCFONT.C)," in *Proc. International Congress of Imaging Science (ICIS'02)*, Tr.4-106(P), pp. 461-462, May 2002.
- A. Kotani, Y. Asai, Y. Nakamura, S. Okada, N. Koyama, K. Yamane, Y. Okano, Y. Mitsuyama, and T. Onoye: "Visibility Font Technology on High Resolution Color LCD 'LCFONT.C,'" in *Proc. The International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2003)*, vol. 1, pp. 535-538, July 2003.
- M. L. Bernard, B. S. Chaparro, M. M. Mills, C. G. Halcomb: "Comparing the Effects of Text Size and Format on the Readability of Computer-Displayed Times New Roman and Arial Text," *International Journal of Human Computer Studies*, 59 (6), pp. 823-835, June 2003
- A. Arditi, J. Cho, "Serifs and Font Legibility", *Vision Research*, 45 (23), pp. 2926-2933, Jun. 2005
- M. Yoshimura and T. Iijima: "A Method for the Design of the Font," *Information Processing in Japan*, vol. 10, no. 0, pp. 77-80, 1970.
- A. Kotani, N. Koyama, Y. Mitsuyama, and T. Onoye: "Gravity center and Readability on 'LCFONT'" for Low Resolution Display", *Journal of IEEEJ*, Vol. 32, no. 5. pp. 621-628, Sept. 2003. (in Japanese)
- A. Kotani, Y. Asai, Y. Nakamura, M. Ootsuka, Y. Mitsuyama, and T. Onoye: "Contour-Based Evaluation Method of Gravity center on 'LCFONT'", in *Proc. Technical Report of IPSJ*, 2004-HI-111, pp. 63-70, Nov. 2004. (in Japanese)
- A. Kotani, Y. Tanemura, Y. Asai, Y. Nakamura, M. Ootsuka, Y. Mitsuyama, and T. Onoye: "Contour-Based Evaluation Method of Gravity center on Characters and Its Application to Font Development," in *Proc. Technical Report of IEICE*, SIS2005-23, pp. 1-6, Nov. 2005. (in Japanese)

JIS X0208-1990, Japanese Standards Association.

# **CHARACTER EMOTION DETECTION**



# Text-to-Emotion Analysis Engines -Theory and Practice

David John

Anthony Boucouvalas

Zhe Xu

Multimedia Communications Research Group

Bournemouth University

Fern Barrow, Poole

Dorset, BH12 5BB, UK.

{djohn, tboucouv, zxu}@bournemouth.ac.uk

## KEYWORDS

Emotional Momentum, Text-to-Emotion, Emotion Extraction, Expressive Internet Communication.

## ABSTRACT

This paper describes the latest developments of a generic prototype a real-time text-to-emotion analyser. The core component of the analyser is the emotion extraction engine. The engine can analyse input text, extract the contained emotion and deliver the parameters necessary to invoke an appropriate image expressing the emotion. The parameters include the interpreted emotion and intensity, both represented by the displayed image. A set of rules are defined to control the behaviour of the emotion extraction engine. The concept of 'Emotional Momentum' is implemented in the average weighted emotional mood indicator to deal with ambiguity or conflicting emotional content. The engine has been applied in a number of different environments including real-time expressive communication systems over the Internet and journal article analysis systems. We present experimental results carried out to test the engine's performance. The results illustrate that the analyser can be used successfully within the designed environments.

## INTRODUCTION

Emotion automation or so-called affective computing automation is one of the great challenges facing the computer community. The benefits of emotion automation are not just for Internet communications e.g. on-line chat rooms, but also for human-computer interfaces and synthetic agents. Picard (1997) demonstrated the potential and importance of emotion in human-computer interaction and researchers such as Elliott (1992), Bates (1994), Koda (1996), Reilly (1996), Andre et al. (1999), Bartneck (2001) and Bianchi-Berthouze and Lisetti (2002) illustrated the important roles that emotion plays in user interactions with synthetic agents.

We have developed a prototype emotion analyser, which can analyse the emotive content within communications over the Internet and automatically display images depicting the appropriate expressions in real time. The intensity and duration of the expressions are also calculated and displayed. Emotions can be detected in various ways: in speech, in facial expressions body languages and in text. We believe

that text is a particularly important means for communicating emotions, as it is still the dominant medium for Internet and computer communication.

This paper presents the latest development of our analyser, where in order to sense textual emotion information, the analyser applies not only grammatical knowledge and keyword tagging, but also takes real-world information, cyberspace knowledge and historical mood analysis into consideration.

## TEXT-TO-EMOTION ANALYSER

The text-to-emotion analyser first discussed by Xu and Boucouvalas (2002), Xu et al. (2002) and Xu (2005) is based on explicit emotional keyword tagging without context consideration. Our latest emotion engine adopted the real-world knowledge approach; it takes emotional momentum (Xu et al. 2006) and context into consideration. With real-world knowledge, inferred emotions can be examined and detected. For example, the sentence "bought a terrible car" contains two emotional words "bought" and "terrible", which belong to conflicting emotional categories. The sentence does not clearly reveal to readers a single emotional category for the sentence. To deal with such ambiguity, we make use of the average weighted mood information derived based on the mood of previous sentences.

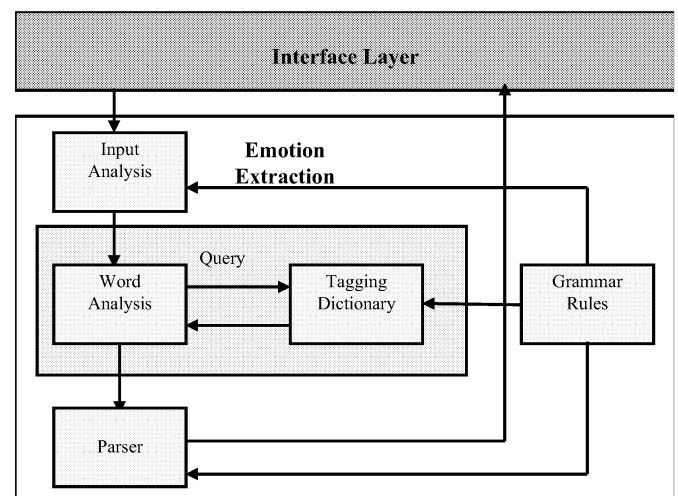


Figure 1: The emotion analyser overview

Figure 1 depicts an overview of the architecture of the emotion analyser. The text-to-emotion analyser engine,

includes three components: input analysis, tagging system and a parser.

The classic rule based architecture was applied in the emotion extraction to direct the analysis process. A set of rules has been set up to specify the engine's behaviour.

### Input analysis

User's input sentences are sent to the input analysis function for initial assessment. The length of a user's input can vary from just a few words to a hundred-thousand word article. An article may contain hundreds of sentences and even in a chatting environment users may type more than one sentence at a time. The output from the input analysis will be passed on to the tagging system.

The tagging system can only handle one sentence a time therefore it is vital to separate sentences correctly. A dot can be not only a terminator, but also a decimal point, e.g. "The index was up 10.3 per cent". Exclamation marks and question marks terminate a sentence in most cases, however in sentences such as "I am really good!!!" or "what???", the first exclamation mark or question mark does not only terminate the sentence, but these marks also represent a feeling of the raising of emotion intensity. In a chatting environment it is common for users to ignore typing a termination mark. A rule is applied to add a full stop when the input analysis function finds a sentence without a termination mark.

The input analysis function generates individual sentences as outputs and sends those sentences to the tagging system.

### Tagging system

The tagging system includes two parts: a tagged dictionary and the word analysis function. The main duty of the tagging system is to speedily locate and determine the emotional information correctly.

Instead of using a general purpose corpus such as The Brown Corpus (2003) and British National Corpora (2003), a dictionary of over 18,000 words has been created solely for this project. Word entries contain the word (including any possible prefix and suffix), a tag which indicates which emotion category they belong (if any), and an indication whether they have different meanings in different contexts. To find all possible emotional words, we manually searched through English dictionaries, and the dialogues in text chatting environments, to identify the emotional words contained in them. The dictionary is constantly growing and testing the reliability of the database remains part of a future project.

A tag set was developed to represent the category of emotion each word in the dictionary belongs to. All emotional words are classified into emotion categories. Ekman's six expression categories (happy, sad, surprise, fear, anger, and disgust) are followed because of the existence of discriminative features between different categories (Ekman et al 1972; Ekman and Oster 1979; Ekman 1999; Ekman et al. 2002). For each expression category three different emotion intensities are defined: low, medium and high.

When receiving the output sentence from the input analysis function, the tagging system will split the sentence into

words, and then check through the tagged dictionary to find each word and the corresponding tag category.

The following rules are applied in the Word Analysis function:

- Emotional words are divided and tagged into two categories: the inferred emotion category and the explicit emotion category. Explicit emotion words directly convey an emotion while inferred emotion words do not directly convey an emotion but describe situations that contain emotional feelings. The intensity of inferred emotion words is lower compared to explicit emotional words. If an explicit emotional word is also presented in the sentence, the intensity of the inferred emotion should be weaker than the explicit emotion.
- In some specific situations, the meanings of an emotional word can be transformed. For example the sentence "I got a pretty car" presents a happy feeling while the sentence "I got a pretty ugly car" will present a feeling with dissatisfaction.
- When emotional words are involved with addresses, names and traditional meanings, the emotional feelings may disappear. For example, "new" is a word with inferred emotion, but the phrase "New York" does not contain any emotional meanings its own.
- In the tag-set, a special tag *NDP* (negative data point) is assigned to the words that have opposite influences on the emotion information. For example, in the sentence "The market halted its slide", the word "halted" will overturn the meaning of the sentence. Another special tag *IGNORE* (ignore word) is assigned to the words following the negative data points.

The outputs of the tagging system are the words and corresponding word-category tags. The output information is sent to the parser for further analysis.

### Parser

The parser receives the output from the tagging system and the initial parsing procedure is accomplished through the use of the rewrite rules (Russell and Novig 1995). The aim is to identify all possible combinable phrases and sentence structures.

If an adjective is found, the parser will examine the subject and tense by analysing the output of the rewrite rule component.

There has been little research into mood (Bianchi-Berthouze and Lisetti. 2002; Egges et al. 2004; Velasquez 1997) and no official definition or accepted mechanism for simulating mood has been produced, therefore for our system we examine mood in the short term (over the course of a number of typed sentences) To represent the momentum of emotion the average weighted emotional mood indicator (AWEMI) of the user calculated as follows:

$$AWEMI = \frac{1}{N} \sum_{i=1}^N \alpha_i E_i \dots\dots\dots(1)$$

where 'i' is an index increasing from 1 to N corresponding to the sentences taken into account, up to maximum N.  $E_i$   $\subset$  corresponding to the sum of the signed emotion intensities (momentum, with a positive or negative sign) of ith sentence and finally  $\alpha_i$  is the weight given to each emotion within a



sentence.  $E_1$  for example, corresponds to the first sentence's total emotion intensity and is the signed sum of all the intensity emotions within the sentence, (happy, very sad, etc).

Similarly  $E_2$  for the second sentence and subsequent sentences up to the previous five are correspondingly labelled and used in our prototype.

It is assumed that the mood of the last sentence should carry more weight to the representation of the user. Therefore greater weight is given to the emotions contained in the most recent sentence with progressively less weight to earlier sentences. For simplicity only the last five sentence emotions are actively influencing the AWEI indicator.

The following rules are applied in the parser:

- In a chatting environment, some emotion symbols, such as :-), are widely used to represent emotional feelings. The engine will search for these emotion symbols and map them to corresponding emotion words e.g. :- is mapped to the word happy.
- In a chatting environment, acronyms e.g. LOL (laughing out loud) are widely used to represent strong emotional feelings. However these acronyms have almost never been used outside the chatting context and life span of their usage is limited. Instead of adding these acronyms into the tagged dictionary, they are stored in the dictionary but are only checked when the parser is applied in a chatting environment.
- In a chatting environment, acronyms and emotion symbols strongly dominate the emotional feelings. For example the sentence "You are ugly :-)" presents an ironic and happy feeling despite the existence of the emotional word "ugly". When emotion acronyms or symbols are found, the parser will set the sentence's mood according to the category of the acronym or symbol and discard other emotional words.
- The parser considers a conditional emotional sentence as an emotional sentence with reduced intensity. The sentence "I will be happy" is assigned as happy with an intensity of two because of word "happy". The sentence "I will be happy if he comes" will be assigned as happy with an intensity of 1.
- If there is more than one emotional word in a sentence and they are connected by a conjunction then the parser will combine these two emotional states, e.g., 'I am surprised and happy about it' will be treated as a sentence with emotions surprise and happy.
- If no subject is found, the engine will refer the subject to the person who is in communication e.g., 'apprehensively happy' will be treated as 'I am apprehensively happy'.
- Some words may not contain emotional meanings themselves. However, when they are constituted into phrases, emotional feelings will emerge, e.g. "cold feet".
- If no adjectives (expressive adjective) are in front of the emotional word then the intensity will depend only on the word's tag category. If there is more than one adjective then the parser will increase the intensity automatically. For example the parser interprets 'very very sorry' as a high intensity emotional phrase.
- If a sentence begins with an auxiliary verb, the sentences are questions and can not provide emotional

information directly from the words. For example, the sentence "are you happy?" does not represent a happy emotion and instead provides a sense of puzzlement. Our engine treats puzzlement as the emotion surprise with lowest intensity.

- If an emotion word is found in negative form, the parser will treat the sentences as no emotional feelings, e.g. I am not surprised. The negation of one emotion does not automatically indicate the presence of the opposite emotion.
- Some emotional words will grammatically fall into the noun category and are used as such. Our parser will treat these words as normal emotional words.
- Phrases such as "rather than" overturn the emotional words that follow. When more than one emotional word is found in a sentence it is possible that the emotion meanings contained in different words may conflict. For example, the sentence "I was fine but my car was damaged a little" includes two emotional words "fine" and "damaged". There are no clues as to know whether this sentence presents a happy or sad feeling within the sentence. To solve this problem, we have applied the average weighted mood indicator calculation rule.
- When a sentence contains both positive and negative emotions, a conflict occurs, e.g., "I am happy, but I am still worried about my future". When conflicting emotions are found, the engine will consult the AWEI mood calculated previously by the average mood calculator. If the result shows that the user was most recently in a specific mood, then the new sentence's emotion will be changed to that particular category with the lowest intensity, otherwise the conflicting emotions will remain unchanged.

When no emotional word is found the following rules are applied:

- It is reasonable to assume that most users are polite or have a low intensity happy mood at the start of a dialogue. When no emotion words are found, the first sentence will be assigned happy emotion with lowest intensity.
- When exclamation marks are found, the intensity of corresponding emotions are increased or when no emotional words are found, the parser will assign a happy emotion with lowest intensity to the sentence.
- Sentences that finish with a question mark often contains puzzlement feeling. In our engine, puzzlement is included in the emotion surprise with the lowest intensity.
- Some specific repeated words present inferred emotional feelings, e.g., "no, no, no, you should go left". When these specific words are found consecutively, the engine will assign an inferred emotion to the sentence according to the word category with intensity equal to the number of consecutive words presented.
- If a sentence does not qualify for all above analyses, the engine will assign to it a neutral emotion with medium intensity.

The output from the parser will be sent to the interface layer. The output contains two parts. One is the current emotion state, that includes the emotion category, the intensity and

the tense. The second part is the average emotion, that includes the emotion category and the intensity.

### **The interface layer**

The emotion analyser is a generic prototype and has been utilised in many applications in different environments.

When applied in a real-time communication system, the emotion analyser will be used to provide an expressive chatting interface. The 'Expressive chatting Interface' is a visual interface for real-time collaboration over the Internet among groups of people. The 'Expressive chatting Interface' application has two variations: the 'Expressive chatting 2D Interface' where users' locations are not important and an 'Expressive chatting 3D maze', in which the users' locations in space are available.

The 'Expressive chatting Interface' is capable of invoking expressions that convey emotion without making use of video. The application utilises discrete images of the participants in order to keep the bandwidth requirements to a minimum yet still provides an elaborate communication tool. The interface allows the viewing at a glance of participants in the system, and those pairs engaged in conversation, as well as the expressive image of the users engaged in the conversation.

Emotions that are detected within the messages typed by users are represented by expressive images from the six expressive categories (happy, sad, anger, disgust, surprise, fear) with three different intensities.

When applied within real-time game system, the emotion analyser will be adopted as expressive on-line game engine. Emotions can be detected and displayed as opponents exchanging text messages during the game in the same manner as the chatting environment. In addition emotions can be assigned to events in the game, such as displaying a happy image for the player taking an opponent's chess piece and displaying a sad image for the player whose piece is being taken.

Emotional content can be detected not only in human to human communication or human-machine communication, but also in other interaction styles. One of the most interesting examples is the stock market. Although frequently used in papers and articles, the term stock market emotion does not have an authoritative definition. Stock market emotions are defined in this research work as the different states of expectation investors derived from their perception of trends in the movement of share prices. The stock market emotions include three different states, i.e. happy, sad and stable.

The happy state matches in the situation where share values are rising. The sad state occurs when share values are falling. If there is no change in the value, the market will be in the stable state. From this definition, the stock market emotion directly reflects the movement of the stock price index. By applying the emotion extraction engine to stock market articles, a stock market emotion analyser can be created.

### **EVALUATION OF THE TEXT-TO-EMOTION ANALYSER**

In order to test the correctness and effectiveness of the emotion extraction engine in different environments, an experiment was carried out. As the characteristics of the various target applications are different, the experiment was divided into two tests. One was the chat environment test, in which text collected from Internet chat rooms were fed into the emotion extraction engine; the other was the article analysis test, in which sentences from different published articles were used. For both the chat environment test and the article analysis test, the measurement was the number of the correctly extracted emotions. This was assessed by comparing the extracted emotional feelings with manually identified emotions.

In the chat environment test, twenty-three pages of chat logs, which includes six hundred and thirty eight sentences, were collected from the Internet. All sentences interpreted as emotional were highlighted manually. The chat sentences were sent into the emotion extraction engine and the engine's output was logged.

In the article analysis test, seven articles, which include three hundred and forty eight sentences articles were collected from the Financial Times website and the BBC web site. In a similar manner to the chat environment test, all sentences interpreted as emotional were highlighted manually. Subsequently these paragraphs were fed into the emotion extraction engine to examine how well its automated estimation correlated with the manual extraction.

For the chat environment test, seventy-eight sentences (87% of the total sentences) were correctly recognised. For the article test, sixty-one sentences (76% of the total sentences) were correctly recognised.

The results of the tests show that the engine produced better results in the chat environment. The reason is that the sentence structures are much simpler in the chat environment and the individual to whom the emotion refers to is also limited. However, the results demonstrated that the emotion extraction engine can cope with the majority of emotional sentences in both environments.

The emotion extraction engine achieved more than 75% accuracy in the experiment. This demonstrates that the engine can be used in the real-time environment. However as the sentences that the emotion extraction engine were not designed to handle were excluded, the accuracy of the engine would be lower in possible field tests.

Even with the limitations presented, the experiment can still demonstrate the emotion extraction ability of the emotion extraction engine and the following conclusions can be made. First, the engine may correctly identify emotions in the majority of sentences when applied in a chat environment. Second, the engine may analyse large numbers of sentences correctly when applied in the article analysis context.

Most textual emotion sensing approaches are based on rewrite rule analysis or statistical correlation information. However, with the increase of complexity of sentence structure, the error rate of the rewrite rule analysis will increase. This will affect the performance of all the sensing approaches. The emotion analyser is based on keyword tagging. This method has numerous advantages (e.g. fast and easy adopt etc.), however the defects are also obvious, which are discussed below.

In some specific situations, a sentence containing emotional words does not present any emotional feelings. For example, the sentence "Research shows that happy images can cause sad feelings." does not provide any emotional feelings although emotional keywords are present. Our engine can detect the emotional keywords and the referred subject, but not the intended meanings in all circumstances.

The meanings of emotional words can be changed according to context and time. One example is the emotional word "thrilled". The word itself used to present a fearful feeling but now this word is equal to happy.

To detect emotions in a sentence, the emotion analyser has to detect emotional keywords first. However, some emotional sentences may not provide any significant emotional keywords, e.g., the sentence "I spent ten days on it and finally I worked it out" presents a relief and happy feeling, but no emotional keyword can be extracted.

The tagged dictionary was created solely for this project. The tagging and rating (intensity decision etc) were agreed by a group of staff members of Bournemouth University (a developer, a linguistic and a user-centre expert) in order to achieve a relatively fair rate. However, future enhancements may involve checking the dictionary against the output of established corpora.

The emotion analyser does detect multi-emotions presented in a single sentence. However, as no discriminative features can be found in mixed expressions (e.g. a mixed "happy and sad" expression), the emotion analyser will always try to display one single expression display by calculating AWEMI. However, in some cases, AWEMI can not provide a single expression display, then the mixed expressions will be displayed separately (e.g. mixed "happy and sad" expression will be shown as a happy expression followed by a sad expression).

Although the disadvantages exist, we believe the emotion analyser still can be applied within Internet communication contexts and achieve considerable performance. As the technology develops, some of the present disadvantages may be solved in near future.

## CONCLUSIONS

We have developed a text-to-emotion detection engine analyser and using the AWEMI and applied this to a number of typical applications. The latest development of the emotion analyser, which is based on keyword tagging, has also been presented. The emotion analyser makes use of real-world knowledge, grammatical knowledge and context. We have discussed the operation and rule book of the emotion analyser in detail. The emotion analyser engine is suitable and has been designed for applications in the Internet communication environment.

Although the engine was developed successfully, some limitations still exist. The performance of the analyser has been tested and the results illustrate that the analyser can be used successfully within the designed environments. Numerous future applications based on the emotion analyser are possible.

The most immediate future work will examine how to apply history mood and context information to negotiate conflicting emotions. Finally a more extensive test of the

emotion detection engine is desirable and will provide a picture of the limitations of the keyword tagging technology and other emotion sensing approaches.

## REFERENCES

- Andre E.; M. Klesen; P. Gebhard; S. Allen; and T. Rist. 1999. "Integrating models of personality and emotions into lifelike characters". In *Proceedings of the workshop on Affect in Interactions - Towards a new Generation of Interfaces*, A. Paiva and C. Martinho (Eds), (3rd Annual Conference, Siena, Italy, October 1999), 136-149.
- Bartneck C. 2001. "How convincing is Mr. Data's smile: Affective expressions of machines". *User Modeling and User-Adapted Interaction*, 11, 279-295.
- Bates J. 1994. "The Role of Emotion in Believable Agents". *Communications of the ACM* 37, No.7, 122-125.
- Bianchi-Berthouze N. and C.L. Lisetti 2002. "Modelling Multimodal Expression of User's Affective Subjective Experience", *User Modeling and User-Adapted Interaction*, 12 No.1.
- British National Corpora 2003. <http://www.natcorp.ox.ac.uk> Accessed 12 August 2003.
- Brown Corpus 2003. [http://clwww.essex.ac.uk/w3c/corpus\\_ling/content/corpora/list/private/brown/brown.html](http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html) Accessed 1 August 2003.
- Egges A.; S. Kshirsagar; and N. Magnenat-Thalmann 2004. "Generic Personality and Emotion Simulation for Conversational Agents". *Computer Animation and Virtual Worlds* 15 No.1 1-13, January 2004.
- Ekman P. 1999. Facial Expressions. In *Handbook of Cognition and Emotion* T. Dalgleish, and M Power. New York: John Wiley & Sons Ltd.
- Ekman P.; W.V. Friesen; and P. Ellsworth 1972. *Emotion in the human face: guidelines for research and an integration of findings* New York: Pergamon Press.
- Ekman P.; W.V. Friesen; and J.C. Hager 2002. *The Facial Action Coding System* Research Nexus eBook. Salt Lake City, UT.
- Ekman P. and H. Oster 1979. "Facial expressions of emotion". *Annual Review of Psychology* 30, 527-554.
- Elliott C. 1992. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*. PhD thesis, Northwestern University. The Institute for the Learning Sciences, Technical Report No. 32.
- Koda T. 1996. "Agents with Faces: A Study on the Effect of Personification of Software Agents". In *Proceedings of HCI'96*.
- Picard R.W. 1997. "Affective Computing". The MIT Press, Mass.
- Reilly W.S.N. 1996. *Believable Social and Emotional Agents*. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA.
- Russell S. and P. Novig 1995. *Artificial Intelligence: A Modern Approach* ISBN 0-13-103805-2.
- Velasquez J. 1997. "Modeling emotions and other motivations in synthetic agents". In *Proceedings of AAAI-97* 10-15. MIT Press.
- Xu Z. 2005 *Real-time Expressive Internet Communications* PhD Thesis, Bournemouth University.
- Xu Z. and A.C. Boucouvalas 2002. "Text-to-Emotion Engine for Real Time Internet Communication". *International Symposium on Communication Systems, Networks and DSPs* 15-17 July 2002, 164-168.
- Xu Z.; D. John; and A.C. Boucouvalas 2002. "Text-to-Emotion Engine: tests of user preferences". *Proceedings of ISCE 2002*, 22-25 September 2002, Erfurt, Germany, B25-B30.
- Xu Z.; D. John; and A.C. Boucouvalas 2006. "Representing Emotional Momentum within Expressive Internet Communication". In *Proceedings of IASTED International Conference on Internet and Multimedia Systems and Applications (EuroIMSA 2006)* 13-15 February 2006.

# A TEXT-BASED SYNTHETIC FACE WITH EMOTIONS\*

Siska Fitrianie  
Leon J.M. Rothkrantz  
Man-Machine-Interaction Group  
Delft University of Technology  
Mekelweg 4 2628CD Delft  
The Netherlands  
E-mail: {s.fitrianie, l.j.m.rothkrantz}@ewi.tudelft.nl

## KEYWORDS

3D synthetic face, emotions, natural language processing

## ABSTRACT

The human face in particular serves not only communicative functions, but it is also the primary channel to express emotion. We develop a prototype of a synthetic 3D face that shows emotion associated to text-based speech in an automated way. As a first step we studied how many and what kind of emotional expressions produced by humans during conversations. The next, we studied the correlation between the displayed facial expressions and text. Based on these results, we developed a set of rules that describes dependencies between text and emotions by the employment of ontology. For this purpose, a 2D affective lexicon database has been built using WordNet database and the specific facial expressions are stored in a nonverbal dictionary. The results described in this paper enable affective-based multimodal fission.

## INTRODUCTION

Emotions play an important role in communication. They are part of communication and control systems within the brain that mobilize resources to accomplish the goals specified by our motives. The instantaneous emotional state is directly linked with the displayed expression (Ekman 1999).

Facial displays as means to communicate provide natural and compelling computer interfaces (Nagao and Takeuchi 1994, Schiano et. al. 2000). At MMI-Group TUDelft, there is a project running on natural human computer interaction. We have developed a synthetic 3D face (Wojdel and Rothkrantz 2005) based on Facial Action Coding System (FACS) (Ekman and Friesen 1975). The system allows average users to generate facial animations in a simple manner. It has a dictionary of facial expressions (FED - de Jongh and Rothkrantz 2004) that stores the facial expressions that naturally occur in face-to-face communication.

Facial expressions do not occur randomly, but rather are synchronized to one's own speech or to the speech of other (Pelachaud and Bilvi 2003, Ekman 1979, Condon and Osgton 1971). The challenge is to find those relations. Human face-to-face conversation involves both language and nonverbal behaviour. We are used to convey our thought through our (conscious or unconscious) choice of words. Some words possess emotive meaning together with their descriptive meaning. The descriptive meaning of this type of words along with a sentence structure informs the interpretation of a nonverbal behaviour and vice versa. Seeing faces, interpreting their expression, understanding the linguistics contents of speech are all part of human communication.

Most existing face animation systems, to the best of our knowledge, have not considered explicitly the emotions existing in the speech content. The difficulty lies in the fact that emotional linguistic content consists of entities of complexity and ambiguity such as syntax, semantics and emotions. The use of simple templates has proven to be useful for the detection of subjective sentences and of words having affective semantic orientation (Hatzivassiloglou & McKeown 1997, Liu et.al. 2003). These simplistic models describe how words with an affective meaning are being used within a sentence, but fail to offer a more general approach. Furthermore, current developments of affective lexicon database (e.g. Ortony et.al. 1987 and Strapparava et.al. 2004) are based on subjective meaning of the emotion words and do not provide information about the (relative) distance between words in regards to their emotion loading context. The lack of a large-scale affective lexicon resource database makes a thorough analysis difficult. As a consequence, although important, an automated emotional expression from natural language is still rarely developed.

Our developed system is able to convey emotions through verbal and nonverbal behaviours. It can reason emotions automatically from natural language text and show appropriate facial expressions as its stimulus response to the emotional content of the text. This research work is an initial step for the development of a system to realize affective-based multimodal fission. In the following sections we will give an overview of the system we currently develop. Further, we will concentrate on emotions analysis from text as well as experiments to generate the reasoning and the affective lexicon database.

\* The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024

## SYSTEM OVERVIEW

Figure 1 shows the communication flows of the developed system. The Emotion Analysis module receives text that contains emotion words. It has a parser that associates the text to emotions. The module exploits ontology and an affective lexicon database. It transforms the input into XML format as the communication language between modules.

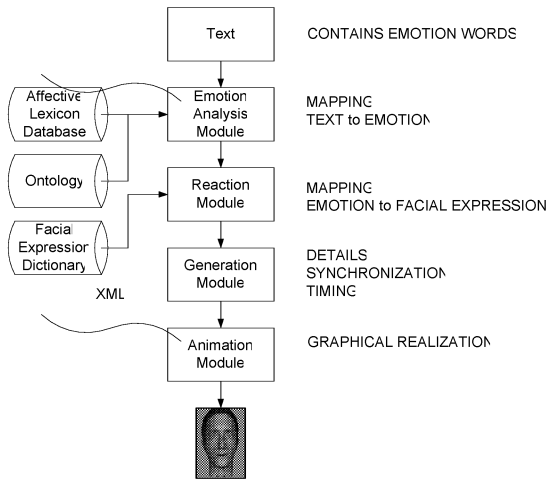


Figure 1: The Architecture Pipeline

The Reaction Module processes the text-emotion mapping result by assigning appropriate facial expressions. It connects to a FED. For every facial expression, the dictionary contains an emblem of the facial expression, a description of which AUs are activated, semantic interpretations, and an example using a synthetic face (de Jongh and Rothkrantz 2004).

The Generation Module plans and annotates the display of the facial expressions synchronized with speech. It has the estimates of word and phoneme timings and constructs an animation schedule prior to execution. This module also considers timings for sentence or clause boundaries. The Animation Module generates facial animation based on a “facial script language”, where basic variables are AUs and their intensity. It has a parametric model for facial animation and a method for adapting it to a specific person based on performance measurements of facial movements (see Wojdel and Rothkrantz 2005).

## DATA ACQUISITION EXPERIMENTS

### Emotion Expressions

Many theorists and psychologists tried to categorize emotions, e.g. (Ekman 1999, Ortony et.al. 1988, Russell 1980). An experiment has been performed to recognize the most expressive facial expressions used in conversations and to determine a list of emotion expressions applied by our system. We recorded dialogs in pairs of two participants. The participants were requested to perform dialogues about different topics and show as many expressions as possible. The video recordings were stored in a database such that video and sound are synchronized. Firstly, three independent observers marked the onset and offset of an expression. Secondly, we labeled the expressions according to the context. The agreement rates between the observers in both

steps were about 73%. Finally, we also collected emotion words used in each expression.

The experimental results indicated that our participants showed most of the time a neutral face. However, we managed to capture in total 40 different facial expressions; about 20-35 different expressions per participant in each dialog, and 140 emotion words. Our experimental results were endorsed by an experiment conducted by (Desmet 2002). He found 41 displayed emotion expressions actually used to appraise a product (see table 1).

Table 1. Emotions in Eight Octants (Desmet 2002)

No	Valence-Arousal	Emotion Expressions
1.	Neutral-Excited	Curious, amazed, avaricious, stimulated, concentrated, astonished, eager.
2.	Pleasant-Excited	Inspired, desiring, loving.
3.	Pleasant-Average	Pleasantly surprised, fascinated, amused, admiring, sociable, yearning, joyful.
4.	Pleasant-Calm	Satisfied, softened.
5.	Neutral-Calm	Awaiting, deferent.
6.	Unpleasant-Calm	Bored, sad, isolated, melancholy, sighing.
7.	Unpleasant-Average	Disappointed, contempt, jealous, dissatisfied, disturbed, flabbergasted, cynical.
8.	Unpleasant-Excited	Irritated, disgusted, indignant, unpleasantly surprised, frustrated, greedy, alarmed, hostile.

### Correlation between Facial Displays and Text

Using the results of the previous experiment, we conducted another experiment to study the relationships between facial expressions and text (Wojdel 2005). We partitioned the dialog (manually) into basic constituents (i.e. a single sentence), and for each constituent into components (i.e. single words and punctuation). Then, for each component, we determined the time of its occurrence (number of frames in which the given word is pronounced). The next, for each selected facial expression, we determined the text that started with the component synchronised with the first frame of a given facial expression and ends with the component synchronised with the last frame of this expression.

The experimental result showed that most of the facial expressions (around 63%) corresponded to the text spoken. For sentences with questions or exclamation marks, distinguishably, “surprise” is the most common facial expression displayed during a question; exclusively in short and single-word question, e.g. “really?”, “sure?”. We noticed that the sentences ending with exclamation mark were usually accompanied by the expression “anger”. In the final experiment, we focused on the mapping of shown expressions to emotional words. The distance of a given facial expression from a particular emotional word was defined as the number of frames with the neutral face, which appear between the facial expressions. The results showed that 54.6% of the emotion words spoken by the participants linked to facial expressions. We also compared results from above with the analogous statistics for non-emotional words. The comparison showed that the use of emotion words, indeed, evoked emotions which were expressed by facial expressions (see table 2). Although, with this experiment, we

still could not draw a direct link between the emotion words with the facial expressions. It is because some words only occurred once or twice and some other words in different context were related to different facial expressions.

Table 2: Statistics of Emotion and Non-Emotion Words in the Input Text Linked to Facial Expression

Words	Total	Linked to Expression
Non-emotion words	2206	1022 (46.3%)
Emotion words	119	65 (54.6%)

## 2D AFFECTIVE LEXICON DATABASE

Russell (1980) and Desmet (2002) brought subjective meanings of the degree of pleasantness and the degree of activation specifically for words that express emotions. The words can be depicted in a 2D space. Although it can categorize emotions, it proved that the approach is not sufficient to differentiate between emotions. For example, “anger” and “fear” fall close together on the circumplex. Moreover, for automatic reasoning, quantitative data would be more efficient. Kamps & Marx (2002) proposed the differences between the relatively objective notion of lexical meaning, and more subjective notions of emotive or affective meaning by exploiting WordNet (Fellbaum 1998), such as: (1) the smallest number of synonymy (SYNSET) steps between two words, e.g.:  $MPL(\text{good, bad}) = 4$  {good, sound, heavy, big, bad}, and (2) the relative distance of a word to two reference words, e.g.  $EVA^*(\text{proper, good, bad}) = 1$  and  $ACT^*(w, \text{active, passive})$ . The latter equations differentiate words that are predominantly used for expressing positive emotions (values close to 1), for expressing negative emotions (value close to -1), or for non-affective words (values around 0).

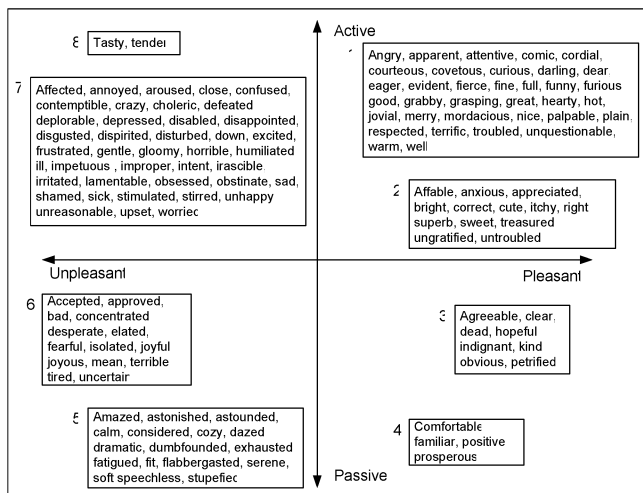


Figure 2: An example of 2D Space-Emotion Words in Octants-Related Pleasantness-Activation

We aimed to use both the labels and emotion words, which were found in the first experiment, as initial records of our emotion words database, by classifying the words by plotting them in a 2D space. The direction and distance of a vector represents the quality and intensity of the emotion words. For this purpose, we combined two approaches. Firstly, we used

(Kamps and Marx 2002)’s relative distance rules to develop a 2D space with bipolar dimension of Pleasant-Unpleasant and Active-Passive. We added a new octant: NEUTRAL with the EVA value = 0 and the ACT value also = 0. Figure 2 shows an example of the plotting. Using 140 emotion words, we found that the degree of correctness by this approach was between 78%.

Secondly, we applied multidimensional scaling (MDS) to represent emotion words also in 2D space using information relative to “similarity” (corresponding meaning) between each couple of emotion words. This procedure finds the configuration or cluster that approximates the observed distances in the best way. As initial input, we still used Kamps et.al.’s relative distances – the MPL of emotion words– to construct an  $N \times N$  matrix as the input matrix. The Euclidian distances among all pairs of points were applied to measure natural distances of those points in the space. By lowering the degree of correspondence between the Euclidian distance among points and the input matrix, the best corresponding MDS map can be achieved. Figure 3 shows an example of emotion words after MDS mapping. We found that the degree of correctness by this approach was between 65%. For both approaches, manual checking was still necessary to all mistaken placed emotion words.

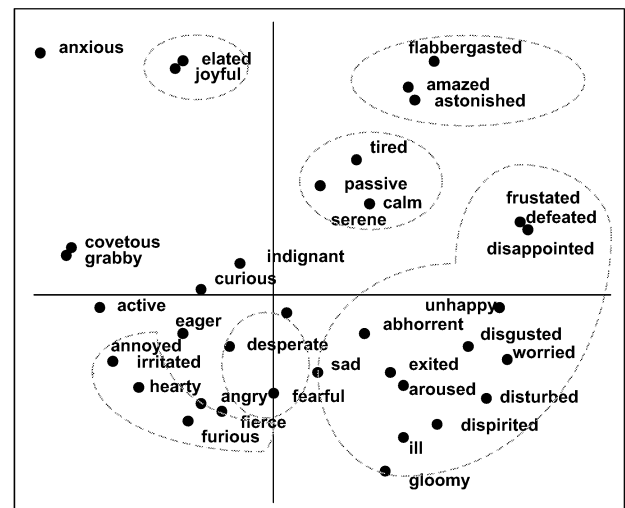


Figure 3: An Example of MDS-Emotion Words Mapping

Our lexical representation consists of an orthographic form, morphological specification, sense definition, and coordinates on 2D emotion circumplex and 2D MDS emotion mapping. This information is stored in our system’s ontology. We develop a heuristic function to manipulate both 2D spaces. The function receives an intended emotion type as input. As initial searching points, the coordinate of six emotion types  $u \in U$  {sadness, happiness, fear, anger, surprise, disgust} on the MDS-mapping space are already stored. The next step is to calculate approximate distances between the adjectives in the text  $T = \{w_1, w_2, \dots, w_n\}$  and other emotion words in  $E$ . Here, we are not interested on  $d_{ij}$  = the relative distance between  $w_i \in T$  and  $w_j \in E$ . However, it uses  $g_{ij}$  = the relative distance of the degree of pleasantness and  $a_{ij}$  = the relative distance of the degree of activation: Negative value means  $w_j$  is less pleasant or less active than  $w_i$ . The octant number of  $w_i$  gives information whether the

meaning of  $w_j$  is not in contradiction with  $w_i$ 's. The function also checks the relative distance between  $w_j$  and  $u_k$ . If the distance is above a threshold  $t_k$  (i.e. the average distance of the nearest two  $u$  values), the process will ignore  $w_j$ .

## MAPPING TEXT TO EMOTIONS

Our developed system extracts emotion indications from text using two approaches. First, the system analyzes the choices of words. It uses both 2D space emotion words. The system exploits WordNet to find matching synonym or antonym of the word, if the exact match cannot be found. The results of this process are the pleasantness degree of the emotion word, its activation degree and its emotion type. Finally, the system analyzes the emotion content of an entire sentence relatively to the entire utterances. For this purpose, it uses three approaches. Firstly, inspired by the language tagging of BEAT (Cassell et.al. 2001), our system also converts the input text into a parse tree. The root of the tree is the UTTERANCE, which is operationalized as an entire paragraph of input. The utterance is broken up into CLAUSES, each of which is held to represent a proposition. Instead of dividing the clauses into the word phrases, the system divides it into its thematic roles. Figure 4 shows an example of a parse tree. The tree also indicates the tag <emotion-object> for the emotion word, and the tag <contrast> the contrast coordinate words (e.g. "but") -, and contrast subordinate words (e.g. "although").

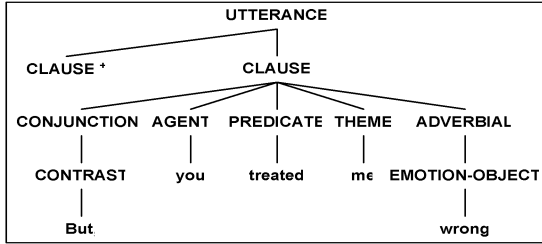


Figure 4: An Example of a Parse Tree

To divide a clause into its thematic roles, we define a case for every verb and follow the frame syntactic analysis used for generating the VerbNet (Kipper-Schuler 2003). The vocabulary is stored in the system's ontology. A lexeme has a link between the thematic roles and syntactic arguments. The definition also defines required and optional roles. Figure 5 a case for the verb "treat". Besides the verbs, we also store vocabularies of nouns, adjectives, pronouns, proper-nouns, conjunctions, and prepositions.

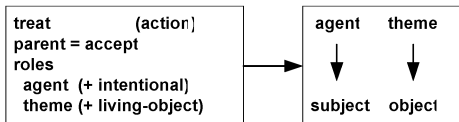


Figure 5: A Case of the Verb "treat"

According to Mulder et. al. 2004, emotions in language can be defined as text having a positive or negative orientation, an intensity, and a direction toward an object or event, which is described using three attributes: the experiencer, the attitude, and the object. Therefore, based on the thematic roles of the sentence, we developed heuristic rules to distinguish five types of emotional intentions: (a) emotionally

active toward an object, e.g. "I am angry with you"; (b) emotionally directed by an object, e.g. "you treat me wrong"; (c) emotions that provoked by an object, e.g. "his letter makes me sad"; (d) emotions that experienced towards an object, e.g. "this picture is really beautiful"; and (e) appraisal toward an object, e.g. "she is ill". Furthermore, other rules are developed to associate the emotion intentions with emotion type. A large number of corpora were analyzed to have more insight how humans express their emotion on a structure of words. The following is an example of rules:

```
If (agent ∈ pronoun) and (theme is "myself")
and (emotion-object is adverbial)
Then intention is "emotionally-directed"
If (intention is "emotionally-directed") and
(pleasantness is "unpleasant")
Then emotion-type is "anger"
```

Table 3: Distance Values between Emotions (Hendrix and Ruttkay 1998)

	Happiness	Surprise	Anger	Disgust	Sadness
Happiness	0	3.195	2.637	1.926	2.554
Surprise		0	3.436	2.298	2.084
Anger			0	1.506	1.645
Disgust				0	1.040
Sadness					0

Finally, we design the "emotion thermometer" to control the intensity of the emotional state in an utterance classified by six Ekman's universal emotions: happiness, sadness, anger, surprise, disgust, and fear (Ekman 1999). We classify emotion expressions in table 1 into these six universal emotions. If an emotion is active, the system will check its correspondence with the universal emotions. It calculates all thermometers (T) using the following equation:

$$T_i(t) = T_i(t-1) + I_i \cdot s$$

$$\forall j \neq i, T_j(t) = T_j(t-1) - d[j, i]$$

Where,  $i$  is the active universal emotion type,  $s$  is a summation factor,  $I$  is the emotion's activation degree, and  $j$  ranges over six universal emotion types. The distance between two universal emotions follows table 3. The highest value of the thermometers is considered as the current emotion state. This emotion state and its value (as the emotion's activation –calm, average or excited) are the end result of this process. A rule-base approach is used to govern the calculation. An example of the rule is:

```
If (tag <contrast> is found) and
(emotion-object is not found)
Then set active emotion is NEUTRAL
```

```
<UTTERANCE>
...
<CLAUSE></emotion emotion-type=neutral>
<CONJUNCTION><CONTRAST>but</CONTRAST>
</CONJUNCTION>
<AGENT>you</AGENT>
<PREDICATE>have treated</PREDICATE>
<THEME>me</THEME>
<ADVERBIAL>
  </emotion emotion-type="anger" degree="average">
  <EMOTION-OBJECT pleasantness="neg" degree="average">
    wrong
  </EMOTION-OBJECT>
</ADVERBIAL>
</CLAUSE>
...
</UTTERANCE>
```

Figure 6: An example of the result XML-Tagged Text

The results of this mapping text to emotions are written as XML-tagged text. Figure 6 shows an example.

## EVALUATION

We conducted another experiment to assess whether or not the developed system showed correct facial expressions for a given text input with emotion words. Instead of using dialogue recordings, we used our developed system by inputting two different texts to the system. As a first step, three independent observers marked the onset and offset of an expression. In the next step, these expressions were labelled according to the context of the text. The agreement rates between the observers in both steps were about 89%.

The experimental results indicated that most of the facial expressions (about 91%) showed by the 3D face correctly correspond to the emotion loading of the text spoken (see table 4). However the results also showed that some expressions were dependent not only on the choices of words but also on the context of the conversation. A word could mean different things according to the context of the dialogue. Thereby, the speaker or the listener might display different facial expressions.

Table 4: Statistics of Correct Agreements for Some Emotions

Expression	Happiness	Surprise	Anger	Disgust	Sadness
Mean	85%	78%	77%	92%	89%

## CONCLUSION

A prototype of a text based 3D synthetic face has been developed. The face is able to show synchronously with the speech. The developed system contains a component that is able to analyze the emotion content in the input text by the employment of ontology and our developed 2D-space affective lexicon database. The animation module of the system uses a facial expression dictionary. The rule-based approach for the emotion reasoning gives opportunities for us to extend both the system's believability and behaviours.

Our approach assumed that the emotion contents shown in the input text are explicit emotions. The system is able to show an appropriate facial expression as its stimulus response to convey an emotion loading in the text. Although our experimental results indicated that these expressions were influenced both by the choice of words and by the content of the conversation, it also showed the relation with emotional word depends mostly on the context, not on the word itself. It is caused by the fact, that a given word used in various situations can have different meaning. Future work will include emotion analysis from discourse information, such as moods, personality characteristics, anaphoric information and background contexts of the dialogue. Furthermore, we will extend the system more modalities, such as synchronized and coherence facial movements and lip movements.

## REFERENCES

- Cassell J., Vilhjálmsdóttir H. and Bickmore T., 2001, BEAT: The Behavior Expression Animation Toolkit. in *SIGGRAPH '01*.
- Condon W. and Ogston W., 1971, Speech and Body Motion Synchrony of the Speaker-Hearer. In D. Horton & J. Jenkins (Eds), *The perception of Language*, Academic Press, 150–184.
- Desmet P., 2002, *Designing Emotion*, Doctoral Dissertation, Delft University of Technology.
- Ekman P., 1999, Basic Emotions, In Dalglish T. & Power M., (Eds). *Handbook of Cognition and Emotion*, UK: John Wiley and Sons, Ltd..
- Ekman P., 1979, About brows: Emotional and Conversational Signals, In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (Eds), *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*,. Cambridge University Press, Cambridge, England; New-York, 169-248.
- Ekman P. and Friesen W.F., 1975, *Unmasking the Face*. Englewood Cliffs, New Jersey, USA: Prentice-Hall, Inc.
- Fellbaum C., 1998, *WordNet: An Electronic Lexical Database*, The MIT Press.
- Hatzivassiloglou V. and McKeown K.R., 1997, Predicting the Semantic Orientation of Adjectives, in *Proc. of ACL'97*, Spain.
- Hendrix J. and Ruttkay Zs. M., 1998, *Exploring the Space of Emotional Faces of Subjects without Acting Experience*, ACM Computing Classification System, H.5.2, I.5.3, J.4.
- de Jongh E. and Rothkrantz L.J.M., 2004, FED: an Online Facial Expression Dictionary, in *Euromedia 2004*, Eurosis, Ghent.
- Kamps J. and Marx M., 2002, Words with Attitude, in *Proc. of Global WordNet CIIL'02*, India., 332-341.
- Kipper-Schuler K., 2003. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis proposal, University of Pennsylvania.
- Liu H., Lieberman H., and Selker T., 2003, *A Model of Textual Affect Sensing using Real World Knowledge*, Technical Report, MIT Media Laboratory, Cambridge, USA.
- Mulder M., Nijholt A., den Uyl M. and Terpstra P., A Lexical Grammatical Implementation of Affect. In *TSD'04*, 171-178
- Nagao K. and Takeuchi A., 1994, Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation, in *Proc. of ACL'94*, USA, 102-109.
- Ortony A., Clore G., and Collins A., 1988, *The Cognitive Structure of Emotions*, Cambridge University Press.
- Ortony A., Clore G., and Foss M., 1987, *the Referential Structure of the Affective Lexicon*, Cognitive Science, 11: 341-364.
- Pelachaud C. and Bilvi M., 2003, Computational Model of Believable Conversational Agents, In *Communication in MAS: Background, Current Trends and Future*, Marc-Philippe Huget (Eds.), Springer-Verlag, to appear.
- Russell J.A., 1980, *a Circumplex Model of Affect*, Journal of Personality and Social Psychology, 39(6), 1161-1178.
- Schiano D.J., Erlich S.M., Rahardja K., and Sheridan K., 2000, Face to Interface: Facial Affect in (Hu)man and Machine, *Proc. of ACM CHI'00 Conference on Human Factors in Computing System*, NY: ACM, 193-200.
- Strapparava C. and Valitutti A., 2004, WordNet-Affect: An Affective Extension of WordNet, in *Proc. of LREC'04*, Portugal.
- Wojdel A. and Rothkrantz L.J.M., 2005, *Parametric Generation of Facial Expressions Based on FACS*, Computer Graphics Forum, Blackwell Publishing, number 4, no. 24, 1-15.
- Wojdel A., *Knowledge Driven Facial Modelling*, 2005, PhD Thesis Dissertation, TU Delft, The Netherlands.



# Using a sparse learning Relevance Vector Machine in Facial Expression Recognition

W.S. Wong, W. Chan, D. Datcu, L.J.M. Rothkrantz  
Man-Machine Interaction Group  
Delft University of Technology  
2628 CD, Delft,  
The Netherlands  
E-mail: L.J.M.Rothkrantz@ewi.tudelft.nl

## KEYWORDS

Facial expression recognition, face detection, facial feature extraction, facial characteristic point extraction, relevance vector machine, corner detection, AdaBoost, Evolutionary Search, hybrid projection.

## ABSTRACT

At TUDelft there is a project aiming at the realization of a fully automatic emotion recognition system on the basis of facial analysis. The exploited approach splits the system into four components. Face detection, facial characteristic point extraction, tracking and classification. The focus in this paper will only be on the first two components. Face detection is employed by boosting simple rectangle Haar-like features that give a decent representation of the face. These features also allow the differentiation between a face and a non-face. The boosting algorithm is combined with an Evolutionary Search to speed up the overall search time. Facial characteristic points (FCP) are extracted from the detected faces. The same technique applied on faces is utilized for this purpose. Additionally, FCP extraction using corner detection methods and brightness distribution has also been considered. Finally, after retrieving the required FCPs the emotion of the facial expression can be determined. The classification of the Haar-like features is done by the Relevance Vector Machine (RVM).

## INTRODUCTION

For the past decades, many projects have been started with the purpose of learning the machine to recognize human faces and facial expressions. Computer vision has become one of the most challenging subjects nowadays. The need to extract information from images is enormous. Face detection and extraction as computer-vision tasks have many applications and have direct relevance to the face-recognition and facial expression recognition problem. Potential applications of face detection and extraction are in human-computer interfaces, surveillance systems, psychology and many more. It is not so hard to imagine the importance of face detection in the means of face and emotion recognition. The importance of this subject can be ratified by the recent

terrorism bombings in London. Face detection and extraction of biometric features helps in the identification of the terrorists. In London, monitoring of people especially in the public places is done by closed-circuit cameras and televisions, which are linked via cables and other direct means. These can too be found in casinos and banks for instance. They are also used to aid in the prevention of calamities using face detection, emotion recognition and crowd behaviour analysis techniques.

Facial expressions are crucial in human communication. Human communication is a very complex phenomenon as it involves a huge number of factors: we speak with our voice, but also with our hands, eyes, face and body. The interpretation of what is being said does not only depend on the meaning of the spoken words. Our body language i.e. gestures modify, emphasize, and sometimes even contradict what we say. Facial expressions provide sensitive cues about emotional responses and play an important role in human communication. Therefore, it is valuable if this aspect of human communication can also be applied for more effective and friendly methods in man-machine interaction. According to Ekman et al. (Ekman and Friesen 1978) people are born with the ability to generate and interpret only six facial expressions: happiness, anger, disgust, fear, surprise and sadness. All other facial expressions have to be learned from the environment the person grows up. Humans are capable of producing thousands of expressions that vary in complexity, intensity, and meaning. Subtle changes in a facial feature such as tightening of the lips are sufficient to turn the emotion from happy to angry. And to think that the eyes and eye brows can also take on different shapes, one may imagine how complex the problem gets. In the past, Morishima et al. (Morishima and Harashima 1993) implemented a five-layered manual-input neural network which is used for recognition and synthesis of facial expressions. In (Zhao and Kearney 1996) a singular emotional classification of facial expressions is explained using a three-layered manual-input back propagation neural network. Kearney et al. (Kearney and McKenzie 1993) developed a manual-input memory-based learning expert system, which interprets facial expressions in terms of emotion labels given by college students without formal instruction in emotions signals. Rothkrantz et al. (Rothkrantz and Pantic 2000) proposed a point-based face model composed of two 2D facial views, namely the frontal- and the side view. Given a characteristic points based face model, expression-classification rules can be converted straightforwardly into the rules of an automatic classifier.

---

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

## RELATED WORK

The online Facial Expression Dictionary (FED) is an ongoing project at the Man-Machine Interaction group of the TU Delft (de Jongh 2002). The goal of the project is to develop a non-verbal dictionary which would contain information about non-verbal communication of people. Resembling a verbal dictionary, instead of words the FED contains facial expressions. This online non-verbal dictionary allows people to issue a query using description of the expression in terms of the expression classes (happiness, sadness, jealousy, etc.) and in terms of characteristics of the expression. Another interesting possibility of FED is the labelling of a picture containing a face. In other words an appropriate facial expression will be matched with the face. In the current state of the system the latter requires the user to select the region of the face and select the characteristic points of the face. These points are predefined consistent with the chosen face model. The face model used is that of Kobayashi and Hara (Kobayashi and Hara 1997). The facial expression recognition model is based on a three-tier framework. The chosen approach splits the FED system in three components, i.e. face detection, facial characteristic point (FCP) detection, tracking and classification. To fully automate the labelling process when inputting a picture, a project has been started on automating the face detection and the FCP detection part. This paper discusses the face detection and the FCP detection module.

## METHODOLOGY

In this section we describe the theoretical background and the methodology of our research. The detection process is based on the detectors described in (Viola and Jones 2001) and (Treptow and Zell 2003).

### Face detection

#### Haar-like features

In order to classify a face, some characteristic features need to be extracted. For this purpose, we used Haar-like features. These features have a rectangular shape and are fairly simple. The processing of this kind of features is computationally very efficient. In our face detection algorithm, five types of rectangular features are used (see Figure 1). Type 1, 2 and 5 are calculated as the sum of all pixels in the dark area minus the sum of all pixels in the light area. Type 3 and 4 are calculated as half the sum of all pixels in both dark areas minus the sum of all the pixels in the light area in the middle.

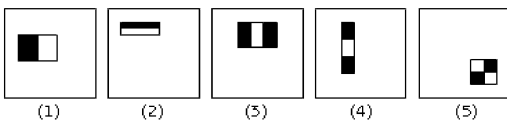


Figure 1: The five basic types of Haar-like features used in our approach

Each of the five basic features is scanned on every possible scale and every possible position within a training sample. Given that the sample's dimension is 24x24, the complete set

of features that can be constructed is quite large, namely 162336. From this set of features, we want the most relevant ones that best characterize the face. The best features are chosen using the AdaBoost learning algorithm (Figure 2).

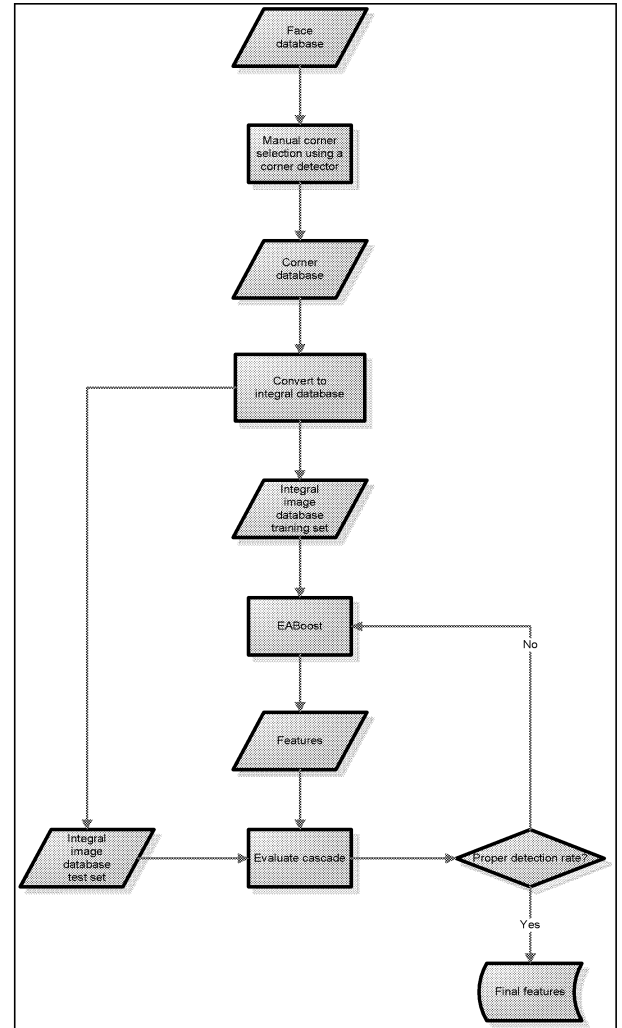


Figure 2: Scheme representing the training of weak classifiers

#### AdaBoost

The AdaBoost algorithm (Freund and Schapire 1995) aims at boosting the classification performance. It is an aggressive and effective algorithm used to select a low number of good classification functions, so called 'weak classifiers', to form a stronger classifier. The final strong classifier is actually a linear combination of the weak classifiers. Each weak classifier is restricted to the set of single feature functions. In the algorithm of (Viola and Jones 2001) the training stage for a single weak classifier involves the computation of a threshold for the feature value to discriminate between positive and negative examples. In our approach, the latter is slightly different. Instead of using a threshold, the chosen weak classifier is the Relevance Vector Machine - RVM for discriminating between the positive and negative examples. This means that for each feature, the weak RVM classifier determines the optimal classification function such that a minimum number of examples is misclassified.

The input of Adaboost is a predefined set of positive and negative training examples. In our case the positive examples are face images and the negative examples are non-face images. At the testing stage of face detection process, a set of scanning windows also called subwindows is extracted from the original image. Each element from the set is used as input for the cascaded classifier. The cascade generated at the training step has the form of a generate decision tree.

The structure of the cascade reflects the fact that within any single image on overwhelming majority of sub-windows are negative. As such, the cascade attempts to reject as many negatives as possible at the earliest stage possible (Figure 3). Every layer consists of only a small number of features. While a positive instance will trigger the evaluation of every classifier in the cascade, this is an exceedingly rare event.

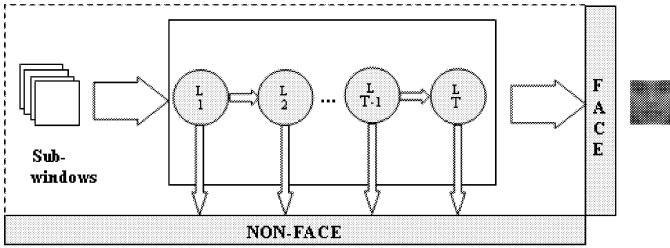


Figure 3: Cascaded classifier with T layers

#### Evolutionary Search

AdaBoost implies a brute force search on the whole space of rectangular Haar-like features. The process of training 162336 features would be time-consuming. Therefore it is beneficial to use GA in combination with AdaBoost (Figure 2).

The purpose of Genetic Algorithms (GA) in our research is to speed up the AdaBoost algorithm. This is done by replacing the exhaustive search of AdaBoost by a genetic search algorithm called Evolutionary Search (ES). The crossover and mutation genetic operators that drive the ES process are used for selecting features from the feature space. The fitness operator measures the performance associated to the use of a certain feature, for the all classification process. The population consists of 250 features. At each stage, a feature is selected so that it satisfies the fitness function for minimum error for all the generated features. The process is similar to the criterion AdaBoost uses to select weak features.

#### Relevance Vector Machine

Tipping (Tipping 2000) proposed the Relevance Vector Machine (RVM) to recast the main ideas behind SVMs in a Bayesian context. A prior is introduced over the weights controlled by a set of hyperparameters, one associated with each weight, whose most probable values are iteratively estimated from the data.

The results have been shown to be as accurate and sparse as SVMs yet fit naturally into a regression framework and yield full probability distributions as their output.

The results in the case of face detection given some kernel functions are presented in Table 1. In the case of three features, the most efficient kernel function is chosen by using

ROC curves. By analyzing Figure 4, it can be concluded that the best classification is obtained by using Laplace 4.0. This kernel function is further used in EABOOST and in the process of constructing the final strong cascaded classifier.

Table 1: 2-fold cross validation results on three weak classifiers for face detection based on Haar-like features

Kernel	Error rate		
	Feature 1	Feature 2	Feature 3
Gauss 2.0	26.30% $\pm$ 0.85	35.45% $\pm$ 7.71	38.60% $\pm$ 1.84
Gauss 5.0	25.35% $\pm$ 5.30	32.40% $\pm$ 2.83	36.55% $\pm$ 3.61
Laplace 0.5	35.20% $\pm$ 14.42	26.70% $\pm$ 0.99	42.70% $\pm$ 9.62
Laplace 2.0	29.25% $\pm$ 9.83	32.00% $\pm$ 7.50	41.60% $\pm$ 7.78
Laplace 5.0	26.20% $\pm$ 2.12	25.90% $\pm$ 1.84	37.45% $\pm$ 7.57

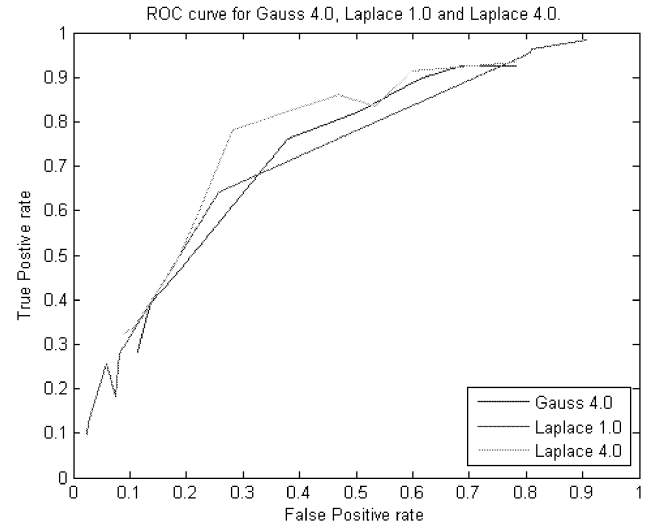


Figure 4: ROC curves of three kernels, obtained by adjusting each classifier's threshold

#### Facial characteristic point detection

Facial characteristic points (FCP) consist of 30 facial points. The detection of FCPs is based on a set of techniques that include corner detection, RVM and hybrid projection methods. The scanning of the whole picture is avoided by the use of a corner detector as a primary step for mouth, eyes and eyebrow FCP detection.

#### Fused corner detector

The corner detector of (Harris and Stephens 1988) takes into account the edge information. It also considers the neighborhood for corner decision, since the gradient swings sharply around the corners. The algorithm of (Sojka 2003) determines what neighborhoods are relevant for deciding whether or not a point is a corner by using a probability function. Where other corner detectors implicitly take into account the corner angle, the Sojka corner detection algorithm explicitly computes the corner angle. This helps to reduce erroneous detection of corners on contrast edges.

A mixed Harris-Stephens and Sojka corner detection algorithm is tuned to select enough corners so that the FCPs are also included.

This is because neither of the two detection algorithms is effective enough in detecting corners, which also includes all corner points of the facial features (mouth corners, eye corners, and eyebrow corners). Tests show that the efficiency of detecting the corner FCPs is increased by using such a combination. Given the selected set of corners, the next step is to identify the FCPs.

#### *Classification of candidate corners*

To classify the candidate corners selected by the fused detection algorithms, a set of RVMs is trained. For every corner point type a different RVM classifier is trained to distinguish the point from other points detected with the corner detector. The training model for corners employs the boosting of simple rectangle features. The set of features is limited to five basic feature types.

#### *Hybrid projection*

Using the combination of corner detectors along with RVM does not enable us to detect all facial characteristic points. To detect the remaining FCPs, a projection method called the hybrid projection (Zhou and Geng 2002) is used. The FCPs can be located at the corresponding boundaries of a face feature. To locate the horizontal boundaries of the features we analyze the horizontal intensity variations in the image containing only the face feature.

The final step aims at deriving the FCPs that were not identified at the previous stages, by using a hybrid projection method. This can be done by calculating a parabolic curve through the detected FCPs.

## IMPLEMENTATION AND RESULTS

### **Face detection**

We designed a learning model to boost the performance of RVM. This model consists of different techniques and algorithms described in the current paper.

The learning procedure is based on the AdaBoost learning algorithm. This algorithm is perfectly suited for the selection of the best features that boost up the performance of the classifier. As known for AdaBoost training, it is slow since it contains a brute force search. In addition, training of the RVM itself is relatively slow. And since they are combined, there is a continuous feedback from RVM to AdaBoost and the other way around.

A genetic search algorithm is added to improve the learning speed. Instead of a training time in the order of weeks/months, this is reduced to hours/days (on an AMD Athlon™ XP 2200+ 1.80 GHz processor with 512 MB RAM). Note that the size of the chosen training dataset is also significant for the speed of the training. After the learning procedure, faces can be distinguished from non-faces using the trained RVMs.

A cascade consists of several layers of classifiers. Each classifier is a combination of a number of RVMs. A practical

problem that we encounter incorporating the cascade technique is that a lot of RVMs need to be trained.

Table 2: RVM test results, both training and testing are performed on MIT CBCL database

Kernel	Nr. of test samples	Detection rate %	Nr. of false negatives	Nr. of false positives
Gauss 5.0	2500	93.92	103	49
Gauss 7.0		95.08	58	49
Laplace 2.0		83.88	339	64
Laplace 5.0		95.04	60	64

Given a few kernel functions, the results of RVM classifier for face detection are presented in Table 2 for the same testing set as the training set, and in Table 3 for different data set for testing stage.

Table 3: RVM test results, the training is done using MIT CBCL, the testing is done on CMU database

Kernel	CMU database consisting of faces only		CMU database consisting of non-faces only	
	Number of test samples	Detection rate %	Number of test samples	Detection rate %
Laplace 2.0	472	22.03	5036	100
Laplace 5.0		51.91		97.34
Gauss 5.0		38.77		96.68
Gauss 7.0		30.30		97.86

However, the test results show that improvement needs to be made. In the current state, the face detector consists of only five layers of classifiers. Recall that in (Viola and Jones 2001) a cascade of 32 layers with over 4000 features is used. Better results are expected by involving more classifiers to the face detector.

### **Facial characteristic point detection**

The same learning model for training the face detection classifier is used for the FCP detection component. Unlike in the case of face detection, no databases of FCPs exist which we can use as our dataset. These databases are extracted manually by us from the BioID and Carnegie Mellon face database. For the detection of the FCPs, a corner detection algorithm is used to filter out the non-FCPs. We have chosen for a combination of the Harris corner detection algorithm and the Sojka corner detection algorithm. Not all of the non-FCPs can be filtered out by these corner detectors. For this, we rely on the corresponding RVMs. The performance of the RVM in the final system is actually determined by that of the corner detectors.

For the FCPs that cannot be detected by the corner detectors, we use the Hybrid Projection technique. This technique is applied on the corresponding facial feature (eye, eye brow and mouth) on which the FCP is localized. Therefore, RVMs are trained to extract these facial features before applying the projection method. The test results of the FCP detector (see Table 4) show that some of the FCPs can be detected better

than others. The explanation for the relative poor performance of some FCPs is probably that the FCP pattern itself is non-stable from the recognition point of view. For instance, the mouth can have different shapes and some associated parameters could exceed the value ranges of the samples used at the training stage, at different expressions. To detect the FCPs we need to account that noise is very probable at corner regions. Taking this into account it means that at the training of the RVM noise is included in the training samples. This affects the final performance of the RVM. It is a trade-off that needs to be made. In the case of invoking the projection method, finding the boundaries is proven to be very robust, except if the feature boundary is distorted.

Table 4: FCP Detection Results

FCP	True positive rate (%)	False positive rate (%)
Right eye inner corner	81.82	6.75
Right eye outer corner	81.82	16.67
Right eye upper corner	88.64	11.63
Right eye lower corner	88.64	11.63
Left eye inner corner	81.82	3.49
Left eye outer corner	63.64	5.94
Left eye upper corner	82.95	17.05
Left eye lower corner	82.95	17.05
Mouth left corner	86.36	3.24
Mouth right corner	90.91	4.71
Mouth upper corner	90.91	9.08
Mouth lower corner	90.91	9.08

## CONCLUSION

We have presented an approach using a sparse learning model as the first step towards a fully automatic facial expression recognition system. This learning model is applied on face detection and FCP detection. The test results reveal that some improvements are still to be made.

In the current situation, a detected face cannot be further processed by the FCP detection module if the face is slightly rotated. Some of the FCPs can be occluded by other parts of the face. The face detection module is trained on a database with unaligned faces. Some of them are slightly rotated to the left, some to the right, some looking up, etc. For the two modules to work together perfectly, the face detector should be trained strictly on full frontal aligned faces. This is because the FCP detection module is designed to work with these faces.

The model may be improved by considering a faster implementation of the training application. Other variants of the AdaBoost may also be considered. They differ in the updating schemes for the weights. In the face detection module, the scanning process can be speed up by other techniques. Using edge detectors, plain backgrounds might be filtered out and pruned from being scanned. This reduces the overall scanning time on different resolutions. The performance of the system can also be improved by using an extended set of the Haar-like features. In our training model, we used only 5 simple features. The detection rate during training may be increased by incorporating the bootstrapping method. This method uses misclassified samples as training

input in the next iteration. This way we can force the learning algorithm to adapt the output results from previous training rounds. We have not implemented this procedure in the current training model because this would certainly affect the training time negatively.

## REFERENCES

- Ekman, P. and W. Friesen. 1978. "Facial Action Coding System." Consulting Psychologists Press, Inc., Palo Alto California, USA.
- Freund, Y. and R.E. Schapire. 1995. "A decision-theoretic generalization of on-line learning and an application to boosting." In *2<sup>nd</sup> European Conference on Computational Learning Theory*.
- Harris, C.G. and M. Stephens. 1988. "A combined corner and edge detector". *Proceedings 4th Alvey Vision Conference*, Manchester, 189-192.
- Jongh de, E.J. 2002. "FED: An online facial expression dictionary as a first step in the creation of a complete nonverbal dictionary." TU Delft.
- Kearney, G.D. and S. McKenzie. 1993. "Machine interpretation of emotion: design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions." (JANUS). *Cognitive Science* 17, Vol. 4, 589-622.
- Kobayashi, H. and F. Hara. 1997. "Facial Interaction Between Animated 3D Face Robot and Human Beings". *IEEE Computer Society Press*, 3732-3737.
- Morishima, S. and H. Harashima. 1993. "Emotion Space for Analysis and Synthesis of Facial Expression". *IEEE International Workshop on Robot and Human Communication*, 674-680.
- Rothkrantz, L.J.M. and M. Pantic. 2000. "Expert Systems for Automatic Analysis of Facial Expressions". *Elsevier, Image and Computing*, Vol. 18, 881-905.
- Sojka E. 2003. "A New Approach to Detecting Corners in Digital Images". Accepted for Publication in *IEEE ICIP*.
- Tipping, M.E. 2000. "The Relevance Vector Machine. *Advances in Neural Information Processing Systems*". Vol. 12, 652-658.
- Treptow, A. and A. Zell. 2003. "Combining Adaboost Learning and Evolutionary Search to Select Features for Real-time Object Detection". *University of Tuebingen, Department of Computer Science, Germany*.
- Viola, P. and M. Jones. 2001. "Robust Real-time Object Detection." *Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling*.
- Zhao J. and G. Kearney. 1996. "Classifying facial emotions by back-propagation neural networks with fuzzy inputs". *International Conference on Neural Information Processing*, Vol. 1, 454-457.
- Zhou, Z.-H. and X. Geng. 2002. "Projection Functions for Eye Detection". *State Key Laboratory for Novel Software Technology, NU, China*.



# **NETWORK MANAGEMENT OPTIMIZATION**





# VARIABILITY OF INTERNET TRAFFIC IN MULTIPLE TIME SCALES AND ITS RELEVANCE FOR QUALITY OF SERVICE

Gerhard Haßlinger

T-Systems Enterprise Services GmbH, ENPS Technologiezentrum

Deutsche Telekom Allee 7, D-64307 Darmstadt, Germany

E-mail: gerhard.hasslinger@t-systems.com

## KEYWORDS

Traffic statistics and measurement; quality of service (QoS); load thresholds for critical QoS conditions

## ABSTRACT

The characteristic of Internet traffic is responsible for temporary overload phases and their impact on variable delay and data loss as main indicators of quality of service (QoS) degradation. We investigate statistical properties of the traffic rate variability on ADSL broadband access platforms, which presently connect 140 million residential users to the Internet [4]. Measurement confirms smooth traffic profiles aggregated via ADSL with less influence of long range correlation than experienced for traffic on Ethernet LANs.

## 1 INTRODUCTION: TRAFFIC IN DIFFERENT TIME SCALES

IP networks measurements usually include 5- or 15-minute mean values of the traffic rate to determine the load of links or a complete traffic matrix of flow intensities between the edges of a (sub-) network. This data forms a basis for network planning and a process of network resource upgrades to adapt to the steadily increasing Internet traffic volume [11]. The 5-/15-minute traffic samples can be collected from standard statistics of IP and MPLS routers without stressing the performance of the routing equipment. They are appropriate to evaluate daily traffic profiles showing the peak rates during busy hours, which are most relevant for network dimensioning.

On the other hand, they do not include all relevant time scales to ensure quality-of-service, since congestion may arise on small time scales e.g. of some seconds being compensated by alternating phases of low load which make them invisible on longer time scales. The impact of traffic variability on QoS is essential even in time frames below 1s. Buffers may be capable to bridge temporary overload on account of delay for the buffered data, whereas long term overload phases cause buffer overflow. Real time applications with strict delay bounds, e.g. less than 0.2s for conversation, limit the waiting time and corresponding buffer sizes.

Since 1990 evaluations of IP traffic measurement revealed long range dependency and self similar pattern over the relevant time scales [10]. While most of this measurement was conducted on Ethernet LANs, ADSL broadband access is presently carrying a major and increasing portion of the Internet traffic. Differences between ADSL and Ethernet

traffic profiles are experienced by [2] based on traffic samples taken at 1s intervals from the digital subscriber line access modules (DSLAMs).

We investigate comparable traffic measurement at the interconnection of ADSL and the IP backbone. In Section 2, we analyse the variability of samples at several time scales starting below 1s. The implications of traffic profiles for waiting times as the main QoS indicator are studied in section 3 in order to estimate load thresholds on transmission links which indicate critical QoS conditions.

## 2 IP TRAFFIC MEASUREMENT

For measurement purposes, we consider the amount of arriving data in a time slotted system, where the time is subdivided into a series of subsequent intervals of length  $\Delta$ . In order to represent the process of arriving traffic in detail, each arriving IP or MPLS packet can be registered with a time stamp as well as its packet size. The storage for measurement traces in this representation is increasing with the line speed, where millions of packets may be counted per second on high speed connections like a 10 Gbit/s link.

On the other hand, knowledge about the amount of data arriving e.g. per millisecond allows to determine waiting times as a main QoS indicator at the same precision of milliseconds. Then a first evaluation step calculates the data volumes  $d_m$  in byte for a series of time slot of length  $\Delta$ . This requires limited storage for  $M = S/\Delta$  integers to represent a traffic trace over  $S$  seconds independent of the speed of the considered transmission line. The traffic rate in each time slot is given by  $R_m^{(\Delta)} = d_m / \Delta$  for  $m = 1, \dots, M$ .

Figure 1 represents corresponding traffic traces with intervals starting at the time scale  $\Delta = 0.01s$ . From a trace on a time scale  $\Delta$ , the traffic rates  $R_m^{(K\Delta)}$  for longer time scales with intervals  $K \cdot \Delta$  are computed by the mean over  $K$  subsequent intervals, ( $K = 2, 3, \dots$ ) such that

$$R_1^{(K\Delta)} = \sum_{k=1}^K R_k^{(\Delta)} / K, R_2^{(K\Delta)} = \sum_{k=K+1}^{2K} R_k^{(\Delta)} / K, \dots$$

Figure 1 includes 3 time scales  $\Delta = 0.01s$ ,  $\Delta = 0.1s$  ( $K = 10$ ) as well as  $\Delta = 1s$  ( $K = 100$ ). It is apparent, that traffic becomes smoother when observed on longer time scales. As a usual measure of variability we consider the coefficient of variation, i.e. the ratio  $\sigma^{(\Delta)} / \mu^{(\Delta)}$  of the standard deviation and the mean, where

$$\mu^{(\Delta)} = \sum_{j=1}^M R_j^{(\Delta)} / M \quad \text{and} \quad \sigma^{(\Delta)} = \sqrt{\sum_{j=1}^M (R_j^{(\Delta)} - \mu^{(\Delta)})^2 / M}.$$

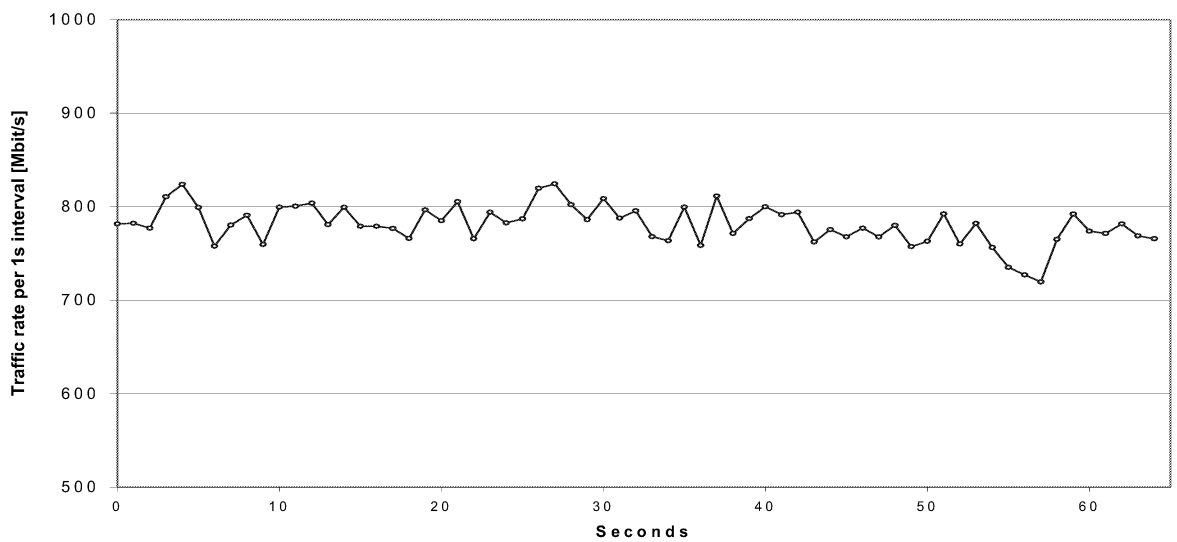
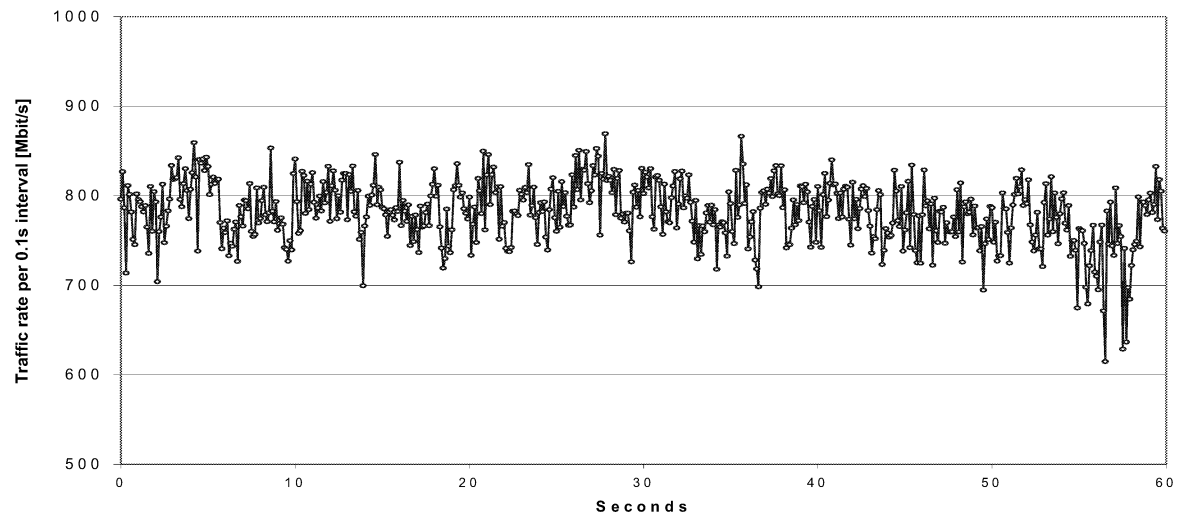
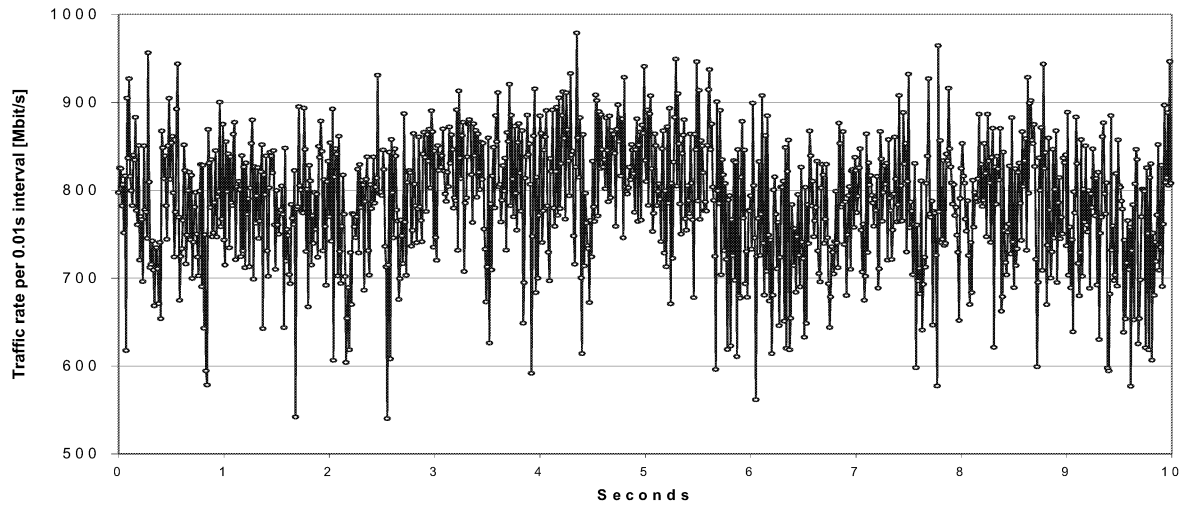


Figure 1: Traffic variability at time scales of 0.01s, 0.1s and 1s

Considering a longer time scale  $K\Delta$ , the mean value is preserved  $\mu^{(K\Delta)} = \mu^{(\Delta)}$ . The standard deviation and the coefficient of variation  $\sigma^{(\Delta)}/\mu^{(\Delta)}$  are preserved if and only if the traffic rate is constant for each sequence of  $K$  intervals

$$R_{nK+1}^{(\Delta)} = R_{nK+2}^{(\Delta)} = \dots = R_{n(K+1)}^{(\Delta)} = R_n^{(K\Delta)} \text{ for } n = 0, 1, \dots, (M/K) - 1,$$

which are comprised on the longer time scale. Otherwise, the coefficient of variation is smaller  $\sigma^{(K\Delta)}/\mu^{(K\Delta)} \leq \sigma^{(\Delta)}/\mu^{(\Delta)}$ . When  $R_m^{(\Delta)}$  is a sequence of independent and identically distributed random variables, then the coefficient of variation is decreasing with  $\sqrt{K}$  since

$$\sigma^{(K\Delta)} = \frac{\sqrt{\sum_{j=1}^K (R_j^{(\Delta)} - \mu^{(\Delta)})^2}}{K} = \frac{\sqrt{K \cdot (\sigma^{(\Delta)})^2}}{K} = \frac{\sigma^{(\Delta)}}{\sqrt{K}}.$$

Traffic measurement has been studied at different time scales by [2] starting from 1s intervals. In addition, this work investigate modeling approaches including M/G/ $\infty$  for the arrival process. The analysis is carried out with different assumptions on the distribution of the flow size, which imply autocorrelated traffic rates  $R_m^{(\Delta)}$ . Therefore Gaussian distributions of the traffic rate are confirmed as asymptotical behaviour [8]. The standard deviation  $\sigma^{(1s)}(R)$  as well as the quantiles  $\gamma_{99\%}^{(1s)}(R)$  of the aggregated traffic rate are shown to increase with the square root of the mean rate  $\mu(R)$ :

$$\sigma^{(1s)}(R) = f_{\sigma}^{(1s)} \sqrt{\mu(R)}; \quad \gamma_{99\%}^{(1s)}(R) = \mu(R) + f_{99\%}^{(1s)} \sqrt{\mu(R)} \text{ where} \\ \Pr \{ R^{(1s)} \leq \mu(R) + f_{99\%}^{(1s)} \sqrt{\mu(R)} \} = 99\%.$$

The factors  $f_{\sigma}^{(1s)}$  and  $f_{99\%}^{(1s)}$  can be determined from the source model with known distribution of the flow size or from available traffic measurement on the network links.

As a criterion for sufficient QoS, the 99%-quantile of the measurement in the time scale of 1s intervals is used by [2] to indicate the demand for capacity  $C$ :

$$C \geq \gamma_{99\%}^{(1s)}(R) = \mu(R) + f_{99\%}^{(1s)} \sqrt{\mu(R)}.$$

The measurement is taken from the DSLAMs in ADSL access platforms as well as for Ethernet traffic. The traffic aggregates of the DSLAMs show a smooth pattern, characterized by the fact that about 99% of the rates  $R_m^{(1s)}$  observed over 1s intervals stay below  $\mu(R) + \sqrt{\mu(R)}$ , where the 5-minute mean value is taken as long mean  $\mu(R)$ :

$$\Pr \{ R_m^{(1s)} \leq \mu(R) + \sqrt{\mu(R)} \} \approx 99\% \text{ with rates in Mbit/s.}$$

This corresponds to a factor  $f_{99\%}^{(1s)} \approx 1$  where traffic rates are represented in Mbit/s. The result is confirmed as a good fit of the statistics at the DSLAM aggregation level with traffic flows with mean  $\mu(R) \leq 20$  Mbit/s [2].

We consider comparable measurement taken at broadband access routers of Deutsche Telekom's IP platform, which connects ADSL access regions to the IP backbone. In Figure 1, the traffic rate variability of a 2.5 Gbit/s link is

captured at time scales starting from  $\Delta = 0.01$ s. It is visible that the variability becomes essentially smaller for longer intervals.

Table 1 summarizes main parameters of our measurement statistics over four time scales, including the coefficients of variation, the 99%-quantiles  $\gamma_{99\%}$  and the maxima in a 15-minute time frame with mean rate  $\mu \approx 781$  Mbit/s. On all those time scales, the distribution of the traffic rate is observed to be close to a Gaussian distribution.

Table 1: Parameters of traffic profiles in multiple time scales

Characteristics of traffic variability in multiple time scales	$\frac{\text{Max} - \mu}{\mu}$	$\frac{\gamma_{99\%} - \mu}{\mu}$	$\frac{\sigma}{\mu}$
$\Delta$ : 10 s ( $\mu \approx 781$ Mb/s)	0.051	0.051	0.021
$\Delta$ : 1 s	0.084	0.067	0.029
$\Delta$ : 0.1 s	0.149	0.102	0.045
$\Delta$ : 0.01 s	0.350	0.200	0.094

The classical results of Internet traffic measurement from the mid of the 1990-ties on the contrary observed long range dependencies over almost any time scale, which motivated the introduction of self-similar traffic models [3][10]. Only a minor smoothing effect of the variability was observed on larger time scales. Two reasons which may account for such a change in ADSL platforms are:

- Most measurements showing self-similar pattern were conducted on Ethernet LANs or on WANs with prevalent Ethernet access, which pose less restriction on the access rate of each user than ADSL platforms. While each Ethernet access is equipped with at least 10Mbit/s, most ADSL access lines are still far below this rate. On the other hand, the user population on ADSL platforms is far larger than it had been in the 1990-ties on Ethernet LANs. Therefore an essentially higher multiplexing level of many small and independent flows can be expected for current ADSL traffic aggregates.
- This trend is strengthened by the dominant traffic volume of peer-to-peer file sharing applications, which subdivide the download of large files into many small data units to be transmitted in parallel TCP connections from different sources [1][7].

Thus a high fluctuation of TCP connections in short term is to be expected for current ADSL broadband access, which may detract from long range dependency. Traffic in Ethernet environments is still experienced to be more variable than over ADSL due to measurement compared by [2] and, as a consequence, the authors do not recommend to transfer the results on QoS criteria from ADSL to Ethernet traffic.

The assumption of uncorrelated i.i.d. rates  $R_m^{(\Delta)}$  per interval yields  $\sigma^{(K\Delta)} = \sigma^{(\Delta)}/\sqrt{K}$  and thus the coefficient of variation  $\sigma^{(K\Delta)}/\mu^{(K\Delta)}$  is decreasing by the factor  $\sqrt{K}$  for increasing interval length  $K\Delta$ . Therefore factors of  $\sqrt{10} \approx 3.17$  between entries in the last column of Table 1 would corre-

spond to i.i.d. samples. The differences from a time scale to the next one are smaller, thus indicating the presence of autocorrelation in the measured traffic.

In order to check the QoS criterion proposed by [2], we took the quantiles of the sequence of traffic rates  $R_m^{(1s)}$  for 1s intervals over a quarter of an hour in the same example shown in Figure 1 and Table 1 and obtained

$$\Pr \{ R_m^{(1s)} \leq \mu(R) + 1.5 \sqrt{\mu(R)} \} \approx 99\% \text{ and thus } f_{99\%}^{(1s)} \approx 1.5$$

when the mean rates are again given in Mbit/s. Including several other traffic flows with mean rates in the 1 Gbit/s we observed factors in the range  $1.4 < f_{99\%}^{(1s)} < 2$ . In addition, a number of MPLS flows with mean rates of 30 – 40 Mbit/s have been evaluated, yielding  $1.7 < f_{99\%}^{(1s)} < 2.5$  for the 99%-quantile. In principle, this confirms the approach taken by [2], where the variability is experienced to be up to 2.5-fold higher in our measurements. Therefore the factor  $f_{1\%}^{(1s)}$  should be carefully determined from measurement.

### 3 LOAD DEPENDENT WAITING TIMES

Traces of the amount of arriving traffic per millisecond or in other time scales  $\Delta$  can be used to determine the course of the waiting time at the same accuracy  $\Delta$ . We presume

- a constant available bandwidth  $C$  in Mbit/s and
- a buffer size  $B$  in Mbit at the router interface.

Initially, the buffer is assumed to be empty and the waiting time is zero. When the data is forwarded in the sequence of its arrival, we can iteratively compute the waiting time after the first, second slot etc. of a considered traffic trace. The amount  $C\Delta$  of data can be served per time slot.

We assume that this capacity is already available for the data arriving in the same time slot. The latter assumption is optimistic, since the data may arrive non-uniformly over the interval. But the difference to a pessimistic assumption that none of the data arriving in the same slot is forwarded, makes a difference of no more than  $\Delta$  in the waiting time. Let

- $A_k$  denote the amount of data arriving in the  $k$ -th slot
- and  $W_k$  denote the waiting time after the  $k$ -th time slot.

Then the amount of buffered data is increasing by  $A_k - C\Delta$  in the  $k$ -th time slot, if the amount  $A_k$  of arriving data is larger. The waiting time  $W_{k+1}$  after the  $k$ -th time slot can be calculated from the waiting time  $W_k$  beforehand:

$$W_{k+1} = \text{Max}(\text{Min}(W_k + A_k/C - \Delta, B/C), 0)$$

The formula accounts for a difference  $A_k/C - \Delta$  in the workload and for a limited range  $[0, B/C]$  of the waiting time, since data is dropped when the buffer size  $B$  is exceeded.

Considering a traffic trace over  $M$  intervals, a corresponding series of waiting times  $W_1, W_2, W_3, \dots, W_M$  after each time slot is determined starting from  $W_0 = 0$ . Next we obtain the relevant statistical parameters including the mean, maximum, and the quantiles of the waiting time. The analysis is applied to the measurement shown in Figure 1, where we assumed a time scale of  $\Delta = 0.01s$  and an infinite buffer, such that QoS degradation becomes visible only through high waiting times. Then the computation of waiting times is simplified:  $W_{k+1} = \text{Max}(W_k + A_k/C - 0.01, 0)$ . The evaluation in the course of a traffic trace is shown in Figure 2.

The analysis can be done for arbitrary capacities  $C$  and corresponding utilization  $\mu(R)/C$ . Figure 3 shows the maximum and mean waiting times again for the example of the measurement trace of Figure 1 for different utilization levels.

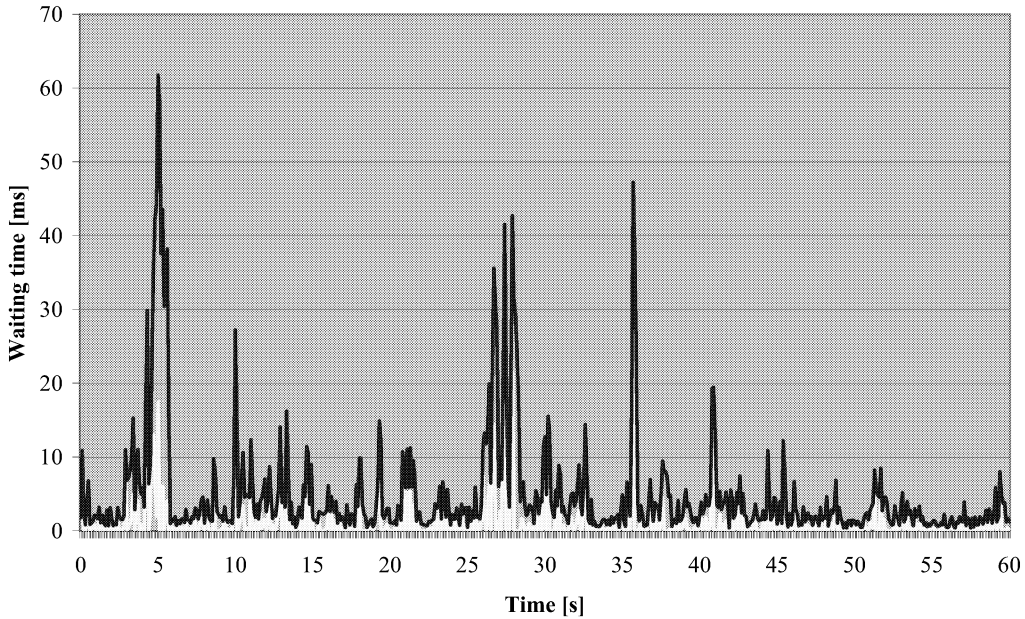


Figure 2: Waiting times in the course of a one minute traffic trace for utilization at a QoS-critical threshold

Up to now, the evaluations are carried out for the complete traffic on a 2.5 Gbit/s link with mean rate close to 1 Gbit/s. For smaller traffic aggregates in the ADSL access area, a higher variability is expected. In the sequel, we investigate single MPLS traffic flows over the measured link, which can be separated by their label in the shim header preceding each IP packet for multiprotocol label switching. An MPLS flows is usually provided for traffic between a pair of edges of the backbone network. MPLS flows with mean rates up to 40Mbit/s are included in the measurement, which is well below the total traffic on a link.

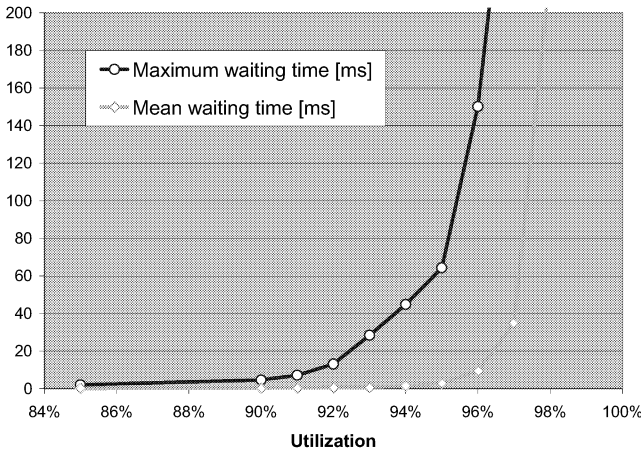


Figure 3: Waiting times as a function of the utilization

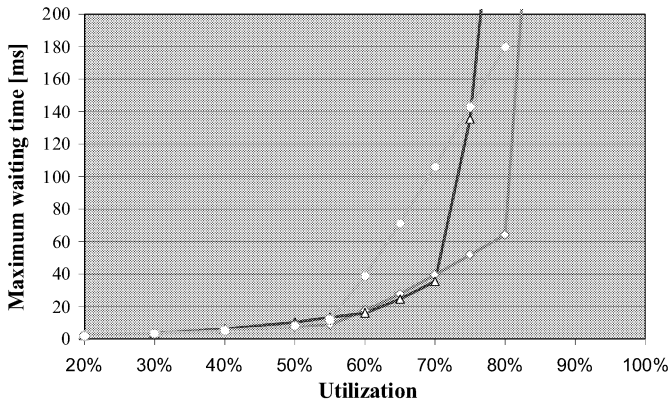


Figure 4: Waiting times as a function of the utilization for three MPLS traffic flows with mean rates of 30 – 40 Mbit/s

Regarding the traffic profiles of the flows, Table 2 shows the parameters of the traffic statistics for a typical example. A comparison with Table 1 reveals that the coefficient of variation and other measures of variability are 3 – 10 times larger on all included time scales.

As a consequence, the QoS-critical thresholds of the utilization become essentially lower for smaller traffic aggregates. The resulting waiting times for different provided bandwidths and corresponding utilization levels are determined in the same way as for the total link traffic. Figure 4 indicates the thresholds, where the maximum waiting time is again represented as a function of the admitted load  $\mu(R)/C$  for three measured MPLS traffic traces with mean rates

$\mu(R)$  in the range 30 – 40 Mbit/s. For those cases, the waiting times become critical already at 50% – 80% utilization.

Table 2: Parameters of traffic profiles in multiple time scales

Characteristics of MPLS traffic flow variability	$\frac{\text{Max} - \mu}{\mu}$	$\frac{\gamma_{99\%} - \mu}{\mu}$	$\frac{\sigma}{\mu}$
$\Delta: 10 \text{ s } (\mu \approx 40 \text{ Mb/s})$	0.34	0.34	0.122
$\Delta: 1 \text{ s}$	0.51	0.42	0.145
$\Delta: 0.1 \text{ s}$	0.73	0.49	0.174
$\Delta: 0.01 \text{ s}$	2.40	0.94	0.322
$\Delta: 0.001 \text{ s}$	14.31	3.00	0.887

## CONCLUSIONS

The evaluation of the variability of traffic generated on ADSL broadband access platforms shows that traffic in the backbone is smooth owing to the statistical multiplexing effect. Although the peaks and phases of possible overload are essentially increasing in smaller time scales, the utilization thresholds indicating QoS degradation are high on backbone links. As a main indicator of the QoS properties we calculate the maximum waiting time as a function of the load based on traffic traces.

Nevertheless, for smaller traffic flows on the first aggregation levels of a broadband access platform, a higher variability is observed allowing for only medium utilization in order to avoid critical QoS conditions.

## ACKNOWLEDGEMENTS

I would like to thank Joachim Mende and Franz Hartleb for their support in evaluating the measurement data.

## REFERENCES

- [1] N. Ben Azzouna, F. Cl  rot, C. Fricker and F. Guillemin, A flow-based approach to modeling ADSL traffic on an IP backbone link, *Annals of Telecom.* 59 (2004) 1252-1255
- [2] H. van den Berg et al., QoS-aware bandwidth provisioning for IP backbone links, *Computer Networks* 50 (2006) 631-647
- [3] M.E. Crovella and A. Bestavros, Self-similarity in world wide web traffic: Evidence and possible causes, *IEEE/ACM Transactions on Networking* 5 (1997) 835-846
- [4] DSL Forum, News Release Q3'05 (2005) [www.dslforum.org/PressRoom/0527\\_Q305dsl\\_figures.pdf](http://www.dslforum.org/PressRoom/0527_Q305dsl_figures.pdf)
- [5] Fraleigh et al., Packet-level traffic measurements from the Sprint IP backbone, *IEEE Network* (Nov./Dez. 2003) 6-16
- [6] G. Ha  linger, Quality-of-service analysis for statistical multiplexing with Gaussian and autoregressive input modeling, *Telecommunication Systems* 16 (2001) 315-334
- [7] G. Ha  linger, ISP Platforms under a heavy peer-to-peer workload, In: R. Steinmetz and K. Wehrle (eds.): *P2P Systems and Applications*, Springer LNCS 3485 (2005) 369-382
- [8] G. Ha  linger, Statistical properties observed for traffic on Internet platforms, *Proc. ESM'2005, Porto* (2005) 324-328
- [9] J. Kilpi, I. Norros, Testing the Gaussian approximation of aggregate traffic, *Proc. Internet Measur. Workshop, Marseille, France*, 2002
- [10] W. Leland, M. Taqqu, W. Willinger, D. Wilson: On the self-similar nature of Ethernet traffic, *IEEE Trans. on Networking* 2 (1994) 1-15
- [11] A. Odlyzko, Internet traffic growth: Sources and implications, *Proceedings SPIE Vol. 5247* (2003) 1-15

# Mill: Scalable Area Management for P2P Network based on Geographical Location

MATSUURA Satoshi  
sato-mat@is.naist.jp

FUJIKAWA Kazutoshi  
fujikawa@itc.naist.jp

SUNAHARA Hideki  
suna@wide.ad.jp

Graduate School of Information Science, Nara Institute of Science and Technology

## ABSTRACT

With the rapid rise in the demand for location related service, communication devices such as PDAs or cellular phones must be able to search and manage information related to the geographical location. To leverage location-related information is useful to get an in-depth perspective on environmental circumstances, traffic situations and/or other problems. To handle the large number of information and queries communication devices create in the current ubiquitous environment, some scalable mechanism must be required. In this paper, we propose a peer-to-peer network system called “Mill” which can efficiently handle information related to the geographical location. To simplify the management of the location related information, we convert two dimensional coordinates into one dimensional circumference. Using this technique, Mill can search information by  $O(\log N)$ . Mill does not adopt any flooding method, and it reduces the amount of search queries compared with other P2P networks using flooding. DHT networks also do not leverage flooding and have good features. Simulation results show that the performance of Mill is good as well as other DHT networks. DHTs support only exact match lookups. The exact match is not suitable for searching information of a particular region. Mill provides an effective region search, by which users can search flexibly location-related information from small regions to large regions.

## Introduction

Today’s mobile devices such as cars, PDAs, sensors, and other devices become powerful. In addition, these devices have connectivity to the Internet and equip positioning devices such as GPS sensors. In ubiquitous computing environment, these devices can immediately collect and provide information anywhere.

If we use a large number of information these devices provide, we can obtain detailed and real time information. Gathering information based on geographical location can be effective for judging traffic situation, weather condition, and other circumstances. For example, if we gather rainfall information based on geographical location throughout a city, we know where rain clouds are exactly. This information is useful to the people riding

a bike, climbing a mountain, and doing other things. However, if we can not immediately obtain this information, the value of information will be lost. Therefore, to immediately obtain some suitable information based on geographical location, a management mechanism which can handle a large number of mobile devices should be required.

Japan Automobile Research Institute (JARI) [1] experimented with IPcars (taxi; has some sensors and connectivity to the Internet). This experiment showed that gathered information from mobile devices is useful to create detailed weather information. In this experiment, a client-server approach was adopted. In the near future, it will be expected that ubiquitous computing environment come out and the number of queries for location-related information will much increase. Then servers will be much overloaded.

To decentralize information and queries, peer to peer (P2P) networks are widely studied. Especially, P2P network with distributed hash table (DHT) are proposed in many studies [5, 7, 8, 9]. DHTs are scalable to the number of mobile devices and are effectively adapted for entry and separation of nodes. DHTs can answer queries even if the network is continuously changing. However, there is serious defect. DHTs support only exact match lookups because of adopting a hash function. If we deal with location-related information, exact match will be disturbance. Despite geographical distance, all information is assigned absolutely different ID by a hash function. If some information is geographically close, assigned IDs are not relevant. Therefore, the exact match mechanism is not suitable for searching a particular region.

There are several P2P networks considering location. However, these P2P networks have some defects in dealing with location-related information. LL-net [6] is location based P2P network. This P2P network defines an area as a square region divided by latitude and longitude. LL-net is optimized for context-aware service, and this P2P network is efficient to find where node is and what services node has. LL-net has two kinds of special nodes (super peer and rendezvous peer). The super peer manages information about all rendezvous peers. All other peers should know the super peer in advance. A rendezvous peer exists per an area. This peer manages normal peers in its area. Besides, LL-net

can not deal with attribute of time and can not gather location-related information such as temperature, speed and other values of sensors, because LL-net manages not location of information but location of nodes.

In this paper, we propose a new approach which can handle information in terms of location. To simplify the management of the location related information, we convert two dimensional coordinates into one dimensional circumference. Using this technique, our P2P network named Mill, which can flexibly search arbitrary region for information. Mill has a good performance as well as DHTs. Mill can search information by  $O(\log N)$  and answer queries in mobile environment and does not require a special node(e.g. central server). In addition, Mill can flexibly search location-related information from a small region to large region. Mill does not adopt a hash function. The strategy of creating ID is quite different from DHTs. Mill can search consecutive IDs at one time. Therefore, Mill reduces the number of queries for a region search.

The rest of this paper is structured as follows. Section 2 describes requirements for information management and retrieval on ubiquitous computing environment. Section 3 presents the mechanism of Mill and explains several of its properties. Section 4 shows Mill's performance through simulations. Finally, we summarize our contribution in Section 5.

## Requirements for ubiquitous environment

In ubiquitous computing environment, there are many devices including mobile phones, desktop PCs, web cameras and sensor devices. If we gather information from these devices, we can obtain valuable information. However, there are several requirements that we have to cope with.

- Scalability  
In the near future, the number of information created by mobile devices and other devices will much increase. And centralized system as like client-server model will much overloaded. It is required to handle information and search queries created by a large number of devices all over the place.
- Region search  
If sensor devices provide location-related information, we try to obtain useful information. For example, someone wants to know weather condition around his/her office and traffic situation between his/her home and office. Therefore, a flexible search mechanism is required, which can be applied to arbitrary size of region.
- Fast search  
If we try to obtain traffic situation or weather condition, the up-to-date information must be provided. Therefore, a search mechanism should be fast.

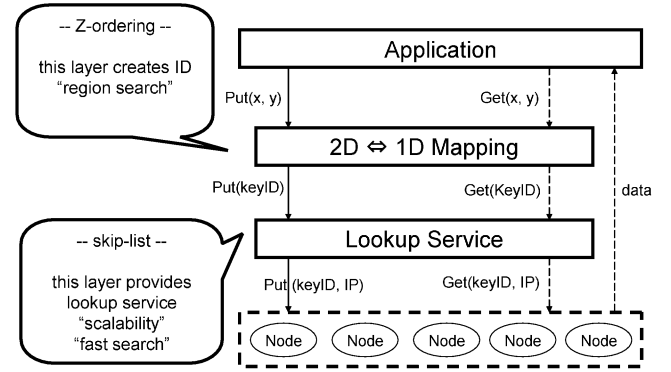


Figure 1: architecture

## Mill: A new Geographical-based peer-to-peer network

To meet the above requirements, we propose a new P2P network system called “Mill” considering geographical location. This section describes the mechanism of Mill. Mill has several protocols, which are join and leave, maintenance overlay network, store and search, optimization of queries on searching several regions, maintenance of routing tables, and other protocols. Due to the space limitation, we explain join, store, and search protocols.

### Overview

If we deal with information based on geographical location, a P2P network system must support region search. In mobile environment, it is difficult to comprehend exact location of mobile devices in advance. Therefore, when searching a particular point, we do not know whether we acquire some information or not.

DHTs support only exact match queries because of adopting a hash function (e.g. SHA-1 [4]). If we search a particular region, we should search all points in the region. For example, if a DHT system expresses a region as 10 bits, we should search 1024 points. Exact match causes the large number of queries on region search.

To support region search, Mill adopts not a hash function but a mapping mechanism optimized for location-related information. Using this mapping mechanism, Mill reduces search queries compared with other DHTs. The architecture of Mill is similar to the DHTs. As Figure 1 shows, the architecture is hierarchy structure. If an application stores or searches location-related information, the application just specifies the latitude (y) and longitude (x). The 2D-1D mapping layer converts x and y into key-ID. This layer corresponds to a hash function of DHTs. The lookup layer searches a particular node based on this key-ID. In case that there are N nodes, a query can be resolved via  $O(\log N)$  messages.

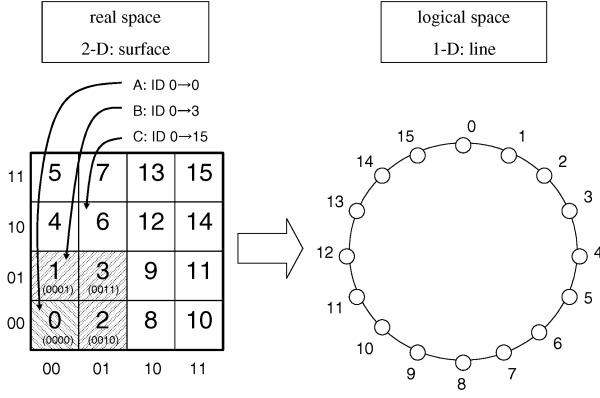


Figure 2: 2D-1D mapping method

## 2D-1D mapping

Mill divides two dimensional space into a grid cell by latitude and longitude. A grid cell is a small square region. If Mill expresses the surface of the earth as 64 bits, a size of grid cell is equals to milli-meter order. As Figure 2 shows, suppose that each cell is assigned a 4bit identifier. Mill manages these IDs as one dimensional circular IDs, and each Mill node is responsible for a part of circular IDs. The number of ID is created by alternating x-bit and y-bit. For example, if an x-bit is '00' and a y-bit is '11', a cell ID is '0101'. This method is called "Z-ordering". Here, ID space is very small (only 4-bit) to explain simply, however in real use Mill's ID space is large 64-bits). A particular region can be expressed as a range between "Start-ID" and "End-ID". For example, in Figure 2 ID range (0, 0) correspond a square cell (region A), ID range (0, 3) correspond quarter of the whole square (region B), and ID range (0, 15) correspond the whole square (region C). In fact, Mill expresses a particular square region as a consecutive of cell IDs and can search range of IDs at one time for information. Mill searches location-related information by a few queries against arbitrary size of region. Because of this feature, Mill can reduce the number of search queries. Here, I summarize the features of "Z-ordering" ( including the features without explanation )

- locality of ID
  - region search, load-balance
- consecutive ID
  - reduce search queries
- create one-dimension ID
  - simple management, fast & simple search

In some cases, altitude is needed. In urban areas, buildings are usually multi-story ones. Then, sensor devices

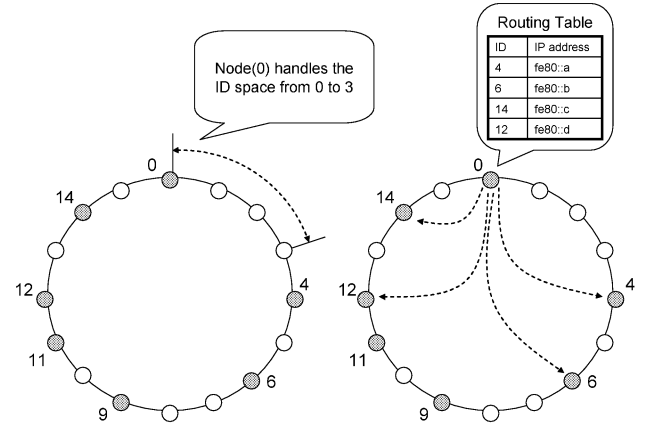


Figure 3: handle ID-space and routing table

are installed on different floors. Therefore, it is necessary to distinguish a sensor on first floor as being different from another sensor on second floor. It is easy to expand Mill network into 3 dimensions. The IDs of Mill network are created by alternating x-bit, y-bit and z-bit in 3 dimensions.

## Join protocol

Each node has a responsibility to handle a part of the circular ID space. And each node communicates with two clockwise side nodes and two counterclockwise side nodes. As Figure 3 shows an example, the overlay network is consist of 7 nodes (0, 4, 6, 9, 11, 12, 14). The node whose ID is 0 handles a part of the circular ID space from ID 0 to 3. This node has 4 connections with other nodes ID 4, 6, 14, and 12.

A new node joins Mill network by the following protocol. Figure 4 shows an example.

1. The new node creates an ID from the actual location (x, y). We define this ID as Node-ID. The new node knows an IP-address of at least one node in advance. We define this node as initial node. And the new node sends Node-ID and IP-address to the initial node. As Figure 4 shows, the new node creates 12 as Node-ID according to its an actual location. Then, the new node sends Node-ID (12) and IP-address to the initial node (Node-ID: 6).
2. The initial node sends this message to clockwise side node, and the clockwise side node sends this message in the same way. As each node send the message repeatedly, finally this message reaches a particular node which handles the ID space including the Node-ID. The initial node (Node-ID: 6) sends the message to the node (Node-ID: 8). And the node (Node-ID: 8) sends the message to the node (Node-ID: 9) which handles the ID space including the new node's Node-ID (12).



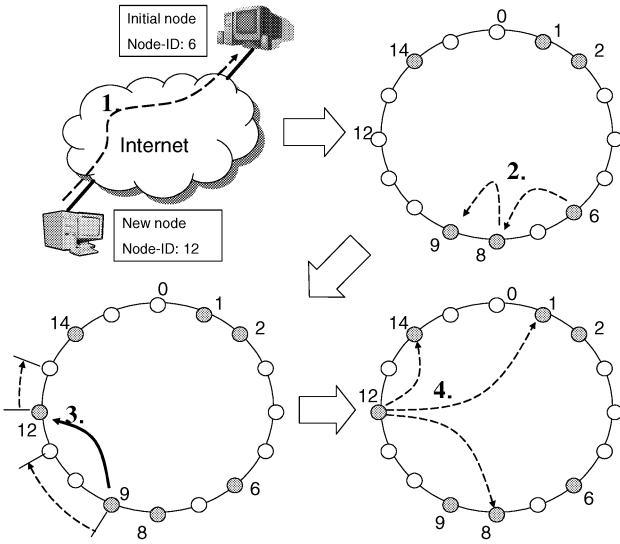


Figure 4: join protocol

3. The node which handles ID space including Node-ID assigns a part of ID space to the new node and reassigns own ID space. And this node also informs the new node about Node-ID and IP-address of neighbor nodes. The node (Node-ID: 9) assigns the ID space (12, 13) to the new node and informs the new node about Node-ID and IP-address of neighbor nodes (Node-ID: 8, 9, 14, 1). And the node (Node-ID: 9) reassigns the ID space (9, 10, 11) by itself.
4. The new node informs neighbor nodes about own Node-ID and IP-address. Then neighbor nodes update their routing table. The new node (Node-ID: 12) informs neighbor nodes (Node-ID: 8, 14, 1) about own Node-ID and IP-address.

Through the join protocol, the new node can be assigned particular ID-space and knows neighbor nodes.

### Store and search protocol

A message flow of store protocol is similar to join protocol. First, if a node gets information, the node records the ID of the location where the information is got. This ID is created by the 2D-ID mapping mechanism. Second, this node sends the ID and own IP-address to clockwise side node. And the clockwise side node sends this message to the next clockwise side node. Sending the message clockwise, a particular node which handles the ID-space including the ID is received this message. This node manages the ID with the IP-address.

The search protocol is similar to the store protocol. If a node wants to get some location-related information, a search query including “StartKey-ID” and “EndKey-ID” is issued. Figure 5 shows an example. The node

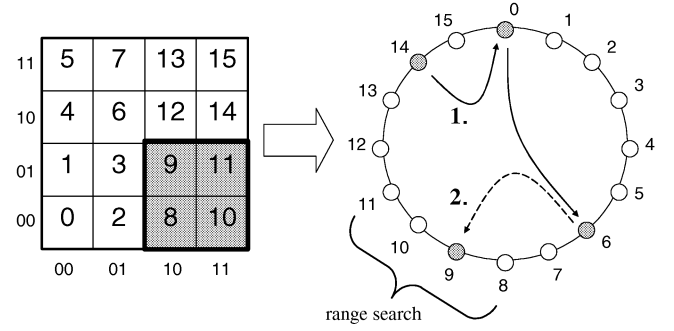


Figure 5: region search

(Node-ID: 14) wants to search the region from ID-8 to ID-11. In this case, StartKey-ID is 8 and EndKey-ID is 11. First, the search query is sent clockwise until the node handles the ID-space including 8 is found. Second, the node (Node-ID: 6) replies to the node (Node-ID: 14) with IDs and IP-addresses related with ID-8. And this node sends the search query to the clockwise side node (Node-ID: 9). The node (Node-ID: 9) replies to the node (Node-ID: 14) with IDs and IP-addresses related with ID-9, 10, and 11. Then the node (Node-ID: 14) knows IP-addresses of nodes which have information related with ID-8, 9, 10, and 11. Connecting to these IP-addresses, the node gets information. If the nodes (Node-ID: 6, 9) do not find any information, they reply to the node (Node-ID: 14) with the message meaning that information is not found.

In practice use, information consists of ID, time, type, and value. Therefore, Mill can supports not only region search and but also time based search and type based search.

### Improvement of routing algorithm

The clockwise liner search is not scalable, because a search query is sent through a sequence of  $O(N)$  other nodes toward the destination. To reduce a searching cost, each node manages information of power of two hops away nodes as like 1, 2, 4, 8, 16 hops away. First, to know the information of a 4 hops away node, each node communicates with a 2 hops away node. Second, to know the information of an 8 hops away node, each node communicates with a 4 hops away node. Repeating this communications, a size of routing table is larger. The maximum size of routing table is no more than  $O(\log N)$ . This routing table is likely to have more entries for closer nodes and fewer entries for further nodes. This list structure is called *skip-list*.

Figure 6 shows a search example and a clockwise routing table of the node (Node-ID: 18). The node gets the information related with ID (34) by the following search protocol.

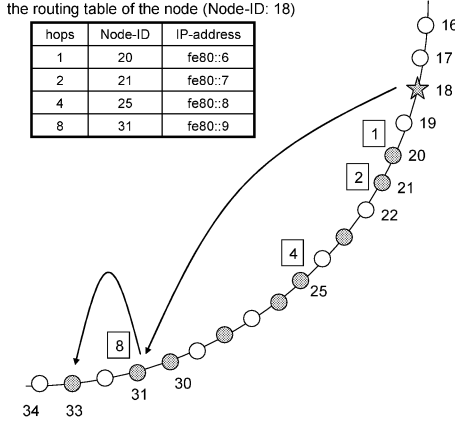


Figure 6: skip-list search

1. The node (Node-ID: 18) compares 34 with Node-IDs on the routing table.
2. On the routing table, the closest Node-ID is 31 ( 8 hops away node )
3. The node (Node-ID: 18) sends the search query to the node (Node-ID: 31)
4. The node (Node-ID: 31) passes the search query to the node (Node-ID: 33)
5. The node (Node-ID: 33) handles the ID-space including 34 and reply to the node (Node-ID: 18) with IP-addresses related with the ID (34).

This routing table reduces the searching cost to  $O(\log N)$ . This routing table also enhances stability of Mill network. Mill network can recover itself to find alive nodes by using this routing table even if several nodes are disconnected at the same time.

### Load balance

DHTs use hash-function for creating IDs. Based on hashed IDs, information generated by nodes is distributed. On the other hand, Mill does not use hash-function but z-ordering algorithm for creating IDs. On the face of things, information is not distributed in Mill network. However, in fact, information is distributed. I explain how to distribute information in Mill network as follows.

In Mill network, each node has responsibility for a part of ID space. The size of IDs each node has is determined by the distance between one node and next node. In the area where density of nodes is high, the distance between two nodes is short. In these areas, lots of information is generated, however the size of IDs each node has is small. The size of IDs is inverse of density.

Table 1: simulation environment	
CPU	Pentium4 2.4GHz
Memory	1GB
OS	WindowsXP SP2
programing language	Java 2 SDK ver1.4.2
the number of node	10 → 2,560
ID-space	$2^{24}$ (4,096 × 4,096)
transfer method	random walk

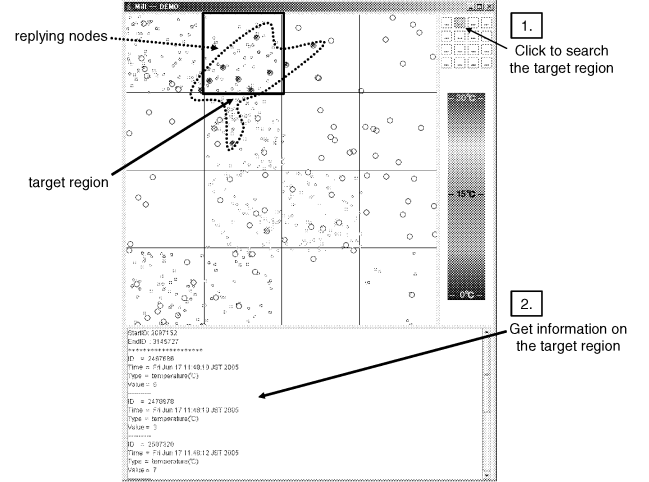


Figure 7: Application Example

The information amount each node has is not effected by density of nodes because of locality of Z-ordering. Every node manages almost same size of information, and load balance is realized in Mill network.

### Evaluation

In this section, we evaluate the performance of Mill system. We have made a simulator to evaluate Mill system by Java 2 SDK. Table 1 shows the simulation environment.

### Application example

We make a sample application on the simulator. This application creates the weather information. In this application 100 mobile devices are moving around, sensing temperature. After running the application, every node communicates with some other nodes and creates Mill network. Figure 7 shows a snapshot of this application. Small circles represent mobile nodes and dots do information of temperature. First, users determine a target region and click the bottom related with the target region. Second, users get information on the target region. The temperature information is plotted on the target region as dots. After we search several region, we can see the atmospheric temperature profile.

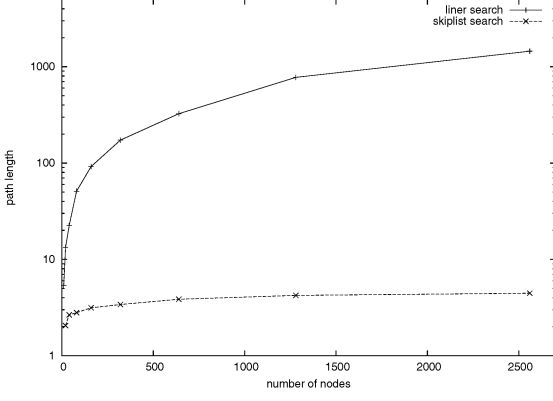


Figure 8: node vs pathlength

### Path length

We define the path length as the number of nodes relaying a search query. As Figure 8 shows, if a searching method is skip-list search, the path length is  $O(\log N)$  and if a searching method is clockwise liner search, the path length is  $O(N)$ . In fact, the path length is almost half of  $\log N$ , because each node has clockwise and counterclockwise information of nodes on a routing table.

Figure 9 shows how many search queries reach the destination. In the 160 nodes system, 80% of search queries reach the destination through 4 hops. In most cases, almost all search queries reach the destination through 8 hops. As the Round Trip Time (RTT) of mobile phones is around 500 (msec), search queries reach the destination by 3 or 4 (sec).

Now, we compare our Mill network system with other P2P network system. We express target region as “i” bits, the number of nodes in the target region as “m” nodes, and the number of nodes in the network as “N” nodes. DHTs only support exact match queries. In a DHT network, a node searches all points in a target region. Then, the search cost is  $2^i \times \log N$ . In the Mill network, a node searches sequential IDs at a time. In the target region, a search query is sequentially sent from a node to a node. Then, the search cost is  $\log N + m$ . Let “i”, “m”, and “N” be 16, 20, and 10000 respectively, each search cost is as follows.

- *DHT* :  $2^{16} \times \log 10000 = 603609$
- *Mill* :  $\log 10000 + 20 = 29$

On region search, DHTs create much larger queries than Mill does. However, if “m” increases, the performance of Mill becomes worse. The performance-degrading factor is sequential search in the target region. It seems that adopting routing table is effective in the target region to solve this degradation. We need to consider the relation between the quality of information, the density of nodes, and the routing table. This optimization is one of the future works.

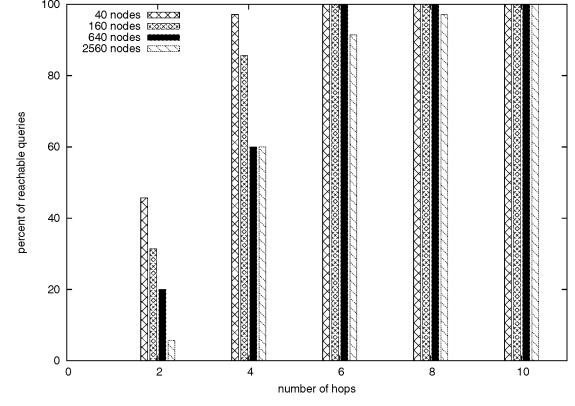


Figure 9: hops vs reachableQueries

LL-net has hierarchical ID spaces to improve search mechanism. However, as the number of hierarchical layers becomes larger, the management cost increases. If one of the layers consists of very small areas, user can search very a small region. On the contrary, as the number of rendezvous peers increases, the super peer should manage a large number of rendezvous peers.

The performance of DHTs and LL-net are directly affected by the size of ID space. One of the Mill’s advantages is that the performance of Mill is not related with the size of ID space.

### Management cost

In a message of Mill, there are 2 types. One type is join-leave message. This message is sent if a node joins or leaves Mill network. Another type is keep-alive message. A node sends this message to recognize a link status of other nodes. We evaluate the number of processed messages per node. As Figure 10 shows, if a searching method is clockwise liner search, the number of processed messages becomes larger as the number of nodes increases. On the contrary, if using skip-list search, the number of processed messages is almost  $O(\log N)$ .

In the Mill network, the order of search cost and maintenance cost is  $\log N$ . In DHT networks, the order of these costs is also  $\log N$ .  $O(\log N)$  is effective for the increasing number of nodes, and Mill and DHTs are scalable to the number of nodes. In the LL-net network, information of every area is centrally managed by the super peer. Therefore, it seems not to be a scaleable to the number of nodes.

### Robustness

We evaluate the robustness of Mill network. It is important to work Mill network even if link status of nodes is continuously changing. We define the normal condition of Mill network as that every node appropriately handles ID-space and circular ID-space is not divided

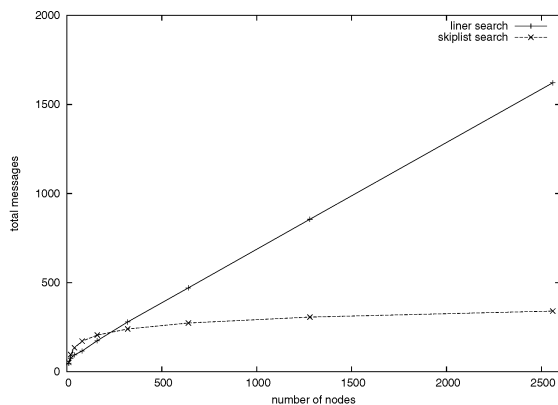


Figure 10: nodes vs messages

anywhere. If Mill network works well, each node can put and get information. In this simulation experiment, a particular percent of nodes is forced to be disconnected at once. Alive nodes try to recover Mill network to find other nodes by adopting a routing table. As Figure 11 shows, Mill network can recover itself perfectly (100%) until disconnected rate is about 15%. Each routing table has information of neighbor and distant nodes, and Mill network can recover itself by this routing table even if several nodes become disconnected at once.

### Concluding remarks

In the ubiquitous computing environment, communication devices can provide information anywhere and any-time. Therefore, information should be shared among communication devices based on geographical location. Mill enables communication devices to share information based on geographical location. In an  $N$ -node network, Mill can search information by  $O(\log N)$  and each node maintains routing information for about  $O(\log N)$  other nodes. Mill can recover the overlay network, even if 15% nodes become disconnected at the same time. In addition, Mill does not adopt any flooding methods, therefore Mill can reduce the number of search queries compared with other P2P adopting flooding mechanism. DHTs also do not adopt flooding methods and have good features. However, DHTs only support exact match queries. Exact match queries is not suitable to searching a particular region, because users should search all points in the region. Mill can search consecutive IDs at one time by 2D-1D mapping mechanism. Mill can also support effective region search and users flexibly search information from a small region to a large region.

### REFERENCES

- [1] JARI : Japan Automobile Research Institute.  
<http://www.jari.or.jp/>

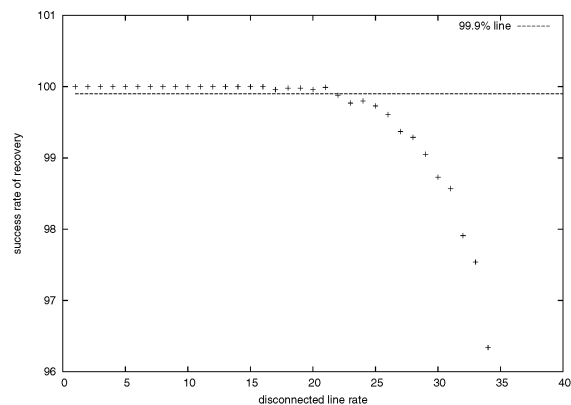


Figure 11: disconnected vs recovery

- [2] Mobility for ipv6. <http://www.ietf.org/html.charters/mip6-charter.html>
- [3] Network mobility. <http://www.ietf.org/html.charters/nemo-charter.html>
- [4] D. Eastlake and P. Jones. Sha-1: Us secure hash algorithm 1. RFC3174.
- [5] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *ACM of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 149–160, 6 2001.
- [6] Y. Kaneko, K. Harumoto, S. Fukumura, S. Shimojo, and S. Nishio. A location-based peer-to-peer network for context-aware services in a ubiquitous environment. In *Proceedings of the The 2005 Symposium on Applications and the Internet Workshops (SAINT-W'05)*, March 2005.
- [7] A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In *IFIP/ACM Int'l Conf. Distributed Systems Platforms (Middleware)*, pages 329–350, 2001.
- [8] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker. A scalable content-addressable network. In *ACM SIGCOMM*, pages 161–172, 2001.
- [9] B. Zhou, D. A. Joseph, and J. Kubiawicz. Tapestry: a fault tolerant wide area network infrastructure. In *Sigcomm*, pages Tech. Report UCB/CSD-01-1141, 2001.

# EFFECT OF QOS ALGORITHMS ON ATM SWITCHES

John D. Garofalakis  
Computer Engineering & Informatics  
Department  
University of Patras,  
Rio – Patras

E-mail: [garofala@cti.gr](mailto:garofala@cti.gr)

Dimitrios S. Goulas  
Computer Engineering & Informatics  
Department  
University of Patras,  
Rio – Patras

E-mail: [goulas@ceid.upatras.gr](mailto:goulas@ceid.upatras.gr)

Vassilis D. Triantafyllou,  
Telecommunication Systems &  
Networks Department  
Technological Educational Institute  
of Messolonghi,  
National Road Antirrio - Nafpaktos  
E-mail: [triantaf@teimes.gr](mailto:triantaf@teimes.gr)

## KEYWORDS

ATM switch, QoS, Routing Algorithms

## ABSTRACT

The term “Quality of Service” is used to define network’s ability to provide specialized services under different traffic models. We can use network backbones with ATM (Asynchronous Transfer Mode) switches as structural elements, in order to provide QoS.

This research examines the effect of quality of service algorithms on ATM switches. It examines algorithms that are applied or used on theoretical level on ATM switches and the way they affect QoS parameters. A set of algorithms are applied on a network simulator and according to network’s behavior we estimate the effect of different algorithms on networks’ performance and behavior.

We derive significant conclusions about the conditions for which an algorithm could be more efficient than another one in a network and the factors that we must consider before taking any decisions about the routing algorithm to use.

## INTRODUCTION

ATM is a cell switching and multiplexing technology. Cell is the main data transferring unit at ATM networks. Cells have the ability to get multiplexed asynchronously in time, so that we have flexible bandwidth distribution for different communication services and have the ability to transmit through virtual paths (VP) and Virtual Circuits (VC). Bandwidth allocation to the services is done on demand. The use of small size cells and of high transmission rates admits the support of high rate services. (M. Naraghi-Pour et al)

There are two main signaling methods at ATM networks: User-to-Network Interface (UNI) and Network-to-Network Interface (NNI). UNI signaling is used to connect ATM end systems to an ATM switch. NNI signaling is used between switches in the same ATM switch network. UNI signaling is converted at the incoming ATM switch to NNI signaling and later it is converted reversely from NNI to UNI at the outgoing ATM switch.

Two routing protocols are used at ATM networks: IISP (Interim Interswitch Signaling Protocol) and PNNI (Private Network-Network Interface). (C.Tham et al, 1996). PNNI provides routing based on quality of service by using a routing protocol based on the current network topology. Knowledge of

the network topology, including the links state and parameters about nodes state, is included in PNNI topology state packets – PTSP.

These packets are exchanged periodically between the nodes or activated by specific events. Moreover, all nodes can synchronize their information databases about the available network resources and accessibility of the network. According this information database, a node can choose the optimal path which satisfies the requested QoS.

The “contract” between the source and the destination in a network is represented by three parts: Traffic parameters, QoS demands and Service category.

The first part of the “contract” (traffic parameters) concerns the traffic load that can be satisfied, the second determines the quality of service and the third defines the service category for the calls served and depends on the two other parts of the “contract”.

The parameters that give us information about traffic and represent the connection traffic descriptor, are the following:

- PCR-Peak Cell Rate
- MBS-Maximum Burst Size
- SCR-Sustainable Cell Rate
- MCR-Minimum Cell Rate
- CDTV-Cell Delay Variation Tolerance

The quality of service parameters, which define the performance of the network are:

- CLR-Cell Loss Ratio
- CTD-Cell Transfer Delay
- CER-Cell Error Ratio
- SECBR-Severely Errored Cell Block Ratio
- CMR-Cell Misinsertion Rate
- MCTD-Mean Cell Transfer Delay
- CDV-Cell Delay Variation

Based on traffic parameters and QoS parameters, we define five service categories for the transmission of the ATM cells:

1. Constant Bit Rate (CBR) service
2. Real time Variable Bit Rate (rt-VBR) service
3. Non real time Variable Bit Rate (nrt-VBR) service
4. Available Bit Rate (ABR) service
5. Unknown Bit Rate (UBR) service

possibility to admit calls with bigger demands to be sent. (S. Alla et al. 2001)

## SIMULATION MODEL

OPNET Modeler provides an auxiliary development environment for the design, simulation and analysis of the performance of communication networks. It can support a wide range of communication networks from a simple LAN to a wide satellite network. Systems performance and behavior can be analysed by using discrete events simulations. (S. Hedge et al. 2002)

The full software packet of OPNET provides a number of tools that allows manufacturers to design models with special specifications, to identify the components of a model, to execute the simulation and analyze the results produced. (X. Chang. 1999)

More specifically OPNET has three main types of tools: development tools, simulation tools and result analyzing tools. (T. Rereira, L. Ling. 2002)

Using those tools we defined different scenarios. We applied different routing algorithms for different traffic loads and we compared the simulation results of each scenario and the performance of different routing algorithms under different traffic loads.

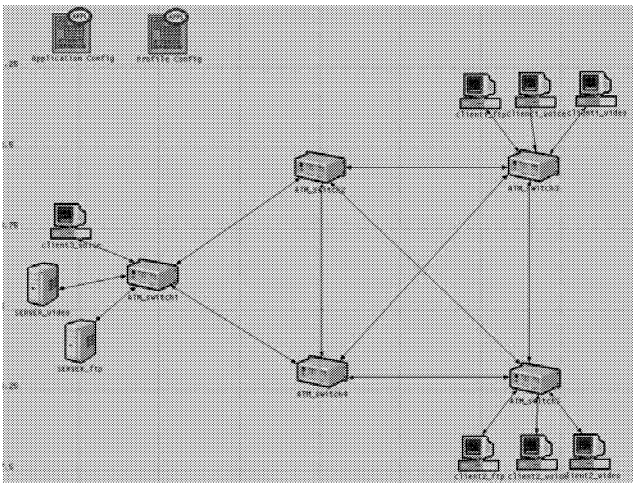


Figure 5: ATM network

We used Solaris SDE V. 1.2 operating system and the simulator was OPNET Modeler 6.0.L. We used a simple network with alternative paths for each destination, using as node component the ATM switch. We compared the results of the use of the different QoS routing algorithms in the cases of low and high traffic load. The network we used was the one in figure 5 and was consisted of 5 ATM switches, 2 ftp applications client and the corresponding server, 2 video conferencing clients and the corresponding server and 3 voice conferencing clients.

We used OC3 links so that the supported bandwidth to come up to 155 Mbps. We also applied the advanced models of uni-clients and uni-servers. The advanced and the intermediate node

models had some extra definitions capabilities. (R. Goyal et al. 1998)

The following traffic categories were considered:

- VIDEO clients under rt-VBR traffic
- VOICE clients under CBR
- FTP clients under ABR
- FTP server accepting only ABR calls
- VIDEO server accepting only rt-VBR calls

In table 1 we present the differences between the two traffic cases, with low and high traffic load.

Table 1: Differences between low and high traffic load

Parameter	Low Load	High load
FTP Average File Size (Bytes)	5.000	50.000
Video Average Conference Duration	5 min	15 min
Video Frame Rate (Frames/sec)	10	15
Voice Encoder Scheme	G.729	G.711

We applied the Weighted Round-Robin Scheme and we defined three queues, one for each traffic category, assigning 50% of the bandwidth to CBR calls, 25% to ABR and 25% to rt-VBR. Applications started running after 10 seconds of simulation and the total simulation time was 60 seconds for each scenario.

We considered watching the ATM QoS parameters and the ftp (ABR), voice (CBR) and video (rt-VBR) traffic parameters.

We implemented the simulations applying on the properly modulated ATM switches two different algorithms and we compared the results of low and high traffic cases.

The different scenarios that we examined were:

1. Scenario A: low traffic load case with LLR algorithm.
2. Scenario B: low traffic load case with MFCR.
3. Scenario C: high traffic load case with LLR.
4. Scenario D: high traffic load case with MFCR.

## SIMULATION RESULTS

A synopsis of the results is presented at table 2. We observed that for low traffic, data delay at the network is smaller using MFCR algorithm but this changes as traffic is growing in favor of the LLR algorithm. Moreover, delay variation is smaller with MFCR when traffic is low but this changes as traffic is growing, in favor of LLR.

In the two traffic load cases, the analogy of the total amount of data LLR/MFCR which satisfy the QoS demands is for low traffic 1,224 and for high 0,985. That means that LLR satisfies the QoS demands of more calls in low data traffic conditions than MFCR and so the data sent are more with the use of LLR and network's performance is better. The opposite happens when traffic is high.

Table 2: Best Algorithm for each parameter

functions that can significantly affect network ability to provide the best QoS.

One of the reasons that QoS becomes lower relies on the fact that the routing table can not be optimum or the buffering method used could be shared by many ports, instead of providing buffering at each port or VC. So, buffering can become deficient and create congestion conditions at the network.

Other reasons which degrade QoS are errors based on the transmission mean, traffic overload, assignment of more bandwidth then we need for a specific traffic category and cell loss because of function break of a port, link or switch.

## FUTURE WORK

Our aim was to show the significance of using a suitable routing algorithm at the switching components of an ATM network in order to ensure Quality of Service.

After many tests with several network structures we ended at the cases we presented above and we had the results that we mentioned and analyzed above.

Our aim was to examine if an algorithm could become more efficient then another one in a network, under different conditions. Moreover we wanted to see which factors we should consider to make a better choice of the routing algorithm we should use. So, we did not examine many different cases of algorithms.

We also did not examine the different switching architectures, mostly because we didn't want queuing methods to affect our results, which would had make our research much more complicated. We only considered the case of heterogeneous sources with no statistical multiplexing, without examining other cases.

We intend to extend our cases studies by applying different network structures and by increasing the number of network nodes. Also we intend to apply more algorithms under the above configurations in order to present a guideline for ATM performance under different network parameters.

## REFERENCES

- Chen Khong Tham, Jianning Mai & Lawrence Wong. 1996. "A QoS-based Routing Algorithm for PNNI ATM Networks". Dept of Electrical & Computer Engineering, National University of Singapore.
- M. Naraghi-Pour, M. Hegde, J. Sanchez Barrera. "QoS-Based Routing Algorithms for ATM Networks".
- Rohit Goyal, Raj Jain, Sonia Fahmy, Shobana Narayanaswamy. 1998. "Modeling Traffic Management in ATM Networks with OPNET". Ohio State University
- Sanjay Gupta, Keith Ross, Magda El Zarki. 1992. "Routing in Virtual Path Based ATM Networks", University of Pennsylvania, Philadelphia.
- Sanjay Gupta, Keith Ross, Magda El Zarki. March 1993. "On Routing in ATM Networks", University of Pennsylvania, Philadelphia.
- Sanjay Hedge, Vicram Mallikarjuna, Cynthia Hood. February 2002. "High Performance Network Simulation using OPNET", Illinois Institute of technology, Chicago.
- Sanjeev Verma. August 1994. "ATM Switch Architectures". Technical report. Department of Electrical & Computer Engineering, Concordia University.
- Sonia Fahmy. 1995. "A survey of ATM Switching Techniques", Department of Computer & Information Science, The Ohio State University.
- Sumanth Alla, Pradeep Reddy Gundlagutta, Manish Gupta, Vijay Kodati, Joe Larosa. November 2001. "PNNI Routing in ATM".
- Tatiana Brito Rereira, Lee Luan Ling. 2002. "An OPNET Modeler Based Simulation Platform for Adaptive Routing Evaluation". FEEC, UNICAMP, Campinas, S.P., Brazil.
- Xinjie Chang. 1999. "Network Simulations with OPNET", Network Technology Research School of EEE, Nanyang Technological University, Singapore
- JOHN D. GAROFALAKIS** obtained his Ph.D. from the Department of Computer Engineering and Informatics, University of Patras and his diploma on Electrical Engineering from the National Technical University of Athens. He is currently Associate Professor at the Department of Computer Engineering and Informatics, and Head of a Research Department at the Computer Technology Institute, Patras, Greece. His research interests include performance evaluation, distributed systems and algorithms, Internet Technologies. He has published over 65 papers in various journals and refereed conferences, including the Theoretical Computer Science journal, the ACM SIGMETRICS Conference, WDAG, EuroPar, Performance Evaluation Journal, WWW Conferences, IEEE, Internet Computing etc.
- DIMITRIOS S. GOULAS** was born in Nafpaktos, Greece and went to the University of Patras where he had studied Computer Engineering and Informatics and obtained his diploma and MSc. After doing his military service, he worked for a couple of years at the Technical Institution of Messolonghi, the Ministry of Defense and the Municipality of Patras where he is working till now.
- Dr. VASSILIS TRIANTAFILLOU** obtained his Diploma and his Ph.D. from the Computer Science and Engineering Department of Patras University (Greece). He is currently an Associate Professor, Department of Applied Informatics in Management and finance of the Technological Educational Institution of Messolonghi, Greece. His main research interests include research in the area of Networks, Telematics and New Services and Computer Integrated Manufacturing Applications. He has extended professional experience in Design and Analysis of Networks and design and implementation of Open Distance Learning Tools. He has published over 25 papers in various well-known refereed conferences and journals. He has participated and technically supervised various R & D projects such as ESPRIT, ISPO, ADAPT / EMPLOYMENT, STRIDE.

# HARDWARE/SOFTWARE CO-DESIGN FOR H.264/AVC INTRA FRAME ENCODING

Jan De Cock<sup>1</sup>, Stijn Notebaert<sup>1</sup>, Peter Lambert<sup>1</sup>, and Rik Van de Walle<sup>2</sup>  
Department of Electronics and Information Systems – Multimedia Lab

<sup>1</sup>Ghent University – IBBT

<sup>2</sup>Ghent University – IBBT – IMEC

Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

+32 9 33 14956

{Jan.DeCock, Stijn.Notebaert, Peter.Lambert, and Rik.VandeWalle}@UGent.be

## KEYWORDS

H.264/AVC, HW/SW Co-design, Intra prediction

## ABSTRACT

The growing use of multimedia resources across a wide range of networks and on a large number of different devices emphasizes the need for more efficient video coding algorithms. In order to fulfil these demands the state-of-the-art H.264/AVC video coding standard was developed, providing very high compression efficiency at the cost of increased computational complexity. This paper presents an architecture for hardware/software co-design for an H.264/AVC intra frame encoder. We determined a partitioning of the functionality for an H.264/AVC intra frame encoder, followed by an examination of the communication between the processor and the dedicated hardware device, which is in this case a Field Programmable Gate Array (FPGA). Using the results from the partitioning and the communication analysis, we implemented an architecture that is able to outperform both the software-only solution and related implementations.

## INTRODUCTION

The H.264/Advanced Video Coding (H.264/AVC) standard (ITU-T and ISO/IEC 2003), developed by the Joint Video Team (JVT) of the ITU-T and ISO/IEC, is gaining widespread acceptance as the state-of-the-art video compression standard. H.264/AVC provides improved compression efficiency over previous standards, at the cost of increased computational requirements. In this paper, we consider an H.264/AVC intra frame encoder, which uses pixels of already encoded macroblocks in the current frame for prediction. In H.264/AVC, the intra prediction is performed on both 16x16 and 4x4 block sizes. Intra prediction can be used in video applications that do not allow inter frame prediction, in order to permit easy viewing and editing of individual frames, such as in broadcasting or production environments. Intra coding can also be used for still image coding, for example as an alternative for JPEG or JPEG2000. As is shown in (Huang et al. 2005), H.264/AVC intra frame coding is competitive with JPEG2000, in terms of both coding performance and computational complexity. For more information on the

algorithms used in H.264/AVC intra frame coding, the reader is referred to (Wiegand et al. 2003).

The computational complexity of the algorithms used for H.264/AVC video compression poses a challenge to the limited capabilities of general purpose processors. In this paper, we examine the possibility of partitioning the workload of an H.264/AVC intra encoder between hardware and software in order to speed up the total encoding process. This implies that we have to identify the functional blocks of the intra encoder that are best suited for hardware implementation, and those that are best executed in software.

In (Amer et al. 2005), an FPGA-based hardware reference model was proposed for the transformation and quantization. The overall integration with the reference software however, resulted in a slowdown of the encoding process. In this paper, we extend the functional blocks that are implemented in hardware. We show that it is possible to obtain a hardware/software partitioning that allows the CPU and the hardware component to function simultaneously, which results in an acceleration of the encoding process. The hardware component used for implementation is a WildCard-II board, developed by Annapolis Micro Systems. This component was also used in (Amer et al. 2005). The WildCard-II is based on a Xilinx Virtex-II XC2V3000 FPGA, and uses a CardBus-interface for communication.

## THE H.264/AVC STANDARD

The overall design of the H.264/AVC specification contains many functional blocks that are also present in previous video coding standards, such as H.261/MPEG-1 Video, H.262/MPEG-2 Video, H.263, or MPEG-4 Visual. By optimizing these blocks and adding new coding tools, significant improvements in coding efficiency are achieved. As a drawback, the algorithms are computationally more complex and are more demanding regarding memory consumption.

The H.264/AVC specification was developed bearing in mind the possibility of a straightforward implementation in hardware. This means that the algorithms avoid the excessive use of costly multiplications or divisions. An



example is the integer variant of the discrete cosine transform, which uses only bit shift operations and additions. Other algorithms are less-suited for hardware implementation, such as the highly efficient entropy encoding algorithms used in H.264/AVC, CABAC (Context-based Adaptive Binary Arithmetic Coding) and CAVLC (Context-based Adaptive Variable Length Coding). Because of the sequential nature of the dataflow in these algorithms, they are less likely to benefit from a parallel execution on hardware.

The supported functionality of the examined intra frame encoder is present in all profiles as defined in the H.264/AVC video coding standard. This means that the encoder consists roughly out of the following blocks (see Figure 1): the (integer variant of the) discrete cosine transform and the quantization (T/Q), the corresponding inverse algorithms (iQ/iT), the actual intra prediction using 4x4 and 16x16 modes, the mode decision, and the CAVLC entropy encoder.

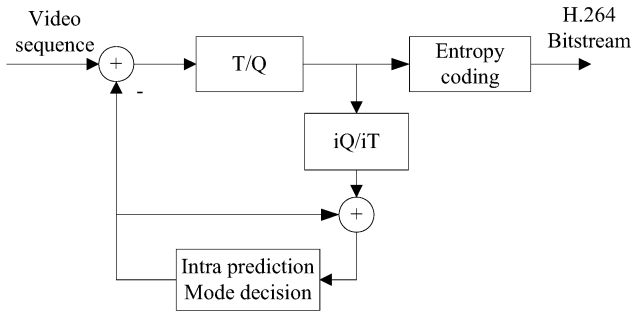


Figure 1: Intra frame Encoder

Profiling tools were used to get an initial view of the shares of these functional blocks in the total execution time, and this for a number of different encoding parameters. In Figure 2 the average results are given. The exact numbers can vary depending on the used parameters (such as the quantization parameter), or on the actual implementation of the software. These results were derived from the H.264/AVC reference software, Joint Model 8.6. Note that in our tests we do not make use of the rate-distortion optimization option of the reference software encoder.

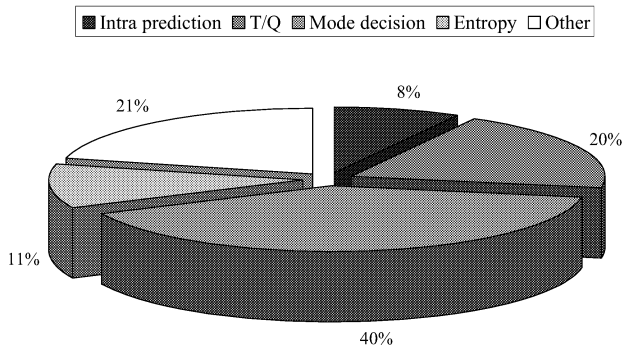


Figure 2: Run-time Percentages of the Functional Blocks

## HW/SW PARTITIONING

By hardware/software (HW/SW) partitioning, we mean the process of identifying the parts of the encoder that are best implemented in hardware and those that are best implemented in software (De Micheli et al. 1997). In order to be able to advance to this partitioning, we have to distinguish the characteristics of and differences between an FPGA and a CPU. FPGAs are composed of a regular array of logic cells, also called Configurable Logic Blocks (CLBs). Each CLB consists of a number of FPGA slices i.e., a grouping of programmable function generators and flip-flops. Because of the large number of distributed logic cells, the FPGA is best suited for applications which can be processed in parallel. Another advantage of FPGAs is that they can be reconfigured on the fly. Typically, the FPGA works at a relatively low clock frequency. A CPU on the contrary works at a much higher clock frequency. Its general purpose character makes it harder to exploit possible parallelism in algorithms. These two characteristics make CPUs more suitable for executing sequential algorithms.

The HW/SW partitioning has to take into account several constraints in order to obtain a design that can outperform the software-only solution. A first constraint is imposed by the algorithms to be implemented. It is clear that not all algorithms or functional blocks in the intra encoder are eligible for hardware acceleration. If an algorithm is essentially a sequential algorithm, the hardware implementation is less likely to outperform a regular CPU. This means that the entropy coding algorithms, CAVLC and CABAC, are less-suited for hardware implementation. Other H.264/AVC algorithms, however, such as the transformation and quantization, contain a large amount of parallelism.

A second constraint is related to the communication. Measurements on the response time of the FPGA show that the minimal overhead for communication runs up to  $70\mu s$  ( $t_c$ ). This implies that the functional blocks transferred from software to hardware have to be large enough in order to obtain a hardware acceleration of the encoding process. As a consequence, transferring only the transformation and quantization to hardware, as was done in (Amer et al. 2005), has a serious negative impact on the overall performance. In our approach, aside from the transformation and quantization, we also transfer the intra prediction (including the mode decision) to hardware. In this way, we also limit the amount of hardware calls needed. The transformation and quantization are performed on 4x4 blocks of pixels. If we extend this with the intra prediction, only one call is necessary for every macroblock of 16x16 pixels. The intra prediction and mode decision in the reference software, combined with the transformation and quantization, require  $280\mu s$  ( $t_s$ ) for one macroblock. As we will see in the results, the time required by intra prediction on hardware ( $t_H$ ) is fast enough, so that  $t_s > t_H + t_c$ .

A third requirement has to be satisfied in order to obtain a maximum degree of coprocessing. It is beneficial to partition the encoder functionality in such a way that both hardware and software can operate at the same time. In other words, we have to minimize the ‘stall time’ of the CPU. Because of this restriction, we did not use memory mapped I/O. In memory mapped I/O, the processor is master of the system bus and invokes all data transfers. Memory mapped I/O is often an inefficient mechanism because the processor stalls while it could be working on other, more important tasks. As an alternative for the exchange of large amounts of data, we used the Direct Memory Access (DMA) mechanism. The DMA controller is programmed by the processor with information about the data transfer. After the DMA controller has been granted access to the system bus, the transfer of data is executed by the DMA controller.

Another advantage of the DMA mechanism is that large transfer speeds are achieved. The transfer speed was tested for the implementation device. As mentioned above, the WildCard-II has a CardBus interface. The CardBus interface (clock speed: 33MHz, buswidth: 32 bits) has a theoretical maximum transfer speed of 132 MB/s. We executed tests to measure the performance of DMA transfers over the CardBus interface (see Figure 3). If we transfer blocks of 10 kB or more, the transfer rate is 70 MB/s (main memory to FPGA) and 90 MB/s (FPGA to main memory). For the tests, we used a laptop with a Pentium-M processor at 1.5 GHz, with 512 MB of RAM, running Windows XP.

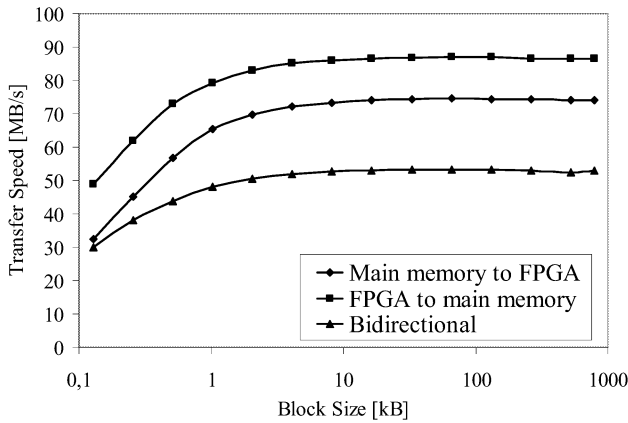


Figure 3: DMA Transfer Speed

Taking into account the previously mentioned restrictions, we obtained the hardware/software partitioning as shown in Table 1.

Table 1: HW/SW Partitioning

Functional block	HW/SW
Intra prediction	HW
Mode decision	HW
T/Q and iQ/iT	HW
Entropy coding	SW
Other	SW

## HARDWARE DESIGN

First, we implemented the H.264/AVC transformation and quantization. The transformation and quantization allow for a fast implementation in hardware. The transformation is applied on 4x4 blocks of residual values and consists of two steps. First, an operation is performed on the rows, followed by an identical operation on the columns. An example of both operations is given in Figure 4, where  $X_{ij}$  are the input residual values,  $T_{ij}$  are the output transformed coefficients, and  $S_{ij}$  represent intermediate values ( $i,j=0,\dots,3$ ). In steady state, the hardware implementation of the transformation and quantization can output one coded block at every clock pulse.

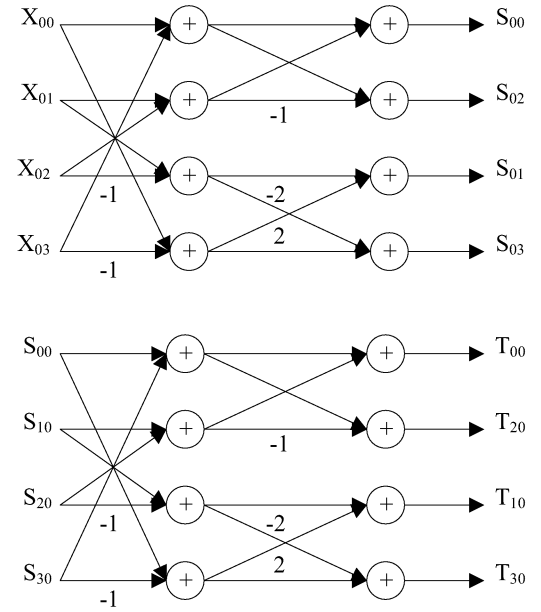


Figure 4: Fast Implementation of the Forward Transform

Since the intra prediction uses prediction pixels from surrounding blocks that were already coded and reconstructed, we also implemented the dequantization and inverse transformation in a similar way. The intra prediction and the mode decision were implemented using a pipelined approach. The intra coding component performs both the 4x4 intra prediction (nine prediction modes) and the 16x16 intra prediction (four modes). All the modes are executed in a pipelined way, in order to use the logic on the FPGA in an optimal way.

As a final step before transmission from the FPGA to the main memory, the residual values were run-level coded. This minimizes the amount of data that has to be transferred, and prepares the residual values for the CAVLC entropy coding. The run-level coded output is sent back to the main memory of the computer where the final encoding steps are performed, including the entropy coding.

The resulting hardware design consists of the implementation of the discussed hardware blocks together with a

controlling unit. The controller is the supervisor on the FPGA leading the processing by passing through a finite state machine. In every state, the controller can interact with the on-board and the off-board memory modules and can operate with the components for intra prediction and transformation. The architecture is visualized in Figure 5. Both the original and the reconstructed frames are stored in the external memory. The intermediate results are stored into the small but easily accessible memory blocks on the FPGA itself. The elementary unit of processing in H.264/AVC intra frame coding is the macroblock, so all processing is done on a macroblock basis.

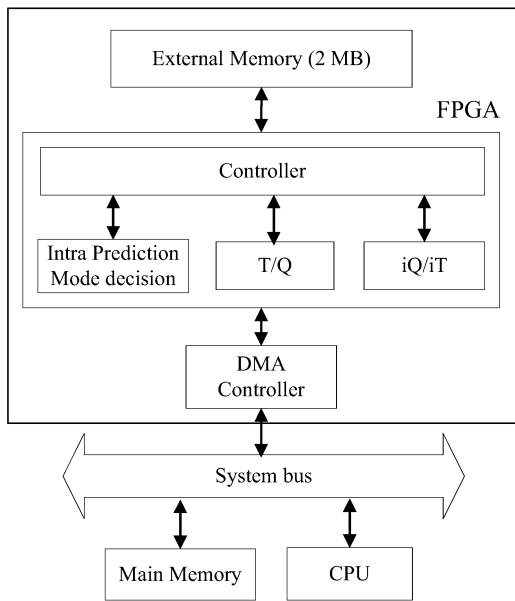


Figure 5: Hardware Design

## HW/SW COPROCESSING

The most efficient architecture for a HW/SW co-design is coprocessing, i.e., the CPU works in conjunction with a dedicated hardware component to deliver a specific application (De Micheli et al. 1997). The degree of parallelism depends on the architecture: the processor can stall when the dedicated hardware component is operational (low degree of parallelism) or the processor and the dedicated hardware component are active simultaneously (high degree of parallelism).

In our design, the CPU controls the encoding of the video and invokes the dedicated hardware component at the right time. During the intra prediction stage of the current macroblock, the processor is working on the entropy coding of the previous macroblock. In this way, every form of parallelism is exploited. The most important dependencies between the different operations on a macroblock are eliminated, and the processor and the dedicated hardware component are active at the same time.

With all mentioned ideas in mind, the following procedure for fast intra coding of a frame was derived and implemented (see Figure 6):

1. The CPU programs the source address, the destination address, the transfer count, and other information to the DMA controller.
  2. A complete frame is transferred from the main memory of the computer to the memory on the FPGA using the DMA mechanism.
  3. After the intra prediction and transformation on the FPGA, the results of the first macroblock are sent back to the main memory of the computer again using the DMA mechanism.
  4. Then the FPGA generates an interrupt to confirm that the results are available in the main memory of the computer.
  5. The CPU processes the new results, they are copied into the structures maintained by the reference software.
  6. The CPU confirms the interrupt, so the FPGA can continue processing the next macroblock.
- This concludes the processing for the first macroblock in the frame. From this point on, steps 3 to 6 are repeated for the subsequent macroblocks until the end of the frame.

## RESULTS

After the implementation of the proposed architecture, we synthesized our design for use on the Xilinx Virtex-II XC2V3000 FPGA. In Table 2 some properties about the design on the FPGA are presented.

Table 2: Used WildCard-II Resources

Clock Frequency	38.12 MHz
CLB Slices	9518
Multipliers	33
Flip-flops	8102
ZBT RAM	50688 bytes

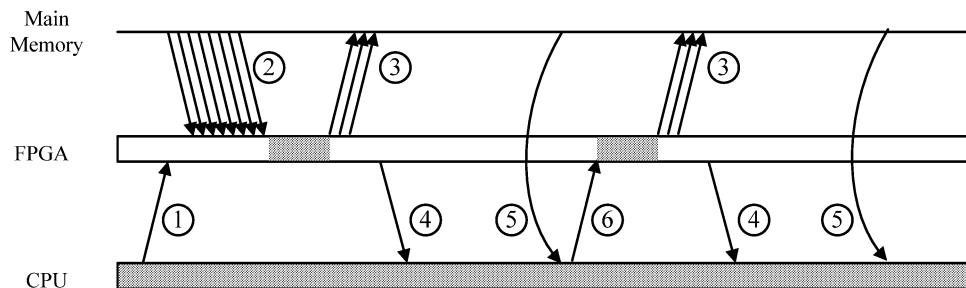


Figure 6: HW/SW Coprocessing

The clock frequency on the FPGA is much lower when compared to the clock frequency of modern CPUs because it is determined by the longest path of operations that are performed in one clock cycle. Another notable fact is the high usage of FPGA slices. The explanation is twofold, i) the hardware implementation of the intra encoder and the forward and reverse transformation require a considerable amount of logic because they represent advanced algorithms and ii) a lot of additional logic is allocated by the controlling components for addressing the memory modules and the interface with the DMA controller. We tested our HW/SW co-design on different test sequences (QCIF, 100 frames). In Table 3 the average execution times are shown. These numbers show the performance of the HW/SW co-design of our hardware blocks and the H.264/AVC reference software, compared to the software-only version. We obtained an acceleration of the total intra encoding process by approximately 20%. Also, the time for intra prediction is seriously reduced by 80%. The resulting execution time in case of the HW/SW co-design is no longer restricted by the communication delay between the computer and the FPGA.

Table 3: Time Measurements

	SW only	HW/SW
Total encoding time	8311 ms	6640 ms
Total time of intra prediction	2775 ms	530 ms
Intra prediction of one MB	280 $\mu$ s	53 $\mu$ s

## CONCLUSIONS

In this paper an architecture for a hardware/software co-design was presented that is able to divide the workload of H.264/AVC intra frame coding by exploiting the benefits of both an FPGA and a CPU. The proposed design satisfies a number of constraints that are essential to the successful cooperation of hardware and software. First, a partitioning of the functional modules of the intra encoder was made keeping in mind the advantages and disadvantages of an implementation in hardware or software. Secondly, the partitioning was made in a way that allows a maximum degree of parallelism between the CPU and the FPGA. This led to a system that allows both hardware and software to be active at the same time, while minimizing the stall time of the CPU. The communication between the components (in this case the host CPU and an FPGA connected via Cardbus) can have a significant impact on the overall performance of the system. Special attention was paid in order to minimize the overhead caused by the interchange of large amounts of data. This was accomplished by using the DMA mechanism instead of more traditional memory-mapped I/O. The performed tests show an acceleration of about 20% and a reduction of the intra prediction process to a fifth of its original execution time. The methodology as explained in this paper will in a later stage be extrapolated to an entire H.264/AVC encoder, including the inter frame motion estimation and motion compensation.

## ACKNOWLEDGEMENTS

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), the Belgian Federal Science Policy Office (BFSPPO), and the European Union.

## REFERENCES

- Amer I., Badawy W., and Jullien G. 2005. "A Proposed Hardware Reference Model for Spatial Transformation and Quantization in H.264". *Journal of Visual Communication & Image Representation*. Article in press.
- De Micheli G. and Gupta R. 1997. "Hardware/Software Co-Design". *Proceedings of the IEEE* 85, No.3 (March), 349-365.
- Huang Y.-W., Hsieh B.-Y., Chen H.-C., and Chen L.G. 2005. "Analysis, Fast Algorithm, and VLSI Design for H.264/AVC Intra Frame Coder". *IEEE Transactions on Circuits and Systems for Video Technology* 15, No. 3 (March), 378-401.
- ITU-T and ISO/IEC JTC 1. 2003. "Advanced video coding for generic audiovisual services". *ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC*.
- Malvar H.S., Hallapuro A., Karczewicz M., and Kerofsky L. 2003. "Low-Complexity Transform and Quantization in H.264/AVC". *IEEE Transactions on Circuits and Systems for Video Technology* 13, No. 7 (July), 598-603.
- Wiegand T., Sullivan G., Bjøntegaard G., and Luthra A. 2003. "Overview of the H.264/AVC Video Coding Standard". *IEEE Transactions on Circuits and Systems for Video Technology* 13, No. 7 (July), 560-576.

# **MEDIA APPLICATIONS IN EDUCATION**



# AUGMENTED INSTRUCTIONS FOR LEARNING MOLECULAR STRUCTURES

Kikuo Asai and Tomotsugu Kondo  
National Institute of Multimedia  
Education  
2-12 Wakaba, Mihama-ku,  
Chiba 261-0014, Japan  
E-mail: {asai,tkondo}@nime.ac.jp

Hideaki Kobayashi  
The Graduate University for Advanced  
Studies  
2-12 Wakaba, Mihama-ku,  
Chiba 261-0014, Japan  
E-mail: hidekun@nime.ac.jp

Norio Takase  
Fiatlux, Co., Ltd.  
1-3-3 Iwamoto-cho, Chiyoda-ku,  
Tokyo 101-0032, Japan  
E-mail: takase@fiatlux.co.jp

## KEYWORDS

Augmented Reality, Printed Learning Material, Molecular Structure, Annotation.

## ABSTRACT

E-learning, which is based on various multimedia contents, has even become popular in higher education. However, printed learning materials are not obsolete – in fact textbooks seem to be preferred for systematic study. Multimedia contents and printed materials have been considered as totally different learning environments where learners can only obtain its alternative merits. Augmented reality (AR) has potential to bridge the gap between multimedia contents and printed learning materials. This paper describes the concept of augmented instruction as a new type of learning environment. The functions of molecular visualization tools are discussed and compared in terms of application for research, presentation, and publication. A VRML export function was developed as an optional module so that molecular scientists themselves can author 3-dimensional contents for augmented instructions.

## INTRODUCTION

Multimedia systems provide learners new ways to interact with various educational resources. Since the Internet in particular has become widely used, learners can obtain a large amount of information from the sources that they look out using search engines. Even traditional printed materials often provide access information to online resources, such as URLs, so that readers can easily find additional relevant information [Dummies]. In molecular chemistry or bioinformatics, up-to-date research results are viewed with high quality graphics that have been prepared using visualization tools.

A multimedia system provides information not only in the form of sounds and images but also displays animated and simulated data. Such a system includes interactive capabilities that allow learners to set parameters for what they are not easily or physically able to experience. For example, a learning tool for system dynamics in control engineering has been developed [Schmid, 1999], that provides a flexible learning environment where learners can make simulations and see their results in animation with computer graphics.

Although multimedia contents have been favored for obtaining complementary knowledge, they have not been widely used for systematic studies. Printed materials, such as

textbooks, are more preferred for systematic studies. Multimedia contents and printed materials have been considered as totally different media that yield distinct learning environments, and learners can only get its alternative merits at each environment.

Our approach is to apply augmented reality (AR) to enhance users' comprehension of the real world by superimposing computer graphics images onto real scenes [Azuma, et al., 2001]. We believe that AR bridges the gap between multimedia contents and printed materials. AR creates a new environment that seamlessly connects a virtual space to the real world, which offers advantages such as spatial awareness [Biocca, et al., 2003] over the real world and tangible interaction [Poupyrev, et al., 2002] over a virtual space.

Various applications have been developed based on AR's potential. However, authoring/editing AR contents currently requires computer graphics expertise and computer programming knowledge, which is one of the reasons that AR has not become widely used. The ARToolkit [Kato, et al., 2000] is a set of open-source libraries used for tracking objects based on markers. The ARToolkit allows users to work with a single camera operating under visible lighting conditions. The advantage of the ARToolkit is that programming skills are not required for just using the runtime sample of attaching virtual reality modeling language (VRML) models to markers.

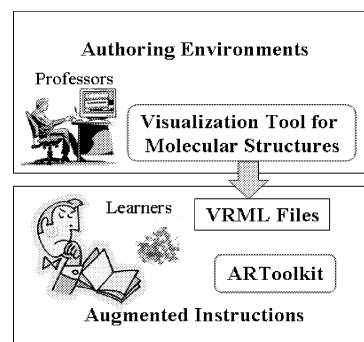


Figure 1: Learning style of augmented instructions

Users only prepare VRML models for presenting molecular structures in an AR environment. However, some explanations should at least be annotated on molecular structures so that three-dimensional (3D) information of the molecular structures works as learning materials. It is difficult for molecular scientists to use 3D graphics tools or 3D modelers, such as 3D Studio Max, Maya, and Softimage, for authoring VRML models. A solution to this problem is to

add a VRML export function to the tools that molecular scientists are familiar with for presenting molecular structures in their researches. Figure 1 shows the learning style of augmented instructions. In this paper, we describe the concept of augmented instructions and a VRML export function added to a molecular visualization tool, Molfeat [Molfeat], to visualize molecular data, such as PDB (Protein Data Bank).

## AUGMENTED REALITY

Users can enhance real scenes with AR by superimposing virtual objects onto real scenes. This improves a user's performance in and perception of the world. The user can see virtual objects in the same position and orientation as real objects in the scene. Unlike virtual reality, AR has the following unique features.

*Awareness of the real world:* Virtual objects that are superimposed onto a real scene function as landmarks of real objects that are important but do not attract a user's attention [Foyle, et al., 1993]. AR may also support spatial cognition by spatially relating information to physical objects in the real world based on suggestions about the relationship between spatial location and working memory [Kirsh, 1995].

*Personalization:* A user is able to see virtual objects which position and orientation are adjusted to keep geometrical consistency in the real scene depending on the viewpoint. That is, AR can present 3D graphics with geometrical information customized to each individual user.

*Tangible interaction:* AR gives us a new approach for flexible interaction between humans and computers. A user can interact with virtual objects in an intuitive and seamless way by manipulating corresponding physical objects and preventing the information space from disrupting the user's sensory perception [Poupyrev, et al., 2002].

Using the above features, many AR systems have been developed for demonstrations, with some systems having targeted education for their applications. Earth-Sun Relationships [Shelton and Hedley, 2002] presents seasonal variation in light and temperature, and the virtual sun and earth are manipulated on a small handheld platform that changes its orientation in coordination with the viewing perspective of the student. Construct3D [Kaufmann, 2002] was designed as a 3D geometry construction tool for mathematics and geometry education and has provided interactive learning environments through various scenarios. Augmented Chemistry [Fjeld, et al., 2003] is a virtual chemistry laboratory where students view simple atoms and acquire their own complex molecules while being bound by subatomic rules. Multimedia Augmented Reality Interface for E-learning (MARIE) [Liarikapis, et al, 2002] has been developed as an application for engineering education to enhance traditional teaching and learning methods.

## MOLECULAR VISUALIZATION

There are many tools to visualize molecular structures, which have similar functions as Molfeat. Here, we introduce

some of the software dedicated for visualizing molecular data and creating graphics materials for presentation or publication. We focus on their visualization functions rather than the theoretical calculation of the molecular properties.

DSViewerPro [Accelrys] provides access to flexible rendering options to examine 3D molecular models. A user can interactively rotate and scale molecules and apply a wide variety of display styles to highlight key structural features. DSViewerPro has a number of add-on modules that extend its computational capability. The software includes html browser-based interaction to access and manipulate molecular structures from a Web page.

PyMOL [PyMOL] is a molecular graphics system designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. Furthermore, PyMOL can edit PDB files. The software is an open source via Python, and users can modify and extend the program to better meet their needs.

CueMol [CueMol] is a freeware software to view molecular structures with a user-friendly interface that enables users to manipulate molecular data with a mouse. This software also has high-quality graphics rendering for publications using POV-Ray and for presentation with Microsoft PowerPoint using the ActiveX Control.

Chime [MDL] has the capabilities to display chemical structures, reactions, and textual content from various databases. This software also allows users to combine chemical structures with other data in HTML format for Web publications or for insertion into Microsoft Office applications. The plug-in is distributed from the MDL Web site.

MOLDA [MOLDA] is a molecular-model building program that supports various data formats, such as Xmol, Chem3D, MDL Mol, PDB, and CIF. 3D molecular structures in VRML can be generated by combining a variety of molecular science programs. Moreover, MOLDA for Java works on various platforms, such as Windows, Mac, and UNIX.

## Molfeat

Molfeat is a 3D graphics tool that enables molecular structures to be visualized for research purposes and to be authored for presentation and publication. There is no specialized function in Molfeat. Although relatively inexpensive and having minimal functions, Molfeat allows users to prepare a series of functions to create molecular structures.

For example, Molfeat provides users with the following functions: visualization of molecules (by balls, molecule surface, and ribbon), visualization of electron density (editing the range, position, and color), annotation in a 3D space (3D layout of characters and manipulation with a mouse), arrangement of molecular models (transformation of coordinates and superimposition of molecules), calculation and visualization of static potential (investigating potential



hydrogen bonds), point mutation (replacement of a residue with another residue), output of high-quality image using ray-tracing (for publication), and ActiveX control (for interactive presentation with the Microsoft PowerPoint).

Figure 2 shows a Molfeat control panel that acts as a graphical user interface. The user can view molecules in their favorite modes and easily control the viewpoint and scale of molecules through mouse manipulations. The user can also edit the molecules using the editing panel, which the user selects with the selection panel. The sequence view function shows the type of atoms, the name of residues, and the 2D structures as a sequence.

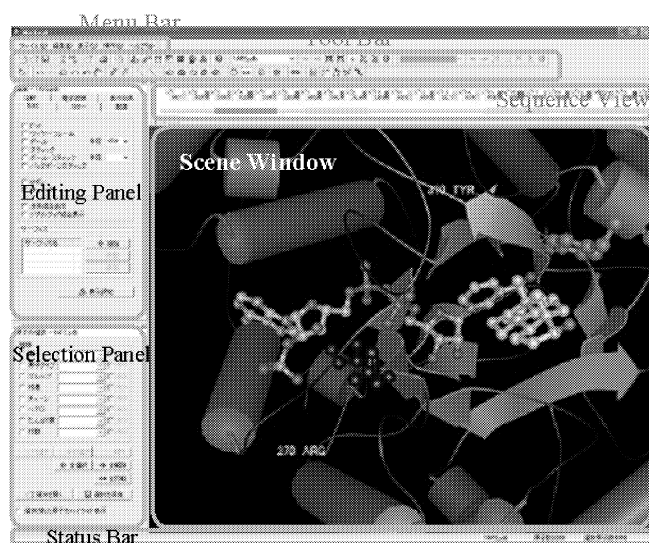
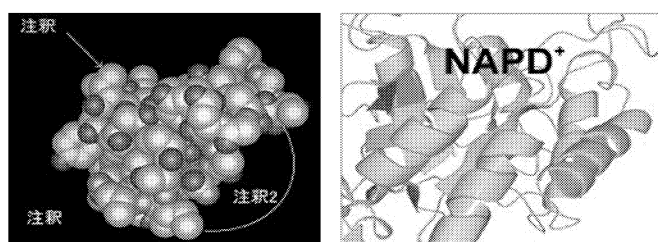


Figure 2: Molfeat control panel

Figure 3 shows examples of annotations to a molecule (a) in Japanese and (b) with uppercase letters. The user can insert letters and characters into the 3D space where required and point out the residue with an arrow. Figure 4 shows an example of an interactive presentation using Microsoft PowerPoint. The user can prepare presentation materials for molecular structures that can be interactively controlled with a mouse.



(a) Japanese annotations (b) Uppercase letters  
Figure 3: Examples of annotations to molecules

Here, we compare the main functions of Molfeat with other software in terms of molecular visualization, as listed in Table 1. The 'O' indicates the function is available, and the '\*' indicates the function requires additional modules or has some limitations. According to Table 1, Molfeat supplies the basic requirements for visualizing molecular structures and creating presentation materials.

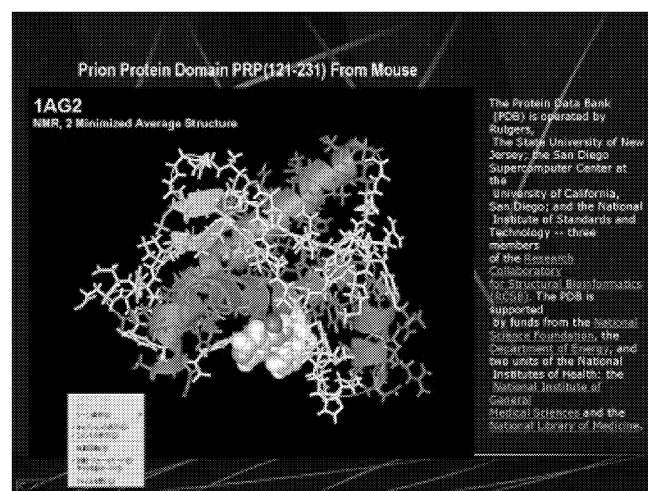


Figure 4: Interactive PowerPoint presentation

Table 1: Comparison of main functions

	1	2	3	4	5	6	7
DSViewerPro			O	O			O
PyMOL	O					O	
CueMol	O		O				O
Chime				O			*
MOLDA		O			O		*
Molfeat	O	O	*	O	O	O	O

- 1: Visualization of electron density
- 2: Annotation in 3D space
- 3: Arrangement of molecular models
- 4: Calculation and visualization of static potential
- 5: Point Mutation
- 6: Output of high-quality image using ray-tracing
- 7: ActiveX control

## AUGMENTED INSTRUCTIONS

Although multimedia systems are being used more often in various fields, traditional printed materials are not obsolete. Textbooks are a very popular medium in lectures and self-learning. There are many teachers who believe that printed materials are the best way to learn and understand essential information.

There are various advantages to using printed materials: 1) they are convenient when serious thought is required, such as solving complicated equations, 2) they are helpful for systematic studies, 3) they aid memory augmentation by relating descriptions to places in the text, and 4) there may be less eye fatigue than when looking at a computer monitor.

Textbooks are generally well organized based on the level of readers. In addition, textbooks involve many elements, such as concepts, rules, analogies, and imagery, so that information may be stored in long-term memory. Augmented instructions enable users to use a learning style based on printed materials while simultaneously accessing additional information using AR.

Supposing flexibility of learning is interpreted that learning methods become more interactive and individual, AR is considered to offer the good option for improving learning flexibility by interactively presenting information at the

user's favorite viewpoint. Augmented instructions would be more flexible if the learning materials were chosen according to the level of the learner's performance.

The basic concept of augmented instructions is the same as the MagicBook [Billinghurst, et al., 2001], which consists of a transitional AR interface that uses a real book to seamlessly transfer users from reality to virtuality. The MagicBook has basically been developed to demonstrate the potential of AR applications. The development of MagicBook did not significantly take into account the benefits that printed media offers as a traditional learning tool.

A head-mounted display (HMD) was used as a device to present information in the MagicBook, though it was adapted for a handheld display. An HMD can present virtual objects at the user's viewpoint, but current HMDs have insufficient resolution and inadequate field of view for users to read the small fonts that are usually printed in materials. Prolonged use of an HMD can tire users, therefore such devices do not commonly suit in learning [Asai, et al., 2005]. Besides, although an inertial tracker was used for tracking head orientation, the devices may not be available for usual learners. Thus, we targeted augmented instructions for general learners and mainly considered supporting their study based on printed learning materials by simultaneously presenting multimedia data.

Markers or tags in augmented instructions are added to text pages to identify information related to descriptions in the text, and are detected with an image-processing tool, ARToolkit. The multimedia contents superimposed over the markers or tags are preinstalled into an AR system. 3D geometric information may be especially effective for a learner's comprehension when used in augmented instructions because texts are limited to two-dimensional presentations.

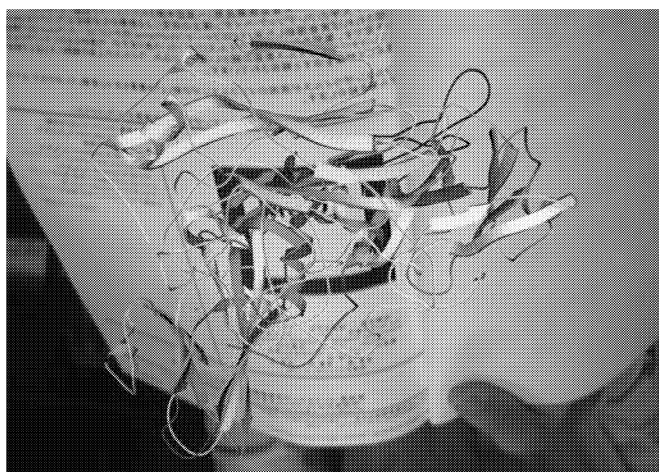


Figure 5: Overlaid 3D molecular structure

When the descriptions in the texts are sufficient for a learner to understand the content and additional information is not required, the learner only needs to use the printed materials. That is, augmented instructions support learning based on printed materials by using AR technology when required. Figure 5 shows an example of 3D molecular structures

overlaid onto a real scene captured by a camera. The user can control the viewpoint of the molecular model by shifting the position and orientation of the printed material.

## Visualizing molecular structures

It is not practical to implement all the functions of molecular visualization tools to augmented instructions, because the user interface becomes complicated as well as the implementation costs a lot. However, there are the minimum requirements of the functions when visualizing molecular structures. For example, one of the basic parameters for the molecular visualization is a visualization mode such as ribbons, ball-and-stick, and space-filled.

The molecular structure in Figure 5 was presented at the ribbon mode. Figure 6 shows examples of the visualization modes for the ball-and-stick and space-filled. The helix structure was presented at the ball-and-stick in the left, and the sheet structure was the space-filled in the right. The ribbon mode is better than the ball-and-stick or space-filled at visualizing the whole structure of thousands of molecules.

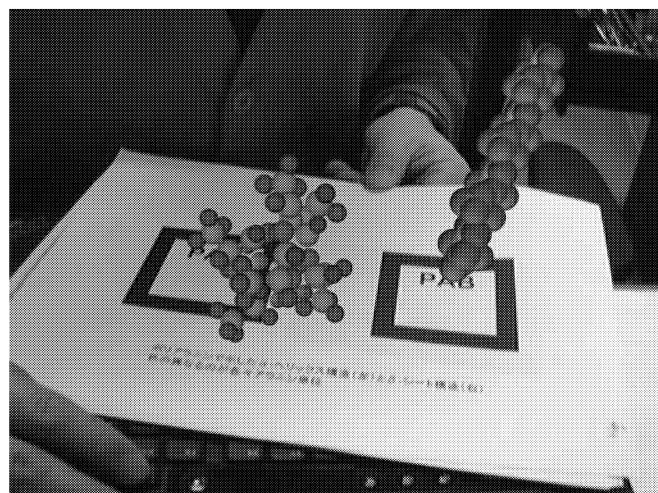


Figure 6: Visualization modes:  
ball-and-stick (left) and space-filled (right)

The other parameters of molecular visualization tools include (1) visualizing secondary structures (for example, a-helix and B-sheet), residues, and ligands and water; (2) changing the setting of colors, labeling, bonds, potentials, and rotation; and (3) classifying amino acids into the properties such as acidic, neutral, and basic categories, charges, polarity, and hydrophobicity. Changing the setting should be done on designating the specific parts of the molecular structure.

The current augmented instructions use the VRML runtime sample of the ARToolkit for visualizing molecular structures, and do not have any functions to control the parameters, even though the VRML files keep the above information. Therefore, the parameter change must be performed through markers, preparing all the VRML files for all sorts of parameters in advance using the visualization tools.

Another key element for efficient use of augmented instructions is that users themselves can author contents.

Embedding all the components of molecular visualization into AR systems requires graphics expertise and computer programming skills. Even adding annotations to molecular models requires users to have proficient skills with a 3D graphics tool or a 3D modeler. Our solution is to provide the current molecular visualization tool, Molfeat, with a function to convert information of the molecular data to the format that gets embedded into the AR system.

## VRML export in Molfeat

The ARToolkit is basically used as a marker-based system for detecting the position and orientation of markers and overlaying VRML models over images of the real scene. If a molecular visualization tool produces VRML files, even molecular scientists themselves can create 3D molecular contents for augmented instructions. Therefore, we developed a VRML export function as an optional module in Molfeat. Another advantage of this function is that it allows users to use a general plug-in for presenting VRML models in a Web browser.

All objects presented in a scene window are output as a VRML file, except for clipping (cutting of the objects). The file format supports VRML 2.0. The text node for annotations is treated as a 3D model, which can vary the font size with zoom-in and -out of the model. The billboard function works for annotations that face a virtual camera all the time. Because of compatibility between Molfeat data and VRML specifications, the following information is discarded in the conversion: the line width of wireframes, broken line of wireframes, size of molecular dots, clip plane, and fog expressions. Illumination is performed with a default light setting (headlight).

The VRML export function is achieved with a single conversion processing of a Molfeat file to a VRML file. Figure 7 shows a schematic of the software architecture. The sequential scenes from presentation data are transferred to the VRML scene builder module from the presentation manager. The scene data is converted to the VRML scene graph data through the VRML scene graph processing. The VRML file is created at the VRML97 parser. The user interface provided allows drag-and-drop manipulation in the application, as shown in Figure 8. When a user drops a FMP file (the Molfeat data format) to the dropping area, the corresponding VRML file is created at the same directory, showing the conversion status in the status window.

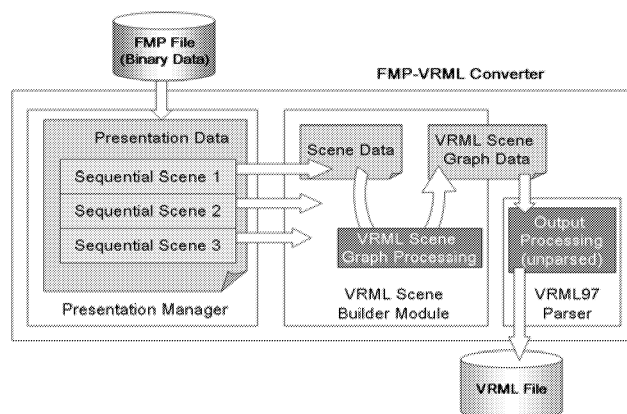


Figure 7: VRML converter architecture

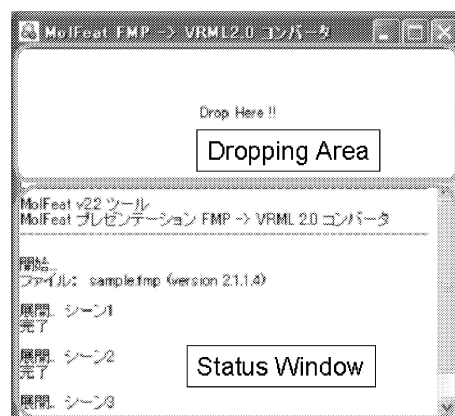


Figure 8: VRML converter user interface

## Implementation

The prototype system was implemented into a laptop PC (IBM, ThinkPad) with a compact USB camera (Creative, WebCam Notebook). Square markers were attached to documents of the printed materials. Figure 8 shows a typical displayed image using augmented instructions. Sample articles were prepared for the demonstration in which the chemical properties were explained. Some words were annotated at the important parts of the molecular structure, helping understand workings of the components. A learner could study the molecular structures of the protein by reading the printed texts and viewing the 3D geometrical model from the learner's favorite viewpoint.

In general, HMDs have been used in AR to present information, but a laptop PC was used here because of its compatibility with printed materials and its ability to reduce mental load that is often associated with prolonged use of HMDs. Despite many educators believing that AR has the potential to be used as an educational interface, limitations associated with HMDs have prevented AR from being widely used. The drawbacks are 1) cost of the required hardware, 2) insufficient resolution of the display, 3) difficulty focusing virtual objects, and 4) motion sickness. A laptop PC may reduce the effects of these drawbacks while still keeping some personalization.

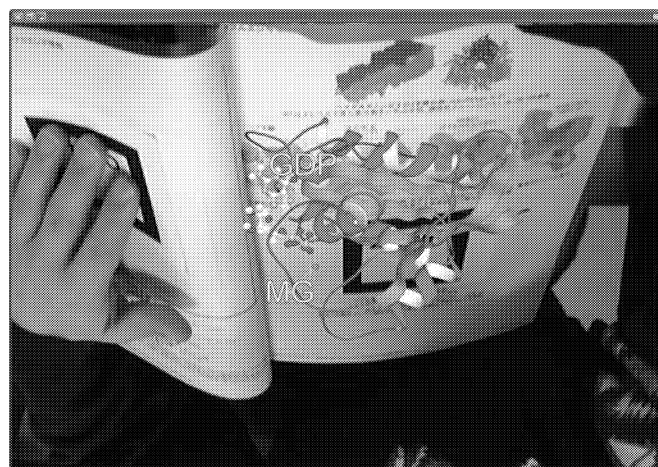


Figure 8: Molecular model with annotations

## SUMMARY

We described augmented instructions for learning molecular structures in which AR was used to bridge the gap between multimedia contents and printed materials. We discussed and compared the functions of molecular visualization tools. A VRML export function was developed as an optional module in Molfeat that enables molecular scientists to author 3D contents for augmented instructions by themselves.

A major benefit of AR is that a user can seamlessly interact with virtual objects by manipulating real objects. Future work includes improving the user interface for augmented instructions with tangible interaction. Acoustic information will also be used to play sounds in augmented instructions.

## Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research (17300283) in Japan.

## REFERENCES

- Accelrys Homepage, DSViewerPro, <http://www.accelrys.com> (As of Jan. 2006).
- K. Asai, H. Kobayashi, and T. Kondo, Augmented Instructions –a fusion of augmented reality and printed learning materials-, *Proc. IEEE International Conference on Advanced Learning Technologies (ICALT2005)*, pp.213-215, 2005
- R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, B. and MacIntyre, Recent advances in augmented reality, *IEEE Computer Graphics & Applications*, vol.21, pp.34-47, 2001.
- M. Billinghurst, H. Kato, and I. Poupyrev, The MagicBook: a traditional AR interface, *Computers & Graphics*, vol.25, pp.745-753, 2001.
- F. Biocca, J. Rolland, G. Plantegenest, C. Reddy, C. Harms, C. B. Owen, W. Mou, and A. Tang, Approaches to the design and measurement of social and information awareness in augmented reality systems, *Proc. Human-Computer Interaction International: Theory and Practice*, pp.844-848, 2003.
- CueMol Homepage, <http://cuemol.sourceforge.jp/> (As of Jan. 2006).
- DUMMIES.com, Homepage, <http://www.dummies.com/> (As of Jan. 2006).
- M. Fjeld, P. Juchli, and B. M. Voegtli, Chemistry education: a tangible interaction approach, *Proc. INTERACT2003*, pp.287-294, 2003.
- D. C. Foyle, R. S. McCann, B. D. Sanford, and M. F. J. Schwirzke, Attentional effects with superimposed symbology: implications for head-up displays (HUD), *Proc. Human Factors and Ergonomics Society Meeting*, pp.1340-1344, 1993.
- H. Kaufmann, Construct3D: an augmented reality application for mathematics and geometry education, *Proc. 10<sup>th</sup> ACM International Conference on Multimedia*, pp.656-657, 2002.
- H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto, and K. Tachibana, Virtual object manipulation on a table-top AR environment, *Proc. ISAR2000*, 2000, pp.111-119.
- D. Kirsh, The intelligent use of space, *Artificial Intelligence*, 73, pp.31-68, 1995.
- F. Liarikapis, P. Petridis, P. F. Lister, and M. White, Multimedia augmented reality interface for e-learning (MARIE), *World Transactions on Engineering and Technology Education*, vol.1, pp.173-176, 2002.
- MDL Homepage, Chime, <http://www.mdl.com/> (As of Jan. 2006).
- MOLDA Homepage, <http://www.molda.org/molda-e/home.htm> (As of Jan. 2006).
- Molfeat Homepage, <http://www.fiatlux.co.jp/molfeat/> (As of Jan. 2006).
- I. Poupyrev, D. S. Tan, M. Billinghurst, H. Kato, H. Regenbrecht, and N. Tetsutani, Developing a generic augmented reality interface. *Computer*, vol.35, pp.44-50, 2002.
- PyMOL Homepage, <http://pymol.sourceforge.net/> (As of Jan. 2006).
- C. Schmid, Simulation and virtual reality for education on the Web, *Proc. EUROMEDIA '99*, pp.181-188, 1999.
- B. E. Shelton, and N. R. Hedley, Using augmented reality for teaching Earth-Sun relationships to undergraduate geography students, *Proc. The First IEEE International Augmented Reality Toolkit Workshop*, 2002.

# THE APPLICATION OF STREAMING MEDIA TECHNOLOGY IN EDUCATIONAL SOFTWARE

Sid Satbhai

Charles A.P.G. van der Mast  
Man-Machine Interaction Group  
Delft University of Technology  
Mekelweg 4, 2628 CD,  
Delft, the Netherlands

E-mail: s.satbhai@ewi.tudelft.nl, c.a.p.g.vandermast@tudelft.nl

## KEYWORDS

Streaming media, educational software, e-learning, web lectures.

## ABSTRACT

The research described in this paper concentrates on reaping the benefits of streaming media web lectures at the university level through various rounds of fine-tuning, customization and expansion. User feedback is crucial in determining advantages and disadvantages. Therefore, web lectures are deployed as part of a university level usability engineering course, students are requested for feedback and this feedback is analyzed. Based on this feedback and a streaming media technology survey, this research strives to find the best solutions regarding web lecture composition, educational framework setup and pedagogical models for deploying streaming media educational software.

## INTRODUCTION

Although the traditional lecture setting is prevalent in academia, there is no solid evidence from learning sciences that supports its effectiveness in the learning process of students. It is a mode of learning that is, unfortunately, till this day firmly entrenched in custom and tradition which has rarely ever been questioned. Like all paradigm changes, making a transition from this traditional manner of teaching to implementing a more efficient and hence superior framework is not only very difficult, but is usually blackballed by the establishment and powers that be.

The development of effective educational software is a highly complex affair. On top of the coordination of multi-disciplinary teams, which is a feat in itself, a pedagogical model must be developed that is tailored to the application area at hand. The various methods for the development of pedagogical models find their basis in a number of different educational theories (Alessi 2001).

The field of educational software relies on a large wealth of previous research on learning and the resulting educational theories. There are numerous approaches that can be taken

towards teaching with technology. Key factors in determining the exact approach are the context, subject matter and available technology (Van der Mast 1995a, b).

Streaming technology offers a completely novel approach to accessing media over the Internet. Instead of waiting for an entire file to download to the user's computer before being able to playback, streaming media playback occurs *simultaneously with file transfer*. The data is first transmitted across the Internet, played back, and then finally it is discarded. Furthermore, streaming media offers the user control over the stream during playback. This is something that is not possible with a standard Web server.

For the purpose of this research, where streaming media technology is the technology of choice, and the web lecture is the actual educational software vehicle that is used to deliver knowledge, a pedagogical model has been chosen whereby a two-layered delivery approach is employed.

The choice of this two-layered delivery has solely to do with certain constants, such as the certain allotted live lecture hours per week for a given course. Instead of squandering precious live lecture time with inefficient propagation of knowledge (traditional lecture), a first layer is introduced whereby a web lecture is administered pre-class in order to deliver synopsisized knowledge on the subject matter to be handled in class. Then, in the second layer, which is the live lecture, a more engaging activity is substituted for the traditional lecture.

## EDUCATIONAL SOFTWARE

The "cone of experience" is a very straightforward model devised by world-renown educational researcher Edgar Dale (1969). Active learning is promoted on the lower levels of the Dale cone by means of involving the student as a participant. These lower levels offer extra stimuli and enhanced natural feedback. In this manner, a purposeful experience is built up by the student, making use of all five senses if need be. Due to this direct, purposeful experience, the learning experience becomes more memorable and so the student is able to retain the subject matter much more easily.

The higher levels of the cone basically endeavor to economize on resources by compressing the educational material into more and more compact forms. This enables faster delivery of the information. However, in these cases, the student has less time to process the information. The highest level on the Dale cone, verbal symbols, is actually the traditional academic teaching method. Being at the highest level, Dale's cone tells us that this is, surprisingly, the most ineffective way in which to teach. At this level, students simply do not have the time, ability or mental framework to effectively process the information being thrown at them.

Technology can support learning in many ways; attempts to use computer technologies to enhance learning began with the efforts of pioneers as early as in the sixties. Two contrasting views on the use of computer technology in schools still exist: the romanticized view is that "its mere presence will enhance student learning and achievement", in contrast is the view that "money spent on technology, and time spent by students using technology, are money and time wasted". The contemporary conception (based on literature reviews by several groups) is that "technology has great potential to enhance learning as long as it is used appropriately" (Bransford 1999). The potential of technology to support learning lies in the possibility to create learning environments that extend the possibilities of books, blackboards, radio and television shows. Also, learning environments implemented in software offer a range of new possibilities.

Van der Mast (1995a, b) presents a new theory for the development of educational software. The emphasis is laid on integrating various disciplines and media to achieve a synergistic effect. Three case studies provide a test bed for the theory, the aim of which is to enhance the practical organization of the development of educational software when working in multi-disciplinary teams. From past experience, the development of educational software is recognized as an extremely difficult problem for the vast majority of subject matters in various contexts. Interactivity is the central issue in educational software. In the past, the focus was on the technological problems in developing educational software and the creation of tools to support the task of programming. Nowadays, the focus has shifted to the pedagogical aspects of educational software because it has been determined that sound didactic strategies that are tailored for the subject matter and context under consideration are key to producing successful educational software.

Web lectures are nothing more than a specific application of streaming media technology for educational purposes. The ultimate goal is to teach a given subject to a student population. Depending on the subject matter and

educational goals, a specific type of web lecture must be chosen. The web lecture type relates mainly to teaching methods and layouts. Various options exist that suit various contexts. Some examples include: sports instructionals, art instructionals, university lectures and various commercial educational products. In Figure 1, a typical example of the web lecture format is given. On the left you see the video and the control buttons. On the right you see a slide of the show running. The learner can control the flow of both the slides and the video (or audio voice-over) independently. The video and the slides are always synchronized.



Figure 1: Example of a Web Lecture Format of the Collegerama System of Delft University of Technology

## STREAMING MEDIA TECHNOLOGY

Streaming media is a real time phenomenon. A hyperlink is used to access the streaming media. A short while after the access, the requested media is played back. This functionality also makes live broadcasts possible, something that cannot even be considered in the case of downloaded files. An important point to note is that because the streaming media packets are discarded after the playback, streaming media offers a very suitable degree of copyright protection.

Streaming media user control over the input stream can be likened to playback control on a VCR or tape deck. This type of control is not available with downloaded files until, obviously, the entire file is downloaded. Intermediate access is not possible in that case since the client can only make sense of the bits once the entire package is received due to the nature of the coding of the file. User interactivity is one of the most salient advantages of streaming media.



Streaming media is realized through the clockwork functioning of a number of different units of software. These units communicate on a number of different levels. A basic streaming media system has three components:

**Player** – The software that viewers use to watch or listen to streaming media

**Server** – The software that delivers streams to audience members

**Encoder** – The software that converts raw audio and video files into a format that can be streamed

These components communicate with each other using specific protocols. They exchange files in particular formats. Some files contain data that has been encoded using a carefully chosen codec. A codec is simply an algorithm, the purpose of which is to reduce the size of files.

Three streaming media production tools were evaluated. The first one, Microsoft Producer, is a free tool. The

following two, WM Recorder and RM Recorder are commercially available. We will describe our chosen tool, Microsoft Producer, briefly here.

MS Producer is a very simple presentation recording tool based on PowerPoint. The basis of each presentation is a collection of PowerPoint slides regarding a given topic. These slides are imported into MS Producer. Once imported, the user can then record an audio and/or video presentation while navigating through the slides. The end result is a web lecture which can then be published via MS Producer. The publishing process involves converting the result into a Windows Media Technology streaming format and uploading this to a streaming server. The server can then be accessed via a browser and the web lecture can be viewed. The entire process is shown in Figure 2. In Figure 3 an overview is given of the main editing screen of MS Producer.

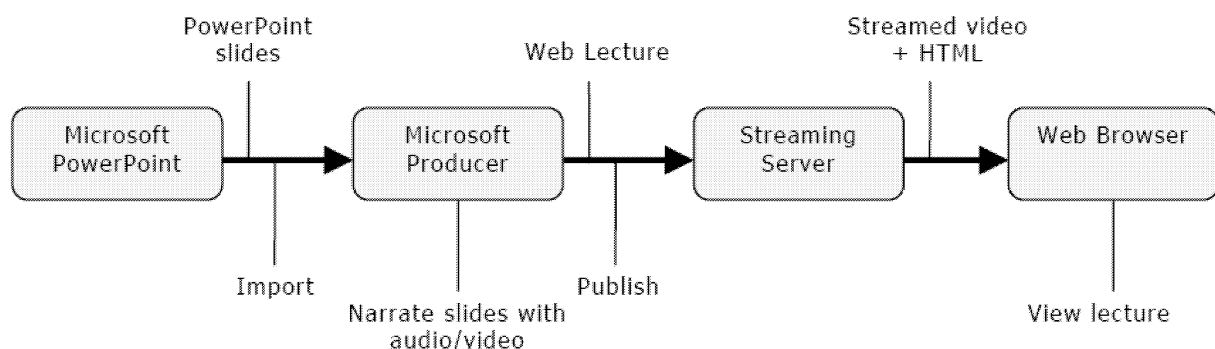


Figure 2: The Web Lecture Production Cycle using MS Producer (Groeneweg 2004)

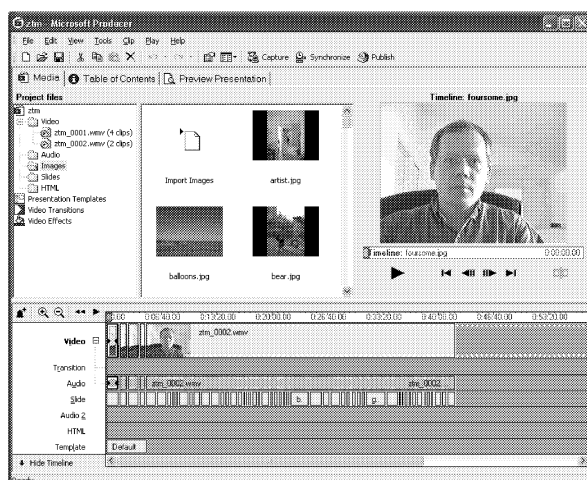


Figure 3: Main Editing Screen of MS Producer

## EXPERIMENT

The empirical component of this research is carried out using the course Usability Engineering (5 ECTS) at the Delft University of Technology as a vehicle for testing the deployment of web lectures.

The human-computer interaction (HCI) research field is quite broad and fragmented. No single unified theory of HCI exists. In introductory HCI courses (usually Bachelors level), students are generally introduced to some basic approaches and are given a foundation with respect to human (especially cognition), computer (especially components) and interaction (especially user interface and dialogue styles).

The Usability Engineering course is a one-semester Masters level course that builds upon previous HCI foundations and provides a coherent approach to HCI. The book used for the course is “Usability Engineering: Scenario-Based Development of Human-Computer Interaction” by Rosson & Carroll (2002). By following the prescribed method, the Usability Engineering course offers a general framework to extend and apply previously acquired usability related knowledge. Using this framework, elaboration is given on current theories, methods and technologies for establishing advanced (intelligent) user interfaces and for interactions with complex information and communication systems.

After some initial setup meetings, a general vision for the web lectures was developed based on Mark’s idea to have each web lecture given by a different guest speaker from TNO Human Factors in Soesterberg, The Netherlands, and Rabobank.

This idea evolved over the course of a series of discussions between us and eventually it was finalized. Various potential guest speakers were approached. The majority of these were his known colleagues and they gladly participated in the project and looked forward to what was, for them, also a new experience – namely, recording a web lecture. This recording was instead of visiting our campus to give a live lecture. This recorded web lecture could be reused.

After having arranged the web lecture side of things, we went back and looked at how the rest of the Usability Engineering course would have to be modified to take into account the introduction of web lectures. It was of critical importance not to overload the students and to keep within the ECTS credit limit of the course. Our ideal web lecture length (and the length we requested all speakers to attempt to stick to) was 15 minutes, give or take a few minutes. On top of this, each speaker was asked to provide a short homework assignment related to the web lecture produced. The aim of this was to give the students some hands-on experience with the specific subject matter dealt with in each web lecture before coming to class.

Due to the introduction of web lectures and homework assignments, the weekly paper summary was scrapped from the roster and instead each project group had to do just one presentation each on a pre-determined paper related to the usability engineering field. Each group was assigned a paper to present.

The eventual proposed educational framework looked like this:

1. The web lectures should be exciting and offer a fresh view into the field of usability engineering.

2. There should be a change of scenery in each web lecture – a different speaker presenting a different topic should be available for each web lecture.
3. The web lecture should ideally be 15 minutes in length.
4. The students should not be overloaded due to the web lectures and homework – i.e. the course should stick to its 5 ECTS limit.
5. The web lectures should serve as pre-class knowledge propagation vehicles with the goal of having students actively use that knowledge during the live lectures in the form of engaging activities such as discussions and/or assignments.
6. Each web lecture should be accompanied with a short homework assignment designed to provide some practical experience with the topic covered
7. In-class engagement should strive to provide the student population with memorable, practical and applicable experience of usability engineering.

It was decided that the final grade would be determined using the following weights: project 50 % of final grade, remaining 50% of final grade, broken down as follows: presentation on article 10%, homework assignments 20%, oral exam 70%.

The recording took place using the same setup as described in (Groeneweg 2004). Most recordings were made in the office of the presenter using a camera connected to the recording laptop via firewire interface, and using some small lights. The camera was a standard home digital video camera. On the laptop MS Producing was running the slideshow under control of the presenter using a remote device in his hand. An extra LCD monitor was positioned just below the camera to show the slideshow to the presenter online.

16 students signed up for the Usability Engineering 2005 course. Four groups of four participants each were formed for this course. Four of the students were on the Erasmus International Exchange Program and were taking this course in order to diversify their academic package. The remaining students were all TU Delft students. Of these, all save one were Media and Knowledge Engineering students. The remaining one student was a Geodetics student and was taking this course as part of his “Free Electives” credits. Usability Engineering 2005 was a truly multicultural affair, with the following student nationalities in the entire mix: Dutch, Turkish, Portuguese, Swedish, Chinese, Swiss, Indonesian and Indian.



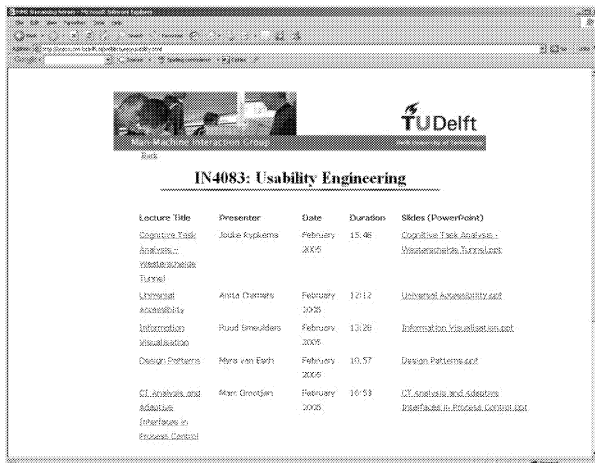


Figure 4: The Menu at the Streaming Server

The web lectures can be started from the left column, the slides can be loaded for later use from the right column.

Also presenter, date and length is shown.

The course followed the schedule described in (Sathai 2005). A web lecture evaluation homework questionnaire (see Table 1) was given at the end of the nine web lectures and focus groups were held at the end of the course (see Agenda in Table 2). In Figure 4 the menu of the streaming server is shown. In Figure 5 a screenshot of a web lecture on cognitive load is depicted.

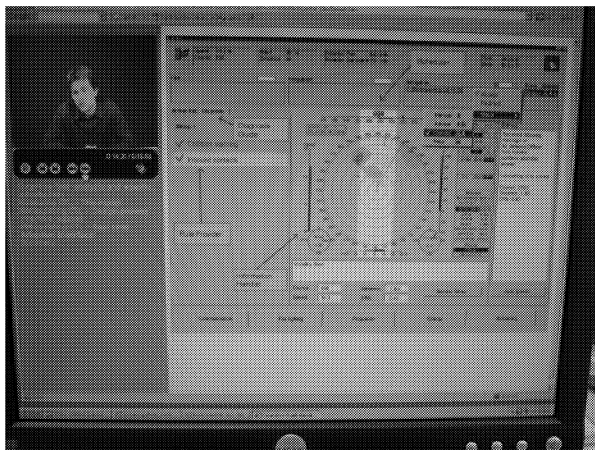


Figure 5: The Web Lecture on Cognitive Load with Complex Tasks

Table 1: The web lecture evaluation questionnaire

<b>Web Lecture Structure</b>
1. What did you think of the overall quality of the web lectures?
2. What did you think of the speakers?
3. What did you think of the PowerPoint sheets?
4. What did you think of the lighting?
5. What did you think of the flow of the web lectures?
<b>Web Lecture Content</b>
1. Do you feel that the web lecture prepares you well for class?
2. Does the speaker explain the material on the PowerPoint effectively?
3. Given the choice, would you prefer a course with web lectures or without?
4. Would you rather view a web lecture in preparation for class, or do you feel an article to read would be more effective?
5. What added features would you like to see in a web lecture?
<b>General</b>
What are, according to you personally, the advantages and disadvantages of the web lectures?

The purpose of the focus groups was to gauge the effectiveness of the web lectures and homework assignments that accompanied the Usability Engineering course this year.

The proposal was to hold 4 sessions: one session for each project group (1, 3, 4 and 5). (Group 2 had been dissolved early on during the course due to the withdrawal of a participant from the course.) This means that the proposal was that 4 focus group sessions were to be held, each session consisting of 4 people. Although sizes of 6-12 are recommended for focus groups, the most important variable in determining group composition for focus groups is homogeneity of the group members. In this case, the members of each project group carried out their project together for 2 semesters and thereby developed a working relationship. In the course of their working relationship, they formed a collective opinion on the entire course based on their mutual experiences and communication. Therefore, retaining this project group status seemed like the best option for focus group composition.

The goal for each focus group was to jumpstart a discussion about their overall experience of the Usability Engineering course, with a focus on the web lectures and homework assignments.

The aim was to record each focus group with the video camera and laptop setup from the Delft lab. The duration of each focus group was not to exceed 20 minutes

Table 2: Agenda for the focus groups at the end of the course

<p>1. How did they like the Usability Engineering course? Would they recommend it to others?</p> <p>Did they find it useful in the framework of their overall education?</p> <p>What did they think of the academic level of the course? Do they think such a course belongs in a technical university?</p> <p>What did they think of the live lectures?</p> <p>How was the overall course workload? Too high? Too low?</p> <p>What did they think of the project?</p>
<p>2. What did they think about the web lectures? Were they interesting?</p> <p>Did they think the web lectures provided a fruitful addition to the course?</p> <p>What would they think of web lectures for every course at the university?</p> <p>In case of negative response – Why?</p> <p>Regarding the composition of a web lecture, would the addition of interactivity be a good idea?</p> <p>How was the accessibility of the lectures?</p> <p>Would they prefer the current streaming media setup or would they rather receive the web lectures on a CD?</p>
<p>3. What did they think about the homework assignments?</p> <p>Were they interesting?</p> <p>Were they too long or short?</p> <p>Were they too easy or too difficult?</p> <p>What would they change about the way the assignments are administered?</p>

## RESULTS

Upon examination of the web lecture evaluations and the focus groups, a few common threads can be identified running through all the results generated during this research.

*The students would prefer web lectures over any other type of pre-class preparation, such as reading articles, making summaries, etc.*

This is overwhelmingly apparent from the results. Everyone feels that reading articles and making summaries takes too long and that it is, in general, more boring and harder to maintain concentration. The web lectures are easier to get into and since they flow naturally and automatically, it is easier to go along and maintain a longer attention span.

*The students were not happy with how web lectures were employed for the Usability Engineering 2005 course.*

Two main reasons for the overall displeasure with the deployment of web lectures for Usability Engineering 2005 were identified:

*The main reason for displeasure with the web lecture system for Usability Engineering 2005 was that most of the in-class time did not make use of the pre-class preparation, but instead repeated what was handled during the web lectures. This defeated the purpose of the web lectures entirely.*

This was the greatest flaw in Usability Engineering 2005. With all the time and effort that went into the production of the web lectures themselves, the second – equally important – in-class component was neglected. The in-class time should have been used far more efficiently and meaningfully. It is really ludicrous that, in some cases, the web lecture material was almost entirely repeated during class.

*The second reason for displeasure with the web lectures system for Usability Engineering 2005 was the varying quality of the web lectures. Most of the web lectures were considered to be lackluster and badly presented. However, other web lectures were enjoyed by all. Furthermore, the inclusion of some Dutch PowerPoint slides in a few web lectures caused irritation among some of the student population who do not speak Dutch.*

A number of complaints were consistently present regarding the deployment of web lectures for Usability Engineering 2005. First and foremost, the quality of the web lectures was deemed substandard on virtually all fronts. The lighting was not to liking. Most of the speakers were not considered good. The sound quality was not up to par. There were some accessibility issues with students who used other platforms (Apple).

*If the pedagogical model was fine-tuned and the in-class structure adjusted to make proper use of the pre-knowledge delivered by the web lectures, the majority of students would not be against web lectures for all courses at the university – provided they were tailored in each case to suit the subject matter (e.g. mathematics differs a lot from usability engineering).*

It is indeed true that designing a specific pedagogical model per subject matter and context is crucial to successfully propagating specific domain knowledge (Satbhai 2004). Therefore, web lectures would need to be suited for any particular context. This tailoring process would rely heavily on previous teaching experience in specific domains, for example mathematics, arts, languages, etc. Using past experience, an effective structure and related content could be devised for the web lectures.

## DISCUSSION

This research has investigated the use of web lectures to assist in the teaching of a university level usability engineering course. The pedagogical model involves a two-tiered setup, in which the web lectures serve as vehicles to deliver introductory knowledge on a specific topic pre-class. In this manner, the intention is to use in-class time for more engaging and hence effective learning activities.

MS Producer has proven to be more than adequate at this evolutionary stage, where the focus lies on reaching a suitable pedagogical model for various university level subjects. This research only investigated application of web lectures to usability engineering and started where [Groeneweg 2004] left off. No doubt, others will continue where this research has ended. After each subsequent pass, a more thorough and solid pedagogical model will emerge until finally a new standard could potentially be reached. At that stage, the focus can turn to cosmetic issues and maybe a commercial streaming media production tool can be investigated. However, up until that point, the focus must remain on researching solid pedagogical models for this teaching setup, and tools should only be secondary. For now, a tool such as MS Producer is fine for the purposes of research.

Making sure of student participation in such experiments is a trivial matter. If participation is simply set as a requirement for the course at hand, participation can be guaranteed. Having the homework assignments, which can only be completed upon watching the related web lecture, count as 10% of the final grade immediately assures student population cooperation.

Quality control regarding the speakers for the web lectures is an important issue and should be planned in advance. In this case, speakers were simply chosen ad hoc and the disparity in quality was apparent. The idea of a different speaker each time was much appreciated by the students due to the variety and freshness for each web lecture, but quality control was lacking. It should be noted that the quality was not lacking because the presenters were bad as such, but because they were not used to recording web lectures. Even experienced speakers took some time to get used to speaking to a camera. The key issue here is sufficient time to develop the web lecture. In this research, most sessions were far too hurried due to time constraints on all involved. In future, more recording sessions with fewer speakers and more takes per session would be a wise move.

Properly developing the in-class phase of the two-tiered approach is absolutely crucial to making any of this worthwhile. This research showed overwhelmingly that the

system failed here because the in-class time was not only not particularly engaging, but a waste of time in quite a few cases because it was nothing but a repeat (in one case exact repeat!) of the web lecture. This is currently the weak link in the chain and must be the focus of any subsequent work in this area.

The overall positive attitude of the students, despite the bugs in the current pedagogical model, is very encouraging and hence this area should certainly be explored further. As mentioned, it was not possible to carry out an extensive, comparative study such as those detailed in (Groeneweg 2004) and (Day et al. 2005). However, in future, with sufficient time and the necessary interest, there is no reason why such a study could not be carried out at Delft. It would be a most beneficial research endeavor. One item to suggest as part of any future research in this area would be to look in depth at the cross-platform compatibility issue, as this issue is currently the biggest bottleneck on the web lecture accessibility front. A fruitful avenue to explore for the solution of this issue would be Macromedia Flash, which is a platform independent multimedia player.

## CONCLUSION

The advantages of using web lectures to modify the existing pedagogical and organizational frameworks within academia lead to a number of very real advantages:

- Modularization of education
- Ease of reuse of educational components
- Immediate, more convenient and persistent accessibility of educational content
- Flexibility on various levels, from the student level to the academic organizational level

Looking at the road slightly ahead, web lectures could potentially be used by students to prepare for any course they may encounter in the future. For example, students could follow various web lectures on their own while following related courses. In this way, they could compare various web lecture sets produced for the same course and choose the web lectures that they like the most. For example, the Delft University of Technology shares web lectures with Georgia Tech. There is an ongoing cooperation in this field between these two institutions. Georgia Tech maintains a great web lecture repository. A student at the TU Delft may get an assignment to watch a locally produced web lecture and another one on the same topic produced at Georgia Tech, and then asked to compare the two. This would further enable the gauging and fine-tuning of web lecture quality.

## REFERENCES

- Alessi, S.M. and S.R. Trollip, *Multimedia for Learning* - Third Edition, Allyn and Bacon, 2001.
- Bransford, J.D., A. L. Brown, and R. R. Cocking, *How People Learn - Brain, Mind, Experience and School*. Washington D.C.: National Academy Press, 1999.
- Dale, E., *Audio-Visual Methods in Teaching*, 3rd ed. New York: Holt, Rinehart, and Winston, 1969.
- Day, J., R. Groeneweg, J. Foley, C. van der Mast, Enhancing the Classroom Learning Experience with Web Lectures, in: C.K. Looi at al. (Eds.) *International Conference on Computers in Education*, Singapore, APSCE, 638-641, 2005.
- Groeneweg, R., Enhancing the Classroom Learning Experience with Web Lectures, Master Thesis, MSC Program MKE, Faculty EWI, Department of Mediamatics, Delft University of Technology, 2004.
- Mast, C.A.P.G. van der, *Developing Educational Software: Integrating Disciplines and Media*. PHD Thesis, Delft University of Technology, 1995a.
- Mast, C.A.P.G. van der, Professional Development of Multimedia Courseware, *Machine-Mediated Learning*, 5(3&4), 269-292, 1995b.
- Rosson, M.B. and J.M. Carroll, *Usability Engineering*, Morgan Kaufmann Publishers, 2002.
- Satbhai, S., Research Assignment: Educational Software, MSc Programme Media and Knowledge Engineering, Delft University of Technology, 2004.

Satbhai, S., The application of streaming media technology in educational software, Master Thesis MSc Programme Media and Knowledge Engineering, Delft University of Technology, 2005.

## BIOGRAPHY

**CHARLES VAN DER MAST** has a PHD Mathematics and Computer Science from Delft University of Technology where he is employed at the Man-Machine Interaction group at the Department of Electronic Engineering, Mathematics and Computer Science as a Associate Professor. He teaches Multimodal Interfaces and Virtual Reality, Developing Highly Interactive Systems, Multimedia, and Usability Engineering, and developed the Bachelor/Master curriculum in Media & Knowledge Engineering. His interests include using various media to improve teaching, VR therapy for phobia treatment.

**SID SATBHAI** was a student of the master programme Media and Knowledge Engineering at Delft University of Technology. He graduated in 2005.

# DESIGN AND EVALUATION OF A VRET SYSTEM FOR AGORAPHOBIA

Charles A.P.G. van der Mast  
Frans S. Hooplott  
Man-Machine Interaction Group  
Delft University of Technology  
Mekelweg 4, 2628 CD,  
Delft, the Netherlands  
E-mail: c.a.p.g.vandermast@tudelft.nl

## KEYWORDS

Virtual reality, phobia treatment, simulation, avatars.

## ABSTRACT

Virtual Reality Exposure Therapy (VRET) involves exposing a phobia patient to a virtual environment containing the feared stimulus instead of taking the patient into a real environment or having the patient imagine the stimulus.

This paper describes how virtual environments for the treatment of agoraphobia are implemented based on requirements from therapists. Research has been done on the requirements for the creation of a valid and anxiety-provoking virtual world for the treatment of agoraphobia. This paper describes the implementation of a prototype of a system. The evaluation by therapists is reported.

## INTRODUCTION

Virtual Reality Exposure Therapy (VRET) is an evolving technique that has attracted a lot of research. Traditional exposure therapies (also called exposure in vivo) expose patients, who suffer from a phobia, to real world environments that contain the feared stimulus (Emmelkamp et al. 2004). Virtual Reality Exposure involves exposing the patient to a virtual environment containing the feared stimulus.

It has been proven that VRET is effective for patients with acrophobia (fear of heights), arachnophobia (spider phobia) and fear of flying (Witmer and Singer 1998). The effectiveness of VRET in other anxiety disorders like claustrophobia, fear of public speaking, fear of driving, posttraumatic stress disorder, and agoraphobia also holds promise for the future (Emmelkamp et al. 2004). At Delft University of Technology a generic system for treatment of phobia has been developed, taking into account specific human-computer interaction issues (Gunawan et al. 2004). This VRET system we developed so far provides interactive virtual worlds including many objects but without virtual humans, and a friendly user interface for the therapist to control a session. We decided to investigate the requirements for worlds to treat agoraphobia, one of the more complex phobias. For this treatment the VRET system had to be enhanced with more complex basic functionality. Several research projects have been

conducted on the efficacy of VRET in treating agoraphobia (Botella et al. 2004). Agoraphobia with panic disorder is the most common form of agoraphobia. It happens when a person that suffers from panic disorder is afraid of being in places or situations from which escape might be difficult, or embarrassing, or in which help might not be available in the event of a panic attack. For example a person that has had several panic attacks in a crowded subway, can avoid the subway in order to prevent the panic attack. In the worst case a person might stay housebound, because he or she has had several panic attacks in several places (like the elevator, a crowded square or a subway) and avoids all of these places and situations. So in fact panic can be seen as the precursor of agoraphobia. Agoraphobia with panic disorder is a combination of panic attacks and avoidance behavior. The core element of agoraphobia is "the fear of the fear". One of the problems with VRET for agoraphobia is that the patients were not able to feel present in the virtual environment.

The goal of this study was: investigate the requirements therapists have for a usable and controllable VRET system for agoraphobia; design and evaluate anxiety-provoking virtual worlds for the treatment of agoraphobia based on these requirements; implement a prototype for the treatment of agoraphobia, and evaluate it with experienced therapists by inspection, as a first step to prepare controlled experiments with many patients.

## THEORETICAL BACKGROUND AND RELATED RESEARCH

### Agoraphobia

Agoraphobia with panic disorder is the most common form of agoraphobia. It happens when a person who suffers from panic disorders is afraid of being in places or situations from which escape might be difficult, or embarrassing, or in which help might not be available in the event of a panic attack. For example a person that has had several panic attacks in a crowded subway, can avoid the subway in order to prevent the panic attack. In the worst case a person might stay housebound, because he or she has had several panic attacks in several places (like the elevator, a crowded square or a subway) and avoids all of these places and situations. So in fact panic can be seen as the precursor of

agoraphobia. Agoraphobia with panic disorder is a combination of panic attacks and avoidance behavior. The core element of agoraphobia is “the fear of the fear”.

One of the most characteristic features of agoraphobia is the avoidance aspect. In the case of agoraphobia with a history of panic attacks, the number of situations that cause anxiety are as diverse as the number of persons that suffer from agoraphobia, because it depends on the persons panic attack history. Agoraphobia is a fear of fear and not so much a fear of certain places. But the central theme is “not being able to leave” or “being stuck”. So for patients to feel present in the virtual world (or place) and really experience anxiety the patients have to get the feeling that they are not able to leave or escape the situation or place.

Another aspect of Agoraphobia that has to be taken in account is that most agoraphobics experience a decreased anxiety when accompanied by a trusted person. So they often avoid being alone.

A few example situations are as follows:

- Standing in a queue
- Being in a large shop or shopping center
- Traveling by public transport (bus, train or airplane)
- Crowds, busy streets, large gatherings
- Driving a car on a motorway (the impossibility of turning on the road)
- Being in a traffic jam
- Crossing a bridge or being on a bridge
- Sitting at the barber's
- Being in conversation with some person on the street

## The VRET

An important aspect of the treatment is exposure. When people that suffer from a phobia are gradually exposed to the anxiety provoking situation, it helps them to create more neutral memory structures that ‘override’ the old anxiety provoking ones.

VRET is a technique that is progressing a lot, because the numerous research projects that have been done. Recently there has been a plenary discussion at the NATO Advanced Research Workshop on novel approaches to the diagnosis and treatment of posttraumatic stress disorder (Van der Mast et al. 2006). Basic functions for a VRET system were provided and the promising future of tele-care were discussed. The former system we developed is described in (Schuemie and Van der Mast 2001). This system is developed since 1999. We learned about the architecture required for optimal VRET.

## ANALYSIS OF THE PROBLEM

In this section we present the results of a requirement and task analysis based on interviews and discussions with therapists who know our current VRET system (Schuemie and Van der Mast 2001). and who are experienced in treating agoraphobia with the traditional methods. It is

emphasized that we try to design a new VRET system using a user-centered method in order to achieve good usability. During initial discussions with the therapists we choose to implement a virtual environment simulating a well known square in the Netherlands: The Market in Delft. We choose the Market for its large surface, historic background, international identity and its diversity. Primarily the large surface and diversity create a lot of opportunities for this research, because these aspects contribute to the anxiety provocation for agoraphobics.

## Requirements analysis

During our research on the parameters that are necessary for creating valid and anxiety-provoking virtual environments for the treatment of agoraphobia, we came up with the following conclusion (Hooplot 2005):

The necessary parameters for creating valid and anxiety-provoking virtual environments for the treatment of agoraphobia can be divided into two groups: one common group of parameters that contribute to the feeling of presence Schuemie et al. (2001), which are suitable for any phobia. For these parameters we used the general theory of presence. All of the five parameters constitute to one of the three forms of presence. A second group of anxiety provoking parameters contributes to the level of fear that the patient experiences. This last group of parameters is phobia specific.

We also concluded that these two types are closely related. For example when the feel of presence is high, but the situation (in the virtual environment) is not provoking any anxiety the effectiveness is low. Also, when the virtual environment is a provoking situation, but the patient does not feel present, the effectiveness is also low.

For the exhaustiveness we give a summation of the two parameter groups for presence and anxiety provocation.

- Level of realism. Of course the more realistic the virtual world is, the more the feel of presence will increase.
- Virtual body. A representation of a user's body in the virtual environment, contributes to the feel of personal presence.
- Number of sensorial modalities. Sensorial modalities are visual, auditory, tactile (feel, touch) and olfactory (sense of smell) The more sensorial modalities the system covers, the more the user feels present in the virtual environment.
- The level of interaction with and existing of other creatures in the virtual environment. The more the patient is able to interact with other (virtual) people in the virtual environment, the more he/she will feel present in the environment. Avatars with expressions and inferred-gaze could have a positive impact on perceptions of communication and thus on the feel of being present (Garau et al. 2003).

- Level of interaction with the environment. This accounts for the degree to which users of the medium can influence the form or content of the mediated environment.

The following parameters contribute to the level of fear experienced by the patient. These parameters are phobia specific. According to (Botella et al. 2004), panic disorder and agoraphobia sufferers usually avoid two different kinds of stimuli. External and introceptive stimuli. So parameters that contribute to the level of fear can also be subdivided into two types. Simultaneously conducting these parameters will show best results.

- Level of possible escape. This parameter accounts for the situation that is simulated by the virtual world. Agoraphobics fear situations from which escape is hard or impossible. So this parameter should be kept as low as possible. Next follows a list of situations where the level of possible escape is very low as we saw earlier on. Standing in a queue; being in a large shop or shopping center; traveling by public transport (bus, train or plane); crowds, busy streets, large gatherings; driving a car on a motorway (the impossibility of turning on the road); being in a traffic jam; crossing a bridge or being on a bridge; sitting at the barber's; being in conversation with some person on the street.
- Level of habituation. By no means may the patient get habituated to the environment, because this causes a decrease in the patients level of fear. The level of habituation should be kept as low as possible. This can be achieved with a virtual environment that has as much variation as possible.
- Number of (agoraphobia specific) bodily sensations. This is the number of agoraphobia specific bodily sensations that can be simulated with the system. (E.g. blurred vision, tunnel vision)

## Resulting requirements

The above parameters have to be taken into account for the implementation of the world. However a trade-off has to be made between the effectiveness and costs. From the above parameters we can easily fill in some of the requirements for our virtual environment that will be built. These requirements are listed in (Hooplot 2005).

We take a look at the anxiety-provoking parameters to extract requirements for the new system. The level of possible escape shows a strong relation with the situation that the virtual environment simulates. In order to keep this parameter as low as possible we choose one of the situations from the list of anxiety provoking situations by Emmelkamp: "Crowds, busy streets, large gatherings". This situation is realized in the form of a large square, with a crowd on it.

As we saw earlier, the level of habituation must be as low as possible. This can be done by making the world heavily adaptable. Examples for our case are the change of weather, day and night, the textures of the houses or the number of people on the square. This way the participant will not get the chance to habituate to the environment. Also using different environments for the same situation is a suitable option.

From the list of bodily sensations, we decided to choose only the ones that could be simulated with auditory- and visual modalities. Thus we have the following bodily sensations: Increasing/decreasing heart rate., Shortness of breath, Blurred vision and Tunnel vision. The first two bodily sensations however, are simulated with auditory modalities and are not the actual heart rate and breath of the participant. In fact they are used to fool the patient. According to (Emmelkamp et al. 2004) the sound of an increasing heart rate, causes the participant to have an increasing heart rate. Also the sound of a decreasing heart rate has the opposite effect. This also accounts for the sound of breathing.

## Task analysis

During our research on the parameters that are necessary for creating valid and anxiety-provoking virtual environments for the treatment of agoraphobia, we interviewed therapist 1<sup>a</sup>. After this interview we came up with a detailed task analysis, that was approved by therapist 1.

## DESIGN

### Modular structure

In our design we tried to have a modular structure. In contraction with our task analysis we came to the following modules, see Figure 1.

In figure 1 we see the modular structure of the virtual environment. The world control objects can be seen as the interface towards the user. These objects can be widgets like sliders, buttons, radio buttons etc. The virtual world module is the actual 3D realization of an environment. Each world is constructed of a few modules. There are two main groups, namely the world controls and the world elements. The world elements are the visual or hearable parts of the world, which are brought to the senses of the user. These world elements have parameters like position, material etc. The world controls make use of these parameters to control the world elements. So the world controls and elements communicate with each other through a two way communication channel. Each world element module has a separate dedicated world control

---

<sup>a</sup> Therapist 1 is experienced in treatment of many kinds of anxiety disorders including agoraphobia. He is also experienced in using our VRET system for acrophobia and fear of flying.

module. These control modules handle all the input from the user and the system itself in order to supply the world elements with the proper actions. World controls do not necessarily have to have a user input. For example people that randomly walk over a square, without any input from the user, still have to have a control that calculates their AI movements (standing still, walking, going to the next destination).

Each world definitely has some world controls and world elements. But the content of these controls and elements may differ for each world.

One special world control object is the navigation, because it has no visual or audio representation in the virtual world, but it actually controls what the participant sees. It

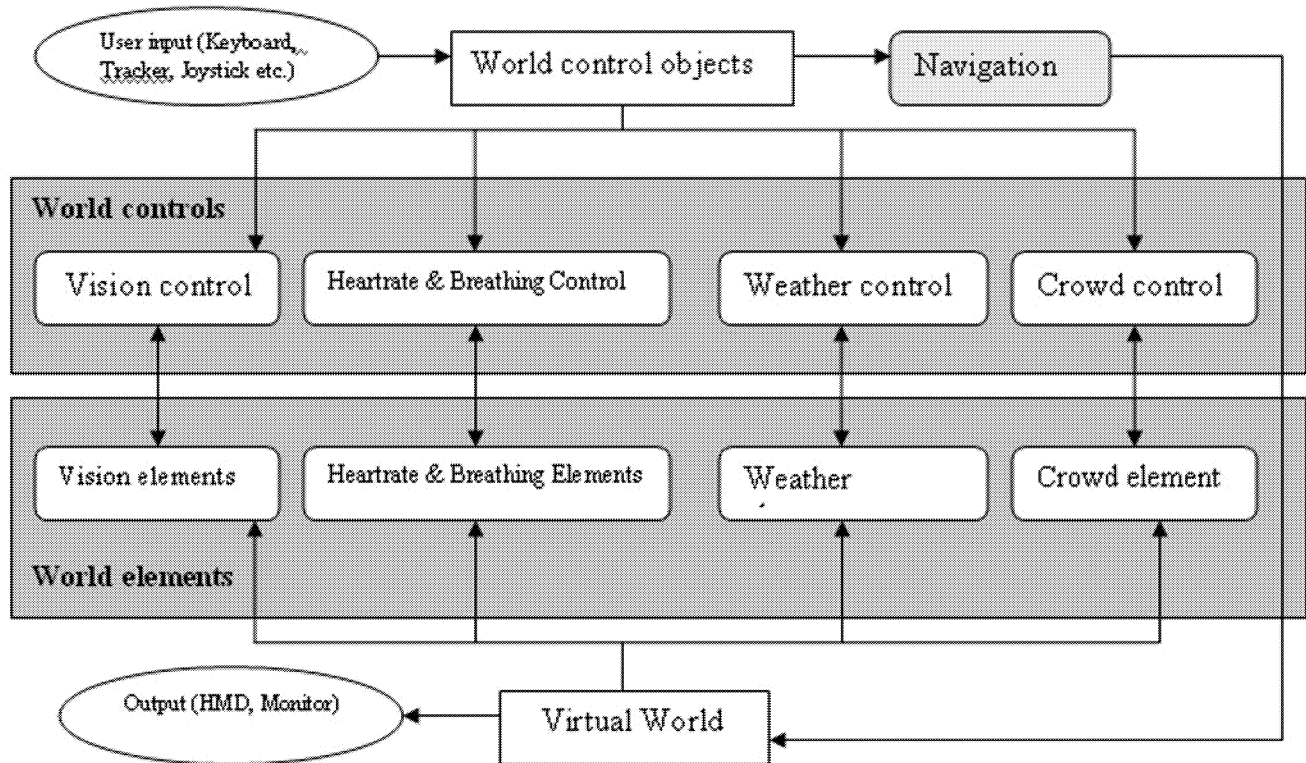


Figure 1: Modular structure of a virtual environment including world elements and world controls. Arrows indicate the direction of dataflow.

is somewhat the control of the viewport, through which the virtual environment can be seen. This module can have user input (E.g. the therapist who is moving the patient in a certain direction). This module however doesn't communicate with any world elements. It only uses external user input and external parameters from the tracker system. The output has effect on the virtual world.

## IMPLEMENTATION

After a comparison of some state-of-the-art VR developing tools (CAVELib, CaveUT, EONReality, MetaVR, MultiGen, Quest3D, Virtools and Vizard, see (Hooplot 2005), we chose for Quest3D as the best for this research project.

But Quest3D is not really object oriented. An average program in Quest3D consists of a complex network of

channels, ordered in directories and sub directories. So it is quite difficult to map the modular design to the working environment of Quest3D. It is up to the programmer to bring as much structure in to his or her program as possible.

Our first concern for the agoraphobia world, was how to make a 3D world of the Market in Delft. There are numerous ways to do this. We could for example measure all the elements and buildings on the market and rebuild them in the 3D world using these measurements. This method is very time consuming and almost impossible, thus we chose to use another method.

We decided not to make a 3D representation of each building. Instead we decided to take four planes (one for each side of the square) and put a large transparent texture on these planes. First we took pictures of all the buildings



on the square. We did this by taking a series of overlapping pictures of the front of the buildings. This resulted in a large set of pictures. These pictures also had perspective and because we wanted to use these pictures as textures for planes we had to remove all the perspective from the pictures. This was done using the crop tool in Adobe Photo shop. The pictures all had to have the same color dept in order to get a smooth transition when we overlap them. Eventually we had to remove everything from the image that is not a part of the building. This was mostly the sky and parts of the ground. This was necessary because these parts were going to be transparent in the virtual world.

After the buildings were constructed we had to place the elements on the square, like the statue of “Hugo de Groot”, the lampposts and more. All these objects were modeled in Maya3D and then converted to Quest3D using the X-file exporter plug-in from Quest3D. This plug-in can be installed in Maya and allows Maya to export 3D objects and 3D environments to the X-file format, that can be imported by Quest3D. The coordinate system of Quest3D differs from that of Maya. So we had to flip the X-axis in order to get the 3D objects right. For the church we choose to make a 3D model. We had to make pictures of each side of the church, in order to get the right textures for the model. We used a UV map to do the texturing of the church. This process was also necessary for the lampposts, the signposts and the statue.

During the implementation we noticed that the ambience sounds show a strong relation with the events in the virtual environment. Therefore we decided to slightly change this aspect within the implementation. We coupled these ambience sounds to the events in the virtual environment. E.g. When the number of people on the square increases, the sound of mumbling people automatically increases.

### Camera setup

To begin with, the virtual environment is constructed of multiple camera-viewports that are mapped onto the screen. In our example we also have a separate camera for the user interface, which can be replaced by an external user interface.

In figure 2 we see the setup that runs with one computer, with two monitors. The therapist view and patient view are actually identical except for their resolution. The therapist view has a smaller resolution in order to appear on the user interface. On top of these viewports there is a separate camera viewport for the simulation of thunder. This is actually a camera pointed at a white surface. This surface

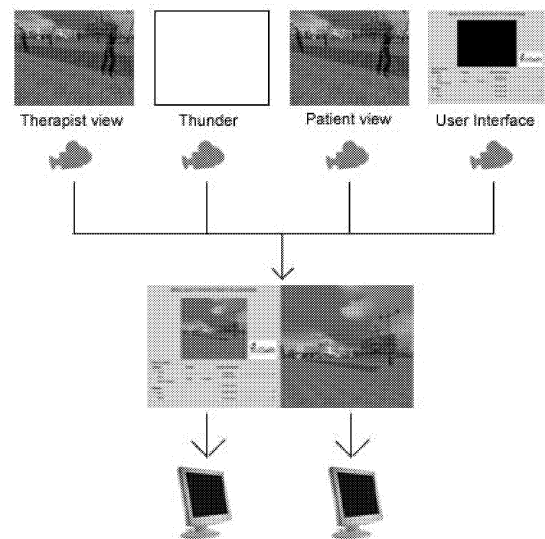


Figure 2: Composition of multiple (camera) viewports changes it's transparency as we will see later on in the weather control and elements. The last viewport is that of the user interface.

### Crowd control and elements

This is one of the more complex modules within the agoraphobia world. The human characters have to wander and walk around on the square, sometimes stand still and randomly choose their own path.

There are two types of human characters in the agoraphobia world: Characters that walk and sometimes stand still on a random position and characters that constantly stand still on a fixed position. First we will describe the implementation of the walking characters, and after that we will explain the implementation of the characters with a fixed position.

#### Walking characters

The human characters are skinned characters. This means that they consist of a bone structure, which is enwrapped by the 3D object of the structure (the skin), eventually a texture is mapped onto the skin. In figure 3 we see an example of this skinning process. By moving these bones, the skin also moves along. Thus the character can be animated by moving these bones. A set of bone movements can be stored as a motion set, containing the animation.

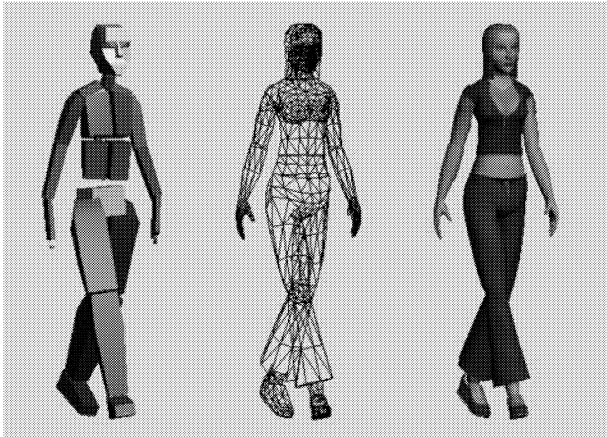


Figure 3: From left to right: 3D bone structure, 3D Skin and the eventual textured 3D result

All characters can move over grid-like network. This is a collection of nodes, which are completely or partially connected. The whole path-finding system revolves around this 3d graph. In our example the environment has a 3d graph that consists of 38 nodes in the range of [0,37]. The motion planning channel gives the human character a destination node (node number) and the human character then automatically calculates a path to that node. Matrix calculations are used to make sure the character faces the destination node that it is heading for.

### Character AI

The human characters have to act automatically and naturally. They have to randomly find their path through this network of nodes and sometimes even stand still. In order to reach the effect of a crowded square the characters all have to circle around the participant (camera). This is implemented in the system.

A problem that arises is when two or more character have the same destination node and all reach this node before the timer reaches zero. We then have two or more characters standing on the same spot. To deal with this problem, we defined the global variable 'stop distance'. In our case the stop distance is one. This means that the character has successfully reached a destination node as soon as it is in a range of one of the nodes' position. This way characters that reach the node coming from a different direction will all stand still on a different position.

### Weather control and elements

The part of the weather control and elements actually consists of two parts. One part controls the actual weather, say clear, rain or thunder and the other part controls the night and day.

### Vision control and elements

The vision control and elements consist of two main groups: The tunnel vision and the blur vision.

The tunnel vision partially uses the same technique as the thunder. We set up a separate camera pointing at a partially transparent surface. This surface spreads across the whole screen and overlaps the complete virtual environment. This surface has one texture that is completely black and four different alpha textures.

We also implemented blurred vision. The therapist can switch all these effects on and off at his user interface.

### Heartrate & Breathing control and elements

According to (Emmelkamp et al. 2004) the simulation of an increasing heartrate and breathing, has an anxiety provoking effect on the patient. This is mainly because the patient recognizes these bodily sensations and thus evokes the anxiety. There was some discussion if the heartrate and breathing should increase and decrease simultaneously. In the interview we had with therapist 1, we came to the conclusion that it would be realistic if the heartrate and breathing increase and decrease simultaneously. In the real world they are also closely related.

### Navigation

In the current system, navigation is done using a fixed path. The therapist can navigate the patient forward and backward on this path using different speeds. We experimented with this type of navigation using Quest3D. We implemented an example of a staircase. In Quest3D this is quite easy to implement. For the virtual world of the square in Delft, we choose to use another navigation. Using a fixed path on a square was not suitable, because there are too many directions you can go in. Instead we choose to use a free navigation, where the therapist can move in all of the four directions and the patient can look around using the HMD.

### EVALUATION

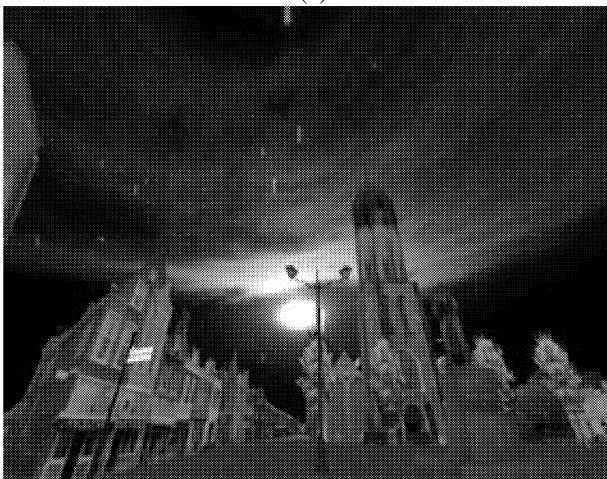
After the completion of the implementation we evaluated the virtual environment with a therapist <sup>b</sup> of a clinic of psycho-medical care in the Netherlands, see Figure 4a and Figure 4b. This therapist 2 was involved in previous projects at the Technical University in Delft regarding VRET. The therapist tested and evaluated the worlds for

<sup>b</sup> Therapist 2 is a professional therapist working in a clinic and is experienced in using VRET for acrophobia.

“fear of flying”, claustrophobia and more. We choose to evaluate this world with therapists instead of patients, because the therapists are specialists on the phobia and are also end users of the application. The therapist has



(a)



(b)

Figure 4 a and b: Screenshots of the virtual environment for agoraphobia, the Market place in the city of Delft. (a) Market place with variable numbers of walking humans.

(b) Market place during night with bad weather.

experience with the current framework and current worlds, but the new agoraphobia world which we evaluated, was completely new to her. We had a set of questions regarding the user interface and the virtual environment itself. It must be said that the user interface that was used for this prototype is not necessarily the same user interface that is going to be used for the new framework. The comments of the therapist can however be used for the design of the new user interface. We asked the therapist to point out what she thought of the different elements that we showed here. We

Module	--	-	++	+	++
<b>World controls</b>					
Vision control				•	
Heartrate & Breathing control			•		
Weather control				•	
Crowd control				•	
Navigation			•		
<b>World elements</b>					
Vision elements				•	
Heartrate & Breathing elements				•	
Weather elements					•
Crowd elements		•			

Table 1: Result of the evaluation by therapist. Legenda: ++ ”outstanding”, + ”sufficient”, ++ ”ok, but could use some adjustments”, - ”needs adjustment”, -- “insufficient”)

first demonstrated the element and subsequently recorded her comments and recommendations. In Table 1 we give an overview of the comments by the therapist, grouped by subject.

## CONCLUSIONS AND RECOMMENDATIONS

During this project we designed and evaluated anxiety-provoking virtual environments for the treatment of agoraphobia based on requirements from therapists. We implemented a “prototype virtual environment” for the treatment of agoraphobia, based on these requirements, which can serve as an impulse towards a new framework for VRET of phobias. Finally we gave an evaluation of the software and techniques that were used to implement the “prototype virtual environment” for the treatment of agoraphobia.

We have done interviews and evaluations with therapists. This resulted in a list of requirements and recommendations. We managed to implement most of the requirements, in order for the virtual environment to serve as a prototype for virtual environments that will be built within the future framework. The level of habituation in this prototype virtual environment could be improved. In the current virtual environment the level of habituation can be decreased by the therapist changing the look of the world, using the world controls.

In this study we also evaluated the software and techniques used to implement this virtual environment. Quest3D, which was used to implement the virtual environment, has many advantages, but also some disadvantages. If we are going to use Quest3D for the implementation of virtual

environments in the future framework, we have to take in to account the complexity that it brings. A simple virtual environment, with a few functionalities can quickly become a complex network of interconnected channels in folders and subfolders. Also the fact that Quest3D is not object oriented is a main drawback.

During our interviews and evaluation we collected a few future recommendations, which will be discussed in the following. For the agoraphobia world it would be a surplus value if the sounds or visuals of an ambulance and a hearse were implemented. According to therapist 1 this provokes a high level of anxiety for agoraphobia patients. It would also be a surplus value if a concentrated group of people was created somewhere on the square. The therapist can then navigate the patient to this heavily crowded spot. This would also save performance costs of the system, because we only crowd a small piece of the square instead of the whole square. The current system consists of a collection of different files on each computer. In order to make the system work these files have to be placed manually in directories and subdirectories on different hard drives. It would be an improvement if the files could reside in one directory or in the optimal case could automatically be placed in a proper destination using an installer. Finally, it would increase the usability of the system, if the user (therapist) could switch between environments, without first having to shut down the application and run a new one. In the optimal case the different virtual environments are interconnected. For example: One can move from a square to take the subway, that brings you to the airport, where a flight can be taken.

## REFERENCES

- Botella, C., Villa, H., Garcia-Palacios, A., Banos, R.M., Perpina, C., Alcaniz, M. (2004) Clinically Significant Virtual Environments for the treatment of Panic Disorder and Agoraphobia. *Cyberpsychology & Behaviour*, Oct 2004, Vol. 7, No. 5: pp. 527-535
- Emmelkamp, P.M.G., Krijn, M., Olafsson, R.P., Biemond, R. (2004) Virtual reality exposure therapy of anxiety disorders: A review. *Clinical Psychology Review* 24:259-281
- Garau, M., Slater, M., Vinayagamoorthy, V. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of

- Communication in a Shared Immersive Virtual Environment. 2003, ACM Press. Pp. 529-536.
- Gunawan, L.T., van der Mast, C.A.P.G., Krijn, M., Emmelkamp, P.M.G., Neerincx, M.A. (2004) Usability of therapist's user interface in virtual reality exposure therapy for fear of flying, in: Jeanne Schreurs & Rachel Moreau, Proceedings of the Euromedia 2004 Conference, April 19-21 2004, Hasselt, Belgium, pp. 125-132.
- Hooplot, F.S. (2005) Thesis Master programme Media and Knowledge Engineering, Delft University of Technology, Department Mediamatics, 2005.
- Mast, C. van der, Popovic, S., Lam, D., Castelnuovo, G., Kral, P. & Mihajlovic, Z. (2006). Technological challenges in the use of Virtual Reality Exposure Therapy, Proceedings of the NATO Advanced Research Workshop on Novel approaches to the diagnosis and treatment of posttraumatic stress disorder, Amsterdam IOS Press. (in press).
- Schuemie, M.J., P. van der Straaten, Krijn, M., van der Mast, C.A.P.G., (2001) Research on Presence in VR: a Survey, *Cyberpsychology and Behavior*, Vol.4, No.2, April 2001, pp.183-202
- Schuemie, M.J. & van der Mast, C.A.P.G. (2001) VR Testbed Configuration for Phobia Treatment Research, in: M.E. Domingo, J.C.G. Cebollada & C.P. Salvador (Eds.), *Proceedings of the Euromedia'2001 Conference*, April 18-20 2001, Valencia, Spain, pp.200-204.
- Witmer, B.G., & Singer, M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7: pp. 225-240

## BIOGRAPHY

**CHARLES VAN DER MAST** has a PHD from Mathematics and Computer Science from Delft University of Technology where he is employed at the Man-Machine Interaction group at the Department of Electronic Engineering, Mathematics and Computer Science as a Associate Professor. He teaches Multimodal Interfaces and Virtual Reality, Developing Highly Interactive Systems, Multimedia, and Usability Engineering, and developed the Bachelor/Master curriculum in Media & Knowledge Engineering. His interests include using various media to improve teaching, VR therapy for phobia treatment.

**FRANS HOOPLLOT** worked in this project for his master thesis of the programme Media and Knowledge Engineering of Delft University of Technology.

# **MEDIA APPLICATIONS IN BUSINESS**



# A WORKFLOW MODEL FOR SUPPORTING LEGISLATION CHANGES

Elias A. Hadziliadis  
IÉSEG School of Management  
Université Catholique de Lille  
3 rue de la Digue, 59000, Lille,  
France  
E-mail: e.hadziliadis@ieseg.fr

## KEYWORDS

Workflow model, E-Government, enterprise modelling.

## ABSTRACT

This paper proposes a workflow model that can be used as a decision support system for legislation changes. The primary objective of this model is to manage the hierarchical decision and information flow that is originated from the primary European Union (EU) legislation. The case study chosen for the demonstration of this model is the legal framework for public procurement in EU, which has caused the interference of European Commission in several countries' decisions. Furthermore, there were major political conflicts in Greece due to the non-conformity of the recently revised Greek Constitution with the corresponding European laws. Two enterprise modelling techniques were selected to carry out this research effort: the IDEF0 function modelling tool to construct the workflow model and the GRAI architecture which contains the legislation components. The e-Government dimension in this workflow model is not related to the use of Internet but lies in the application of Information Technology for administration purposes..

## INTRODUCTION

In order to construct a modern democratic society, all levels of governance must accelerate their functional transformation and implement administrative innovation, with the core objective to provide better public services to citizens. All governmental decisions depend on the information grasp: proper decisions must be based on abundant and accurate information.

*Ignorantia legis non excusat* – ignorance of the law does not excuse – is a centuries-old law maxim familiar to everyone: all are presumed to be familiar with all the laws that concern them or face the costs of their ignorance. Nowadays, European laws and regulations exist at all levels: at the town, province, region, nation level up to the EU level. Despite many promises of reduction, the efficiency of the legislative machinery is remarkable and corpora are constantly growing.

E-Government is the use of information and communication technologies in public administrations to improve public services and democratic processes and to strengthen support to public policies (EUROPA 2005). It is a

way for public administrations to become: more open and transparent, and to reinforce democratic participation, more service-oriented, providing personalised and inclusive services to each citizen, and productive, delivering maximum value for taxpayers' money.

E-Government has proven to be a durable and popular public management reform option over the last decade, attractive to elected officials and stakeholders who see its political benefits (Coursey and Killingsworth 2000). Public management scholarship on e-Government focuses on the beneficial effects of new technologies and examines the reasons why and how technology adoption occurs (Fountain 2001, Moon 2002, Ho 2002).

Research in the e-Government field has not directly questioned the basic premise that using Information Technology (IT) in public administration is a positive and inevitable route to improvement, progress and cost effectiveness. Certain e-Government proponents argue that declining rates of trust in government can be reversed through the use of technology, either indirectly because of greater citizen satisfaction with more convenient services or directly through enhancing civic participation in the public sphere. Fountain (2003) argues that IT can enhance democracy by making public information more accessible and by enabling a range of civic discourse that otherwise would not occur: from facilitating citizen-initiated contacts through the Web (Thomas and Streib 2003), to enabling a representative and meaningful discourse that replaces complicated administration processes (Shi and Scavo 2000). However, the potential of e-Government in this area has remained largely unfulfilled (West 2004).

This paper addresses the issues of legislation changes in national level that should take place because of the introduction of European laws. It has been often observed that the latter create conflicts during the harmonisation process of the legal frameworks. Furthermore, there is another phenomenon in the opposite direction: laws are introduced at a national level and afterwards they are challenged by the European Commission (EC) due to their content and nature. In the context of this research, we propose a workflow model based on IDEF0 method that identifies the steps for handling legislation changes and the GRAI architecture, which is comprised of the laws under consideration. The case studies that served as the motivation of this research concerning public procurement comprise the second section of the paper, which is followed by the

description of the proposed workflow model along with the presentation of the tools used. In the last section, the conclusions and further research are discussed, as well as issues on the advantages and disadvantages of the proposed workflow model.

## **MOTIVATION FOR THIS STUDY**

In this section, we describe the current situation in the public procurement domain at the Member States that should apply EU law according to the EC guidelines (EC Reports a, b, c, d 2005). EC decided in several cases against seven Member States either to refer those Member States to the European Court of Justice (the Court) or to formally request them to correct breaches of EU public procurement law. Requests of this kind take the form of “reasoned opinions”, the second stage of the infringement procedure under Article 226 of the EC Treaty.

### **Case study 1: Germany**

Several infringement cases were pursued against Germany over the award of service contracts and concessions without competition. In December 1999 the municipality of Hinte in Lower Saxony awarded a service concession to the Oldenburgisch-Ostfriesischer Wasserverband for the provision of waste water disposal services. No transparent award procedure was carried out as required under EU law as interpreted by the Court (C-324/98, Telaustria). Germany argued that the municipality of Hinte had not procured a service on the market but rather that the service had been transferred between public bodies, which, it maintained, are not covered by the EU rules on public procurement. This view was not accepted, since the Court has established that contracts concluded between public bodies are covered by the obligations of EU law. Thus, EU law was broken by the award of the service concession and the Commission therefore decided to refer the case to the Court.

### **Case study 2: Greece**

The Hellenic Public Power Corporation launched a call for tenders for the construction of a thermoelectric plant in Lavrio. The two companies that reached the last phase of the procedure (submission of financial bids) did not meet the conditions set out in the call for tenders, despite the fact that in the announcement of the call and the invitation to tender it was explicitly stated that any bid not meeting the specific requirements would be rejected.

### **Case study 3: Spain**

Spain was brought before the Court in connection with a case of incorrect implementation of Directive 89/665/EEC on the application of review procedures to the award of public supply and public works contracts. The EC considered that Spanish law is not in line with the Directive on the grounds that by allowing the award to coincide with the conclusion of the contract it denies unsuccessful tenderers the possibility of challenging, in good time, the

validity of the award decision and taking legal action against it at a stage when infringements can still be rectified. It was considered that making a declaration of invalidity subject to an exception for the protection of the public service could also render the provisions of Directive 89/665/EEC ineffective, since, under Spanish law, the scope is very broad, covering, in addition to cases of (automatic) absolute invalidity of decisions, the pure and simple cancellation of illegal decisions.

### **Case study 4: Italy**

Italy was brought before the Court for infringement of Community law on public procurement. The EC considered that this provision was contrary to the Community rules to the extent that it authorised contracting authorities in Italy to renew a public supply or service contract without any tendering procedure. A reasoned opinion was sent to Italy on the direct award, without prior competition at Community level, of the construction and operation of the motorway linking the Ospitaletto toll-area (A4), the new Poncarale toll-area (A21) and Montichiari Airport in Lombardy. This direct award constituted a infringement of Directive 93/37/EEC, which stipulates that contracting authorities wishing to conclude a public-works concession contract must announce their intention by means of a notice published in the Official Journal of the European Union.

### **Case study 5: Austria**

In 2001, the City of Villach concluded a waste disposal service contract for a minimum period of 15 years after selecting a service provider from a limited number of companies operating in Austria that already had an establishment in the Austrian State of Carinthia. The Austrian authorities claimed that the contract concerned a service concession and did therefore not fall under the scope of the specific rules on public service contracts set out in Directive 92/50/EEC. However, the contract was covered by Directive 92/50/EEC and should have been advertised in accordance with the rules applying to public service contracts. But even if it did qualify as a service concession, the selection procedure applied by the City of Villach would breach the general principles of the EC Treaty, and in particular the principle of non-discrimination on grounds of nationality. A reasoned opinion has therefore been sent to Austria.

### **Case study 6: Portugal**

Two cases of incorrect implementation by Portugal of Directives 93/38/EEC and 92/13/EEC were brought before the Court. The first of these Directives concerned the coordination of procurement procedures in the water, energy, transport and telecommunications sectors, while the second was aimed at ensuring effective application of the first by providing suppliers, entrepreneurs and service providers with effective remedies in the event infringement of Community law in that field. Portuguese law was not in conformity with Community legislation, particularly as



regarded its scope and application thresholds, competition and abnormally low bids.

### Case study 7: Finland

A reasoned opinion was issued against Finland concerning a decision by the Ministry of Finance to award a framework contract for air-travel services for government officials using discriminatory award criteria and thus infringing the public services Directive 92/50/EC. The Ministry of Finance had awarded the contract on the basis of non-published criteria, compared ticket prices that were not based on equal or similar terms, and included a destination among the routes to be served that was already reserved for a certain Finnish airline company thus making it impossible for others to tender for this route. The estimated value of the contract was € 30 million.

### EU Primary law versus Constitutional processes: Greece

The EC decided to ask Greece for its observations on the compatibility with EU law of Greek national law preventing companies “interconnected” with Greek mass media businesses from obtaining public contracts. This request took the form of a letter of formal notice, the first stage of the infringement procedure under Article 226 of the EC Treaty. EU public procurement markets are worth over € 1.500 billion, more than 16% of total EU GDP. The existing EU public procurement Directives have increased cross-border competition in procurement markets and reduced by around 30% the prices paid by public authorities for goods and services.

The conflict with the EU primary and secondary legislation lies on the Article 14(9) of the Greek Constitution and the implementing law 3310/2005, which declare a total and absolute incompatibility between any activity or shareholding above a certain level in mass-media companies and the performance of public contracts. It is considered that this is contrary to both secondary Community law (the Directives on public procurement), in that it lays down exclusion criteria that are not provided for in the Directives, and primary Community law (the EC Treaty), in that it lays down measures that impede, or render less attractive, the exercise of almost all the fundamental freedoms acknowledged by the EC Treaty.

### THE PROPOSED WORKFLOW MODEL

The workflow model under consideration attempts to address the unconformities that exist in the multiple legislation levels. The two complementary methodologies are firstly described, the IDEF0 and GRAI.

#### The IDEF0 modelling tool

The IDEF0 modelling method is designed to model the decisions, actions, and activities of an organisation or system. It is not only the most widely used, but also the most field proven function modelling method for analysing and

communicating the functional perspective of a system (Malhotra and Jayaraman 1992). IDEF0 was derived from a well-established graphical language, the Structured Analysis and Design Technique – SADT (Marca and McGowan 1988). The IDEF0 modelling method establishes the scope of analysis either for a particular functional analysis or for future analyses from another perspective.

The basic activity element of an IDEF0 model diagram is represented by a simple syntax. A verb-based label placed in a box describes each activity. Inputs are shown as arrows entering the left side of the activity box while the outputs are shown as exiting arrows on the right side of the box. Controls are displayed as arrows entering the top of the box and mechanisms are displayed as arrows entering from the bottom of the box. Inputs, Controls, Outputs and Mechanisms (ICOMs) are all referred to as concepts.

An IDEF0 model diagram is then composed of several activity boxes and related concepts to capture the overall activity. IDEF0 not only captures the individual activities, but also reveals the relationships among activities through the activities’ related concepts. For example, the output of one activity may in turn become the input, control or even a mechanism of another activity within the same model.

A strategy for organising the development of IDEF0 models is the notion of hierarchical decomposition of activities. A box in an IDEF0 model represents the boundaries of an activity. Inside that box is the breakdown of that activity into smaller activities, which together comprise the box at the higher level, as shown in Figure 1.

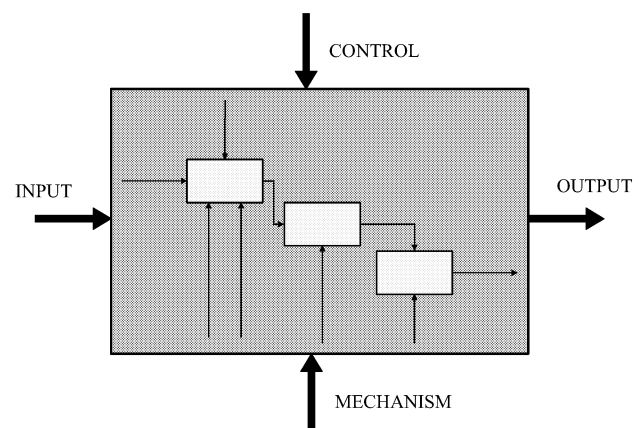


Figure 1: Hierarchical decomposition of an IDEF0 diagram

#### The GRAI methodology

A tool for handling the decisional and informational flow in a system is the GRAI Grid, which is a part of the GIM (GRAI Integrated Methodology) architecture (Girard and Doumeingts 2004). The GRAI Grid provides a general view of a selected process of the system, from a high-level perspective. The main objective of the GRAI Grid is to analyse the functional areas of the system and to identify the most important decision centres that are responsible for the decisions that are taken. The GRAI model proposes a framework in which any system can be described. It is based

on the system theory, the hierarchical theory and the activity theory. The system is divided into three parts: the operating system which converts input entities into output entities, the decision-making system which controls the above conversion and the information system which links up the operating system and the decision-making system and also supports the link with the environment.

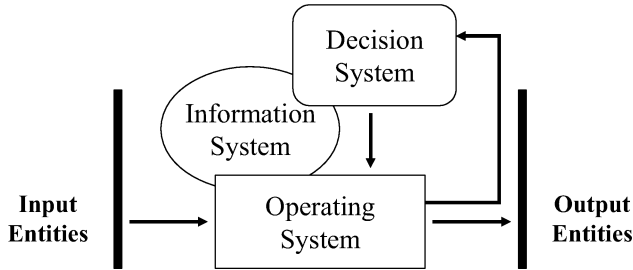


Figure 2: The conceptual model of a system

The GRAI architecture, was developed twenty years ago in order to design the manufacturing system and, more recently, the enterprise system. It is based on the GRAI model which acts as a conceptual reference model that describes basic enterprise concepts on a conceptual level. To give concrete expression to these concepts, the GRAI architecture uses a set of formalisms combined with graphic representation in order to facilitate use of the concepts by the modeller. Based on a structured approach, several modules are defined in order to provide an answer to various problems such as re-engineering, the choice and implementation of software, performance indicators and business plans.

### Combining IDEF0 and GRAI

The scenario on which we will build our model is based on the introduction of a new law at a national level. The workflow for checking the compliance of the new Member-State legislation with the existing EU legal framework is the following:

1. Construct the GRAI Grid with the three levels of legislation: EU primary legislation (treaties), EU secondary legislation (Regulations, Directives, Decisions) and Member-State legislation (Constitution, Laws). The columns are classified in the following categories (Agriculture, Biotechnology, Civil Society, Culture, Economic and monetary union, Education and Training, Employment and Social affairs, Energy, Enterprise, Environment, Food Safety, Technology and Innovation, security and justice, Information Society, Public Procurement, Regional policy, Research, Development, Space, Sport, Taxation, Trans-European networks, Transport, Youth, Budget, Fight against fraud, Grants).
2. Identify the decision and information flows among the decision centres of the GRAI Grid. The decision flow defines the top-down relation among the three levels of legislation and imposes the rules for the hierarchical organisation of the decision centres.

3. Classify the new law in the GRAI Grid. There is high possibility that the new law falls into more than one of the categories described in step 1. In this case, there is a detailed scanning of all the different decision centres, which are related to the lowest hierarchically level under consideration.
4. Recognise unconformities with existing legislation in the different categories of the GRAI Grid.
5. Rearrange the Member-State introduced law so that there is compliance with the primary and secondary EU legislation.
6. Test the revised law in the GRAI Grid to validate its conformity. In case minor non-conformities are traced, they are eliminated in this step.
7. Update the GRAI Grid at all levels with the new legislation corresponding to the appropriate decision centers.

Concerning the case of Greece where there is a conflict between EU Primary and the Constitutional processes, we propose the GRAI Grid that is shown in Figure 3.

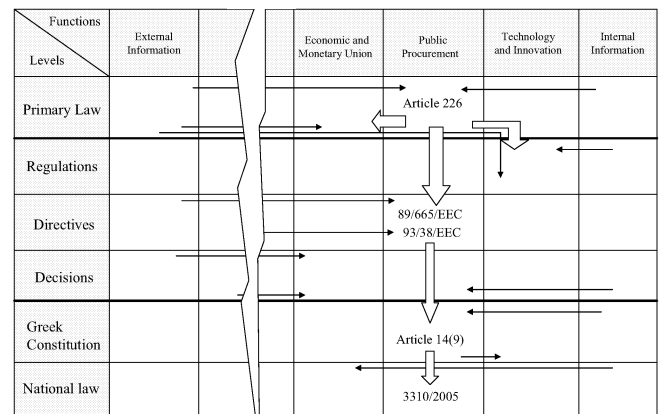


Figure 3: The GRAI Grid for the public procurement case

### IMPLEMENTATION OF THE MODEL

The proposed model was realised with the software GraiTools 1.0 (GRAISOFT 2005) to test the coherency among the decision centres and to demonstrate it as an e-Government application for the top level administration. This software is composed of four graphical integrated editors which allowed us to present the various models of our study: the IDEF0 editor was used to present the functional view (workflow) and the model of the physical system, the GRAI Grid editor described the global model of the decisional system, the GRAI Net editor was used to present a more detailed view of the decisional system, and the Entity/Relationship editor was used to model the informational view. The coherence links between the various models are managed by the software. GraiTools also proposes a dictionary in order to associate terms specific to the legislation level with more generic terms usually used in

the area of enterprise modelling. By this way, the legislation framework can be used within the models and the user is not constrained to use the generic terms.

## CONCLUSIONS

The combined IDEF0/GRAI workflow model examined the case where EU Member States are dealing with legislation change, when introducing a new law. It is an e-Government application destined to be used by the government officials and public servants who make decisions on various issues. As shown above, it is a usual phenomenon that EC acts on national legislation that is not according to EU legislation by bringing Member States to the Court and imposing fines in case of non-conformity.

The advantages of adopting this workflow model for managing law change lie on the premises of making the right decisions at first instance. This e-Government information system can save valuable resources from the national parliaments that vote for legislation, by checking the conflicts of the decision centres of GRAI Grid. Even though it is not web-based as the majority of e-Government applications, it serves one of the basic objectives of e-Government, which is to deliver maximum value for taxpayers' money.

The disadvantage of the proposed model is that only Member States with statute law, that is, written law enacted by a legislative body, are able to exploit the features of this workflow model. For EU countries with common law, which is a system of jurisprudence developing under the guidance of the courts so as to apply a consistent and reasonable rule to each litigated case, is not possible to implement this model.

Future research will be focused on the use of software agents (Artificial Intelligence) aiming at the faster processing of the steps of the proposed model. However, the integration of agents with the IDEF0/GRAI tools is not an easy task, since there are deterministic factors which constrain the behaviour of agents.

## REFERENCES

- Coursey, D. and J. Killingsworth. 2000. "Managing Government Web Services in Florida: Issues and Lessons". In *Handbook of Public Information Systems*, Garson (Ed.). Marcel Dekker, New York, 331–344.
- EC Report a. 2005. "European Commission: Public procurement - Commission acts on Greek legislation excluding certain companies from public contracts", <http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/05/361&format=HTML&aged=0&language=en&guiLanguage=en>, accessed 22.03.2005
- EC Report b. 2005. "European Commission: Public procurement - Commission acts to enforce EU law in Germany, Greece, Spain, Italy, Austria, Portugal and Finland", <http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/05/44&format=HTML&aged=0&language=en&guiLanguage=en>, accessed 22.03.2005
- EC Report c. 2005. "European Commission: Public procurement - Commission acts to ensure six Member States apply EU rules",

<http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/03/1763&format=HTML&aged=0&language=en&guiLanguage=en>, accessed 15.02.2005.

- EC Report d. 2005. "European Commission: Public procurement - Commission acts to enforce EU law in Italy, the Netherlands, Spain, Finland and Denmark", <http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/04/951&format=HTML&aged=0&language=en&guiLanguage=en>, accessed 15.02.2005.
- EUROPA. 2005. Europa Information Society portal, [http://europa.eu.int/information\\_society/soccul/egov/index\\_en.htm](http://europa.eu.int/information_society/soccul/egov/index_en.htm), accessed 15.02.2005.
- Fountain, J.E. 2001. *Building the Virtual State: Information Technology and Institutional Change*, Brookings Institution, Washington.
- Fountain, J.E. 2003. "Electronic Government and Electronic Civics". In *Encyclopedia of Community*, Wellman (Ed.). Great Barrington, Berkshire, 436–441.
- Girard, P. and G. Doumeingts. 2004. "Modelling the engineering design system to improve performance". *Computers & Industrial Engineering*, 46, 43–67.
- GRAISOFT. 2005. <http://www.graisoft.com/>, accessed 15.02.2005.
- Ho, A.T-K. 2002. "Reinventing Local Governments and the E-Government Initiative". *Public Administration Review*, 62, No.4, 434–445.
- Malhotra, R. and S. Jayaraman. 1992. "An Integrated Framework for Enterprise Modeling". *Journal of Manufacturing Systems*, 11, No.6, 426–441.
- Marca, D.A. and C.L. McGowan. 1988. *SADT: Structured Analysis and Design Technique*, McGraw-Hill, New York.
- Moon, M.J. 2002. "The Evolution of E-Government among Municipalities: Reality or Rhetoric?". *Public Administration Review*, 62, No.4, 424–433.
- Shi, Y. and C. Scavo. 2000. "Citizen Participation and Direct Democracy through Computer Networking". In *Handbook of Public Information Systems*, Garson (Ed.). Marcel Dekker, New York, 331–344.
- Thomas, J.C. and G. Streib. 2003. "The New Face of Government: Citizen-Initiated Contacts in the Era of E-Government". *Journal of Public Administration Research and Theory*, 13, No.1, 83–101.
- West, D.M. 2004. "E-Government and the Transformation of Service Delivery and Citizen Attitudes". *Public Administration Review*, 64, No.1, 15–27.

## AUTHOR BIOGRAPHY

**ELIAS HADZILIAS** is a Senior Assistant Professor at the IÉSEG School of Management of the Catholic University of Lille. After obtaining his Diploma in Mechanical Engineering (1997) and his Doctorate of Engineering (2003) from the National Technical University of Athens in Greece, he assumed his current position in France. He is also the Academic Director of the Master of International Business since 2005 at the IÉSEG School of Management. His research interests are enterprise modelling, virtual enterprises, strategic management, management information systems and e-government. In parallel to his academic activities, he has been involved in consulting projects with private enterprises and public organizations.

# Enhancing Knowledge Management with Business Intelligence – A Case Study

Dipl.-Inf. Kay Grebenstein  
Prof. Dr.-Ing. Stephan Kassel  
Westfälische Hochschule Zwickau (FH) – University of Applied Sciences  
P.O. Box 201037  
D - 08012 Zwickau  
E-mail: Stephan.Kassel@fh-zwickau.de

## KEYWORDS

e-Commerce, Public Market Places, Business Intelligence, Market Intelligence, Knowledge Management, Expert Systems.

## ABSTRACT

Knowledge management provides companies with a means to support their business goals by providing an explicit representation of the knowledge of the staffers. This knowledge can then be used by co-workers or by expert systems to solve future problems.

A great disadvantage lies in the large expenses and difficulties to make implicit knowledge accessible and distributable. Furthermore, great parts of the knowledge and the authority of the employees are purely based on empirical experiences, therefore leading to uncertainties whether this experience-based knowledge is correct.

In the project RESOV, the company AGETO and the University of Applied Sciences Zwickau have represented the experiences and the knowledge for the sale of products on public market places in an expert system, in order to provide forecasts for the best placement of a commodity.

Additionally, the available sales data and possible influencing factors on customer behaviour were collected with Business Intelligence techniques. Based on these data, classification numbers are currently developed to provide retailers with easily useable information, thus enhancing the quality of the forecasts.

## INTRODUCTION

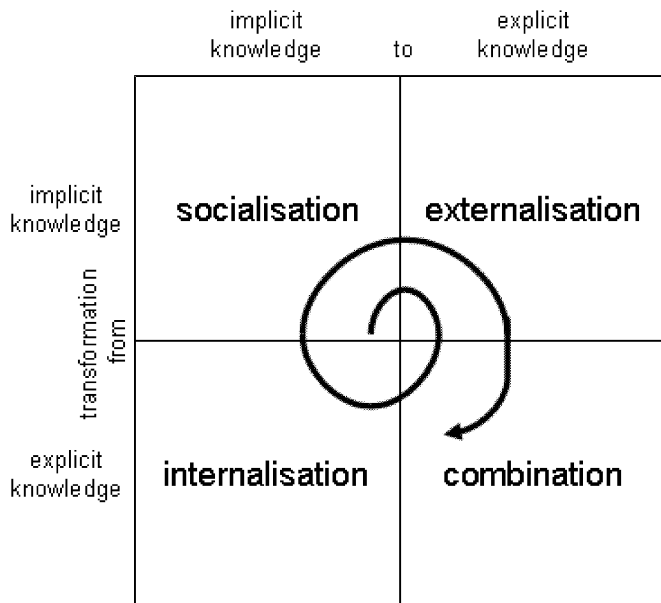
Decisions in enterprises are based on internal and external information. The project RESOV is settled in the domain of e-commerce and uses Knowledge Management to handle the decisions of the best placement of an auction on virtual marketplaces like eBay.

The change to information society leads to increasing challenges for enterprises. On the one hand, there is an ever-growing amount of information which could be the base for successful decisions. On the other hand, there is a lack of time to analyze the information due to hustling decisions.

Enterprises acting on a global market become more and more dependant on specialized knowledge of their employees. In the 90's the necessity of "Knowledge Management" was perceived.

Major tasks of Knowledge Management consist of the extraction of knowledge from different sources, the transformation and storage of generated knowledge and the supply of knowledge according to requirements. (Kuppinger and Woywode 2000)

As the main knowledge carriers are the individual employees, Knowledge Management leads to a fundamental change of the corporate and management philosophy. Therefore, technologies are used for saving knowledge and making it available to the organisation. But to be successful, it is even more important to motivate each employee to share his knowledge with his co-workers to integrate it into the work process. To fully understand knowledge sharing, you need to analyze the sharing process. As seen in figure 1, knowledge can be shared by socializing in the organisation. If you are part of the organization, you will learn what the other members of the organization already know. But this process is very slow and inefficient. In a time of quick organizational changes, the structures are changing too rapidly. It is more efficient to transform the knowledge into explicit knowledge, which can be shared more easily. This step is also leading to discussions with other experts doing their work (slightly) different. So the explication of knowledge leads to discussions about the knowledge artefacts and therefore to organizational learning. And this is rather fruitful to really make mistakes only once. The key factor of this learning process is the explication of the implicit knowledge of the employees. To encourage the transformation, the employees revealing their knowledge should be honoured and their position in the enterprise should be strengthened. Often, the employees are anxious, that their knowledge could be used by colleagues and they will be discriminated with their career. Above all, within global enterprises, cultural differences should be paid attention to, as they are of high importance and thus a great challenge (Grebenstein et al. 2003). In this way the knowledge of the employees become the human capital of the enterprise.



Figures 1: Knowledge flow model (Nonaka and Takeuchi)

The company AGETO GmbH developed a multi-channel sales system for e-commerce, the AGETO eBay Web service, which facilitates a simple and fast access to online marketplaces without in-depth knowledge of e-commerce processes. The solution shall utilize the potential of online marketplaces to gain a measurable economic success for the enterprises.

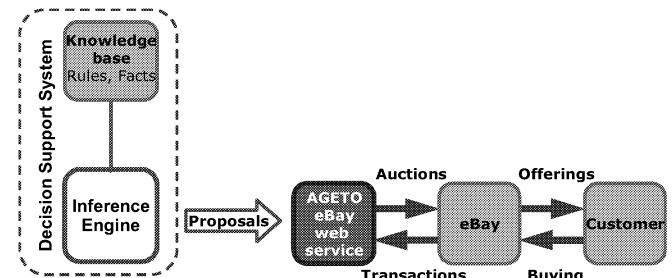


Figures 2: Architecture of AGETO eBay WebService

The AGETO eBay Web service establishes channels to online marketplaces to enrich existing e-commerce systems of retailers and manufacturers, or provide completely new business opportunities for enterprises which are not yet using such systems. This Web service permits a simple integration into existing corporate processes and legacy systems and can be adapted easily to individual needs. The product descriptions, which are part of the product representation in the legacy systems, are extracted automatically. The user of the AGETO eBay Web service can choose for which products offerings on the electronic markets should be created automatically. He can also choose between different kinds of offerings, depending on the market places. After ordering the product (either by direct selling or after an

auction), the orders are routed to the legacy system of the user for further processing (payment process and delivery of products).

In the initial system, auctions are placed automatically at the public market by setting the parameters placement time, total quantity and minimal distance of two auctions. There is no warranty of success for the placed auctions because the placement doesn't necessarily fits to the search and buying patterns of the buyers on the market place.



Figures 3: Architecture of AGETODSS (phase I)

One result of the project RESOV is the AGETODSS, a knowledge-based decision-support-system for e-commerce (see figure 3). In the future, the AGETODSS will optimise the placement times and thereby gain sales advantages. This is achieved by the inclusion of knowledge, like the general sales behaviour of the customers or special influences of weather and holidays, in a decision support system. Beyond that, the sellers can determine whether the auctions should optimise achieved prices or sold quantities.

## PROBLEM DEFINITION

In the project RESOV, the practical knowledge of power sellers is adopted into the knowledge base of the expert system, which is based on the public domain shell Mandarax. The supplied rules and facts enable forecasts for an optimal placement time on public market places like eBay. In the first phase of the project, this knowledge was represented by explicitly writing rules and facts reflecting the knowledge of the power sellers. One example for this knowledge is the customer behaviour on eBay, which periodically changes over the week. So the following rule was entered into the system:

"If the product you are selling is of category Antiques the end time of the auction should be on weekends."

But the knowledge of power sellers is usually based on empirical observations and experiences. Verification or quantification of the predication hardly can be done by the experts themselves.

In addition, electronic commerce channels are not as trend steady as the normal retail trade. They are characterised by a higher frequency of changing customer behaviour. This is

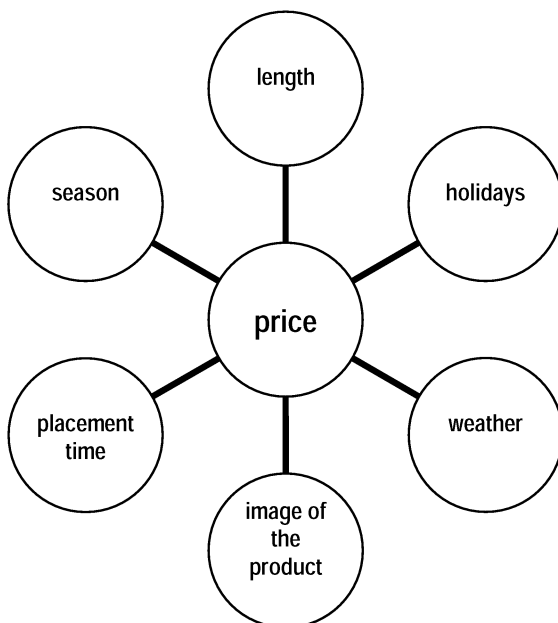
caused by a high number of buyers and sellers on the electronic markets and by a greater transparency. Products are more comparable than on classical markets. Besides, the special experience of a power seller for computer articles cannot be simply transferred to the sales of wines.

The sellers can be comfortably supported by decision-support systems. But if these systems use outdated knowledge, suggestions for an optimal placement of products are not helpful at all. It is necessary to update the knowledge base frequently, based on current trends (which can be product specific).

Fortunately, knowledge stored in an expert system, can be rather easily changed and enhanced in comparison to the implicit knowledge of the salespersons. After the extraction of the implicit knowledge of the human capital, the explicitly formalized knowledge can be easily verified and adapted to new situations as soon as the changes become remarked.

## SOLUTION

The knowledge base of the expert system was extended by adding additional empirical information, and quantifying the impact of influences - such as weather, determined holidays or events - on sales numbers (see figure 4).



Figures 4: Influences on the development of prices

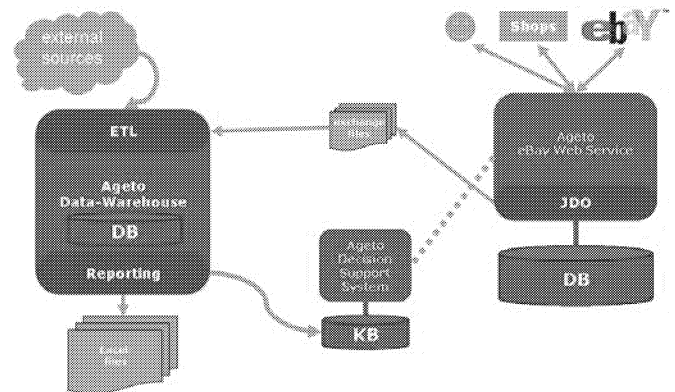
This should be briefly explained for the influence of weather to the price. It is quite obvious that there are a lot of products whose attractiveness depends on weather conditions. As example use ski suits, which are rather unattractive when there are no winter sports conditions. So the prices for ski suits are directly dependant on the weather. (Similar

observations could be derived for sun glasses or fans.) Normally these products are known as seasonal products, but in reality they are often bought at the moment when it is quite obvious that they soon will be used.

If this observation can be included into the rules for the knowledge base, higher prices or higher sales numbers could be achieved easily than without this knowledge.

To include this behaviour in the knowledge base, data from the weather forecast were collected in the system and a correlation analysis was made on the weather conditions and the sales numbers.

A database was built using techniques and strategies of Business Intelligence, to derive information on the different influences on customer behaviour at public market places. The data originate from the transaction and auction data bases of eBay and external sources providing information on the external influencing factors.



Figures 5: System architecture of the project “RESOV”

The task of the last phase consisted of a combination of the results: Classification numbers were generated from the stored information of the Data Warehouse by visualization of the connections with reports and statistic procedures. These classification numbers are dependant on products as well as on markets.

## IMPLEMENTATION

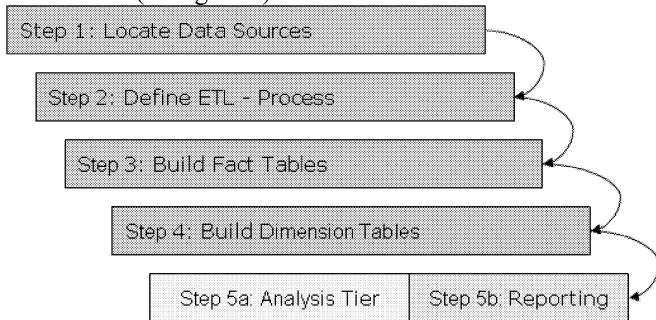
The concepts of the Data Warehouse, a central data collection, were developed as a solution to the problem of inconsistent and redundant data sets. At the beginning of the 90's, the term Business Intelligence was defined to denominate software tools extracting enterprise-internal and external data on one hand and transforming them into a database model and finally producing decision-relevant information on the other hand. (Gadatsch 2002)

Improvements in the process of extraction, transformation and loading the data into the Data Warehouse (ETL process) and, even more recently Enterprise Application Integration tools, have accelerated data collection. OLAP technologies are enabling a faster analysis of the data. Business

intelligence has emerged as the art of sieving through large amounts of data, extracting information and turning that information into applicable knowledge.

In the project, a data warehouse was built using a PostgreSQL database. It served as a starting base for the application of analysis tools. PostgreSQL was chosen because of its offered functionality of stored procedures (written in an own programming language called PL/pg-SQL), and low license costs.

An approach with five steps for building the data warehouse was chosen (cf. figure 6).

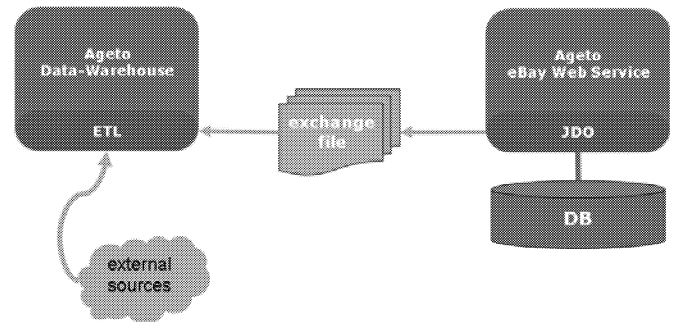


Figures 6: Procedure Model of Phase II

The first step comprises the identification of the necessary data sources and the analysis of the source data concerning quality and quantity. One result of this analysis was the special format of the external data from the weather service, because the meteorological categories and technical classification of the weather phenomena that are used have to be adjusted to the project data base. The weather service provides weather forecasts for a variety of local towns and regions. These data cannot be included in their entirety, but a transformation rule has to be applied to the weather data to get some greater categories of weather which are suitable for greater regions. Goal of this transformation is the derivation of something like a general weather situation for a country. The problems in this phase are similar to the problems of classification and we had to use some ideas from fuzzy theory to achieve useful results for the next steps.

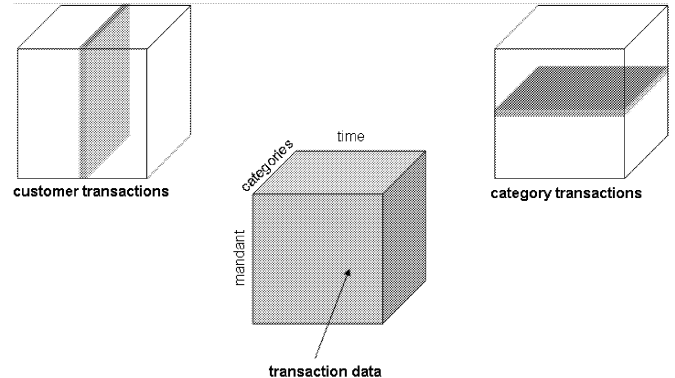
This already leads to the second step, where the ETL process is performed. This process is transferring the data from the operational data bases and external sources into the Data Warehouse (see figure 7). In the ETL process the fact tables are filled. To provide an optimal format for a later analysis, the data have to be converted by de-normalizing and by compressing the data base in an appropriate way. The fact tables are the central data base of the concept of a central Data Warehouse. As an example for this step, all the different bids on one auction on eBay are collected and condensed to one entry in the fact table. With this step, the details of the bids are lost. So it is very important to collect different information in different facts of the database. So one bid is not only used as a counter for the number of bids on a special product, but it is used as well for a statistic of

bidding times on a product category. This can be used for customer analysis purposes.



Figures 7: Data sources and ETL-interfaces

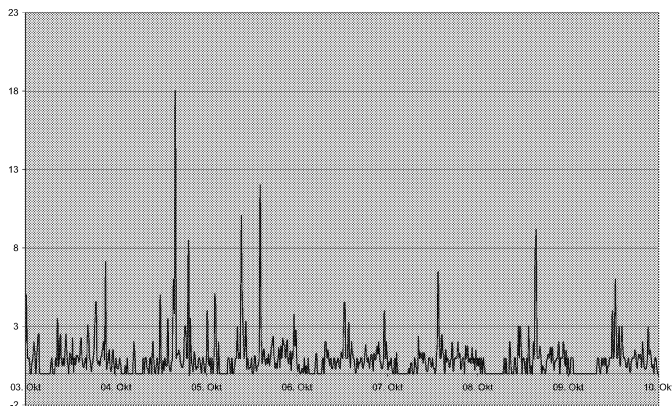
PostgreSQL is a relational database system. So it is necessary to provide beside the fact table so-called dimensional tables. They have to be built to accelerate a later production of reports and to perform on-line analysis. The dimensional tables are providing the effective multi-dimensional access to the data (cf. figure 8). So they are used to provide slices along interesting dimensions for analyses. For example, monthly reports on different products are very common, so it is useful to provide the dimension of the month in a dimensional table. This can be used to partition the data.



Figures 8: Multi-dimensional access in the Data-Warehouse

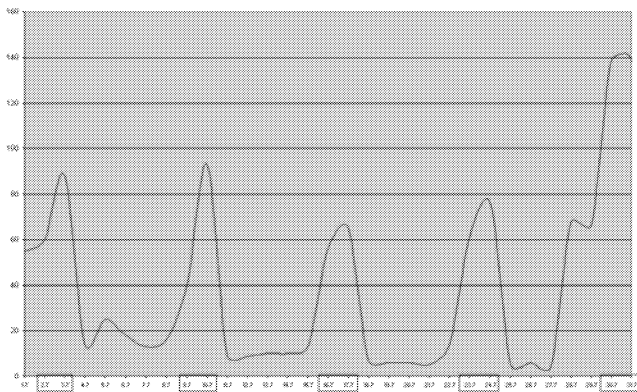
The reports produced by the summarized data of the Data Warehouse provide an accurate visualization and analysis of sales figures or attendance of auctions. The transactions and auctions are summarised by category (cf. figure 9). They can be enhanced with a representation of possible influence factors on the success of the auction, like weather or special events. If the reports are showing evidence of associations between influencing factors and sales numbers, a hypothesis could be formulated and further statistical methods like regression analysis could be used to verify the hypothesis. If such correlations between external factors and sales numbers are found, new rules could be introduced into the expert system reflecting those connections.





Figures 9: Average hits of the auctions of the category “collecting & rare”

The experience-based rules for best placement times on public market places depending on specific times or weather conditions could be confirmed and even quantified with statistic measures. These dependencies were verified for example for the correlation between the period weekend and sales numbers for wine.



Figures 10: Bids-quotient of the category “wines”

The weekly distribution of a quotient of the number of bids for the auctions and the actual number of auctions for the eBay category wine is represented in figures 10. Even without statistic analysis it is possible to diagnose a higher rise of bids at the weekends (marked days). The significance was statistically proven by evaluation with the t-test.

The characteristic numbers won by statistic procedure are stored in a specific table in the Data Warehouse. A table is computed in the Data Warehouse to serve as an equivalent to the facts of the knowledge base. It consists of the time (represented as fixed-length time slices) together with the possible sales margins, and is re-computed frequently to provide actual numbers. Additionally, tables are available quantifying the influence of the weather on sales numbers by characteristic numbers. In a following step the characteristic numbers are transferred into the knowledge base of the expert system.

## SUMMARY

Software solutions like AGETO DSS enable a simple and fast access to decision knowledge for special applications like online marketplaces by using technologies and procedures of knowledge management and artificial intelligence.

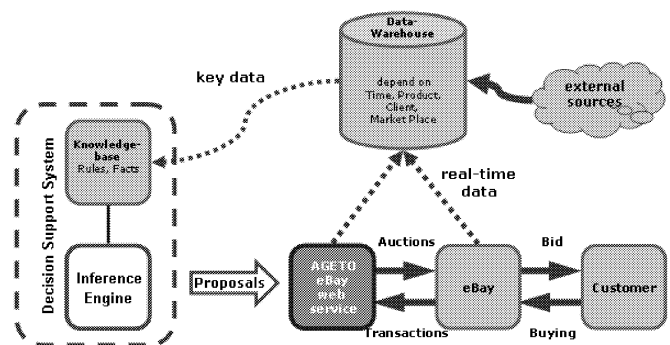
On the other hand Business and Market Intelligence facilitate a current view of the condition of the enterprise and the markets. The operational and strategic management are supplied with decision supporting facts by techniques like OLAP or reporting. But the evaluation of the supplied numbers and diagrams can only be done with domain competence and is rather time-consuming.

Nowadays, small and medium-sized enterprises also have a need for decision support systems reacting fast and automatically to constantly changing conditions of the markets they are acting upon.

By the combination of the measures of knowledge management and artificial intelligence, a system for forecasts of “optimal” sales times on public virtual marketplaces like eBay could be developed. This is performed with explicit knowledge of power-sellers.

Via the procedures of Business Intelligence current sales numbers and influence factors on the behaviour of customers can be collected and provided in a database.

In the end of the project “RESOV”, the results of Knowledge Management / Artificial Intelligence and Business Intelligence / Data Warehousing were combined in a feedback process (see figure 11). Characteristic numbers were computed from the information stored in the Data Warehouse, thus describing auctions of specific products on different market places.



Figures 11: Final architecture of project RESOV

These characteristic numbers were integrated into the explicit knowledge of the expert system to enable more exact and trend-dependent forecasts. The static rules of the expert system have been changed to dynamic rules through the integration of the characteristic numbers. Then the expert system could include actual changes of the behaviour of the



customers on the market into its decision process. This makes the expert system reactive to new trends on the market and changing preferences of customers.

## REFERENCES

- Abramowicz, W. 2003. "Knowledge-Based Information Retrieval and Filtering from the Web". Kluwer Academic Publishers.
- Aebi, R. 2000. "Kundenorientiertes Knowledge Management". Addison Wesley Verlag.
- Deutscher Managerverband e. V. 2003. Die Zukunft des Managements. vdf Hochschulverlag AG. Zürich
- Dietrich, J. 2004. "A Rule-Based System for eCommerce Applications". In *Proceedings of Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference* (Wellington, New Zealand, Sep. 20-25). Springer LNCS 3213 / 2004, Heidelberg, 455-463.
- Gadatsch, A. 2002. "Management von Geschäftsprozessen". Vieweg Verlag, Braunschweig/Wiesbaden.
- Grebenstein, K.; Schumann, C.A.; Tittmann, C.; Tsering, G.; Weber, J.; and Wolle, J. 2003. "Globale IT-Infrastruktur für die interkulturelle Kommunikation". In *Kommunikation in der globalen Wirtschaft 2003*, Bleich, S.; Jia; and W.; Schneider, F. (Eds.) Peter Lang Verlag, Frankfurt a.M., 135-153.
- Hannig, U. 2002. "Knowledge Management und Business Intelligence". Springer Verlag. Berlin – Heidelberg – New York.
- Horton, F. W. and Marchand, D. A. 1982 "Information Management in Public Administration". Information Resources Press. Arlington.
- Kassel, S. and Grebenstein, K. 2004 "AI-based integration of business intelligence and knowledge management in enterprises", *Proceedings Conference AIAI 2004*. Toulouse
- Kassel, S., Schumann, C.-A., Grebenstein, K., Tittmann, C. 2005 "A Knowledge-Based Decision-Support-System for e-Commerce", *Proceedings Conference Euromedia 2005*. Toulouse
- Kuppinger, M. and Woywode, M. 2000. "Vom Internet zum Knowledge Management Veränderungen in der Informationsgesellschaft". Carl Hanser Verlag München Wien.
- Lehner, W. 2003. "Datenbanktechnologie für Data-Warehouse-Systeme". dpunkt verlag GmbH. Heidelberg.
- Ortmann, G.; Sydow, J. 2001. "Strategie und Struktur". Gabler Verlag. Wiesbaden.
- Puppe, F.; Gappa, U.; Poeck K. 1996. "Wissensbasierte Diagnose- und Informationssysteme". Springer Verlag. Berlin – Heidelberg – New York.
- Sachs L. 1992. "Angewandte Statistik". Springer Verlag. Berlin – Heidelberg – New York.
- Salcedo L. Market Forecast Report European Commerce, 2003–2009. 2004. JupiterResearch. Jupitermedia Corp.

Sol, H. 2002. "Expert Systems and Artificial Intelligence in Decision Support Systems". Kluwer Academic Publishers.

Vitt, E.; Luckevich, M.; Misner, S. 2002, "Business Intelligence: Making Better Decisions Faster", Microsoft Press, Redmond.

## AUTHOR BIOGRAPHY

KAY GREBENSTEIN was born in Plauen, Germany. He studied computer science at the University of Applied Sciences Zwickau. After obtaining his degrees, he worked at several projects in the areas of Knowledge Management, Distance Open Learning, Artificial Intelligence and Information Systems.

STEPHAN KASSEL was born in Bad Kreuznach, Germany. He studied computer science at the University of Kaiserslautern. After obtaining his degrees, he worked at several universities in Germany in the areas of Distributed Artificial Intelligence and Information Systems. He earned his Ph. D. at the Technical University of Chemnitz, in 1998. In 1999 he started to work for Intershop, an E-Commerce vendor, in international e-commerce projects. Since 2003, he holds a chair for Information Systems at the University of Applied Sciences in Zwickau, Germany.

# A Formal Method for Modeling and Evaluation of Protocols of Electronic Documents Transfer and their Security on the Web

Gilles Eberhardt, Ahmed Nait-Sidi-Moh, Maxime Wack  
Laboratoire Systèmes et Transports  
Université de Technologie de Belfort-Montbéliard  
90000 BELFORT Cedex  
[ahmed.nait-sidi-moh@utbm.fr](mailto:ahmed.nait-sidi-moh@utbm.fr)

## KEYWORDS

Electronic signature, Security, Electronic Data Interchange, Discrete Event Systems, Information systems, Petri Nets,  $(\max, +)$  algebra.

## ABSTRACT

Digital signature and its associated protocols are a new area of interest and many standards have emerged. Indeed, these technologies offer several advantages: identification, authentication and non-repudiation capabilities during Internet transactions. This article deals with the modelling and the evaluation of document transfer and their security on the web. The study discusses the protocol of electronic signature and its associated processes. For this study, we use some tools and formal methods that have already proved their efficiency in the framework of modelling and analysis. We use, Petri Nets (PN) and the theory of linear systems in  $(\max, +)$  algebra. We introduce these two formalisms with the aim to describe the graphical and analytical behaviours of studied process. The resolution of  $(\max, +)$  model that describes the system enables us to evaluate the process performances in terms of occurrence dates of various events that compose it (authentication, hash coding, signature, time stamping, storage).

## INTRODUCTION

The security of the information systems presents a paramount task because of a strong growth of the technology. Actually, the treatment technique of the information occupies a dominating place in the data-processing applications. With the growth of the electronic interchanges, the security becomes increasingly crucial (Kae, 1999). To improve the quality of service (QoS) of the electronic interchanges, it is necessary to have the means of analysis and the appropriate methods. This QoS must be validated with specification of models that repose on formal properties like robustness, reliability, etc.

The information systems constitute a particular class of dynamic discrete event systems (DDES) whose dynamic is governed by various phenomena as synchronisation and parallelism (Bac et al, 1992). These systems result essentially from the human design, according to a certain point of view, their activities are due to asynchronous occurrences of discrete events. It is thus possible to model them via various adapted and dedicated tools.

The objective of this paper, which enters within the framework of the modelling and evaluation of electronic documents transfer and electronic signature, is to study processes evolving in a space of discrete states. Our contribution concerns, more precisely, the adaptation of the concepts and the theoretical results which formalized graphically by PN and mathematically by  $(\max, +)$  algebra (Baccelli *et al*, 1992), (Nait, 2003). This is in order to model and analyse the electronic data interchange between various confidence actors in the process of electronic delivery of signature (Cott et al, 03). Our study is thus concretized by the development of formal models able to bring solutions to the problems of safety and improvement of service quality of electronic signature. These models facilitate the structural and behavioural analysis of signature process. Before proceeding to the modelling and evaluation of this process, we give some elements and definitions about the electronic signature and its associated processes.

## DIGITAL SIGNATURE AND MESSAGE SIGNING BASICS

### Digital Signature

A digital signature makes it possible at the reception of a message to ensure its origin. It represents the coding of the digest of the message by the private key of the author. It is the result of many researches on asymmetrical cryptography and hash coding (NIST, 1995, 1999, 2000).

When a sender entity (a person, a server, etc.) needs to securely send a message to a receiver entity, it encrypts the message using the receiver's public key (RSA, 1993). This key is published so that any sender can make use of the receiver's public key to encrypt data. The encrypted message is then unintelligible and can not be decrypted without the corresponding private key. This key must be securely stored by the receiver that does not publish it. Only the receiver would be able to decrypt the encrypted message. Asymmetric key cryptography achieves privacy and confidentiality. The most widely used asymmetric key cryptography algorithms are RSA (R.S.A, 1993), triple-DES (NIST, 1999).

One in the manners for guaranteeing the integrity is to provide with the secrecy a concise digest of the sent message. At the reception, a concise digest of the same type is extracted from the message, and it is compared with the one received. The

generation method of this digest is called "hash coding function".

Digital signature generation is the simple application of asymmetric key cryptography over streams hash codes. Unlike data encryption, digital signature's purpose does not consist of data confidentiality but rather in providing (Kaeo, 1999):

- *Data integrity*: digital signatures enable to detect data modification sources, i.e. unauthorized data modification;
- *Authentication*: as the signature key is (theoretically) owned by the signer only, it is impossible for anyone else to generate the sender's signature on a given data stream. The data are authenticated by comparing the signature with the signer's corresponding verification key;
- *Non-repudiation*: this service based on authentication is a proof of transaction effectiveness. The signature entity can not deny being the signature author because nobody could create such a signature on a data stream.

Digital signature is generally computed on hash codes rather than directly on the data. Although digital signature makes it possible to authenticate the received data, it does not make it possible to identify the data origin (the signer). Thus, any irrefutable link exists between the signer and its signature key. Such identification is provided by electronic certificates. The following section gives some elements about the electronic certificates.

### Electronic Certificates

A qualified electronic certificate is the identification link between the public key and the identity of his owner. It is a normalized document, duly filled out, joins to the public key of the certificate holder, and signed by the private key of the certification authority (CA). That makes it possible to ensure the membership of the public key used for the coding of the message, and vice versa: the membership of the private key used to affix the signature. The certification is ensured by a Certification Authority who is accredited by governments or their representatives. This authority delivers certifications by affixing its signature with its private key. Certificates can be then used for the identification. The structure of a certificate is normalized by the X.509 standard of the ITU (International Telecommunication Union). The contained informations in the certificate are:

- The name of the certification authority with its signature;
- The name of the certificate owner;
- The validity date of the certificate;
- The encoding algorithm used (MD5 or SHA-1);
- The public key of the owner.

Certificates are valid until they are revoked or until their expiry (Figure 1.). In both case, a new certificate can be re-emitted by the CA. A certificate revocation intervenes when its owner is informed that its certificate is corrupted, or that an unauthorized entity could use it. It is also possible for a government or the CA to revoke a certificate in the case where its owner made a fraudulent use of it.

So the CA proposes a service of revocation which results in the publication of a List to Revoked Certificates (LRC) (Housley, 2002) consultable on-line or locally after download.

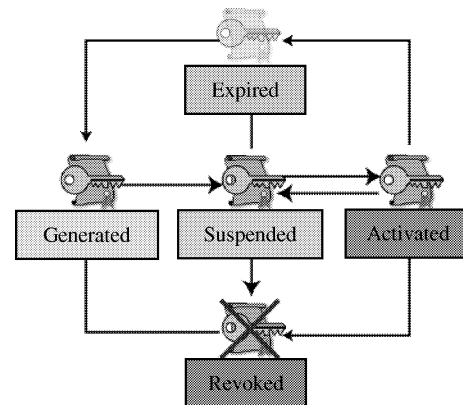


Figure 1. Certificate life cycle.

### Confidence actors

#### PKI (Public Key Infrastructure)

A PKI, also called IKM (Infrastructure of Keys Management) is a whole of physical components (computers, smart cards), human procedures (checks, validation) and software (system) who participate to create and to manage the life cycle of the numerical certificates. The creation of these certificates is relatively simple; what makes the complexity of a PKI, is the set of the mechanisms of internal audit for the various operations of creation, renewal, revocation of the certificates. These precautions are related to the fact that the numerical certificates can have a legal value by allowing an electronic signature which is equivalent to a handwritten signature.

A PKI provides four functionalities:

- Generation of bi-keys;
- Certification of public key;
- Certificates Revocation;
- Management of the function of certification.

A PKI is constituted of three authorities:

- The Local Recording Authority (LRA): entity indicated by a CA to help other entities in certificates requests, revocation (or suspension if the entity is entitled), or both, like in the approval of the requests. The LRA is not the representative of the certificate applicant. An LRA cannot delegate its power of approval of the request for certificate.
- The Principal Recording Authority (PRA): entity entitled to register other entities and to assign them a relative distinctive value such as a distinctive name or a condensed value of a certificate. The inscription procedure of each domain warrants that each registered value is unambiguous inside the domain.
- The Certification Authority (CA): confidence organisation which emits suspends or revokes a certificate. CAs are identified by a distinctive name on all certificates and Certificate Revocation List (LCR) which they emit.

#### Signer

It is the principal entity in the signature process. It is the person receiving the benefits to affix its signature on a document. The signature process starts while affixing the signature by the signer.

#### *Time stamping actor*

The certification of the date, even of the precise hour, of a document under electronic form can be carried out by a person receiving benefits of time-stamping. This person presents the guarantees of independence and necessary competences. Such services lie on reliable systems founded on the asymmetrical cryptography. The time-stamping permits to prove that a person signed a document or a message at a defined date.

#### *Storage actor*

Electronic storage enables to preserve electronic writings in order to use them as proof later, or with aim of constituting an informational patrimony of the company. This function, taking into account its complexity, is often entrusted to the storage actor within the framework of a specific contract. The stored Data, protected against major risks, are restored in the technical state of origin where those were received.

Obligations of the Storage actor are:

- Conformity to Z 42-013 standard;
- Confidentiality: users identified by contract and encoding;
- Not use of data or programs on a purely personal or individual basis;
- Restitution of the stored files in the state where they have been entrusted by the client;
- Transmission of the stored files of any other Storage actor or the client at the end of the contract;
- Destruction of the documents at the end of the contract or on request of the client.

#### **The signature process**

The signer needs a numerical certificate; he makes a certificate request near a Recording Authority with a strong authentication (request for rent receipts or phone invoice). During the request, two keys are generated: a private key stored on a smart card which will be sent by the CA to the signer and a public key which is sent to the Authority of Certification. The Recording Authority records the request and checks the identity of the person. The PRA sends its green light to the CA as well as necessary documents. The CA generates the certificate and publishes it in its public repertories (LDAP: Lightweight Directory Access Protocol) then sends the smart card containing the certificate and the private key to the PRA or LRA. Also it sends the password as well as convocation to the signer to recover his smart card at the PRA or LRA.

To sign a document, the signer must have a locked signature key and stored on a personal and confidential support (CD-card, smart card, token, key USB). The security is in general ensured by a secret code, but other means can be used according to the desired level of security (retinal identification, digital, thermal).

The document to be signed is « hashed » in order to obtain a single digest (NIST, 1995). These digest “will be signed” by the private key of the user. One thus obtains a signature which is referred to a single document. This signature is then subjected to one or more time-stamping actors which affix their signatures (dates) and preserve a trace of the transaction as well as the digest of the document.

When the time-stamping actor is not online, the time-stamping operation cannot take place. Nothing prevents the user from affixing his signature, but this one will be valid only at the time when the time-stamping actor will be available and will have delivered a fixed time.

Delivered dates by different time-stamping actors can differ from a few moments (network, answer times of time-stamping machines, queues and breakdown). From where the need of defining a generation and validation protocol of time-stamping calling upon several time-stampings machines (Cottin, 2003)

Once the time-stamping validated, the document can be stored by the storage actor if this one is available. It can also be sent directly to the recipient if the storage act is not needed.

## **PROCESS MODELLING**

### **Graphical Modelling**

#### *Basic Petri Nets Definitions*

In this section, we recall the basic Petri net tool that we will use in our modelling. A complete introduction can be found in (René and Alla, 1992). A PN is a graph composed of two nodes: places and transitions. The oriented arcs connect certain places to certain transitions, or conversely. With each arc, we associate a weight (non negative integer). In a formal way, a PN is a 5-uple  $PN = (P, T, A, W, Mo)$  where:

- $P = \{P_1, \dots, P_n\}$  is a finite set of places;
- $T = \{T_1, \dots, T_m\}$  is a finite set of transitions (represented with line segments);
- $A \subseteq (P \times T) \cup (T \times P)$  is a finite set of arcs;
- $W = A \rightarrow \{1, 2, \dots\}$  is the weight function associated with arcs;
- $Mo = P \rightarrow \{0, 1, 2, \dots\}$  is the initial making of graph.

An important class of PN that we will use in this paper is Timed Event Graph (TEG). In this class, each place has only one input and one output transition. The aim of using this tool is to be able to describe, in a simple way, the behaviour of the process by mathematical and linear equations in (max, +) algebra (Baccelli, 1992) and (Nait, 2003).

#### *Petri net Model*

In order to comprehend the process working, we represent it by a graphical model with using a PN. In this model, the transitions model the events (authenticity, hash coding, signature, etc.) and their firings model the occurrence of these events. The places (resp. associated times) model the transit (reps. necessary times for transit) from one operation to another. The figure 2 presents the associated model to the considered process.

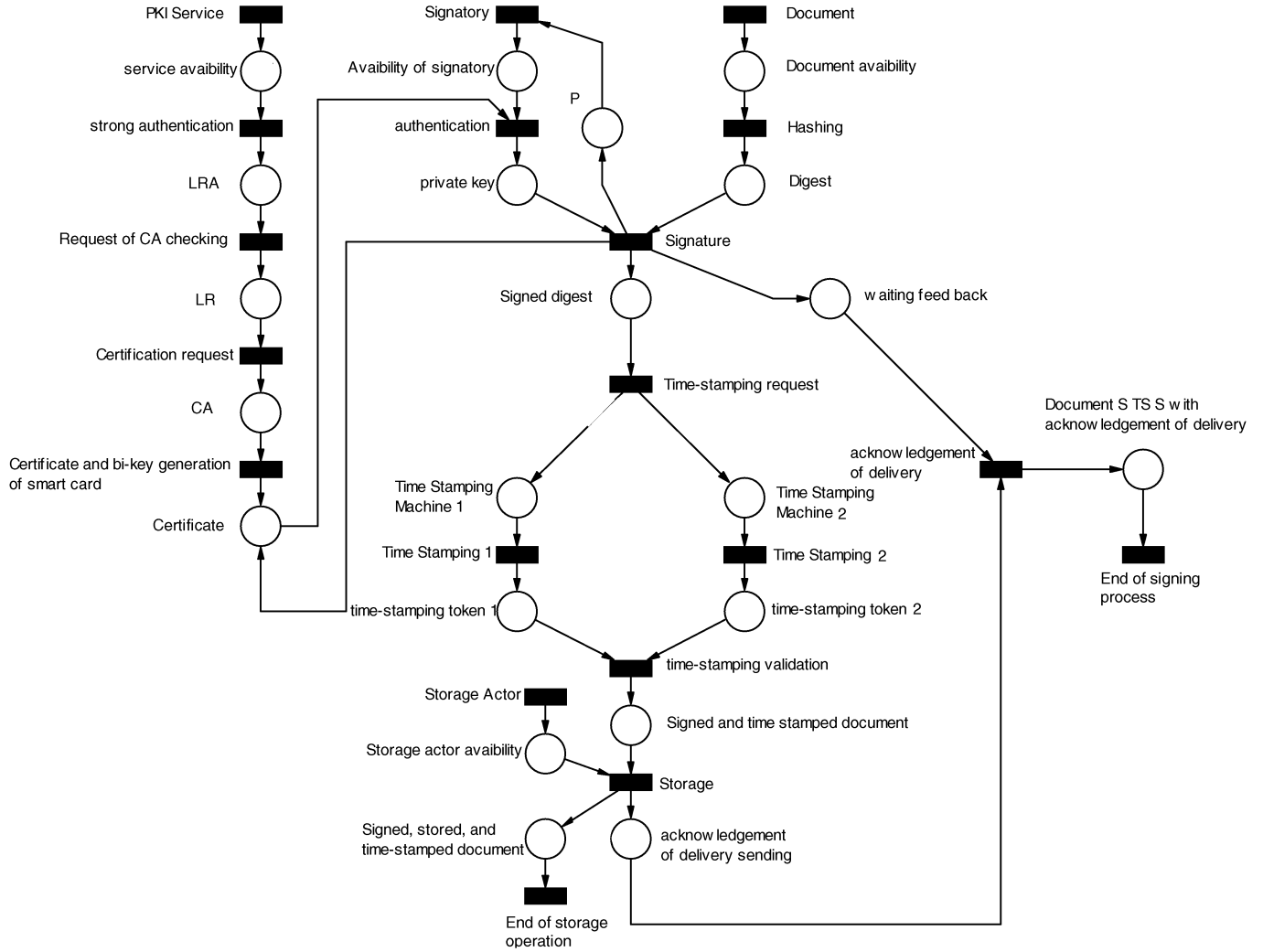


Figure 2. Petri net model of the global signature process.

**Notice 1:** The place named “Certificate” is associated with two upstream transitions and one downstream transition. This behaviour prevents the obtained PN model to being a GET. But knowing that only one certificate is sufficient to sign all documents, then the transitions that model the PKI are fired only once during the process. This lets us to consider that the transition “Certificate” is associated only with one upstream and one downstream transition; what gives the TEG behaviour to the obtained model.

The presence of a token in the upstream place “P” of the transition “Signatory” models the fact that the signer is available to sign a next document.

### Mathematical Modelling

In the second time, we describe the considered signature process by a state representation in  $(\max, +)$  algebra. First of all, we give some elements of this algebra (Baccelli, 1992 and Gaubert, 1992).

#### Some elements of $(\max, +)$ Algebra

We denote by  $\mathbb{R}_{\max}$  the dioid  $(\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$  where the operators “ $\oplus$ ” and “ $\otimes$ ” are respectively “max” and “the

usual addition” ( $\forall a, b$  in  $\mathbb{R}_{\max}$ ,  $a \oplus b = \text{Max}(a, b)$  and  $a \otimes b = a + b$ ). The neutral elements for the operators  $\oplus$  and  $\otimes$  are respectively  $\varepsilon = -\infty$  and  $e = 0$  ( $\forall a \in \mathbb{R}_{\max}$ ,  $a \oplus \varepsilon = a$ ,  $a \otimes e = a$ ). Like other algebraic structures, the  $(\max, +)$  algebra have a properties and characteristics, such that the associativity of addition, and the multiplicativity, the commutativity of addition, the distributivity of multiplication, existence of zero element (denoted  $\varepsilon$ ), etc. Let us consider  $a$  and  $b$  two elements in a dioid  $D$ , the quantity  $a^* \otimes b$  is the smallest solution of the equation  $x = a \otimes x \oplus b$ , where the expression of  $a^*$  (called Kleene star) is given by:  $a^* = e \oplus a \oplus a^2 \oplus \dots$ . In addition, each solution  $x$  satisfies  $x = a^* \otimes x$ . The greatest solution of  $f(x) = a \otimes x \oplus b$  is:  $x = b / a$ . The operator “/” (resp. “\”) represents the subtraction in the right (resp. the left) in the  $(\max, +)$  algebra. Finally, from a scalar dioid  $D$ , let us consider  $A \in D^{m \times n}$  and  $B \in D^{m \times p}$ , the sum and product of matrices are defined conventionally from the sum and product of scalars in  $D$ . In what follows, and without no risk of ambiguity, we omit “ $\otimes$ ” and let us write  $a \otimes b$  like  $ab$  or  $a.b$ .

#### $(\max, \text{Plus})$ linear model

In order to interpret the obtained graphical model with the mathematical equations in  $(\max, +)$  algebra, we associate a

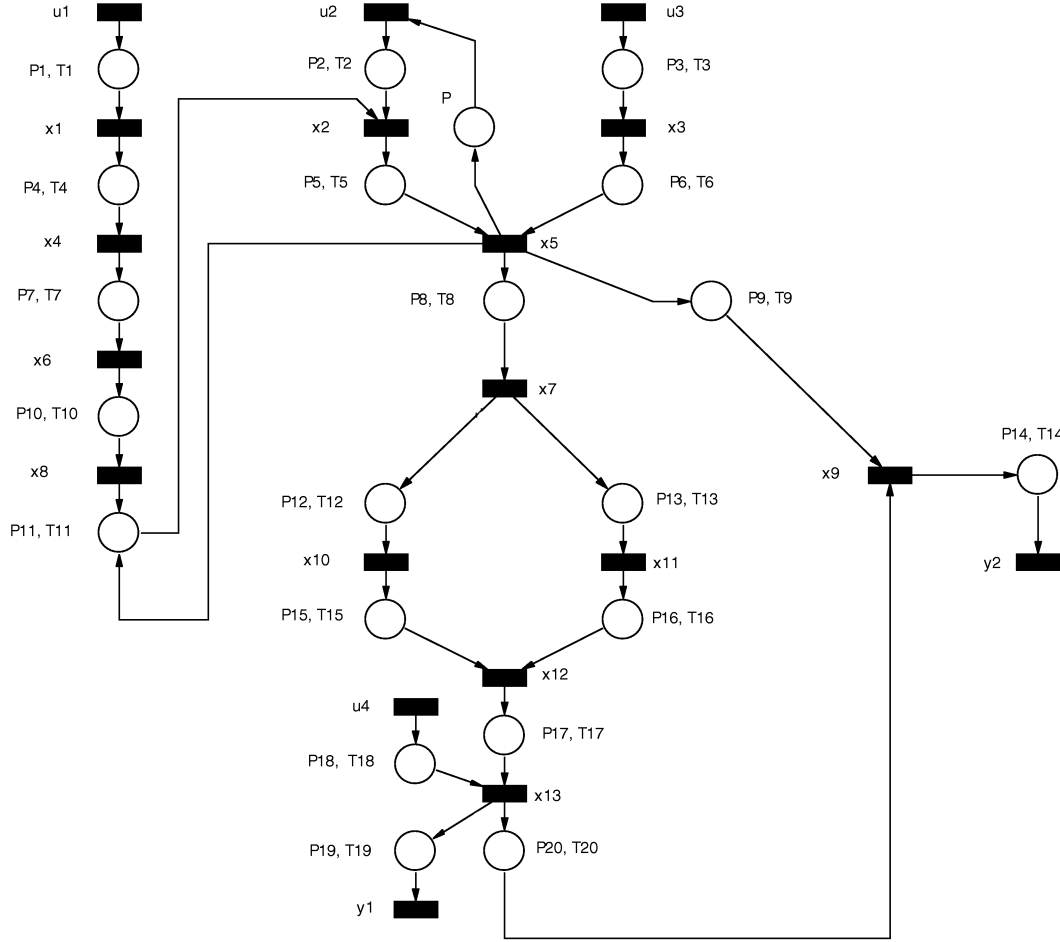


Figure 3. Graphical Model of the global signature process: introduction of variables and temporisations.

variable to each model transition (Figure 3). Thus, we associate with input transitions the input variables (denoted  $u_1$ ,  $u_2$ ,  $u_3$  and  $u_4$ ), with internal transitions the state variables ( $x_1$ ,  $x_2$ , ...,  $x_{13}$ ), and finally we associate output variables (denoted  $y_1$  and  $y_2$ ) with the output transitions. Also, we affect a temporisation, which represents the necessary average time between the end of an operation and the beginning of another, to each model place. By introducing these new variables and temporisations, we obtain the graphical model of the figure 3.

We denote by  $x(k)$ , called dater, the time of the  $k^{\text{th}}$  firing of the transition  $x$ . By using the associated dater to each model variable, we can express all equations that describe the analytical behaviour of the considered process.

**Notice 2:** In the level of  $(\max, +)$  equations, the content of the notice 1 is expressed by the “freeze” of the firing of all transitions which model the PKI starting from  $k = 2$ . Formally,  $\forall k \geq 2, u_1(k)=x_1(k)=x_4(k)=x_6(k)=x_8(k)=\varepsilon$ .

The system (1) represents the  $(\max, +)$  model that represents this signature process.

$$\forall k \geq 1 \quad \begin{cases} x_1(k) = T_1 \otimes u_1(k) \\ x_2(k) = T_2 \otimes u_2(k) \oplus T_{11} \otimes x_8(k) \\ x_3(k) = T_3 \otimes u_3(k) \\ x_4(k) = T_4 \otimes x_1(k) \\ x_5(k) = T_5 \otimes x_2(k) \oplus T_6 \otimes x_3(k) \\ x_6(k) = T_7 \otimes x_4(k) \\ x_7(k) = T_8 \otimes x_5(k) \\ x_8(k) = T_{10} \otimes x_6(k) \\ x_9(k) = T_9 \otimes x_5(k) \oplus T_{20} \otimes x_{13}(k) \\ x_{10}(k) = T_{12} \otimes x_7(k) \\ x_{11}(k) = T_{13} \otimes x_7(k) \\ x_{12}(k) = T_{15} \otimes x_{10}(k) \oplus T_{16} \otimes x_{11}(k) \\ x_{13}(k) = T_{18} \otimes u_4(k) \oplus T_{17} \otimes x_{12}(k) \\ y_1(k) = T_{19} \otimes x_{13}(k) \\ y_2(k) = T_{14} \otimes x_9(k) \end{cases} \quad (1)$$

In order to write this system in the matrix form, we define the following vectors:

$U(k) = [u_1(k), u_2(k), u_3(k), u_4(k)]^T$  : input vector;  
 $X(k) = [x_1(k), x_2(k), \dots, x_{12}(k), x_{13}(k)]^T$  : state vector;  
 $Y(k) = [y_1(k), y_2(k)]^T$  : output vector.

With using these vectors, we obtain the following matrix form:

$$\begin{cases} \forall k \geq 1, \\ X(k) = A X(k) \oplus B U(k) \\ Y(k) = C X(k) \end{cases} \quad (2)$$

With  $A \in \mathbb{R}_{\max}^{13 \times 13}$ ,  $B \in \mathbb{R}_{\max}^{13 \times 4}$  and  $C \in \mathbb{R}_{\max}^{2 \times 13}$  are the characteristic matrices of the model and whose elements represent the date of system.

#### Resolution of the state model

The equation (2) is an implicit equation. In order to solve it, and find all dates where various operations of process occur, we replace in the second member of (2), and in iterative way,  $X(k)$  by its expression, thus we obtain:

$$\begin{aligned} \forall k \geq 1, \\ X(k) &= A (A X(k) \oplus B U(k)) \oplus B U(k) = A^2 X(k) \oplus A B U(k) \oplus B U(k) \\ &= \dots \\ &= A^n X(k) \oplus A^{n-1} B U(k) \oplus A^{n-2} B U(k) \oplus \dots \oplus A B U(k) \oplus B U(k) \\ &= A^n X(k) \oplus (A^{n-1} \oplus A^{n-2} \oplus \dots \oplus A \oplus E) B U(k) \end{aligned}$$

Where  $n$  is the order of the matrix  $A$  (in our case  $n = 13$ ).  $E$  is the identity matrix in dioid algebra (it composed of the element “e” on diagonal and “ε” elsewhere).

In the expression of  $X(k)$ , it figures the quasi-inverse of the matrix  $A$  (denoted  $A^* = A^{n-1} \oplus A^{n-2} \oplus \dots \oplus A \oplus E$ , and called Kleene star). As  $A^n = \varepsilon$ , (because the graphical model in figure 3 does not contain no strongly connected component), then the smallest solution of the equation (2) is given by:

$$\forall k \geq 1, \\ X(k) = A^* B U(k) \quad (4)$$

From (4), the expression of the output system that expressed by (3) is given by:

$$\forall k \geq 1, \\ Y(k) = C A^* B U(k) \quad (5)$$

With the given solutions by the expressions (4) and (5), it is possible to evaluate all occurrence dates of various process operations: authenticity, signature, time stamping and storage of documents. Moreover, it is possible to reuse the graphical model in order to recover the equations (min, +) which enable to evaluate the number of documents signed, time-stamped and stored at a given moment “t”. This will be the subject of a next contribution.

## CONCLUSION

After describing the electronic signature as well as the various operations of its process, we have treated a problematic related to the modelling and the evaluation of process behaviour. For this problematic, two complementary tools are used to represent this process. We have used then Petri nets and (max, plus) algebra, and showed the possibility and the efficacy of these tools to model the considered process. In this study, we are interested in the case where the synchronisation between the process actors is considered (synchronisation between: document and signer, document and time-stamping actor or storage actor, etc.). It has been shown that the use of

the (max, plus) algebra enables to describe and analyse the behaviour of the graphical tool then the considered process. By analysing the obtained (max, plus) model, we could evaluate various dates where different operations of process occur.

This study has shown the feasibility of use of Petri nets and (max, plus) algebra to model and analyse the information systems and particularly the electronic signature. This demonstration of feasibility must enable us to develop a more complete model and introduce a control policy for the good management and the control of the electronic interchanges.

Among the objectives of this study, we aim to apply the proposed models for developing a decision-making aid software applying to the public markets like invitation to tender. This software will enable to the users a better temporal management of steps to follow in order to answer these processes. It will also be used to enable the users to take better decisions at the solicited moments.

## REFERENCES

- Baccelli F., Cohen G., Olsder G. L. and Quadrat J. P., 1992. *Synchronisation and linearity: an algebra for Discrete Event Systems*. Wiley.
- Cottin N., Sehili A., Wack M., Juillet 2003. Time-stamping Electronic Documents and Signatures. *AICCSA'03: ACS/IEEE International Conference on Computer Systems and Applications*, Tunis.
- Dobbertin H., Bosselaers A., Preneel B., 1996. RIPEMD-160, a strengthened version of RIPEMD. *Fast Software Encryption, LNCS* vol. 1039, D. Gollmann Ed., pp. 71-82.
- Gaubert S., Juillet 1992. *Théorie des systèmes linéaires dans les dioïdes*, Thèse de doctorat. Ecole Nationale Supérieure des mines de Paris.
- Housley R., Polk W., Ford W., Solo D. (April 2002). *RFC 3280: Internet X.509 Public Key Infrastructure, Certificate and Certificate Revocation List (CRL) Profile*.
- Kao M., 1999. *Designing Network Security*. Macmillan Technical Publishing, USA, ISBN 1-57870-043-4.
- Kaliski Jr B. S., Janvier 1992. *RFC 1319: The MD2 Message-Digest Algorithm*. RSA Laboratories.
- Menezes A. J., Van Oorschot P. C., Vanstone S. A., Février 2001. *Handbook of Applied Cryptography*. CRC Press, USA, ISBN 0-8493-8523-7.
- Nait-Sidi-Moh A., Wack M., Mai 2005. Modelling of process of electronic signature with Petri Nets and (max, plus) algebra. *LNCS, ICCSA*, Vol. 3485, Issue IV, pp. 792-801, ISBN. 3-540-25863-9.
- Nait-Sidi-Moh A., Décembre 2003. *Contribution à la modélisation, à l'analyse et à la commande des systèmes de transport public par les réseaux de Petri et l'algèbre (Max, Plus)*. Thèse de doctorat en Automatique et Informatique. Université de Technologie de Belfort-Montbéliard et Université de Franche-Comté.
- National Institute of Standards and Technology (NIST), Avril 1995. Secure Hash Standard (SHS). *Federal Information Processing Standards Publication*, FIPS PUB 180-1.

- National Institute of Standards and Technology (NIST),  
Octobre 1999. Data Encryption Standard (DES). *Federal Information Processing Standards Publication*,  
FIPS PUB 46-3.
- National Institute of Standards and Technology (NIST),  
Janvier 2000. Digital Signature Standard (DSS). *Federal Information Processing Standards Publication*,  
FIPS PUB 186-2.
- Preneel B., Bosselaers A., Dobbertin H., 1997. The cryptographic hash function RIPEMD-160. *CryptoBytes*,  
vol. 3, No. 2, pp. 9-14.
- René David et Hassane Alla, 1992. *Du grafset aux réseaux de Petri*. Série Automatique, hermes, Paris.
- Rieupet D., Wack M., Cottin N., Assossou D., Mai 2004. Signature électronique multiple. *Atelier Sécurité des Systèmes d'Information, XXII<sup>ème</sup> Congrès INFORSID*.
- Rivest R. L., Avril 1992. *RFC 1320: The MD4 Message-Digest Algorithm*. MIT Laboratory for Computer Science and RSA Data Security.
- R.S.A., Date Security Inc., 1993. *Public Key Cryptography Standards, PKCS 1-12*, available on-line at [ftp://ftp.rsa.com/pub/pkcs](http://ftp.rsa.com/pub/pkcs).



# MOBILE AGENT BASED LARGE-SCALE COLLABORATIVE VIRTUAL ENVIRONMENT SYSTEM

Qingping Lin, Liang Zhang, Irma Kusuma, Norman Neo  
Information Communication Institute of Singapore,  
School of Electrical and Electronic Engineering,  
Nanyang Technological University, Singapore 639798  
Tel: (+65) 67904688 Fax: (+65) 67922971  
Email: iqplin@ntu.edu.sg

## Abstract

The scalability of the existing Collaborative Virtual Environment (CVE) systems is limited due to the constraints in computer processing power and network bandwidth. To address the scalability issues, we have developed a mobile agent-based large-scale CVE (LCVE) system to support a large number of concurrent participants in a CVE with a large amount of evolving virtual entities. In our approach, the system tasks /services are modeled as mobile agents, which are not bound to any fixed nodes as the traditional CVE architectures do. The mobile agents can migrate between different system nodes. Thus, nodes can be treated as a kind of system computing resource, which is independent of the services they provide. The mutual independence of services and nodes provides large freedoms for the LCVE system to utilize the computing resource efficiently. Furthermore, any overloaded service mobile agent can autonomously reproduce itself and transfer the cloned one to the less-loaded nodes to share its workload, i.e. the communication architecture of the service will be extended autonomously at run-time to ensure the system scalability. This paper is an extension and refinement of our previous work to include our latest work in system prototype and experimental results of the proposed approach. Our experiment of the proposed architecture in supporting real-time interaction among 1000 concurrent users in LCVE demonstrated the effectiveness of our method.

**Keywords** *collaborative virtual environment, scalability, mobile agent, CVE architecture*

## 1. Introduction

Collaborative Virtual Environment (CVE) systems support multiple geographical dispersed human-to-human & human-to-machine communication and real-time interaction in a shared virtual environment. CVE provides users a three-dimensional shared virtual world over the networks, where participants can interact with each other naturally. One of the main issues in CVE is its scalability in supporting large number of participants geographically distributed over the Internet to interact in a common virtual world with large number of virtual entities including static, dynamic and evolving objects. Extensive research work in LCVE has been done by both academia and industry. There exist numerous systems supporting LCVE, such as AVIARY [1], BrickNet [2], DIVE [3], MASSIVE [4], NPSNET [5], PARADISE [6], RING [7],

SPLINE [8], VLNET [9] and VELVET [10]. These systems they adopt specific mechanisms or technologies for some aspects of a particular application to achieve the scalability. For example, MASSIVE adopts Awareness Management to reduce the unnecessary communication traffic; NPSNET adopts IP-Multicast technology to improve the transmission efficacy; SPLINE adopts Locale notion to confine the information exchange in a small group, etc. However, the scalability of the existing system architecture is still an issue in an unpredictable network traffic environment over the Internet or an ever-evolving VE system. In this paper, we address this scalability issue from the aspect of system architecture by proposing a mobile agent-based LCVE.

Mobile agent is an autonomous entity that can migrate from one machine to another in a heterogeneous network. Compared to traditional distributed computing schemes, mobile agents promise (at least in many cases) to cope more efficiently and elegantly with a dynamic, heterogeneous, and open environment which is characteristic for today's Internet [11] [12]. Thus, we apply mobile agent paradigm to LCVE to support a large number of concurrent participants in a VE with a large amount of evolving virtual entities.

This paper is organized as follows. Section 2 will review the related work; Section 3 will discuss the design of the proposed system; Section 4 will present system prototype and experimental results; Section 5 will draw conclusions and outline the future work.

## 2. Literature Review

A number of existing CVE systems have attempted to address the issue of scalability. In this section, we review some well-known LCVE systems.

### NPSNET

NPSNET [13] (Naval Postgraduate School Networked Vehicle Simulator) is a 3D networked virtual environment system developed at the Computer Science Department of the U.S. Naval Postgraduate School in 1993. It is designed to support large-scale military training and simulation exercises. NPSNET-IV is the most successful system to support LCVE. It complied with the Distributed Interactive Simulation (DIS) protocol to interoperate with other simulation system, and incorporates the dead-reckoning algorithms and IP Multicast protocol to reduce network traffic to achieve scalability.

NPSNET-IV is the first CVE system to adopt the multicast. And it is also the first successful LCVE system. In NPSNET-

IV, it logically partition virtual environments by associating spatial, temporal, and functionally related entity classes with network multicast groups. An entity can belong to several groups. For example different media communication can belong to different multicast group, say, position message and audio message. Multicast can save the bandwidth, filtering different kinds of traffic in the network interface hardware and does not consume processor cycles. What's more, it makes communication architecture easy to realize.

However, NPSNET is running at dedicated hosts and networks; the requirement of NPSNET system is extremely demanding in terms of network and computational resource [14].

#### **MASSIVE**

- MASSIVE [15] aims to be a large-scale multi-user virtual environment with rich facilities to support user interaction and cooperation. Though, it does not fully address the scalability issue, MASSIVE propose a spatial model of interaction that is useful for the later research. In MASSIVE, the clients communicate through peer-to-peer connections and the spatial trader serves as an aura manager to check for aura collisions. If aura manager detect a collision it will notifies any objects concerned to set up a peer-to-peer connection and find out about each other. When the peer-to-peer connection is enabled, the communication is managed according to mutual levels of awareness which are negotiated through the use of focus and nimbus. The calculation of mutual awareness levels is the responsibility of the peer objects. This model can decrease the unnecessary communication traffic.

However, the communication between spatial trader and each peer is still a client/server mode. The spatial trader will become a bottleneck of the whole system as the number of participants increases to a certain level.

#### **SPLINE**

SPLINE [16] (Scalable Platform for Large Interactive Networked Environments), developed at the Mitsubishi Electric Research Laboratory in 1996, is a software platform suitable for implementing multi-user interactive environments. SPLINE aims to make the multi-user Virtual Environment large in spatial extent, large in number of objects, and large in numbers of users interacting with the environment.

One of the most interesting features of SPLINE is the definition of locales. The concept of locale is based on the idea that even in a very large virtual world, most of what a single user can observe at a given moment is nevertheless local in nature. Locales divide a virtual world into chunks that can be processed separately. This division is purely an implementation issue—it is not apparent to the user. A user sees several locales at once—generally the locale containing the user's point of view and those neighboring it.

Each locale is associated with a separate set of multicast addresses, which guarantees the efficiency of the communication. Each locale has its own coordinate system, which guarantees precise position and velocity information

about objects that are arrayed across a large volume of space. Each locale can has arbitrary shape, size, and orientation which allows designers to create parts of a virtual environment separately and combine them together later [17]. The locales notion can make VE spatially scalable to an arbitrary extend. However, similar to NPSNET, supporting heterogenous LCVE is still an issue.

#### **DWTP**

- DWTP (the Distributed Worlds Transfer and communication Protocol) [18] is an application layer protocol proposed by *Broll* from GMD. It provides a flexible and scalable network architecture for LCVE, which is not based on Client/Server approach. It uses a set of different daemons to provide particular services to the participants. Each daemon is responsible for a duty.

These daemons can either be combined in a single program to provide all the services required to realize a single shared virtual environment or they maybe split among several hosts. The basic communication architecture of DWTP is based on IP multicasting groups [19]. The heterogeneous architecture of the approach also allows non-multicast capable participants to join shared virtual worlds [20].

This architecture is flexible that the individual daemons can be distributed on several hosts to reduce the load at each host. Additionally most daemons (except the reliability daemon) can be replicated to split the load of a particular daemon between several hosts. However run-time autonomous adapting system architecture to support evolving virtual entities and participants is not possible in this approach.

### **3. Design of Mobile-Agent based LCVE**

We attempt to enhance the existing system architectures by proposing a mobile agent based LCVE system [21, 22]. In our proposed approach, each mobile agent in a LCVE system will be responsible for an independent task for the LCVE, such as the management of the VE contents, maintenance of consistency and persistency of a LCVE. Mobile agents will execute in system-registered nodes, and when necessary, agents will move from time to time to other system-registered nodes. The features of mobile agents (e.g. cloning, autonomous decision making and computing, and dynamic movement in a network) will be exploited in order to achieve scalability and extensibility of LCVE. The following sections will present the system architecture and design of the mobile agent based LCVE.

#### **3.1 Classification of LCVE Nodes**

In a LCVE, there are a large number of computer hosts (nodes) connected to the virtual environment. In this system, the participating nodes are divided into two types: *User Nodes* and *Service Provider Nodes*. The classification is made based on the functional roles played by each of the nodes. User Nodes are the participating nodes that logged on to the system. Service Provider Nodes are hosts owned by the LCVE owner. User Nodes and Service Provider Nodes are both further divided into sub-types. User Nodes are divided into Normal User Nodes and Trusted User Nodes:

**Normal User Nodes** are nodes that perform only basic actions such as navigating inside the virtual environment and interacting with virtual entities and other CVE participants.

**Trusted User Nodes** are nodes that have a higher computing ability that can be used to assume some of the task in the LCVE management. Thus, besides performing basic actions as Normal User Nodes, these nodes also take part in system management. Because of this, it is very important for these nodes must meet the requirement set by the LCVE owner, and notify the system when they are logging off from the system such that the computing task assigned to them can be re-assigned accordingly.

Service Provider Nodes are also divided into two types of nodes:

**Controlling Nodes** are nodes owned by the system owner that performs some tasks for LCVE system management and maintenance.

**DB Nodes** are nodes owned by the system owner that provides database support for the system.

Trusted User Nodes and Controlling Nodes both are used to share the workloads in performing computing tasks for system management. Therefore, both of them are further classified together as **System Controlling Nodes (SC Nodes)**.

### 3.2 System Architecture

The proposed mobile agent-based LCVE is a fully distributed system. It comprises of mobile agents, each mobile agent is a software component that performs one of the management, maintenance, and control task by collaborating with other agent(s). The agents will be deployed to different participating nodes and able to move and reproduce themselves when necessary. The system architecture is visualized by Figure 1.

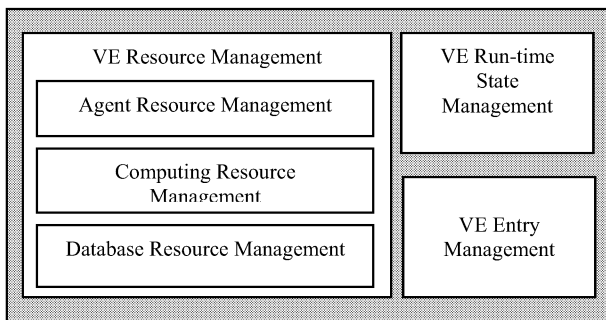


Figure 1 Mobile Agent-Based LCVE Architecture

It can be seen from Figure 1 that the system architecture consists of three parts: VE Resource Management, VE Run-Time State Management, and VE Entry Management. Mobile agents are allocated to each of the three to perform the corresponding management tasks as described in the following paragraphs.

**VE Resource Management** is the *resource provider* that handles all tasks related to system resource management, and it is further divided into:

**Agent Resource Management (ARM):** responsible for managing all mobile agents. In particular, it is responsible in the registration of the mobile agents, allocating them to members of the nodes for them to perform their tasks, and upgrading the agents with new functionalities. ARM is also responsible for allocating SC Nodes to different groups for different functionalities, monitoring them, and balancing workload among them.

**Computing Resource Management (CRM):** responsible for managing the computing resources (processing power, bandwidth) which are shared among the SC Nodes.

**Database Resource Management (DRM):** responsible for managing the database resources in the system. The database is distributed among the DB Nodes in order to support the large size virtual environment. DRM is responsible for registering active DB nodes, monitoring the storage and fetching capacity of the active DB Nodes, and performing load balancing among the DB Nodes. This will ensure that the database management is done efficient and effectively.

**VE Run-Time State Management** is the *content provider* that maintains and manages the run-time state of the virtual environment. It is responsible for sending relevant part of the virtual environment to participating nodes and maintaining the virtual world states’ persistency and consistency. The run-time management is achieved by the collaboration of Region Agents, Cell Agents, Persistency Agents, and Consistency Agents .

**VE Entry Management** is the *entry service provider* that handles the joining of new participants into the system. It is responsible for user authentication and directing the user node to the VE Run-Time State Management. This is done by way of Gateway Agents to which users will be registered when first joining the system.

In this system, each of the management aspects mentioned above is achieved by one or more type(s) of mobile agents, several of them have been mentioned briefly. So mobile agents working in this system are divided into several types and each agent performs their tasks accordingly. The classification of agents is shown in Table 1.

Management Aspects		Agents
VE Resource Management	Agent Resource Management	ARM Agent
	Computing Resource Management	CRM Agent, Group Manager Agents, Node Agents
	Database Resource Management	DRM Agent, DB Agents
VE Run-time state Management		Region Agents, Cell Agents, Persistency Agents, Consistency Agents
VE Entry Management		Gateway Agents

Table 1 Classification of Mobile Agents based on Logical Aspect

### 3.2.1 VE Resource Management

#### *Agent Resource Management (ARM)*

ARM is achieved by ARM Agent. When LCVE first constructed, all agents need to register to ARM Agent. Likewise, when there is new agent joining the system at run time, it will need to register to the ARM Agent. After ARM Agent has received the information regarding the existing agents, it is responsible to distribute the mobile agents to different nodes based on the current states of the system. This is to achieve a distribution of tasks, load balancing, and to enable system extensibility.

Task distribution is especially important during system's initialization and evolvement. During LCVE system initialization, ARM Agent will distribute the task of managing the run-time state by allocating the agents of the VE Run-Time State Management to the nodes to do the necessary system setup. When the CVE content evolves, ARM will distribute mobile agents to manage the newly extended environment. Another case is when new SC Nodes are joined to the system, CRM can ask ARM to allocate Group Manager Agent to manage the new group of SC Nodes.

ARM also has the responsibility to upgrade the mobile agents. Examples of upgrading scenarios are when there is new management mechanism introduced to the system or when there is a newer version of mobile agent. In those cases, ARM Agent will be the one responsible to replace/upgrade the agents currently running in the system.

#### *Computing Resource Management (CRM)*

In this system, workload is distributed between Service Provider Nodes and Trusted User Nodes. Therefore, Trusted User Nodes have to be part of User Nodes that have the computing resource above certain requirement and are willing to contribute its resource for the management task of the system. Besides avoiding bottleneck, this mechanism also enhances the system's reliability. The responsibility of CRM is to manage the computing resources such that a sufficient amount of resources can be allocated to each computing task.

Each node belongs to the SC Nodes has Node Agent running in it. This Node Agent is responsible for the local computing resource management of its host. To make the system nodes more manageable, the SC Nodes are divided into groups, and each group is controlled by a Group Manager Agent which runs in a node called Group Manager Node. Group Manager Agent is responsible for monitoring and ensuring balancing of workload between nodes in its group. When there is one node whose workload has exceed its capability, Group Manager Agent has to first try to find within the group a suitable node with which the workload can be shared. If the ideal node is not found locally, Group Manager Agent will communicate with the CRM Agent to negotiate with other Group Manager Agent to search for the ideal node in other groups.

In this system, hosts belong to the same subnet are most likely to be put into the same group. This is under the assumption that those belong to one subnet tend to have a

higher communication speed. Thus, they can communicate faster to facilitate task transferring.

CRM Agent manages the node groups. When a SC Node first joins the system, it will register to CRM Agent, and CRM Agent will allocate it to one of the node groups. Node groups can be created dynamically, for example when existing node groups have reached their limit. If that need arises, CRM Agent will request ARM Agent to send a new Group Manager Agent to manage the newly formed node group.

Group Manager Agent (as explained previously) is responsible for managing the computing resource management within its group.

Node Agent monitors its host's computing load and network traffic. When the load exceeds certain limit, the Node Agent can negotiate with other Node Agents to share and balance the load. This can be done by cloning or transferring mobile agents to other nodes.

#### *Database Resource Management (DRM)*

The database of LCVE consists of a large number of VE content and configuration data. The database is distributed among the DB Nodes, and new DB Nodes can be added to the system at run time to support extensibility. DRM has the responsibility to register the DB Nodes, monitor DB Nodes' computing and fetch capacity, and manage the data storage. To achieve this, there are two types of agents involved in DRM: DRM Agent and DB Agent. DRM Agent is responsible to manage all the DB Nodes in the system. All DB Nodes need to register to DB Agent, and their storage and fetch capacity will be continuously monitored by DRM Agent. DRM Agent maintains a table containing the DB Nodes addresses for each *cell* (the smallest spatial unit of VE). DRM Agent also responsible in allocating DB Nodes for extending VE content and for load sharing when there is a DB Node which has reached its limit. When a DB Node is allocated to share storage load with another DB Node, there might be times when data retrieval must be done from both of the DB Nodes.

DB Agent resides in each of DB Nodes and is responsible to continuously monitor its own node's computing and storage capacity to be reported to DRM Agent. DB Agent is also responsible for the reading and writing of data to the database portion stored in its node.

### 3.2.2 VE Run-Time State Management

The virtual environment is spatially divided spatially to a number of manageable regions. Each VE region can be further sub-divided into more child regions and cells if required. The child regions can also have their own child regions and cells. So the division of regions can be recursive and forms nested regions. As mentioned before, cell is the smallest spatial unit of virtual environment. It is the basic unit for virtual entity downloading, communication, persistency and consistency. Communication within a cell is achieved using multicast. Figure 2 shows the example of a VE region's composition where a region, namely Region A, which consists of 1 child region and 4 cells. As explained before, a child region can be further sub-divided into child

regions and cells. That is what happened to Child Region 1. As shown in the figure, Child Region 1 is sub-divided into 1 child region and 1 cell. This architecture is just like the file management system in a computer Operating System, a cell is like a file while a region is like a folder. Region is used to manage cells just as a folder is used to manage files. By having this kind of spatial division of

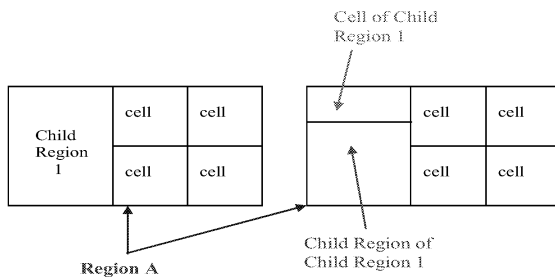


Figure 2 Example of Region Composition

virtual environment, the system will be easier to manage, ensuring that there will not be any region or cell that is so crowded beyond certain limit.

The VE Run-Time State Management is achieved by the collaboration of Region Agents, Cell Agents, Persistency Agents, and Consistency Agents. The relationships among those agents are shown in Figure 3.

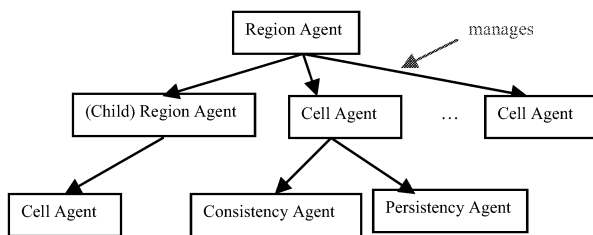


Figure 3 Structure of Agents for VE Run-Time State Management

Figure 3 shows Region Agent is in the top of the hierarchy. Region Agent can manage several other Region Agents and Cell Agents, depending on how the region is divided (refer to Figure 2). There are also Consistency Agent and Persistency Agent running in each cell and managed by the corresponding Cell Agent. Below are the functionalities of each agent:

Region Agent does not directly manages the VE content. Instead, it is responsible for managing the hierarchy of VE content. It monitors the relationship between the child regions and cells and it also stores the information regarding the network address of the (Child) Region Agents and Cell Agents respectively.

Cell Agent tasks are to download the VE content of its cell from the database, maintain the cell VE content, maintain the information of users who are active in the cell, and deliver the current cell VE content to new users.

Persistency Agent is responsible for cell persistency. It maintains the cell VE content by sending the cell history and the latest status of cell content to the database periodically.

Consistency Agent is responsible in maintaining cell consistency. It ensures that any interaction occurs in the cell are sent to those users who need to be aware of the changes (e.g. users active in the cell). Consistency Agent also ensures that there is no ownership conflicts of virtual entities among the cell members.

It is important to note that the regions and cells can be divided or merged dynamically. For example, Region Agent can merge several cells and child regions if all of them have very few active users. After the merging, Region Agent can re-divide the newly merged region into child regions and cells in a different manner according to the need.

During navigation time, users in a cell may only see a portion of the whole virtual environment at one point of time depending their viewpoint. This portion may include the current cell and its neighboring cells. So each time a User Node joins a cell, it needs to download the VE content of its own cell and also the neighboring cells. When the user moves to another cell, the Cell Agent of the new cell in which the user currently residing will contact its parent Region Agent to get the neighboring cells so that the User Node can download the new VE content correspondingly. To ensure smooth navigation, contents of the neighboring cells can be pre-fetched based on user's location and moving vector.

### 3.2.3 VE Entry Management

VE Entry Management is responsible in authenticating users and directing users into the VE. This management is done by the use of Gateway Agent. When users first connected to the system, they will need to provide their profile/authentication information to the nearest Gateway Agent along with the destination VE that they would like to go to. Gateway Agent then perform authentication based on the user information. If authentication succeeds, the user will be directed to the destination. Thus, Gateway Agent is also responsible in maintaining the record of region names and addresses. If a user is identified as a Trusted User Node, Gateway Agent will direct it to CRM Agent for registration and allocation to a Node Group.

When a User Node leaves the CVE, the Gateway Agent is responsible to store the user's login history in the system database. If the User Node is a Trusted User Node, before leaving, the current tasks assigned to the node needs to be transferred to other SC Nodes. The CRM Agent will also be informed of its leaving the CVE.

## 4. System Prototype and Experimental Results

As the existing mobile agent platforms is designed for general purpose which is not suitable for real-time multi-agent communication required in LCVE, we have implemented our own mobile agent platform with our own mobile agent communication protocol designed for LCVE application. Figure 4 illustrate a sample screen capture of mobile agent environment. Figures 8 and 9 show sample screen captures of a LCVE user interface developed based on our platform. It can be seen from the interface, the display of the virtual world itself is complemented with interface to facilitate users' collaborative interaction of adding, removing,

or modifying the objects within the virtual world to evaluate the effectiveness of the proposed approach. There are also small display areas for monitoring the state of the entities and avatars in the currently displayed cell. Other features like chatting and file transfer between users in the cell have also been implemented in the system. In our experiment, 1000 concurrent users (including real users, and simulated users which generate random user movement and interaction messages based on the scene contents) can interact with each other in real-time in a large-scene with various interactive virtual entities with which the users can manipulate and add or remove as shown in Figures 5 and 6. The scene consistency is maintained by the Consistent Agents to provide the sense of CVE while its persistency managed by Persistent Agents to allow the virtual world to evolve without interruption.

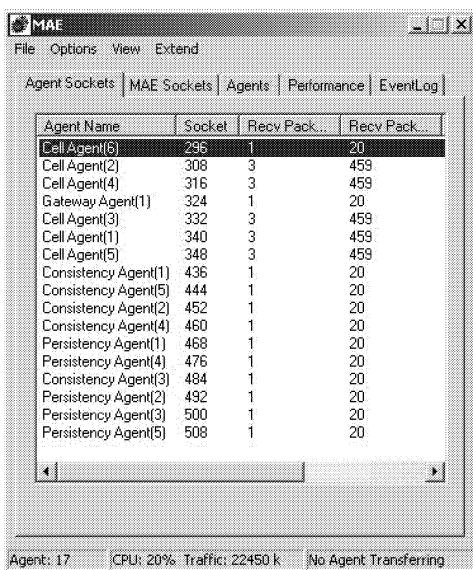


Figure 4. Screen Capture of Mobile Agent Environment

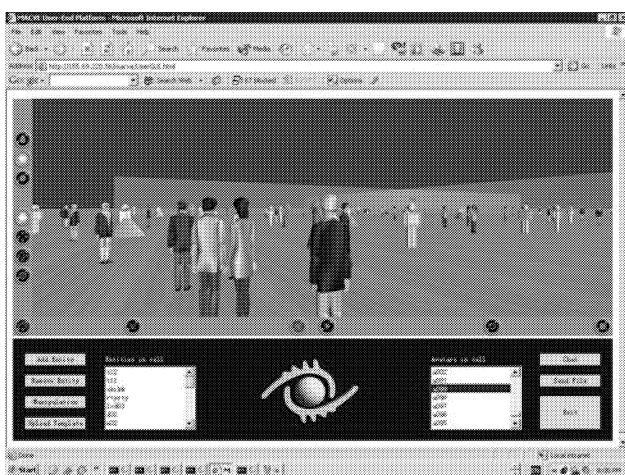


Figure 5. Sample Screen Capture of Mobile Agent based LCVE User Interface with 1000 concurrent users in a large scene with various interactive virtual entities.

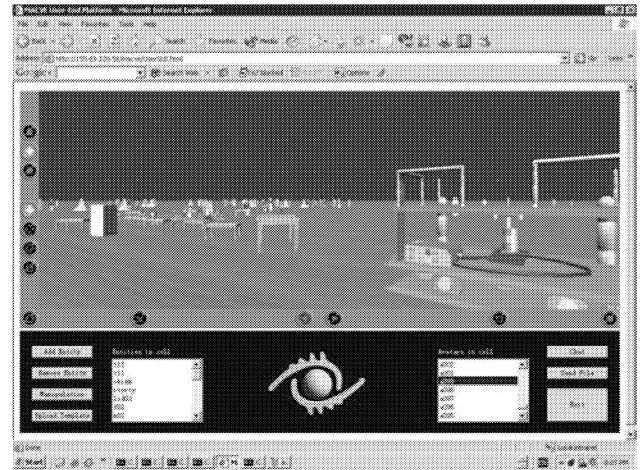


Figure 6. Sample Screen Capture of Mobile Agent based LCVE User Interface with 1000 concurrent users in a large scene with various interactive virtual entities.

To study the agent migration impacts on CVE users, we do experiments to study the effect of number of concurrent users and the number of entities in a cell on the migration time spans of a Consistency Agent. The reason we choose to study Consistency Agent migration is that its migration is most complex among all types of agents and it has most impacts to users.

The migration of a Consistency Agent includes 3 steps: (1) transferring the agent code to the destination; (2) synchronizing the agent state; (3) shifting the agent control. Transfer & Synchronization Time measures the time required to transfer the agent code and synchronize the agent state. The Transfer & Synchronization process will not directly affect users' collaborative interaction in the CVE as it can be done in a separate thread while the CVE system tasks are performed by the current mobile agents. Handover Time measures the time required to shift agent control which will affect users' collaborative interaction in the CVE. Thus it is particularly important to evaluate the delay caused by the Handover Time.

As shown in Figure 7, we observe from our experiment that the Transfer & Synchronization Time increases with the increase of the number of current users, whereas the Handover Time is relatively stable which is 425.0 ms when the concurrent user number reaches 1000 in a cell. In Figure 8, we observe that the Transfer & Synchronization Time increases with the increase of the number of entities, whereas the Handover Time is relatively stable which is 344.4 ms when the number of entities in the cell reaches 3000. A temporary short delay of less than half a second in updating scene state caused by the Handover Time will have little impact on the performance of a CVE with a large number of concurrent users and virtual entities. We found from the experiment that the users would not feel the migration of the Consistency Agent as the user interaction with the LCVE will not be affected during the agent transfer and synchronization period while the agent control handover time is short enough to avoid apparent interruption.

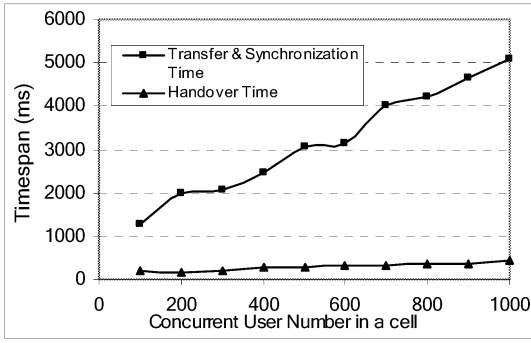


Figure 7: Consistency Agent Migration Time on Concurrent User Number

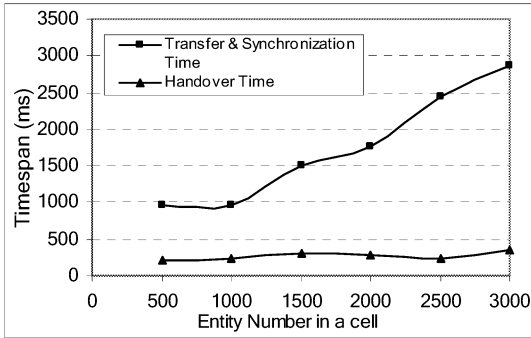


Figure 8: Consistency Agent Migration Time on Entity Number

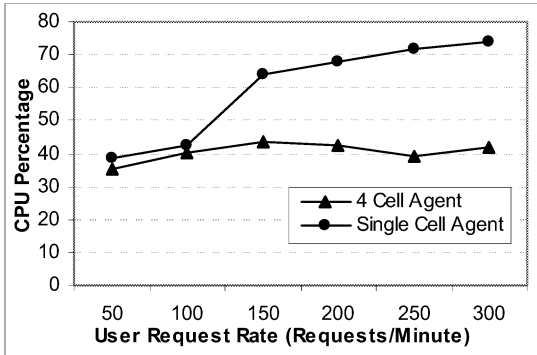


Figure 9: CPU Utilization at Controlling Node

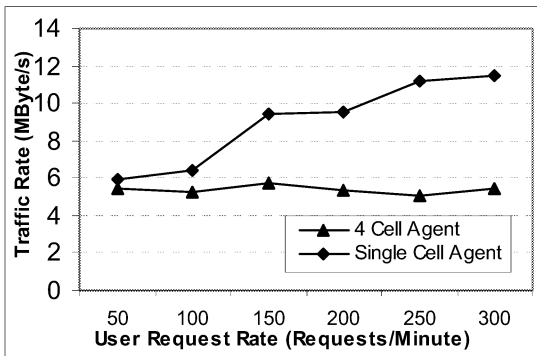


Figure 10: Outgoing Traffic at Controlling Node

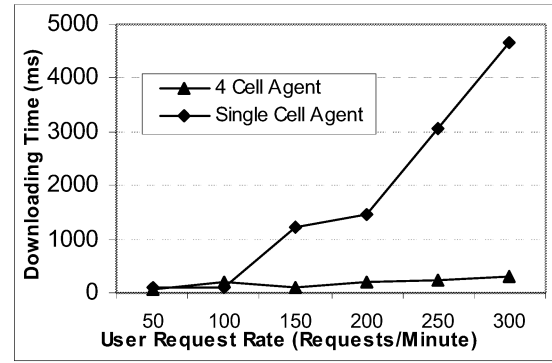


Figure 11: Scene Data Download Time

Our experiments show that Consistency Agent migration does not affect the real-time interaction of the CVE users, and so do other types of agents. Therefore, our proposed MACVE will improve the scalability of LCVE without compromising its performance.

To study effect of Cell Agent clone on the scalability of LCVE system, we do experiment to compare the scene delivery performance of a cell having only one Cell Agent with that of a cell having one Master Cell Agent and 3 Cloned Cell Agents. In the experiment, Master Cell Agent runs at a Controlling Node and the 3 Cloned Cell Agents run at 3 different Trusted User Nodes. Under the same density of a CVE scenario, more concurrent users result in more user requests to the Cell Agent to cache the scene data. Therefore, we measure the CPU utilization at the Controlling Node, outgoing traffic rate at the Controlling Node, and the basic scene data (the 3D geometric description of the cell) download time over ever increasing user request rates.

As shown in Figure 9 and 10, when there is only one Cell Agent, the CPU utilization and outgoing traffic rate at the Controlling Node increase steadily as the user request rate increases. Whereas, when the Master Cell Agent spawns 3 Cloned Cell Agent at Trusted User Nodes, the CPU utilization and outgoing traffic rate at the Controlling Node increase at significantly slower pace. This indicates that pervasive shared caching can alleviate the workloads of delivering scene data at the Controlling Node.

As shown in Figure 11, when there is only one Cell Agent, the time for downloading the 3D geometric description of the cell increase rapidly as the user request rate increases. At the number of 300 for the user request rate, the downloading time reaches 4,673ms. This delay will interrupt user's continuous navigation. Whereas, when there is 4 Cell Agent work together, the downloading time is relatively stable. And the delay is only 294ms when the rate reaches 300. This indicates that pervasive shared caching can significantly reduce the time for users to download the scene data.

From Figure 9, 10 and 11, we also observe that when the user request rate is below 100, the difference between the two methods is marginal; whereas, when the rate is above 100, pervasive shared caching present a much better performance. This indicates that pervasive shared caching is more effective for large number of concurrent users.



This experiment shows that by cloning Cell Agent at Trusted User Nodes, MACVE can significantly improve the scene delivery performance for cells.

## 5. Conclusions and Future Work

In this paper, we have proposed a novel fully distributed mobile agent-based architecture for LCVE. This architecture is flexible and autonomous in ensuring the scalability of LCVE. Experiments of our prototype system of the proposed architecture demonstrated its effectiveness in supporting large number of concurrent users with real-time interaction in LCVE.

In our future work, we will develop real-time self-learning intelligent decision making methods for different mobile agents in our system. These self-learning mobile agents will be able to make decision in response to the evolving CVE contents or changing activities in the CVE, and adapt to the requirement of VE size and complexity and the number of concurrent users. The system will also provide autonomous and intelligent scheduling of mobile agents to ensure the scalability and extensibility of LCVEs. We intend to make our platform available for free to academic non-commercial use to allow further research and development in this area.

## 6 Reference

- [1] Snowdon, D. and West, A. "AVIARY: Design Issues for Future Large-Scale Virtual Environments", *Presence* Vol. 3 N.4, 1994.
- [2] Singh, G., Serra, L., Pang, W. and Ng, H., "BrickNet: A Software Toolkit for Network-Based Virtual Worlds", *Presence*, MIT Press, Vol. 3, No. 1, 1994, pp. 19-34.
- [3] Hagsand, O. "Interactive Multiuser VEs in the DIVE System", *IEEE Multimedia*, Spring 1996, Vol. 3, N. 1, IEEE Computer Society, pp. 30-39.
- [4] Greenhalgh, C. and Benford, S., "MASSIVE: A Collaborative Virtual Environment for Teleconferencing", *ACM Transactions on Computer Human Interfaces*, Vol. 2, N. 3, pp. 239-261, September 1995.
- [5] Macedonia, M., Zyda, M., Pratt, D., Barham, P. and Zestwitz, S., "NPSNET: A Network Software Architecture for Large-Scale Virtual Environments", in *Presence*, Vol. 3, No. 4, Fall 1994, pp. 265--287, MIT Press.
- [6] Singhal, S. and Cheriton, D., "Using Projection Aggregations to Support Scalability in Distributed Simulation", *Proc. International Conference on Distributed Computing Systems (ICDCS'96)*, 1996.
- [7] Funkhouser, T., "RING: A Client-Server System for Multi-User Virtual Environments", *Proc. Siggraph Symposium on Interactive 3D Graphics*, Monterey, April, 1996.
- [8] Abrams, H., Watsen, K. and Zyda, M., "Three-Tiered Interest Management for Large Scale Virtual Environments", *Proc. VRST'98*, Taipei, September, 1998.
- [9] Capin, T., Noser, H., Thalmann, D., Pandzic, I. and Thalmann, N., "Virtual Human Representation and Communication in VLNet", *IEEE Computer Graphics and Applications*, Vol. 17, N. 2, 1997, pp. 42-53.
- [10] Oliveira, J.C., Georganas, N.D., "VELVET: An adaptive hybrid architecture for very large virtual environments". *Proc. 2002 IEEE International Communication Conference (ICC'02)*.
- [11] Fnfrocken, S., and Mattern, F. "Mobile agents as an architectural concept for internet-based distributed applications: the WASP project approach." *Proc. KiVS'99 ("Kommunikation in Verteilten Systemen")*. 1999 (29 October 2001).
- [12] Cheng-Zhong Xu, Brian Wims, "A mobile agent based push methodology for global parallel computing". *In Concurrency - Practice and Experience* 12(8): 705-726 (2000).
- [13] Michael R. Macedonia, Michael J. Zyda, and others, "Exploiting Reality with Multicast Groups: A Network Architecture for Large-scale Virtual Environments", *Proc. Virtual Reality Annual International Symposium (VRAIS'95)*, March 11 - 15, 1995.
- [14] Manuel Oliveira, Jon Crowcroft, Mel Slater, "Component framework infrastructure for virtual environments", *Proc. the third international conference on Collaborative virtual environments*, p.139-146, September 2000.
- [15] Greenhalgh, C. and Benford, S., "MASSIVE: A Distributed Virtual Reality System Incorporating Spatial Trading", *Proc. the 15th International Conference on Distributed Computing Systems (DCS'95)*, Vancouver, Canada, 1995, pp.27-34.
- [16] Barrus, J.W., Waters, R.C., and Anderson, D.B. (1996), "Locales: Supporting Large Multiuser Virtual Environments", in *IEEE Computer Graphics and Applications*, 16 (6), Nov., 50-57.
- [17] Barrus, J.W., Waters, R.C., Anderson, D.B. "Locales and beacons: efficient and precise support for large multi-user virtual environments". *Proc. 1996 IEEE Virtual Reality Annual International Symposium (VRAIS'96)*, pp. 204--213.
- [18] Broll, W., "DWTP - An Internet Protocol for Shared Virtual Environments", *Proc. 3rd Symposium VRML'97*, Monterey, February, 1998, pp. 15-24.
- [19] V. Kumar, "MBone: Interactive Multimedia on the Internet". *New Riders Publishing*, Indianapolis, Indiana, 1995.
- [20] Broll, W., "SmallTool - A Toolkit for Realizing Shared Virtual Environments on the Internet", *Distributed Systems Engineering, Special Issue on Distributed Virtual Environments*, No. 5, pp. 118-128 (1998).
- [21] Zhang L., Lin Q., Choo T. F., "Mobile Agent-Based Architecture for Large-Scale CVE", in proceedings of CW2003, December, 2003, Singapore.



# **SEMANTIC WEB**



# ON THE REQUIREMENTS TO THE METHODS FOR WEB SERVICE COMPOSITION

Hristina Daskalova, Tatiana Atanasova

Institute of Information Technologies -BAS, Acad. G. Bonchev 2

1113 Sofia, Bulgaria

E-mail: [daskalovahg@abv.bg](mailto:daskalovahg@abv.bg) [atanasova@iinf.bas.bg](mailto:atanasova@iinf.bas.bg)

**KEYWORDS:** Web services, Semantic Web Services, Composition Methods

## ABSTRACT

For successful composition of web services there are several key requirements. They may serve for comparison when selecting the most appropriate one from the existing methods and may be useful in developing of new methods.

## INTRODUCTION

The composition of web services is based on different kinds of knowledge (Lee et al 2004). The compositional knowledge consists of syntactic, semantic and pragmatic knowledge. The syntactic knowledge relays on syntactic constraints that concerns to correct I/O conditions. The syntactic constraints guarantee the service data flow. The semantic knowledge defines the proper order of composed services as semantic constraints forms rules for service integration. The semantic composition rules often require extensive domain expert knowledge. The pragmatic knowledge may be expressed by sets of rules. Pragmatic knowledge concerns to the goals of the consumer of the service flow and includes QoS or preferences refinement. Mechanism of business process management together with semantic web and semantic web services technologies can provide suitable techniques for these kinds of knowledge and their representation (Hepp et al 2005).

All kinds of knowledge are used in different manner in different types of web service composition methods to ensure useful result for the given user request.

## WEB SERVICES COMPOSITION METHODS

The wide ranges of developed composition methods take a variety of forms depending on realization of every one of the following phases. The phases are (Rao et al., 2003):

- presentation of single service;
- translation of the languages;
- generation of composition process model;
- evaluation of composite service;
- execution of composite service;

Building composite Web services with an automated or semi-automated tool is critical to the success of the Web service applications.

The web service composition methods could be considered in three groups:

workflow-based composition; Web Service Composition via AI Planning ; Service Composition using Program Synthesis. (Rao et al., 2003)

Using **workflow-based composition** methods, one should distinguish the static and dynamic workflow generation. The static one means that the requester should build an abstract process model before starting the composition planning.

The abstract process model includes a set of tasks and their data dependency. Each task contains a query clause that is used to search the real single Web service to fulfill the task. In that case, only the selection and binding of single Web service are done automatically by program. The most commonly used static method is to specify the process model in graph. On the other hand, the dynamic composition both creates process model and selects services automatically. This requires the requester to specify several constraints, including the dependency of services, the user's preference and so on.

**Web Service Composition via AI Planning** -many research efforts tackling Web service composition problem via AI planning have been reported. DAML-S (also called OWL-S in the most recent versions) is the only Web service language that announces the directly connection with AI planning This kind of methods have been reported frequently in recent years, they classify the methods into five categories, namely, the situation calculus, the Planning Domain Definition Language (PDDL), rule-based planning, the theorem proving and others.

**Service Composition using Program Synthesis** **Program** synthesis is a method of software engineering used to generate programs automatically. There are three different approaches to program synthesis: transformational, inductive and deductive program synthesis. Deductive program synthesis is based on the observation that proofs are equivalent to programs because each step of a proof can be interpreted as a step of a computation. This transforms program synthesis into a theorem-proving task. The key ideas of this approach, namely the correspondence between theorems and specifications and between constructive proofs and programs are presented in (Manna et al., 1992).

There are several languages proposed for service composition. The first generation of composition languages - IBM's Web Service Flow Language (WSFL) and BEA Systems' Web Services Choreography Interface (WSCI), were not compatible, so the second generation was developed as Business Process Execution Language

for Web Services (BPEL), which combines WSFL and WSCI with Microsoft – XLANG specification, but regardless of that there is no standards for aggregation, choreography and composition on process level still. By this reason some developments suggested for web services compositions will be considered in the paper. After that their comparison will be made according how they correspond to the requirements for the web service composition.

Several proposals for Web services composition are examined here comparing how they meet some requirements.

## SERVICE-ORIENTED WORKFLOW

The workflow-based composition is based on the Business Process Management (BPM). Business process can be described as a set of activities provided certain business goals, and business applications represent these activities as business process. BPM performs integration of individual applications to form business processes (<http://www.ilog.com/products/jviews/workflow/>). The wide accepted standards for such web services composition are BPEL (Web services Business Process Execution Language) (Andrews et al., 2003), WSCDL (Web services Choreography Description Language) (Kavantzias et al., 2004) and BPML (Business Process Modelling Language) (Aalst et al., 2002).

The BPM methods use WSDL (declared operations and data-typed messages) for extending and supplement of composition scheme. All of these standards are in a category of static composition (design-time). In industry the workflow-based composition is used for service integration within the enterprise (Enterprise Application Integration) and during B2B cooperation.

In the workflow management the application logic is realized by composing autonomous applications. This allows defining of increasingly complex applications by progressively aggregating components at higher levels of abstraction. So the workflows are a collection of coordinated tasks designed to carry out a well-defined complex process. The process modeling in workflow management can be divided into two basic categories:

- activity-based and
- communication-based.

The communication-based approach focuses on the communication between people and the commitments that follow from it; the activity-based approach focuses on the activities people engage in and the results of those activities. Five most relevant workflow aspects are as follows: the functional, behavioral, informational, organizational, and operational aspects.

The workflow management is analogous to composing of web services where the business logic of the client is implemented by several services.

An approach in (Jaeger et al., 2004) identifies abstract composition patterns, which represent basic structural elements of a composition, like a sequence, a loop, or a parallel execution.

The patterns are used to model the structure of a composition. The composition patterns are based on structural elements as used in workflow descriptions. Translation process has to transform the composition into a target web service-oriented workflow language as workflow patterns. The composition model of service composition supports nesting of one composition into another one.

The end-user requirements for the composite service can consist of functional as well as non-functional requirements.

## BPEL

BPEL ([www.ibm.com/developerworks/library/ws-bpel](http://www.ibm.com/developerworks/library/ws-bpel)) is an XML language that supports process oriented composition of web services (Narayanan et al., 2002). It was developed by BEA, IBM, Microsoft, SAP, and Siebel. With BPEL the composition is accomplished by interoperation with a set of web services with the aim to solve some specific task. The result of the composition is a *process*, participated services are *partners*, and message exchange or intermediate result forms an *activity*. In such a way the process contains a set of activities, and its relation with external partners is realized through WSDL interface. For determination of the process the following is used:

- a BPEL source file (.bpel), which describes activities;
- a process interface (.wsdl), which describes ports of a composed service; and
- an optional deployment descriptor (.xml), which contains the partner services' physical locations (a partner service's implementation and location can be changed without modifying the source file).

BPELJ as combination from BPEL and Java (<http://www-128.ibm.com/developerworks/webservices/library/ws-bpelj>) has been realized subsequently which allows to developers to include Java code inside the BPEL file. The Java snippets can be used to implement internal transformation such calculations of different values in documents, make up and decompose documents by using information forms from other documents and variables and useful calculations for the flow controls. They can perform sideward activities without creating Web services.

Interdependences between BPEL and Web-services standards are shown in Gartner Group, 2004. Accepting the BPEL as a technology for web service composition reflects on developing of Enterprise Application Servers, for example Oracle Application server (Bultan et al., 2003), Microsoft BizTalk server (Berardi et al., 2003), and stand-alone tool from IBM, BPWS4J (<http://www.alphaworks.ibm.com/tech/bpws4j>).

Limitations in BPEL rely on expressivity of XML / XML Schema and it is insufficient for task automating.

## SEMANTIC WEB-SERVICES COMPOSITION METHODS

The composition of Web Services is also an issue for contributions about describing services with semantic information.

The semantic description of services can be used as a criterion for the selection or dynamic binding of services. The fundamental precondition for the semantic web is the extending of current web interface into the format that is accessible to software programs. Applying to the web services composition this brings along with automatic selection and execution. Automation is achieved with semantic description of web services that allows advertising of different services and accomplishment of all decisions connected with service composition both from client side and from system side. These decisions include selection of appropriate services, their real combination and define how the composition reflects on criteria specified by the client.

In the literature the process of composing services based on semantic annotations has taken two paths:

- similar annotations using domain ontologies, meta-data, rules, etc.
- methods to combine services whose annotations match based on some notion of similarity.

DARPA's OWL-S (Ontology Web Language for Web services) is a leader in investigation of semantic composition realizations. OWL-S (Dean et al., 2004), (Ankolekar et al., 2001) ontologies give mechanisms for description of functions of web services and make them possible for automatic discovering and integration. In (Sirin et al., 2003) the OWL-based method for dynamic composition is described, where semantic service description is used for discovering of the required from the client service at every stage during the composition as its execution is accomplished through direct access to the service (grounding). In another method the list of tasks is created by using Artificial Intelligence planning technique, which combines service selection and their appropriate integration according to the client request.

In (McIlraith and Son, 2002) the Golog – AI planning Reasoner is used for automatic composition while in the same situations other methods (Wu et al., 2003), (Nau et al., 1999) realize automatic service composition through Hierarchical Task Network (HTN) planning paradigm.

At present despite the enthusiasm of many semantic web research groups the way for creating of flexible frame of the interoperation of intelligent agents and proposed tools for delivering of meaning in Semantic Web Services is only outlined.

The goal of semantic services composition languages is to allow discovering, composing and invoking be satisfactory when the complex services are used. DARPA Agent Markup Language (DAML) extends XML and Resource Description Framework (RDF) to ensure set of concepts for creating of machine understandable ontologies and information mark-up. This language is matched with Semantic Web and is a Web Ontology Language for Services ([www.daml.org/services](http://www.daml.org/services)). OWL-S (known before as DAML-S) can provide for automatic discovery, invocation, composition, interoperation, and execution monitoring (Ankolekar et al., 2002).

The OWL-S models ontologies are divided in three parts:

- service *profile* – describes what the service requires from the clients and what it can give to them;
- service *model* – specifies how the service acts;
- service *grounding* – give information how to use the service.

The process model is a sub-class of the service model, which is in terms of inputs, outputs, preconditions, postconditions and is necessary in terms of its own sub-processes. In the process model it is possible to describe composition of processes and their processes, dependencies and interactions. In OWL-S processes are grouped into three types: *atomic*, - the process that does not contain subprocesses; *simple* to which it is no direct addressing, but it can be used as abstract element in atomic process and in composed process; and *composite* process that contains sub processes. The compound process is specified by using flow-control constructs:

- sequence, split, split+join,
- unordered, choice, if-then-else, iterate
- repeat-until.

In (McIlraith and Son, 2002) methods for transformation of OWL-S descriptions in Prolog, and in (Narayanan and McIlraith, 2002) – into Petri-net based notation for preparing verification analysis are examined. The transformations into Prolog are made manually, which give the possibility for discovering for appropriate solution (plan) for web services composition according the goals of the description. With given repository of web services the inference rules can be used for automatic service selection for the given task. The transformation to Petri-net is automatic and is suitable for the tasks of simulation, validation, verification, composition and performance analysis.

In <http://www.w3.org/Submission/OWL-S-related> the relationship of OWL-S to selected Web services and Semantic Web technologies is discussed, in the hope that it may clarify the possible using of OWL-S, and perhaps it gives some guidance to its potential place in the realm of activities at W3C. The technologies and languages discussed are: WSDL, SOAP, UDDI, BPEL4WS, CDL, ebXML, OWL, WSMO.

A number of initiatives based on the WSMF framework (Fensel and Bussler, 2002) have started. WSMF is an extension of the UPML framework (Fensel and Motta, 2001) revised to integrate fully with web services and to support ecommerce.

Three related initiatives associated with WSMF have recently begun. These are WSMO (WSMO, 2004) which is developing ontology for WSMF, WSML (WSML, 2004) that aims at providing a formal language for representing the WSMO based descriptions and WSMX (WSMX, 2004) developing a reference implementation. The IRS-III (Domingue et al., 2004) ontology is basically an implementation of WSMO standard D2v02 (WSMO, 2004) with some additional attributes.

## WEB COMPONENTS

The method of Web components (Yang and Papazoglou, 2002) treats services as components that support the main principles of software development as reuse, specialization, and extension. The basic idea is to find the composition logic of information that represents the method of Web component. The general interface of the Web-components can be published and used for discovery and reuse. The composition logic comprises composition types and corresponding messages. The composition types can have two forms:

- Order – defines how the component performs the compound services: consecutively or in parallel.
- Alternative execution – shows when the given component can invoke alternative service for reaching the goal.

Corresponding messages determines input and output of composed message and are formed in three types:

- Synthesis generates a composed service's output message by combining the output messages of constituent services.
- Decomposition binds the composed service's input messages into the input messages of constituent services.
- Message mapping allows custom mapping between constituent services' inputs and outputs.

The Web component approach supports several basic composition constructs: *sequential*, *sequential alternative*, *parallel with synchronization of results*, and *parallel alternatives*. They are augmented with *condition* and *while-do* constructs. Composition integrates (through calls) several pre-built components or software parts.

## ALGEBRAIC PROCESS COMPOSITION

Algebraic Composition of resources aims at achieving a description like other methods. It models services as mobile processes to ensure verification of such properties as liveness (correct termination, for example) and resources management. The theory of mobile processes (Milner et al., 1993) is based on  $\pi$ -calculations which main essence is

- the process itself (it can be empty);
- selection of several I/O operations and their duration;
- parallel composition;
- recursive definitions or recursive addressing.

I/O operations can be input (receive) or output (send). Some processes can be accomplished in parallel and they can use compatible channels. Describing the services in such a way is necessary for checking of composition correctness (Meredith and Bjorg, 2003). With algebraic process composition, the general question is what information to type. Typing too little can make it impossible to verify some properties, such as security. On the other hand, typing too much creates a complexity that renders verification unusable or impractical.

## PETRI NETS

Petri nets are proved method for process modeling. They represent an organized two side connected graph in which the nodes are places; transition and symptoms have these places. When at least one symptom is present in every connected place for transition then the transition is possible. One possible transition can start by putting of one symptom (symbol) to every input place and putting every symptom to every output place. The services can be modeled by Petri nets through appointment transition being systems and state places (Hamadi and Benatallah, 2003). Every service is connected with Petri net, which describes the behavior of the service and possesses two ports - input and output. In every moment the service can be in one of the following states: *not instantiated*, *ready*, *running*, *suspended*, or *completed*. After the defining the net for **every** service the compositional operators perform sequence, alternative (choice), unordered sequence, iteration, parallel with communication, discriminator, selection, and refinement. The correct ending of the composed service is from great significance, so the confirmation is necessary.

## MODEL CHECKING AND FINITE-STATE MACHINES

This method for web service composition consists of controlling and modeling of the composition as Mealy machines and automatic composition by using finite-state machines (FSM). The model verification is used for checking of the formalization of finite-state coordinated systems. The temporal logic is used for describing the system specification, following by refutation and verification for model checking to see whether the specification is supported (Fu et al., 2002). It is recommended also conversation *specification* for web service composition (Bultan et al., 2003). The method models services as Mealy machines, which are FSM with input and output. The services are connected by sending of asynchronous, and every service in the queue waits for further instructions. A global observer follows the course of events and keeps in touch with the situation with all messages. The interaction begins as a sequence of messages. By investigating and understanding of interoperations properties of this method it is possible to provide new means for analysis and design of good accomplished service composition. In (Berardi et al., 2003) a framework for describing of web services behavior as a tree on implementations is presented. After that it produces transformation to the FSM. The algorithm for verification of existing composed web services and retrieving those of them that are proved to be suitable is proposed. During composition process the correct composition and complex calculations are ensured which lead to automatic composition with confidence that it will be accomplished by finite number of steps.

## HYBRID METHODS

Great efforts for realizing of automatic service composition by their semantic description are in progress in parallel but isolated from the development of workflow-

based standards, which are preferred from the industrial organizations. These organizations would rather use ready made composition techniques that support their business needs and are oriented to their specific requirements than immature dynamic service composition that is oriented first and foremost to the automation of the service composition process. The hybrid solutions are presented in (Mandell and McIlraith 2003; Traverso and Pistore 2004) that connect the two methods and combine their advances by introducing semantics to workflow-based composition. This alliance of semantic with business is shown in (Akkiraju et al., 2005; Sivashanmugam 2003), where the semantic annotation is accomplishing inside WSDL-files to make the service discovery and selection easier. Another hybrid method (Osman et al., 2005) uses ontologies together with WSDL-files for describing the service domains and defining incrementally any mismatches in the provider's services. The logical execution for the unification of domain specific WSDL and domain specific ontologies elements is supported by using ontology reasoner and coding in membership verification module. The membership module scans the service participant's ontology files for equivalent properties. In (Mandell and McIlraith 2003) the method of interpretation of web services in BPEL is proposed; the OWL-S based description is used for connecting partners' services in run-time. The execution contains OWL-S-profiles from given repositories and uses semantic profiles for estimation of needed partners during selection of desired properties. The method ensures selection of partners in run-time; in contrast to the corresponding selected in design-time BPEL model of the process. The execution includes SDS (Semantic Discovery Service) module, which works as a broker for the semantic service discovery. SDS is a connection between BPEL process engine (BPWS4J) and Web-service partners. The method uses Semantic Web Technology for automatic meaningful selection of services. However, the problem for realization of real automatic composition is not discussed because of the composition logic is constructed manually for the selected services. In (Osman et al., 2005) the composition produced by the service composer is considered. This composer categorizes possible partners according to their domains and provides to them domain specific interface (WSDL+OWL). This interface works as necessary condition in relation to the domain. By this method first the expected requirements are declared and after that the domains are fulfilled with appropriate services in contrast to the method described in (Osman et al., 2005), which uses OWL-S profiles for the selection of partners provided services through service description. The method (Osman et al., 2005) allows constructing general program framework for service selection from particular domains and their automatic composition.

## REQUIREMENTS FOR WEB SERVICES COMPOSITION

The access point to documentation or code (either source or binary) for service-oriented computing (SOC), for developed applications and for users is realized only

through basic (primary) WSDL functional description. The services are performed in different containers and the composition mechanism has to satisfy at least the following requirements: connectivity, nonfunctional quality-of-service properties, correctness, and scalability (Milanovic and Malek 2004). Every method for service composition has to guarantee connectivity. Through checking the connectivity it is possible to define which services are composed and what are the reasons for input/output messages. This is because of the web services are based on message exchange but during development it is also necessary to address nonfunctional QoS properties, such as timeliness execution, safety and security. Composition correctness requires properties verification of composed services. Beside that the complex business transactions may cause very complicated services in comprehensive access chain. The composition framework has to be coordinated with possible large total number of composed services.

## METHODS COMPARISON

### *Connectivity and Nonfunctional Properties*

All methods ensure connectivity. Though the services themselves are modeled in different manner then the interaction at the low level is reduced to mapping and orchestrating of input and output messages between ports of different partners services. Most methods do not take notice of nonfunctional QoS properties, such as security, dependability, or performance characteristics. Only OWL-S allows to set some nonfunctional properties (namely, quality of service), but these potential possibilities have not been completely specified yet.

### *Composition Correctness*

Correctness verification depends on the services themselves and specification of the composition. BPEL and OWL-S have no possibilities for correctness checking. BPEL is a Turing-complete language related more to operation (execution), than to specification so it is difficult to provide formalism for correctness verification of BPEL flows. Other methods provide means for verification in other ways. OWL-S, when is combining with Prolog or Petri nets, ensures estimations for correctness but the methods for such checking are different for Prolog and Petri nets.

The method of Web components proposes means for compatibility and coordination verification.  $\pi$  – calculations provide for powerful algebraic proofing for determination of liveness, security and quality of services. Applying of such proofing however depends on when and how the services are modeled as processes. Petri nets use complex algebra for verification which is compatible with  $\pi$  – calculations. There are methods that can confirm that specification of service composition come nearer to the model although their using may be connected with some difficulties on calculation and time-consuming testing.

## Automatic Composition

Methods for service composition are oriented to automatic composition that promises faster application development and flexibility during interaction of users with the complex set of services. Using automatic composition the end user or application specifies the goal as the main problem is how to identify candidates for services, to compose them and to verify them so they replay to the user request. The FSM method is considered as the most promising for automatic composition.

## Composition Scalability

The methods for composition support connectivity through messages passing via the ports. Accomplishment of composition of two services differs from composing of 10 or 100 services. So it is interesting how the proposed methods relate to the number of services being integrated. Using BPEL in composition of many services the process is prolonged, because XML-files became bigger. Unfortunately BPEL has no standard graphical notation yet. Some orchestration proposes UML-like notations. There are some such publications for OWL-S. The Web component method provides for good scalability by class of definition, but the additional time is needed for combination and synchronizing of definition class and XML. The  $\pi$ -calculus method proposes short annotations with powerful reducing mechanisms that facilitate complex service specification. The Petri net are not very scalable modeling technique. FSM models depend on the checker type and machine state operations and can be better in this respect from Petri net and compatible with  $\pi$ -calculus method.

## CONCLUSION

The methods to be preferred for service composition are those that cover the all four requirements, outlined in (Milanovic and Malek 2004), to them. The main problem for methods aiming at becoming industrial standards (like BPEL and OWL-S) is correctness verification. From the other side the formal (abstract) methods do not reflect industrial needs. From the correctness verification point of view they have a certain advantage, but it is necessary to pass from WSDL and SOAP to elegant mathematical solutions formalized in abstract methods. It is expected that the perspective research works will be oriented to adapting and jointly using of the most proved investigations among abstract and industrial methods for the web services composition.

## ACKNOWLEDGEMENT

This work is carried out under EU Project INFRAWEB - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

## REFERENCES

Aalst, van der W.M.P., Dumas, M., ter Hofstede, A.H.M., Wohed, P., "Pattern-Based Analysis of BPML (and WSCI)", *QUT Technical report, FIT-TR-2002-05*, Queensland University of Technology, Brisbane, 2002.  
Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., and Verma, K. "Web Service Semantics - WSDL-

S, "A joint UGA-IBM Technical Note, version 1.0, April 18, 2005  
Andrews, T., Curbera, F., Dholakia, H., Golland, Y., Klein, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., and Weerawarana, S., "Business Process Execution Language for Web Services, Version 1.1", <http://www-128.ibm.com/developerworks/library/wsbpel/>  
Ankolekar A. et al., "DAML-S: Web Service Description for the Semantic Web," *Proc. Int'l Semantic Web Conf. (ISWC)*, LNCS 2342, Springer-Verlag, 2002, pp. 348-363  
Berardi D. et al., "Automatic Composition of E-Services that Export Their Behavior," *Proc. 1st Int. Conf. Service-Oriented Computing (ICSOC 03)*, LNCS 2910, Springer-Verlag, 2003, pp. 43-58.  
Bultan T. et al., "Conversation Specification: A New Approach to Design and Analysis of E-Service Composition," *Proc. Int. World Wide Web Conf. (WWW 2003)*, ACM Press, 2003, pp. 403-410.  
Dean, M., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P. F., and Stein, L. A., "Semantic Markup for Web Services", OWL-S version 1.1, Nov 2004, <http://www.daml.org/services/owl-s/1.1/>  
Domingue J, L. Cabral, F. Hakimpour, D. Sell, and E. Motta. IRS-III: A Platform and Infrastructure for Creating WSMO-based Semantic Web Services. *In Proc. of the Workshop on WSMO Implementations (WIW 2004)*, 2004  
Fensel and Bussler, 2002 The Web Service Modeling Framework WSMF. *Electronic Commerce Research and Applications* 1(2): 113-137  
Fensel and Motta, 2001 Structured Development of Problem Solving Methods, *IEEE Transactions on Knowledge and Data Engineering*, 13(6):9131-932  
Fu X., T. Bultan, and J. Su, "Formal Verification of EServices and Workflows," *Proc. Workshop on Web Services, E-Business, and the Semantic Web (WES)*, LNCS 2512, Springer-Verlag, 2002, pp. 188-202.  
Hamadi R. and B. Benatallah, "A Petri-Net-Based Model for Web Service Composition," *Proc. 14th Australasian Database Conf. Database Technologies*, ACM Press, 2003, pp.191-200.  
Hepp M., F. Leymann, J. Domingue, A. Wahler, and D. Fensel, Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management, Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management, *Proceedings of the IEEE ICEBE 2005*, October 18-20, Beijing, China, pp. 535-540  
Jaeger, Michael C., Gregor Rojec-Goldmann, and Gero Muehl, QoS Aggregation for Web Service Composition using Workflow Patterns, [user.cs.tu-berlin.de/~michi/resources/edoc04-jaegeretal-qosagg.pdf](http://user.cs.tu-berlin.de/~michi/resources/edoc04-jaegeretal-qosagg.pdf)  
Kavantzaz, N., Burdett, D., Ritzinger, G., and Lafon, Y., *Web Services Choreography Description Language (WS-CDL) V. 1.0*, <http://www.w3.org/TR/2004/WD-ws-cdl-10-20041217/>  
Lee Y, et al, Compositional Knowledge Management for Medical Services on Semantic Web, *WWW 2004*, May 17-22, 2004, New York, New York, USA.  
Manna Z. and R. Waldinger. Fundamentals of deductive program synthesis. *IEEE Transactions on Software Engineering*, 18(8):674-704, 1992  
McIlraith S. and T.C. Son, "Adapting Golog for Composition of Semantic Web Services," *Proc. Int. Conf. Principles of Knowledge Representation and Reasoning (KRR 02)*, 2002, pp. 482-493.  
Meredith L.G. and S. Bjorg, "Contracts and Types," *Comm. ACM*, vol. 46, no. 10, 2003, pp 41-47.



- Milanovic N. and M. Malek, "Current Solutions for Web Service Composition", *IEEE INTERNET COMPUTING*, [www.computer.org/internet/](http://www.computer.org/internet/) NOV DEC 2004
- Milner R., "The Polyadic  $\pi$ -Calculus: A Tutorial," In: *Logic and Algebra of Specification*, F.L. Bauer, W. Brauer, and H. Schwichtenberg, (eds.), Springer-Verlag, 1993, pp. 203–246.
- Narayanan S. and S. McIlraith, "Simulation, Verification and Automated Composition of Web Services," *Proc. Int. World Wide Web Conf. (WWW2002)*, 2002, pp. 77–88.
- Nau, D. S., Cao, Y., Lotem, A., and Muñoz-Avila, H. "SHOP: Simple Hierarchical Ordered Planner". In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp.968—973, 1999
- Rao J. and X. Su. Toward the composition of Semantic Web services. In *Proceedings of the Second International Workshop on Grid and Cooperative Computing, GCC'2003*, volume 3033 of *LNCS*, Shanghai, China, December 2003. Springer-Verlag
- Sirin, E., Hendler, J., and Parsia, B. *Semi Automatic Composition of Web Services using Semantic Descriptions*. Proc. ICEIS-2003 Workshop on Web Services: Modeling, Architecture and Infrastructure. Angers, France, April 2003
- Sivashanmugam, K., The METEOR-S Framework for Semantic Web Process Composition. *Master's thesis, Department of Computer Science, University of Georgia*, 2003
- Taha Osman, Dhavalkumar Thakker and David Al-Dabass, "Bridging the Gap between Workflow and Semantic-based Web services Composition" *WWW Service Composition with Semantic Web Services* (wscomps05) France, September 19, 2005, in conjunction with the 2005 *IEEE/WIC/ACM International Joint Conference on Web Intelligence (WI 2005) and Intelligent Agent Technology (IAT 2005)* pp. 13- 24
- Traverso, P., Pistore, M. "Automated Composition of Semantic Web Services into Executable Processes", in *Proceedings of Third International Semantic Web Conference (ISWC2004)*, November 9-11, 2004, Hiroshima, Japan, pp.380-394
- WSML, 2004 Languages for WSMO, <http://www.wsmo.org/2004/d16>
- WSMO, 2004 Web Service Modeling Ontology – Standard, <http://www.wsmo.org/2004/d2>
- WSMX, 2004 Overview and Scope of WSMX, <http://www.wsmo.org/2004/d13/>
- Wu, D., Parsia, B., Sirin, E., Hendler, J., and Nau, D. "Automating DAML-S web services composition using SHOP2", In *Proceedings of 2nd International Semantic Web Conference (ISWC2003)*, Sanibel Island, Florida, October 2003
- Yang J. and M.P. Papazoglou, "Web Component: A Substrate for Web Service Reuse and Composition," *Proc. 14th Conf. Advanced Information Systems Eng. (CAiSE 02)*, LNCS 2348, Springer-Verlag, 2002, pp. 21–36.

# AN APPROACH FOR SEMANTIC WEB SERVICE COMPOSITION

Tatiana Atanasova, Hristina Daskalova,  
Institute of Information Technologies -BAS, Acad. G. Bonchev 2  
1113 Sofia, Bulgaria  
E-mail: [atanasova@iinf.bas.bg](mailto:atanasova@iinf.bas.bg) [daskalovahg@abv.bg](mailto:daskalovahg@abv.bg)

**KEYWORDS:** Semantic Web Services, Design time Composition, Run time Composition, Service-oriented workflow, WSMO

## ABSTRACT

In the paper the Semantic Web Services (SWS) composition is considered as a process of manipulating of complex service-oriented ontology-based workflows.

The task for the composition is formulated as combining of existing semantic web services into package of services using service templates as composition patterns. Fulfillment of this objective is achieved by developing of workflow-based tool for graphical creation of SWS compositions using WSMO services and domain knowledge represented as WSML ontologies. The main aim of the proposed approach is to facilitate the reusing of software services.

The Semantic Web Services Composer is a part of Semantic Web Unit (Atanasova et al, 2005) within the INFRAWES project. The SWS INFRAWES Composer is considered in the Design time and in the Run time. The Design-time Composer is intended to construct templates of semantic services combinations. On the base of these pre-defined plan templates and logical discovery the Run-time Composer will realize binding of concrete web services and will assist user in constructing its own composite service. Both Design-time and Run-time SWS Composers use the previous service compositions.

## INTRODUCTION

The area of web service composition attracts the attention of numerous researches. Current solutions for service composition based on business web services (using WSDL, BPEL, SOAP etc.) or semantic web services (using ontologies, goal-directed reasoning etc.) are both piecemeal and insufficient (Au et al, 2005). There is still no general solution. The web services composition is not standardized yet. The main requirements to which every method has to be compared to ensure stable and reliable solution for web services combining are not defined too.

It can be pointed that the information modeling for the semantic web services composition is challenging (Kumar et al, 2005) in the following aspects:

- planning methods in representation of complex actions,
- handling of rich typed messages,
- dynamic object creation,
- specification of multi-partner interactions

- visual representation of the service composition is also an aspect of future development
- up to now (d'Aquin et al, 2005) there are no case-based realizations in the area of the semantic web.

It is agreed that semantically rich descriptions are necessary preconditions for user-friendly discovery automatic management, and composition of entities in web applications.

In the paper the Semantic Web Services (SWS) composition in the ongoing EU/FP6 INFRAWES project is considered. The composition has two aspects – design time and run time. The design-time composition is performed from the service provider side and constructs service templates using business logic and abstract service definition. The run time composition takes place on the client side and provides binding services to selected abstract services templates assisting user in creating its own package of services. As a framework for the Semantic Web Services description the WSMO (WSMO Specification, 2005) methodology is accepted.

## BUSINESS USE CASE REQUIREMENTS

In the paper the semantic web services composition will be considered on the base of e-business like classical travel agency and e-government use cases.

The business scenarios can be described with following characteristics:

- The scenarios are based on business logic;
- The processes are spanned across several organizations (hotels, airplane companies, car rent, etc. and different authorities);
- The scenarios are independent on the technology realization;
- The catalogue/registry should contain established packages;
- The scenarios combine complex and heterogeneous services.

## WORKFLOW-BASED COMPOSITION

Workflow provides a formal method for expressing the tasks that need to be completed in order to fulfill some goal. Workflow has been proven method for describing metamodel for business processes. It is closely connected to the composition problem (van der Aalst et al, 2003). Traditional workflow provides very basic model for the resources involved in a business process. A resource is an

entity that can execute the task. Task execution is outside the workflow management system. The traditional workflow only partially is compatible with service-oriented engineering of resources.

Workflow composition is based on the idea that a workflow as sequence of tasks makes a plan implicitly defined by the pre-conditions and post-conditions of all the tasks in the domain. The initial and final states of such a plan are determined by the business goal requirements. Composition of tasks can transform an initial world state into a final required world state.

Encoding business rules into the workflow template is possible for many business applications where the steps in business process are fixed. But partners in transactions are selected at run-time and services used for a specific instance may vary.

It is pointed that there are different consideration about using workflow templates for web services:

- the activities in the workflow may be bound to an existing service in design time or
- the activities may be defined in terms of abstract definitions that will be matched with available services at run-time.

Each service provider has its own business rules for processing of web services that can be encoded into workflow. Autonomous and heterogeneous units have to be coordinated and data flow of compositions has to be managed which makes the implementation of web services composition hard.

A promising solution for processing of the complex model generated via composition of particular services is involving semantic. Past knowledge may be reused in the workflow design and synthesis of new solutions. The reuse of existing, verified compositions may help to avoid errors during semantic service composition. But the representation of previous composition sets some questions that have to be resolved:

- how to refer specifically to the services used inside this composition;
- how to classify a template's suitability to the user for the adaptation process;
- how to quantify the appropriateness of a structure.

## APPROACH DESCRIPTION

INFRAWEBS SWS Composition deals with the combination of different semantic services to obtain a new complex Semantic Service. The Semantic Web Services (SWS) Composer is a part of Semantic Web Unit (SWU) (Atanasova et al, 2005) within the INFRAWEBS project (Nern et al, 2004). The Composer is only one module in the INFRAWEBS framework. The main INFRAWEBS project focus and objective is the development of an application-oriented software toolset for creating, maintaining and executing WSMO-based Semantic Web Services (SWS) within their whole life cycle. The specifically developing tools aim at providing such

functions as design, discovery, storing and maintenance of semantic descriptions, monitoring and execution.

SWS use ontologies to describe web service. The semantic description provides information about input/output types and logical constraint, internal structure of the process and encoded business rules.

The Case-based memory (Agre et al, 2005) is used in INFRAWEBS Composer to support workflow model reusing during the workflow design. The composed semantic service is a service again; it is described semantically and it is published in a semantic repository. The preconditions and effects of semantic web services functionality may be used for reasoning.

The INFRAWEBS Composition model retains overall structure of traditional workflow, but instead tasks the semantic web services are involved. It should be mention that:

- The tasks from a traditional workflow are equivalent to the service description in the Composer;
- The interaction between different services realized specific activity is independent from providers that cover these activities - flight booking is an independent as a role from the company that proposes the service;
- The roles of participated partners are presenting in the INFRAWEBS Composer as abstract services.
- The granularity and the complexity of involving services can vary.

The SWS INFRAWEBS Composer is considered in the Design time (from complex service provider side) and in the Run time (from the client side). The INFRAWEBS Design-time Composer is intended to construct plan templates of semantic services combinations. On the base of these pre-defined plan templates and logical discovery the INFRAWEBS Run-time Composer will realize binding of concrete web services and will provide the intelligent assistance to the user.

Both INFRAWEBS Design-time and Run-time SWS Composers use the previous service compositions. Such compositions are represented by Service Composition Templates (SCT), which semantically describe control and data flows between several sub-services.

The INFRAWEBS Design-time Composer first produces abstract semantic service descriptions that represent the roles of the services. The abstract semantic services are composed by the provider as a semantic service-oriented ontology-based workflow that satisfies the desired functionality.

The INFRAWEBS Run-time Composer is a part of semantic service composition oriented to finding of corresponding service instances to the given process template. The aim of the Run-time Composer is to bring forward process of dynamic cooperation between independently developed services. This is a non-trivial problem and involves a number of issues related to QoS optimization, satisfying non-functional requirements, data

flow orchestration, data type matching and invocation protocol matching. While some of these issues can be resolved in an automated manner, others might require human interaction from a developer supervising the composition process.

The Run-time composition involves following activities:

- discovery of appropriate services
- accommodate inter service dependencies
- selecting suitable services to bind for a given process.
- assistance to the user on the base of SCT

## DESIGN TIME COMPOSITION

The task of the design-time composition is formulated as preparing of composition templates on the base of workflow management systems that implement the desired functionality and publishing such semantic Service Composition Templates (SCT) in the local repository of semantic services.

Workflow in INFRAWEBs Composer is considered as goal-oriented process that is reflected on expected user behavior. There is a template that corresponds to the predefined goal. The templates are base for discovering and binding of web services that implement the desired functionality.

As a result of design-time composition new semantic web service description is produced. The consideration the complex composite service as only a service allow us to avoid the problem of defining of orchestration in WSMO (that is not specified yet) and to use a choreography specification to link different services.

The general flow of INFRAWEBs composition process is defined as:

First the desired functionality is identified. Then a query is formulated to the case-based memory. In the result of the similarity matching between the query and service description the appropriate services is retrieved from semantic service repository. The selected services are used for creating of workflow that encodes the desired business logic. The generated complex service is publishing in the repository.

The composition process in the design time consists of the following main steps:

- Design of SCT of composed services according to the desired functionality;
- Associating semantic services with corresponding ontologies to the SCT components.

In the INFRAWEBs Composer there is a difference between service roles and instances. In the beginning the composition of services is produced as a general plan by Design-time Composer using service roles (for example, travel plan: flight booking, hotel reservation and renting a car, followed by confirmation and payment). Then the plan is concretized into an executable plan by selecting the

appropriate services instances that is governed by the Run-time Composer.

So the Composer needs:

- repository of services with their semantic descriptions;
- description of goals as requirements for the composed services;
- tools for graphical designing a composition of available services with desired functionality;
- tools for indexing and publishing of the composite service.

The SCT are based on description of services combined into designed workflow and corresponding ontologies from the application domain. The components that contribute to a composite service can be distributed, but application provider (travel agency, for example) employs a central control point.

The INFRAWEBs Design-Time Composer separates the workflow logic and the implementation technology that means the indirection between a capability described by abstract service and web service that implements this capability. The following consideration discusses this in more details.

### Creating of abstract service description

For generating an abstract template we need first to create abstract descriptions from the service instances before composition. The abstract service description has no instance variables nor is it grounded. The variables are subsumed by ontologies concepts.

The proposed algorithm is:

```
while  $\exists \forall$  instance variable
  refer to corresponded ontology
  replace instance variable with the
  ontology concept
end
```

All variables instances in SWS description are replaced with concepts from corresponding ontologies (Figure 1).

### Abstract components retrieving/indexing/publishing

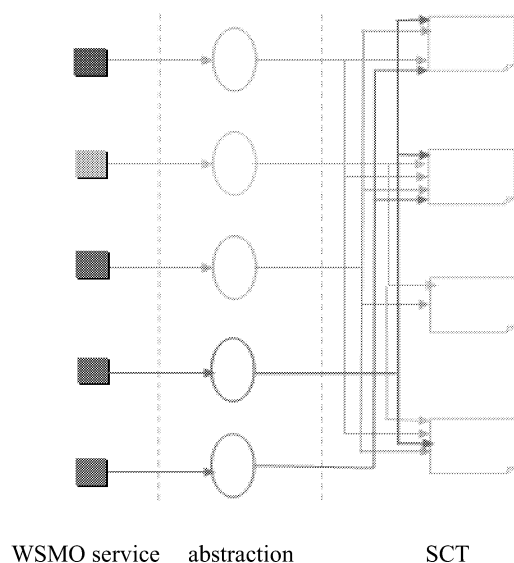
According to WSMO, a template is described by its non-functional properties, a set of imported ontologies as well as its capability. The abstract template is considered as a service again, so it is natural to suppose that all INFRAWEBs steps for publishing, indexing and querying can be applied to it.

### Creating of abstract service description

Abstract service definitions (ASD) should be generic enough and some of the features should be fixed to allow dynamic discovery and automatic accomplishing of possible matching. ASD represents a role that the semantic service should play in the constructed workflow.

When an existing composition has to be used to achieve some goal, part of or all of the composition is retrieved, appropriately abstracted and then stored in repository. It

can be retrieved a combination of services that can provide the required functionality.



**Figure 1. SCT creation from abstract service descriptions**

### Defining composition patterns

A SCT as semantic workflow is an abstraction of a business process in the INFRAWEBS Composer. It comprises a number of logic steps, dependencies among services, routing rules, and participants.

In WSMO framework a service state is described by an ontology (Roman et al, 2005), and transition rules express changes of states. For every composition template (workflow pattern) the abstract WSMO service is associated with its Non-Functional Properties (NFP), Capabilities and Interface.

Templates can be indexed and retrieved by:

- type of service that they present;
- main ontological concepts used for their definition;
- keywords used for natural language description.

### Creating of Composition Pattern (SCT)

A specific set of workflow patterns defines an orchestration among the integrated components. A workflow of business processes in INFRAWEBS Design-time Composer is constructed according to the use cases. When creating a new SCT (travel package), the identification of the functionality originates from the problem that needs to be solved.

Composition patterns creation is based on Aalst work (Aalst et al, 2002). The approach is modified by involving of components technology to make it possible to include semantics in the workflow description. The SCT (semantic workflow) is constructed in such a way that every workflow pattern is binding with one operation from the given operation set. So the abstract components are used

according to the most suitable parameters. The service availability in run time is not considered during the design time composition.

The composite service SCT (Package) is modeled by starting from scratch, designing of graphical view of the process as a workflow, assigning for the each workflow pattern appropriate semantic abstract service and then publishing the constructed template (Figure 2).

The template is composed of one or more semantic services which can be situated in own semantic repository or in partner's repository.

The WSMO descriptions are queried from the repository and the matching is applied. Messages and operations of services are associated with workflow domain concepts.

### Publishing composition patterns

The SCT (Package Description) is stored in the semantic repository and indexed in distributed registry.

### Choose a Stored SCT

Choosing of stored SCT is made using Case-Based Memory. Indexing of stored SCT is provided according service ontologies and taxonomies and similarity measuring between packages reflects both semantic and structural similarity.

SCT that use the same template may differ in the set of constraints - part of the axioms and functions of the ontologies in which the services are defined.

The existing template SCT can be modified and adjusted.

### Characteristics of Design-time Composition process

- Elements for composition are semantically annotated services
- The composed process has an explicit control flow
- The service templates are grouped in taxonomies
- The template may be selected from a list
- The template is creating by graphical tool

### Case-based Memory Usage during Design Time Composition

The INFRAWEBS Design-time SWS Composer is intended to explore the reuse of existing semantic services descriptions. The reusing will lead to more efficiency in the composition process. The application development can be accelerated and the user's needs can be more complete satisfied (Limthanmaphon and Zhang 2003). Case-based memory can provide advices for service composition at the process- and component-related level.

Semantic Web Services are available and usable by advertising their capabilities. Describing services in terms of their capabilities allows SWS services be located on the base of service ontologies. Expressing services as blocks of components (WSMO) is another point in the Composer. For reusing of templates the descriptive and functional service aspects has to be taken into account: NFP and domain ontologies, operations, parameters, concepts in

transitional rules, etc., because the goal of the template is to define a sequence of appropriate services.

The minimum of the required information is the operation name, its input and output. All of them may be described using ontologies concept that allows ontological querying. Only metadata describes all service aspects for abstract service description. There are no implementation details.

### Queries formulation

The Semantic Web Services have NFP, Capabilities and Interface. The Capabilities and Interface are logical terms and expressions, and may be used for reasoning about the retrieving of the process template too.

### Matching and ranking

Finding the requested service is based on the matching process and ordering of the services by their ranks.

Two-step Discovery process used in INFRAWEBs is exploited here. Ranking and ordering is also provided by the discovery mechanism.

### Case-Based Semantic Web Services Composer

The presumption is that Service provider maintains a repository of semantic web services. When a new complex service requirement arises, it can be expressed in the context of domain ontology. The service composition environment can then be used to generate potential workflows for achieving the desired functionality by reusing existing web services, aggregating several simpler services into a higher-level service – meta-service.

Composition of multiple semantic web services into an integrated service consists of following steps:

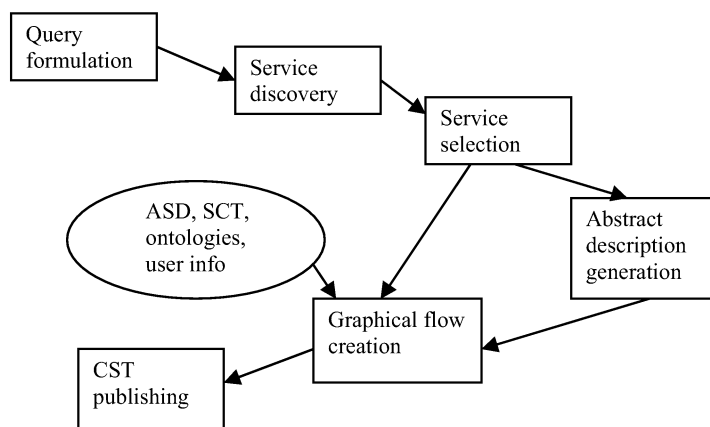
1. services are described semantically, classified into categories and put into registry.
2. the composed “meta” service is specified; the most suitable combination of semantic web services is defined across several businesses and shared interfaces.
3. the composed service is published in the repository.

In the case-based memory the web services are organized in taxonomies that provide means for an ontological organization of the domain available service space.

The case-based memory is a meta-meta-model layer, WSMO ontology represents meta-model, WSMO element (service instantiation) is a model. Defining a service as  $S = \{Name, Description, NFP, Choreography, [Orchestration]\}$  allows representing a composite service template as feature vector.

The process of service composition is case-based, which means that the user may ask the Case-based Memory for providing of some services and templates similar to that with expected functionality.

Asking for the assistance is implemented as a process of satisfying the user query describing what will be the service composition about.



**Figure 2. Steps for SCT creation**

### General Requirements to the Design-time SWS Composer

Ontologies and semantic descriptions are basic elements in the abstract service templates (patterns). They allow reusing services.

The existing semantic web services are categorized in taxonomies and they are registered in SWS repositories. For each category of services there are several different services. These services may have different structures and interaction models. Any service is implemented as WSML-file with variables and concepts of ontologies.

Abstract activities (tasks, services, operations) in templates can be used to specify the characteristics of services and services can be discovered and used to generate semantic workflow. The abstract service definition is not connected to the concrete grounding.

A composition template (workflow) describes the activities that are needed to perform the task. Generic template can be configured and customized through parameterization with variables for a specific instance based on the user requirements in run-time.

WSML capabilities with associated ontologies are used in defining of composition patterns. The composition pattern describes the abstract services; the concrete services are selected on the base of user input.

As a result of design-time composition new semantic web service description will be produced. The composition templates (ontology-based workflow patterns) can be publishing and discovering.

This abstract workflow is passed to the Run-time Composer that then matches all the component services in the workflow with available implementations.

### RUN-TIME COMPOSER

The role of the Run-time Composer is to ensure environment for combining semantic web services in run time that respond to user needs. The full automation of the

composition process is still the objective of ongoing research activities. In the proposed approach a human holds control of the definition of the composition, but the discovery and invocation of services according to an abstract representation of users' requirements is carried out by the machine. However, the approach introduces additional features such as the intelligent assistance for the user. The tool guides users in a step-by-step composition process. The composition starts with the selection of the desired service from the list received in the result of user's query.

The start point is that the semantic services and SCT already exist. Once the ontology and semantic contents are well prepared and shared, they can be reused to provide knowledge support.

The Run time Composer is an intelligent assistant that help user to select needed service template or construct the desired set of services together with formulating of necessary constraints for their composition.

The interactive workflow composition in Run-time includes 1. semantic web services descriptions with associated ontologies; 2. discovery mechanism for matching of abstract descriptions to concrete services; 3. retrieving of matching templates; 4. suggestions on needed data to be provided; 5. intelligent assistance for the user to select needed service template or construct the desired set of services together with formulating of necessary constraints for their composition.

The intelligent assistance is needed to avoid errors during manual creating of workflow, to make suggestions about additional constraints, to automate the workflow construction by generation of its parts and to support in such a way the development of valid composition with specific data.

According the business use case the user may start from top-down selection of services with their abstracting representation or from selecting already prepared by provider composition of services from different partners and specifying then the concrete data. The semantic descriptions of the WSMO-based workflow components allow reasoning and analyzing on it as whole and in parts according to the input/output parameters, organizing the services in taxonomies, discovery and matching.

#### **Generating potential workflows SCT from composition patterns for achieving the desired functionality and reusing existing semantic web services**

A package of semantic web services is created using the templates and specific parameters for each service obtained from the customer interaction with the interface. We consider Package as a complex Service, but a service again, that will be instantiated in run-time.

#### **Different instance selection**

Concrete semantic services are found by means of Case-based memory and using discovery mechanism for selecting suitable services for the package.

#### **Characteristics of Run-time Composition process**

- A specific instance for every control activity is finding by analyzing constraints
- In order to satisfy the goal a composition process will locate the right set of services combining existing own services with foreign services
- In the proposed approach a human holds control of the definition of the composition
- The intelligent assistance is supported by SCT organized in Case Based Memory

#### **RELATED WORK**

The creation of composite services is investigated in wide range of works. The research on web services (WS) composition is based on such scientific areas as process algebra, theory of automata, temporal logic, situation calculus, AI planning and so on.

In the industry track, initiatives are focused lately on BPEL. The most practical approaches for web services composition use the theory for controlling of workflows as a model of composition process to achieve the formalization of control and data flows. The main shortcoming of this approach is that the static composition is performed in which the service selection and control flow is accomplished manually and preliminarily. Another possibility is to use composition techniques based on the Semantic Web (Akkiraju et al, 2005; Cardoso and Sheth 2005), for example, OWL-S (The OWL Services Coalition) in which ontologies are used to ensure web services description mechanism in machine understandable form for their automatic discovery and integration. This allows realizing dynamic integration of compatible web services and possibility of their discovery in real time in the composition scheme. Developing of these methods is still at the starting phase and unfortunately largely is far behind from the results achieved in static (design-time) composition.

The composition problem is naturally connected with planning. Planning problem is formulated in terms of tasks (task is an abstract activity). The resulting plan is a sequence of operations. The work of (Kumar et al, 2005) differentiates web services types from services instances and introduces the notion of roles to help avoiding of ambiguity in input/output parameters. Hierarchical network planning (HNP) is an example of template-based composition (Sirin et al, 2005). But classical HNP has a lot of restrictions. Both investigations consider OWL-S described services and propose to extend the current OWL-S ontology with additional kinds of services.

The OWL-S composer described by (Sirin et al., 2004) supports users in the composition by narrowing the list of Web services based on the match of their inputs and outputs or by applying filters over their non-functional properties. This approach is, however, too restrictive.

A similar approach is used in CAT (Kim et al 2004). CAT also integrates planning techniques to track relations

among the composition components, but as the OWL-S composer, CAT does not support the use of mediators and control operators.

While OWL-S provides a service composition model the WSMO specification for composition is not specified in details yet. In spite of that the IRS-III (Domingue et al 2004) is a framework and implemented infrastructure which supports the creation of semantic Web services according to WSMO. Three models for orchestration are developed on different paradigms, namely, state-machine model, structured model and dataflow model. The models extend WSMO ontology and represent its orchestration component. Users of IRS-III directly invoke Web services via goals. The capability-driven service invocation extends the WSMO goal and Web service concepts.

Using goals provides a certain level of dynamism for compositions, in a sense that, suitable Web services are discovered during the execution time to fulfill every component Goal. Another important advantage of models and tool is the use of mediators. Mediators are essential features of WSMO and IRS-III. But it is necessary to create specific mediators and reasoning on them. As WSMO is still intensively developing (WSMO Choreography and Orchestration, 2005) its orchestration model may adopt any of the approaches or a combination of them.

## CONCLUSION

At present in spite of enthusiasm of many semantic web research groups the way for creating of flexible frame of the interoperation of intelligent agents and proposed tools for delivering of meaning in Semantic Web Services is only outlined

Web service composition is a very complex and challenging task. Mechanism of workflows as business process management together with semantic web and semantic web services technologies can provide suitable techniques for the knowledge representation that can facilitate the composition.

At present:

- composition of Web Services is static; it is made in design-time;
- dynamic service discovery, run-time binding, analysis and simulation are not supported directly;
- the composition structure is not addressed in the way that reflects the occurring patterns in workflows

We can summarize that SWS Composition within INFRAWEBs:

- is a task of combining and linking of semantic web services;
- is based on workflow-based templates which ensure structure that allows to integrate services, to change parameters without changing the structure of the process;

- should not react to any changes from affiliated partner to the service provider, but use the power of distributed registries;
- needs more investigation on the optimization of quality constraints.

Human interaction is needed in defining the actual logic of service template in the composition model presented. The automation of this task would require a complex planning and reasoning facilities according to specification of WSMO orchestration.

The composer architecture includes development environment with a graphical user interface to compose semantic web service components and it is associated with WSMO-described services.

The SWS Composer uses the previous service compositions that form the abstract services. The similarity-based retrieval of an appropriate semantic services and templates is supported by Case-based memory.

## ACKNOWLEDGEMENT

This work is carried out under EU Project INFRAWEBs - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

## REFERENCES

- Aalst, van der W.M.P., Dumas, M., ter Hofstede, A.H.M., Wohe, P., "Pattern-Based Analysis of BPML (and WSCI)", *QUT Technical report, FIT-TR-2002-05*, Queensland University of Technology, Brisbane, 2002
- Agre, G., T. Atanasova, H. J. Nern, "Case Based Designer and Composer", *Proc. of the 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005*, April 2005, IRIT, Université Paul Sabatier, Toulouse, France
- Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., and Verma, K., "Web Service Semantics - WSDL-S," *A joint UGA-IBM Technical Note*, version 1.0, April 18, 2005.
- Atanasova T., G. Agre, J. Nern. Infraweb Semantic Web Unit For Design And Composition Of Semantic Web Services, *Proc. of the 1st Workshop for "Semantic Web Applications" at the EUROMEDIA 2005*, April 2005, IRIT, Université Paul Sabatier, Toulouse, France
- Au, Tsz-Chiu, Ugur Kuter, and Dana Nau, Web Service Composition with Volatile Information, *Proceedings of 4th International Semantic Web Conference, ISWC 2005*, Galway, Ireland, November 6-10, 2005, LNCS 3729, pp. 52
- Cardoso J. and A. Sheth, Introduction to Semantic Web Services and Web Process Composition, in *Semantic Web Process: powering next generation of processes with Semantics and Web Services, Lecture Notes in Computer Science, Springer*, 2005
- d'Aquin, M., Jean Lieber, and Amedeo Napoli, Decentralized Case-Based Reasoning for the Semantic Web, *Proceedings of 4th International Semantic Web*



- Conference, ISWC 2005*, Galway, Ireland, November 6-10, 2005, LNCS 3729, p.142
- Domingue, J. et. al, IRS-III: A Platform and Infrastructure for Creating WSMO-based Semantic Web Services. *Proc. of the Workshop on WSMO Implementations (WIW 2004)* Frankfurt, Germany, September 29-30, 2004
- Kim, J., Spraragen, M., Gil, Y. An Intelligent Assistant for Interactive Workflow Composition. *In: Proceedings of the International Conference on Intelligent User Interfaces (IUI-2004)* Madeira, Portugal, 2004
- Kumar, A., Biplav Srivastava, and Sumit Mittal Information Modeling for End to End Composition of Semantic Web Services, *Proc. of 4th International Semantic Web Conference, ISWC 2005*, Galway, Ireland, November 6-10, 2005, LNCS 3729, p. 476
- Limthanmaphon B., and Y. Zhang, Web Service Composition with Case-Based Reasoning, *Proc. of the 14<sup>th</sup> Australasian database conference on Database technologies* 2003, Vol. 17
- Nern, H.-Joachim, G. Agre, T. Atanansova, J. Saarela. System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEBs II. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, ISSN 1790-0832, Issue 1, Vol. 1, July 2004, 286-291
- OWL-S, The OWL Services Coalition: OWL-S: Semantic Markup for Web Services, version 1.0; available at <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>
- Roman, D. H. Lausen, U. Keller, J. de Bruijn, Ch. Bussler, J. Domingue, D. Fensel, M. Kifer, J. Kopecky, R. Lara, E.I Oren, A.I Polleres, M. Stollberg. D2v1.1. Web Service Modeling Ontology (WSMO), 10 February 2005 <http://www.wsmo.org>
- Sirin, E., B. Parsia, J. Hendler, Template-based composition of semantic web services, [www.aaai.org](http://www.aaai.org), 2005
- Sirin, E., Parsia, B., Hendler, J. Filtering and selecting semantic Web services with interactive composition techniques. *In IEEE Intelligent Systems*, Vol. 19, Issue 4, 2004, 42-49
- WSML, Web Service Modelling Language, <http://wsmo.org/wsml/index.html>, accessed 2005.
- WSMO Choreography and Orchestration, Roman, D.; Scicluna, J., Feier, C. (eds.): Ontology-based Choreography and Orchestration of WSMO Services, WSMO Working Draft D14, 01 March 2005.
- WSMO Specification, <http://wsmo.org/>, accessed 2005

# USING CASE-BASED REASONING FOR CREATING SEMANTIC WEB SERVICES: AN INFRAWEBS APPROACH

Gennady Agre

Institute of Information Technologies – Bulgarian Academy of Sciences

Acad. G. Bonchev 29a

1113 Sofia, Bulgaria

E-mail: agre@iinf.bas.bg

## KEYWORDS

Semantic Web Services, Case-based reasoning, WSMO framework, INFRAWEBS Project.

## ABSTRACT

The paper presents an approach for using case-based reasoning to facilitate the task of creating complex logical description of a semantic Web service capability according to WSMO Framework. The case-based memory provides the service designer with descriptions of existing semantic Web services similar to her request represented by a query form. Such a form allows expressing the user's intentions on how the different part of the service under design should look like using both natural language and ontology-based words. Some possible approaches for measuring similarity between a query (a description of a problem to be solved) and a case (a description of another, already solved problem) are discussed and several similarity functions are proposed. All of them are asymmetrical, which is argued by a principal difference in the amount of known features used for describing objects to be matched. Evaluation of similarity is based on a specific, non-logical representation both of a query and a service. Such a representation allows an arbitrary WSMO object to be treated as a specially constructed text document.

## INTRODUCTION

Case-based reasoning (CBR) is an AI paradigm that can be synergistically combined with other approaches to facilitate a broad array of tasks (Aha and Daniels 1998). The main intention of CBR is to reuse previous experiences for actual problems. It was the main reason to include CBR into the process of designing and composing of semantic Web services in INFRAWEBS Project (Nern et al. 2005).

The INFRAWEBS Framework consists of coupled and linked INFRAWEBS semantic Web units (SWU), whereby each unit provides tools and system components to analyze, design and maintain Web services realized as semantic Web services within the whole life cycle. All SWU components may be split in the following three groups:

**Information structures** – intended for storing and retrieving semantic and non-semantic information:

- *Distributed repository of semantic Web services (DSWS-R)* is aimed at effective storing and retrieving all elements of the Semantic Web according to the WSMO Framework (WSMO objects) (Roman et al. 2005): goals, ontologies, SWS and mediators written on

WSML language (Bruijn et al. 2005). Each DSWS-R consists of two parts: *Local Repository* – a place where all WSMO objects created in this Unit are stored; they are accessible only from components of this Unit, and *Local Registry* – a place where the *advertisements* of WSMO objects are stored. Such advertisements may be a part of service descriptions (e.g. service capability) or the whole descriptions of WSMO objects (e.g. ontologies) that have been published by some tools belonging to this local Unit as well as by some components of *external friendly* SWUs. It is expected that such “friendliness” will be actually defined by business contracts between SWUs in the real life.

- *Similarity-based OM* contains non-logical representation of the WSMO objects stored in DSWS-R as well as some additional non-semantic data e.g. graphical models of SWS and “natural language” templates for WSMO Goals needed for modifying or using these objects. OM is a Web service implementation of a case-based memory.
- *Semantic Information Router (SIR)* contains description of Web services (WSDL files) used for execution of semantic Web services created in a local SWU. Sir also proposes some methods for Web services annotating and categorization.

**Tools** – intended for storing and retrieving semantic and non-semantic information:

- *Case-based Service Designer* is aimed at designing a WSMO-based semantic Web service based on the description of an existing non-semantic Web service.
- *Case-based Service Composer* is aimed at creating a Semantic Web service through composition of existing WSMO-based semantic Web services.
- *Case-based Goal Editor* is aimed at creating predefined WSMO-based goals and their “natural language” templates needed for designing SWS-based applications.

**Environment** – responsible for communicating with different users, agents and other SWUs. The INFRAWEBS Environment contains Discovery, Executor, User Interface Agent and Security components.

The INFRAWEBS Conceptual Model reflects a novel approach for solving problems occurring during creating SWS applications - the tight integration of similarity-based (CBR) and logic-based reasoning. The similarity-based reasoning is used for fast finding an approximate solution, which is farther clarified by the logic-based reasoning.

The main purpose of this paper is to describe the usage of the case-based (CB) memory in the Capability Editor – a

sub-module of the INFRAWEBS Designer responsible for creating capability description of WSMO-based semantic Web services. The structure of the paper is as follows: the next section starts with an analysis of methods for creating the appropriated description of the problem to be solved (new case). Such a description is represented as a query, which the user sends to the CB memory. After the analysis, some methods for query formulation as well as some concrete representations for queries and cases are proposed. In the next section several methods for evaluating similarity between a query and a case are discussed. Some asymmetrical similarity functions useful for measuring similarity between different sections of case and query descriptions are considered. Representation of query results and a way for communication with the CB memory are presented as well.

## QUERY TYPES

According to the WSMO Framework, a semantic service description consists of the following main parts (Roman et al. 2005):

- *Top level concepts* – describing service name spaces, used ontologies and mediators as well as service nonfunctional properties.
- *Service capability* – describing what the service can do. The description is separated into four sections – assumptions, preconditions, postconditions and effects represented as WSMML logical expressions (axioms). The connections between them are the common set of ontologies used as well as a set of global (so called “shared”) variables (optional). In the INFRAWEBS Designer an axiom is created in a graphical way by means of a tool called Axiom Editor (Agre et al. 2005).
- *Service interface* – describing how the service works. The service interface is represented by service choreography and service grounding. The choreography defines how the user can communicate with the semantic service while the grounding is responsible for communication with the corresponding Web service represented as a WSDL file.

Creation of the service capability description (WSMML axioms) consists of sequential designing its four sections and specifying the service shared variables. The order of section creation depends only on the user – the semantic service provider. Moreover, the user can start creation of an axiom without initial specification of its role, i.e. whether it will be served as service preconditions or postconditions etc. Such a specification may be done after completing the axiom. The process of creating an axiom in the INFRAWEBS Capability Editor is case-based, which means that the user may ask the case-based memory for providing her with axioms similar to the one she is going to construct. Asking for assistance is implemented as a process of satisfying the user’s query describing how the desired axiom should look like. The more detailed the user’s query is, the more similar will be a found axiom proposed to the user as a template for creating a new axiom. That is why just the diversity of the user-supplied information available during an axiom construction may be used for classifying the types of different queries to the memory.

## Query by loaded ontologies

This is the simplest type of queries, which ask for existing services described by the same (or almost the same) set of ontologies as ones imported by the service under construction. The rationale behind this is an assumption that services using the same set of ontologies are likely to belong to the same problem domain. Practically such a query is equivalent to browsing capabilities of available semantic services from the same problem domain and may be useful only for receiving some initial ideas of how the desired service capability should look like. It is not expected to receive a good selectivity from such type of the query.

## Lexical or “natural language” queries

“Natural language” queries are intended for searching existing semantic services, whose descriptions contain natural language words specified in the query. Since natural language is used mainly for describing nonfunctional properties of a semantic service (e.g. title, publisher, description, etc.) such a query will be matched against these properties of a service. The queries may be unstructured or structured. In the first case the matching is performed against all nonfunctional properties of the service including nonfunctional properties of service axioms, while in the second - the similarity is evaluated only for service nonfunctional properties whose names are specified in the query. For example, one can search for the similar services created by a concrete organization (service publisher) or written in a concrete natural language or even created by a concrete person. In order to guarantee the compatibility of the found service descriptions with the already selected set of ontologies, such “natural language” queries are normally (by default) considered in conjunction with the first type of queries (queries by loaded ontologies).

## Query by ontological terms

The only information available to the user before creating a description of the first service capability axiom is an initial set of ontologies, which was selected as the most appropriate for semantic describing of the Web service. So the only way to represent more clearly the “meaning” of the new axiom is to express it as a query containing a set of concepts and relations that user intends to include into the axiom description. Since an axiom can exist only as a part of a semantic service capability description, the meaning of such a query is to find a set of existing semantic Web services, whose capability descriptions contain the set of ontological concepts specified in the query.

### Unstructured queries

Even in the situation, when only a set of ontologies to be used are known, it is possible to formulate queries with different meanings:

- A query for a semantic service, whose capability description as *a whole* contains (is similar to) a specified set of ontological concepts.

- A query for a semantic service containing one or several axioms, whose descriptions *separately* contain (are similar to) a specified set of ontological concepts.

In the first case the similarity is measured based on all ontology concepts used for representing *all* axioms participated in the service capability description. In other words the query is matched against a compound axiom constructed by merging all service capability axioms. The most similar service will be a service whose capability axioms as a whole have the *highest average similarity* with the query.

In the second case the query is matched against *each capability axiom separately*. The most similar service will be the one containing an axiom which has the highest similarity to the query.

### Structured queries

The next natural step is to allow the user to construct structured queries. In such queries the user can specify not only the ontological concepts the desired service capability description should have, but also *which part* of such a description (i.e. postconditions, assumptions etc.) should have a specified set of ontological words.

For example, it is possible to search for an existing semantic service whose preconditions are the most similar to the set of ontological words specified in the query; or to find such a service whose preconditions are similar to one ontological word set *and* postconditions are similar to another set of such words defined in the query. In such case the most similar service will be the service with the *highest average aggregated similarity* to the structural query.

### Query by example

As it has been already mentioned, the creation of a service capability description is a sequential process of constructing the axioms. Thus, an already constructed axiom (or axioms) may be used for formulated more precise request for the existing service containing similar axioms. The main assumption is that a service containing similar axioms in the corresponding service description sections (e.g. preconditions) will probably have the similar axioms in its other parts (e.g. postconditions). That is why after the completion of an axiom and assigning to it a given role the user can query the CB memory for finding existing semantic services, which have similar axioms playing the same roles as specified in the query.

In fact, such type of query is a kind of the structured queries in which structured parts are constructed from already prepared axioms of the service under construction.

### Complex queries

It is possible to construct complex queries combining ontological-based and natural language-based sets of words (either structural or unstructured). For example, one can ask for similar services containing specified ontology words in their postconditions *and* created by a specified organization.

## FORMULATING QUERIES

A way for formulating a query depends on a concrete situation. The following situations are possible:

1) *Neither ontologies nor some axioms are known*: the main approach for converting a Web service into a semantic Web service using INFRAWEBS Service Designer is “bottom-up”, i.e. starting from selection of a desired Web service, selecting a set of appropriate ontologies and applying them for annotating the service in the Grounding Editor, and then creating service choreography and capability. So in such a scenario an initial set of ontologies used in the Capability Editor has been already selected before designing a first capability axiom. However, in order to increase the flexibility of our tool and to ensure its compatibility with other WSMO-based tools for semantic service creation (e.g. WSMO Studio (Smirnov et al. 2006), which main paradigm for semantic service creation is “top-down” (i.e. starting with creating a service capability description first), we do not impose on the user any restrictions on the way she is going to construct a new semantic Web service. It is exactly the situation, when the user is going to construct her first axiom but does not know precisely neither what ontologies are needed for this nor how the axiom should look like.

However, even in this situation the user is able to describe in “natural language” what the desired axiom should be about and use the constructed “description” for finding “similar” existing axioms. Such a request may be accomplished by *formulating a natural language query*, which is the only type of queries available for use in such a situation.

Formulating a query is implemented as a process of filling a *query form* containing named text fields describing possible nonfunctional properties (NFP) of the desired service according to the WSMO Framework (e.g. title, creator, type, etc.). A query form will contain also a “functionality description” field, in which the user can describe the desired functionality of a service. This description will be matched against NFP of the existing service axioms.

If the user has filled only the “functionality description” field, the corresponding text will be used for matching against all NFP occurring in a semantic service description.

If some of structured text fields of the query form have been filled, services matching such a structured query will be searched.

2) *Only an initial set of ontologies has been already loaded*: this is a “normal” situation observed before starting the process for creating the first capability axiom description. Now the user can facilitate her work by attempting to find “similar” existing services which use:

- Similar set of imported ontologies. The user is able to do this by *formulating a query by loaded ontologies*, which is the first of the two types of queries available to her in such a situation. As in the previous case, the query is represented by its query form describing NFP of the desired service. However, now such a “natural language” description will be used *only for ranking* existing services, which capability descriptions are constructed by the terms from the same (or similar) set of imported ontologies.
- Similar set of ontological terms in the axiom descriptions. The user is able to do this by *formulating a query by ontological terms*. In this situation the query is represented by the query form describing NFP of the desired service *extended by additional five fields*:
  - *service assumption*
  - *service precondition*

- *service postcondition*
- *service effect*
- *service functionality*

Each of these fields may be filled by an unrestricted number of ontological terms selected from the loaded ontologies. The first four fields are used when the user searches for similar services containing the same ontological words in the specified sections of the service capability description. The last one (“service functionality” field) is used for indicating that it does not matter for the user in which section of the capability description of a service the specified terms are used. In other words, the content of this field will be matched against the aggregated content of all capability description sections of a service except those that have been filled in other ontology-related query form fields.

The information provided in the NFP fields of the query form now will be used *only for ranking* existing services similar to the query formulated by means of ontological terms.

*Both types of queries (by loaded ontologies and by ontological terms) may be implemented as a single query by ontological terms:* if the additional (ontology-related) fields of the query form have left unfilled, the query will be treated as a query by loaded ontologies, otherwise – as a query by ontological terms.

3) *Some of the capability axioms have been already constructed:* this is a normal situation reflecting the sequential character of the capability description construction process. Finding a “template” for the next axiom the user is going to create may be accomplished by:

- Formulating a query by ontological terms. The user is able to repeat again a query by ontological words by specifying the content of the corresponding capability description section by means of words selected from the loaded ontologies.
- Formulating a query by example. In this situation the similar services are searched by comparing descriptions of already constructed axioms as a template for finding the existing services containing similar axioms in the corresponding sections of the service capability descriptions.
- Combining query by example with query by ontological words, i.e. describing a content of a desired axiom by means of some ontological terms *and* using the descriptions of the previously constructed axioms as an additional template for restricting the search.

A query by example may be represented as a query by ontological terms in which some of the ontology-related fields are filled automatically from the description of the corresponding axioms. *That is why all three types of queries mentioned above in this section may be implemented uniformly as a single query by ontological terms:*

- if some of the ontology-related fields are filled *manually* (by selection from the set of loaded ontologies) the query will be treated as a simple query by ontological terms;
- if the query contains the only ontology-related fields filled automatically (or specified as to be used as an

example) than such a query will be treated as a query by example;

- a query containing both fields filled will be treated as a complex type query.

Summarizing, it is possible to conclude that *formulating a query may be implemented in a uniform way as a process of filling a unified query form.* Depending on the current situation, such a form will consist of one or two parts: *textual* – to be used for formulating structural (or unstructured) “natural language” queries, and *ontological* – to be used for formulating queries (also structural or unstructured) based on ontological words. Content of the both parts may be used for formulating rather complex queries. The proposed schema for query formulation allows implementing a very flexible process of finding similar services and personalizing the Capability Editor for working with the user having different level of knowledge and skills.

## REPRESENTATION OF CASES AND QUERIES

### Representation of cases

In order to be reused, WSMO descriptions of semantic Web services and other objects (goals, ontologies, mediators) are considered in the INFRAWEBs Framework not only as logical (WSML) representations of these objects but also as *a special class of text documents*, containing natural language and ontology-based words. Such special “text” representation serves as a basis for constructing case representation of such WSMO objects.

The main object of a case-based memory is a case. Generally *a case* is a pair  $\{P, S\}$ , where  $P$  is a description of a problem, and  $S$  is its solution. Since the main objective of the INFRAWEBs case-based memory is to provide the user with a set of WSMO objects similar to that the user is going to use, each case may be considered as a special description of *a WSMO object* (semantic Web service, WSMO goal, etc.):  $P$  is a special structured feature vector representation of this object, and  $S$  is a pointer to the place in a local DSWS-R where the WSML description of the object is stored. More precisely, since the memory provides an access to different types of WSMO objects, an INFRAWEBs case is a triple  $\{T, P, S\}$ , where  $T$  is a type of the WSMO object stored. This parameter determines the structure of the corresponding feature vector representation of the object. For example, for cases of type  $T = \text{service}$ , the *structured feature vector representation* (SFVR) is  $P = \langle \text{NFP}, \text{O}, \text{Cp}, \text{I} \rangle$ , where  $\text{NFP}$  is a SFVR of nonfunctional properties of the service,  $\text{O}$  is a feature vector of names of ontologies imported by the service,  $\text{Cp}$  is a SFVR of service capability and  $\text{I}$  is a FVR of service interface. In its turn  $\text{Cp} = \langle \text{As}, \text{Pr}, \text{Ps}, \text{Ef} \rangle$ , where  $\text{As}$  is a feature vector corresponding to the service assumptions,  $\text{Pr}$  – a feature vector corresponding to service preconditions,  $\text{Ps}$  is a feature vector corresponding to service postconditions and  $\text{Ef}$  – a feature vector corresponding to service effects.

### Representation of queries

A query is represented by a triple  $Q = \{T, Pq, Sn\}$ , where  $T$  is the type of an object requested for searching in the CB

memory (e.g. service, goal, etc.),  $Pq$  is a SFVR of the desired object and  $Sn$  is a type of a query. The query type may take two values – *design-time* and *run-time* and determines the range of objects to be searched by the memory for satisfying the query. In the first case the memory would return similar WSMO objects stored in the whole DSWS-R, while in the second case – only in the “public” part of DSWS-R – its Registry.

## EVALUATING SIMILARITY

In his famous work “Features of Similarity” (Tversky 1977), which is a classic book for all contemporary work on studying the similarity, Tversky claimed that the similarity between two objects is a function of three variables:

- A number of features common for both objects;
- A number of features unique for the first object;
- A number of features unique for the second object.

Most of existing similarity functions using these three variable (e.g. Jaccard coefficient) are symmetrical and defined on the interval  $[0, 1]$  or  $[0, \infty]$ .

Unfortunately, we can not apply such an approach directly to our problem. The main reason for this is that one of the objects to be compared (the query) has a *principally incomplete description*, thus the value of similarity between such a partially described object and a fully described object is *only an approximation* for “real” value of similarity of these objects described at the same level of granularity.

The main consequence of this fact is that our similarity function should be *asymmetrical*, since the importance of features unique for the first and for the second object now is not the same. Indeed, the uniqueness of some features belonging to the second object (the case) detected during the matching now is not so important since it may be caused simply by *incompleteness of the description of the first object* (the query) rather than explained by real dissimilarity between the objects. From another side, the uniqueness of some features belonging to the first object (the query) now becomes more important since they really reflect the dissimilarity between the objects under comparison.

The concrete form of the similarity function depends on the level of completeness of the query description and that is why should be explicitly defined for different types of queries.

### Calculating similarity for ontological words queries

A similarity between an ontological query  $Q = \{service, Pq, Sn\}$  formulated by ontology keywords the user expects to be occurred in the service capability, and a case  $C = \{service, Pc, Sc\}$  is calculated as average similarity between the corresponding parts of capability descriptions specified in the query and in the service:

$$\begin{aligned} Sim(Q, C) &= Sim(Pq, Pc) = \\ &= Sim(< As_Q, Pr_Q, Ps_Q, Ef_Q >, < As_C, Pr_C, Ps_C, Ef_C >) = \\ &= \frac{1}{\sum_i W_i} \times \sum_{V_Q \in P_Q, V_C \in P_C} W_i \times sim_i(V_Q, V_C) \end{aligned} \quad (1)$$

where  $W_i$  is a coefficient weighting similarity between corresponding sections of the query and the case description,

and  $sim_i$  – is a function evaluating the similarity between such sections.

A feature vector  $V_i \in Cp = \{As_C, Pr_C, Ps_C, Ef_C\}$  of a section of service capability  $Cp$  is represented as a vector of weighted ontological words  $w_j$  occurring in the corresponding section:  $V_i = < k_{1i}w_{1i}, \dots, k_{ni}w_{ni} >$ , where  $n_i$  is a number of different words in the section  $i$ . In the simplest case, when all ontological words have the same weights, the coefficient  $k_{ij}$  is the number of occurrences of the word  $w_{ij}$  in the “ontological text” extracted from the logical description of section  $i$ . In more complex situation such a coefficient may reflect not only the number of the word occurrences but the importance of the concrete word as well. For example, one can assign lesser importance to the names of attributes of ontological concepts than to the concepts themselves.

When the query is constructed by selection of words from available ontologies, it is expected that each section of such a query will not contain repetitions since the meaning of such a query is to find a service, which capability description contains specified ontological words in the corresponding capability description section. That is why the similarity between feature vectors of the corresponding sections of a query and a case is a normalized number of identical words calculated according to the following formula:

$$\begin{aligned} sim_i(V_i^Q, V_i^C) &= \\ &= sim_i^{Onto}(< w_{1Q}, \dots, w_{nQ} >, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >) = \\ &= \frac{1}{n_Q} \times \sum_{j=1}^{n_Q} \delta^{Onto}(w_{jQ}, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >), \end{aligned}$$

(2)

where:

$$\delta^{Onto}(w_{jQ}, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >) = \begin{cases} 1, & \text{if } w_{jQ} = w_{l_C} \\ & \text{and } l_C = \{1_C, \dots, n_C\} \\ 0, & \text{otherwise} \end{cases}$$

In practice, such a similarity function calculates the size of intersection between two sets of ontological words (without repetitions) and normalizes it by the size of the set of words mentioned in the query ( $n_Q$ ).

It should be mentioned, that such a function is *not symmetrical*, i.e.  $sim_i(V_i^Q, V_i^C) \neq sim_i(V_i^C, V_i^Q)$ . The asymmetry is caused by the normalizing factor – the number of ontological words used for formulating the query. This similarity function may be seen as an asymmetrical variant of Jaccard coefficient with neglecting features unique for the second object (the case).

### Calculating similarity for ontological queries by example

The situation, when the user is looking for a service by means of a query by example, is rather different since such a query can contain several occurrences of the same ontological words in each its section. For example, it will be the case, when the “example” is a logical expression containing several variables (attribute values) that are different instances of the same ontological concept. That is why an existing service description containing the closest number of such words in the corresponding section of the

capability description should be selected as the most similar to the query.

In this situation the elements of the feature vectors are numerical values (integers) and hence the similarity between such vectors may be calculated by means of a function “inverse” to the distance between the vectors.

However, as in the previous situation, the new similarity function *should not be symmetrical* again. Since the query is used as a template, the service, which capability section description contains more occurrences of a given word than the specified in the query, should be ranked as more similar than another case containing lesser number of occurrences of the same word. The rationality behind such asymmetrical definition of similarity is the difference in the *degree of utility* of the found service description, which in this case may be interpreted as the *simplicity for further adaptation of the solution found* (i.e. logical expression). It is simpler to remove some “unnecessary” additional terms from an expression rather than to extend (or refine) it by addition of some new terms.

That is why we propose the similarity between feature vectors of the corresponding sections of a query by example and a case to be calculated according to the following formula:

$$\begin{aligned} \text{sim}_i^{\text{Exp}}(V_i^Q, V_i^C) &= \\ &= \text{sim}_i^{\text{Exp}}(< k_{1Q}w_{1Q}, \dots, k_{nQ}w_{nQ} >, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >) = \\ &= \frac{1}{\sum_{j=1}^n k_{jQ}} \times \sum_{j=1}^n \delta^{\text{Exp}}(k_{jQ}w_{jQ}, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >), \quad (3) \end{aligned}$$

where

$$\delta^{\text{Exp}}(k_{jQ}w_{jQ}, < k_{1C}w_{1C}, \dots, k_{nC}w_{nC} >) = \begin{cases} 1 - \frac{k_{1C} - k_{jQ}}{k_{1C} + k_{jQ}}, & \text{if } w_{jQ} = w_{1C} \\ & \text{and } k_{1C} \geq k_{jQ}, l_c = \{1_C, \dots, n_C\} \\ 1 - \frac{k_{jQ} - k_{1C}}{k_{jQ}}, & \text{if } w_{jQ} = w_{1C} \\ & \text{and } k_{1C} < k_{jQ}, l_c = \{1_C, \dots, n_C\} \\ 0, & \text{if } w_{jQ} \neq w_{1C} \end{cases}$$

Of course, the proposed atomic similarity function  $\delta^{\text{Exp}}$  is an example of an asymmetrical function defining similarity between two numerical values ( $k_Q$  and  $k_C$ ). In general case, such a function may be defined via the distance between the values:

$$\delta(k_Q, k_C) = \begin{cases} 1 - w_L \times |k_Q - k_C|, & \text{if } k_C < k_Q, 0 \leq w_L \leq 1 \\ 1 - w_R \times |k_Q - k_C|, & \text{if } k_C \geq k_Q, 0 \leq w_R \leq 1 \end{cases} \quad (4)$$

where the weighting coefficients reflects the degree of asymmetry of the function.

### Calculating similarity for queries by loaded ontologies

A query by a set of loaded ontologies may be seen as a mix of queries by ontological words and queries by example. As in the case of queries by example this type of query is formed automatically – the names of all ontologies loaded to the system are used as an example for searching similar services. However, the feature vector constructed from the names of these ontologies does not allow any repetitions and

can be treated as a “normal” feature vector constructed by ontological words. Since the set of ontologies imported by a semantic service can not also contain repetitions, the similarity function for such kind of query may be naturally defined as an intersection of both sets as it has been defined for queries constructed by ontological words:

$$\begin{aligned} \text{sim}_i(V_i^Q, V_i^C) &= \\ &= \text{sim}_i^{\text{Import}}(< w_{1Q}, \dots, w_{nQ} >, < w_{1C}, \dots, w_{nC} >) = \\ &= \frac{1}{n_Q} \times \sum_{j=1}^{n_Q} \delta^{\text{Import}}(w_{jQ}, < w_{1C}, \dots, w_{nC} >), \quad (5) \end{aligned}$$

where  $\delta^{\text{Import}}(w_{jQ}, < w_{1C}, \dots, w_{nC} >) = \begin{cases} 1, & \text{if } w_{jQ} = w_{1C}, \\ & \text{and } l_c = \{1_C, \dots, n_C\} \\ 0, & \text{otherwise} \end{cases}$

and  $n_Q$  is a number of imported ontologies defined in the query description.

There is no sense to punish the presence of additional imported ontologies in the description of a found service since:

- *The modular structure of WSML ontologies:* an ontology may refer to some concepts described in other external ontologies. Such external ontologies can be loaded in the INFRAWEBS Capability Editor on-demand, i.e. when in the process of constructing a WSML logical expression (axiom) the full structure of a concept belonging to such an ontology is requested by the user (see (Agre et al. 2006) for details). That is why there is no guarantee that such “additional” (in this moment) ontologies will not be requested by the user in the future, when she will start creating a new axiom.
- *The relative simplicity of possible adaptation of the found service description:* even if the presence of additional ontologies is caused by using “unnecessary” concepts from such ontologies, removing of such concepts from the capability description of newly created service will automatically remove the corresponding ontologies when such a description will be stored (see again (Agre et al. 2006) for details).

From another side, the lack of some of the requested ontologies in the service description is the sign that the corresponding service axioms may significantly differ from that the user is going to create.

### Calculating similarity for lexical queries

The lexical similarity between a natural language query and a case representing a WSMO object is carried out in two stages: 1) Preprocessing the query and 2) Matching the preprocessed query to the correspondent section of the object feature vector. The goal of the query’s preprocessing is to unify the words used in the query with those used in the object’s feature vector. As opposed to the ontology-based similarities where only a limited number of terms is used, in the case of a natural language query the user can freely choose the words to search by. This is why a query’s preprocessing is needed. The preprocessing includes non-alphabetical letters excluding, composite words tokenization, stop words exclusion, synonyms substitution and similar words substitution. The preprocessed query is matched against the “lexical” part of the object feature vector.

The following formula is used to calculate the similarity between the feature vector corresponding to a NFP section of a natural language query and a feature vector of the corresponding NFP section of a stored case:

$$\begin{aligned} \text{sim}_{NFP}(V_{NFP}^Q, V_{NFP}^C) &= s \\ &= \text{im}_{NFP}(<k_{1Q}w_{1Q}, \dots, k_{nQ}w_{nQ}>, <k_{1C}w_{1C}, \dots, k_{nC}w_{nC}>) = \\ &= \frac{\sum_{j=1}^n \delta_{NFP}(k_{jQ}w_{jQ}, <k_{1C}w_{1C}, \dots, k_{nC}w_{nC}>)}{\sum_{m=1}^n k_{mQ} + \sum_{j=1}^n \sigma_C(k_{jQ}w_{jQ}, <k_{1C}w_{1C}, \dots, k_{nC}w_{nC}>)}, \quad (6) \end{aligned}$$

where:

$$\delta_{NFP}(k_{jQ}w_{jQ}, <k_{1C}w_{1C}, \dots, k_{nC}w_{nC}>) = \begin{cases} k_{jQ} + k_{lC}, w_{jQ} = w_{lC}, \\ \text{and } l = \{1, \dots, n\} \\ 0, w_{jQ} \neq w_{lC} \end{cases},$$

$$\sigma_C(k_{jQ}w_{jQ}, <k_{1C}w_{1C}, \dots, k_{nC}w_{nC}>) = \begin{cases} k_{jC}, w_{jQ} = w_{jC}, \\ \text{and } l = \{1, \dots, n\} \\ 0, w_{jQ} \neq w_{jC} \end{cases}$$

The result is given as the sum of weights of co-occurring words in the query and case feature vectors normalized by the sum of all weights of words presented in the query's feature vector and the weights in words common for the case and the query. The neglecting weights of words unique for the stored case makes this similarity function asymmetrical again.

As it can be seen, the formula is similar to the formula (3) but it takes into account the weights of the co-occurring words from both sides – the query's feature vector and the case feature vector. The rationale behind this is that an object (case) having more occurrences of a word given in the query is more relevant as result for this query.

### Calculating similarity for complex queries

A complex query is represented by a query form in which both NFP and "ontological" sections are filled. Each of these sections may be described either by structural or unstructured "simple" query. Since in WSMO Framework NFP of a semantic service are *optional*, we have decided to use "lexical" similarity only for ranking the services matched the ontological part of a complex query.

That is why a similarity of a complex query  $Q$  to a service  $S$  is measured by two values – *ontological similarity* and *lexical similarity*:

$$\begin{aligned} \text{Sim}_{\text{complex}}(Q, S) &= \text{Sim}_{\text{complex}}(<NFP_Q, C_Q>, <NFP_S, C_S>) = \\ &= \text{Sim}_{\text{lexical}}(NFP_Q, NFP_S), \text{Sim}_{\text{ontological}}(C_Q, C_S) > \end{aligned} \quad (7)$$

where  $NFP$  and  $C$  with the corresponding indexes are nonfunctional properties and capabilities of the query and the service participated in the matching.

The ontological similarity is the main element evaluating the utility of the service found and the lexical similarity is used only for ranking services with the same degree of ontological similarity.

### QUERY INTERFACE

### Query Result

A result of a query is a list of existing semantic Web services ordered by the value of ontological similarity of a service to the query. Each entity in this list is represented as:  $\langle Title, \text{Sim}_{\text{Ontological}}, \text{Sim}_{\text{Lexical}}, \text{Pointer} \rangle$ , where:

- *Title* is the name of a service (if available). This value is the filler of the service NFP "dc:title"). If the corresponding filler is unfilled, the title of such a service will be "unknown".
- $\text{Sim}_{\text{Ontological}}$  is a real number in the range [0, 1] evaluating the ontological similarity between the query and the service found. In a case of a pure lexical query, this value is "undefined";
- $\text{Sim}_{\text{Lexical}}$  is a real number in the range [0, 1] evaluating the lexical similarity between the query and the service found. In a case of a pure ontological query, this value is "undefined";
- *Pointer* is a pointer to the place in the DSWS-R where the semantic description of this service is stored.

The first three items are used for visualizing the query results; while the fourth item is used for loading the corresponding service description if it has been selected by the user.

The number of entities in the list forming the query result is controlled by two user-supplied parameters:

$N$  – is a maximum length of a list containing matched services ( $N \geq 1$ ),

$B$  – is a similarity threshold ( $0 \leq B \leq 1$ ); only services which similarity (ontological – in the case of ontological or complex queries or lexical – in a case of pure natural language queries) greater or equal than this threshold will be included into the list with the query results. The default values of these parameters are set during the system initialization; however, the user can change them by specifying the desired values in each query.

### Communication with the case-based memory

The communication with the similarity-based OM, which is a Web service implementation of the case-based memory, is carried out through the SOAP protocol (Nern and Boyanov 2006). The OM Web service provides a method called *SearchForWSMOObjects* accepting as input a query represented according to the rules described in the previous section. The output of the *SearchForWSMOObjects* method is of type *SearchResult* and contains result elements represented by an object title and identifier in the form of IRI. Lexical and ontological similarity coefficients are also given for each element and used for ordering the elements. An example of such query result in XML form is as follows:

```
<SearchResult>
  <SearchTime>20</SearchTime>
  <ResultSet>
    <ResultElement>
      <Title>WSMO Object 1</Title>

      <OntologySimilarity>0.85</OntologySimilarity>

      <LexicalSimilarity>0.78</LexicalSimilarity>
      <Identifier>http://dsws-
r.infrawebs.org/1.wsmo</Identifier>
    </ResultElement>
```



```

<ResultElement>
  <Title>WSMO Object 2</Title>

  <OntologySimilarity>0.77</OntologySimilarity>

  <LexicalSimilarity>0.86</LexicalSimilarity>
    <Identifier>http://dsws-
      r.infrawebs.org/44.wsmo </Identifier>
  </ResultElement>
</ResultSet>
</SearchResult>

```

## CONCLUSION

In the present paper we have discussed possible ways for using a case-based memory for creating semantic service descriptions in the INFRAWEBS Framework. The case-based memory provides the service designer with descriptions of existing semantic Web services that are similar to the user requests represented by means of a query form. Such a form allows the user to express her intentions on how the different part of the service under design should look like using both natural language and ontology-based words.

We have also discussed some possible approaches for measuring similarity between a query (a description of a problem to be solved) and a case (a description of another, already solved problem). Several similarity functions have been proposed. All of them are asymmetrical, which has been argued by principal difference in the amount of known features used for describing objects to be matched.

In INFRAWEBS Framework the case-based memory may be used not only for facilitating the process of constructing a semantic service description, but also for service composition, as well as for goal and service discovery (Agre and Boyanov 2006). For example, during service discovery the memory is able to restrict the search space of potential services by retrieving a set of such services that are the most similar to the formulated goal represented as special text documents.

In the INFRAWEBS Framework the case-based memory will be implemented via INFRAWEBS similarity-based Organizational Memory (OM). A detailed specification of OM describing the organizational structure of the case based memory can be found in (Nern et al. 2006).

## ACKNOWLEDGEMENTS

The research has been partially supported by INFRAWEBS - FP62003/IST/2.3.2.3 Research Project No. 511723

## REFERENCES

- G. Agre and A. Boyanov 2006. *INFRAWEBS Deliverable D5.4.2.2. Realization of Case-Based Composition in Design Time and Test Composer - Self Organizing Case Memory*.
- G. Agre, P. Kormushev and I. Dilov 2006. *INFRAWEBS Axiom Editor User's Guide Version 1.0.5*; available at: [http://www.fh-bochum.de/infrawebs/?menue=dissemination&site=open\\_software](http://www.fh-bochum.de/infrawebs/?menue=dissemination&site=open_software).
- G. Agre, P. Kormushev and I. Dilov 2005. INFRAWEBS Axiom Editor – A Graphical Ontology-Driven Tool for Creating Complex Logical Expression, *International Journal "Information Theories and Applications"* (in print)
- Bruijn, J.; Lausen, H.; Krummenacher, R.; Polleres, A.; Predoiu, L.; Kifer, M.; Fensel, D. 2005. D16.1 – The Web Services Modeling Language (WSML). WSML Draft.
- D. Aha and J.J. Daniels (Eds.) 1998. Case-Based Reasoning Integrations: *Papers from the 1998 Workshop*, Menlo Park, Calif., USA, AAAI Press.
- H.-J. Nern and A. Boyanov 2006. *INFRAWEBS Deliverable D2.2.2 Realized Full Model of the General OM & Coupling to SIR (Specification)*.
- H.-J. Nern, G. Agre, T. Atanasova, A. Micsik, L. Kovacs, T. Westkaemper, J. Saarela, Z. Marinova, A. Kyriakov 2005. Intelligent Framework for Generating Open (Adaptable) Development Platforms for Web-Service Enabled Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems. *W3C Workshop on Frameworks for Semantics in Web Services*, Digital Enterprise Research Institute (DERI), Innsbruck, June 9-10, 2005, available at: [http://www.w3.org/2005/04/FSWS/Submissions/60/Infrawebs\\_paper.PDF](http://www.w3.org/2005/04/FSWS/Submissions/60/Infrawebs_paper.PDF)
- D. Roman, U. Keller, H. Lausen (eds.) 2005. *Web Service Modeling Ontology, WSMO Final Draft, version 1.2*.
- A. Smirnov, M. Dimitrov and V. Momchev 2006. *WSMO Studio Guide – part 2*, available at: <http://www.wsmostudio.org/>
- A. Tversky 1977. Features of Similarity. *Psychological Review*, Vol. 84, 327-352.

## BIOGRAPHY

**GENNADY AGRE** is an associate professor in the Artificial Intelligence department of the Institute of Information Technologies Bulgarian Academy of Sciences. His current research interests are semantic Web services, machine learning, data mining and case-based reasoning.

# COMBINATION OF SEMANTIC WEB SERVICES BY THE CONTRIVANCE OF THE CURRENT WSMO SPECIFICATION

Vladislava Grigorova

Institute of Information Technologies -BAS,

Acad. G. Bonchev, 2

1113 Sofia,

Bulgaria

E-mail: v.grigorova@abv.bg

## KEYWORDS

Modeling, Web Services, Semantic Web Services, Composition.

## ABSTRACT

The possibilities for creation of web services composition through WSMO specification are investigated in this paper. WSMO is in its intensive development and a snapshot for the combination of semantic web services on its current state is made. The most likely problems are marked.

## INTRODUCTION

Semantic Web Services are integrated solution for realizing the vision of the next generation of the Web. WSMO (Roman et al. 2005) and OWL-S (Martin et al. 2004) are main initiatives to describe Semantic Web Services, aiming at describing the various aspects related to Semantic Web Services in order to enable the automation of Web Service discovery, composition, interoperation and invocation.

WSMO is examined in its capacity of European initiative in the semantic web technologies sphere. On account of the fact that the specification is in progress of development an investigation for the potentialities for composition and interaction between semantic web services is not trivial problem.

WSMO defines four top-level modeling elements for describing aspects of SWS:

- Ontologies provide the terminology used by other elements; define an agreed common terminology by providing concepts, and relationships between the concepts;
- Goals specify the problems that should be solved by services (queries) and the potentially satisfy desires that to be reached by service execution;
- Service descriptions depict services that are demanded by service requesters, covered by service providers, and agreed between service providers and requesters;
- Mediators are connectors between components with mediation facilities for handling heterogeneities ( on data

mediation level, protocol level, process level - ooMediators, ggMediators, wgMediators, wwMediators). (Roman et al. 2005) (de Bruijn et al. 2005)

## SEMANTIC WEB SERVICES

WSMO web services represent the non-functional (non-functional properties), functional (capabilities) and behavioral/usage (interface) aspects of the web services.

Service description definition includes: namespace, keyword webService and its identifier, non-functional properties (Accuracy, Contributor, Coverage, Creator, Date, Description, Financial, Format, Identifier, Language, Network-related QoS, Owner, Performance, Publisher, Relation, Reliability, Rights, Robustness, Scalability, Security, Source, Subject, Title, Transactional, Trust, Type, Version), imported ontologies , used mediators (ooMediator, wwMediator), functional capability and interfaces.

Capability definition shows the functionality: non-functional properties, imported ontologies, used mediators (ooMediator, wgMediator), all-quantified shared variables, preconditions (necessary conditions before the execution), assumptions (describe the state of the world before the execution), postconditions (result of service execution delivered to the user), effects (describe the state of the world after the execution). (Roman et al. 2005)

The interfaces are defined by specifying non-functional properties, imported ontologies, used mediators, choreography and orchestration and provide how the functionality of the Web service can be achieved.

The choreography defines the communication from the client's point of view. It includes non-functional properties, the state signature and the transition rules. A (world) state is represented by ontology which defines the state signature over which the transition rules are executed, ooMediators, a set of statements defining the modes of the concepts and relations (static, in, out, shared, controlled and lists of entries – as the meaning specifies itself from that who can change/read/write the instances of concept or relations - the choreography execution, the environment or both). The grounding information is defined by the withGrounding keyword followed by a list of IRIs. Transition rules are if-then rules, forall-with-do rules, choose-with-do rules that specify transitions between states. A set of fact modifiers which are of four different

types  $(add(fact), delete(fact), update(fact^{new}), update(fact^{old} \text{ } fact^{new}))$  is used to change instances to/from concepts and relations. Usually it models a single transition rule for each of the service's operations.

The orchestration describes how the overall functionality of the Web service is achieved by means of cooperation of different Web service providers but it not fully specified yet. (Feier 2005; Scicluna et al. 2006)

## COMPOSITION

Stepping on the current level of WSMO specification development an attempt at bringing out principles for automatic SWS composition from existent semantic web services is made.

Composition is the process of selecting and combining web services to achieve a user's objective. The target is to aggregate web services into a complex functionality.

Available semantic web services in the travel area are single for every aspect (booking of a hotel, booking of air tickets, etc.) and for every company. But to achieve the costumer request moreover faster it is necessary at this stage to pool selected semantic web services. And they stand sub-services of the new one. The union will be fulfilled on the level of every structural part from WSMO specification for web services – namespaces, non-functional properties, imported ontologies, used mediators, capabilities, interfaces. The elements, which are doubled, are not allowed. The others are added. The identifiers of the keywords and the definitions will be changed because they have to be specified by means of synonymity.

In the interface part problems are more complicated. A choreography model describes collaboration between a collection of services to achieve a goal. It captures the interactions in which the participating services engage to achieve this goal and the dependencies between these interactions, including: causal and/or control-flow dependencies (i.e. that a given interaction must occur before another one, or that an interaction causes another one), exclusion dependencies (that a given interaction excludes or replaces another one), data-flow dependencies, interaction correlation, time constraints, transactional dependencies, etc. (Bussler 2005). That's why the suggestion is to keep transition rules of the choreography not changed. The orchestration is not investigated yet.

The approach for composition is valid for services with similar functionality.

When the functionality which have to be achieved is more complex, rules that govern behavioral characteristics relating to how a group of Web services interact can be applied as a workflow generation method. The approach demands a composite service to be defined in an abstract manner, e.g. with regard to offered functionality and constraints. It models an abstract process that defines the order of the business logic. The abstract process model includes a set of tasks, their data dependency, conditions that have to be carried out, different interaction scenarios etc. The process can be showed by a graph as the nodes automatically are bound up with services. These services have to be abstract described too. The principles of composition should be similar like the above.

WSMO discriminates between abstract and concrete services – a concrete service provide different concrete (value) services to the user by means of different types of invocations of itself, while an abstract service is a set of concrete services each of which providing the use with actual values by respective service invocation to achieve the given goal (request). (Roman et al. 2005)

An abstract service abstractly describes the behaviour of the service with respect to an invocation instance of a service, independent of implementation details.

The approach for abstraction is to substitute concrete instances and values for concepts from related ontologies which are imported into the semantic web service and are known preliminary. The element "State Signature" with its parts: a set of statements defining the modes of the concepts and relations and a set of update functions and respective grounding have to be omitted because they give some kind of concreteness. The non-functional properties are generalized.

After abstraction the concrete service provided by abstract service should be covered by its (abstract) capability.

The abstract service has its own instance store linked to this ontology in which there are instances specific to the web service.

These semantic abstract services may be called: abstract service templates for concrete functionality. They allow some of the features to be fixed in all possible suitable to the occasion services as well as they are generic enough so all relevant services to be found.

The Figures 1 and Figures 2 show a concrete postcondition from semantic web service and an abstract postcondition from the same semantic web service.

### postcondition

#### nonFunctionalProperties

*dc#description* **hasValue** "The output of the service is a NH hotel room booking. An instance of a MH hotel room booking has been created. This booking has an identifier, a ticket, and the hotel stay info has been obtained from the service providers data base and satisfy the booking buyer preferences. The hotel stay info has: checkin and checkout date, room or rooms, the hotel."

#### endNonFunctionalProperties

#### definedBy

?HotelRoombookingRequest **memberOf**

hb#hotelRoombookingRequest **implies**

exists {?HotelRoomBooking, ?Buyer, BuyerContactInformation, ?HotelStay, ?MhotelServiceProvider}

(

?HotelRoomBooking **memberOf**

hb#hotelRoomBooking[

bookingIdentifier **hasValue** ?BookingIdentifier,

bookingTicket **hasValue** ?HotelRoomBookingTicket,

hotelStay **hasValue** ?HotelStay,

buyer **hasValue** ?Buyer,

seller **hasValue** ?MhotelServiceProvider,

paymentMethod **hasValue** ?Points]

and

?Buyer **memberOf** cu#customer[

contactInformation **hasValue** ?ContactInformation]

and

?BuyerContactInformation **memberOf**

cu#customerContactInformation [

emailaddress **hasValue** ?InputBuyerEmail]

and

```

?HotelStay memberOf hb#hotelStay [
    checkIn hasValue ?CheckIn,
    checkOut hasValue ?CheckOut,
    room hasValue ?Room,
    hotel hasValue ?Hotel]
and
?MhotelServiceProvider memberOf hb#HotelServiceProvider [
    hasCompanyID hasValue "MH Hotels",
    allowPets hasValue _boolean("false"),
    allowGroupsReservation hasValue _boolean("true"),
    minStars hasValue 4,
    hotels hasValue {?Hotel}]
).

```

Figures: 1

```

Postcondition
nonFunctionalProperties
    dc:description hasValue "The output of the service is
a hotel room booking. An instance of a hotel room booking have
been created. This booking has an identifier, a ticket, and the
hotel stay info has been obtained from the service providers data
base and satisfy the booking buyer preferences. The hotel stay
info has: checkin and checkout date, room or rooms, the hotel."
endNonFunctionalProperties
definedBy
?HotelRoombookingRequest memberOf
hb#hotelRoombookingRequest implies
exists {?HotelRoomBooking, ?Customer, ?Customer
ContactInformation, ?HotelStay, ?HotelServiceProvider}
(
?HotelRoomBooking memberOf hb#hotelRoomBooking[
    bookingIdentifier hasValue ?BookingIdentifier,
    bookingTicket hasValue ?HotelRoomBookingTicket,
    hotelStay hasValue ?HotelStay,
    buyer hasValue ?Buyer,
    seller hasValue ?HotelServiceProvider,
    paymentMethod hasValue ?Points]
and
?Customer memberOf cu#customer[
    contactInformation hasValue ?ContactInformation]
and
?CustomerContactInformation memberOf
cu#customerContactInformation [
    emailAddress hasValue ?InputCustomerEmail]
and
?HotelStay memberOf hb#hotelStay [
    checkIn hasValue ?CheckIn,
    checkOut hasValue ?CheckOut,
    room hasValue ?Room,
    hotel hasValue ?Hotel]
and
?HotelServiceProvider memberOf hb#HotelServiceProvider [
    hasCompanyID hasValue ?Hotel,
    allowPets hasValue _boolean,
    allowGroupsReservation hasValue _boolean,
    minStars hasValue ?minStars,
    hotels hasValue {?Hotel}]
).

```

Figures: 2

## CONCLUSION

The composition of abstract service templates allows paying attention to the functionality and to the important constraints and features i.e. to control flow – not to the specific data of the concrete services and it is more

suitable in a dynamic environment where the available services and the requirements are constantly changing. (Sirin et al. 2005)

And like one farther step it is possible to use these abstract service templates to build the WSMO goal template which will represent the some eventual desire of the customer. The goal templates ease the usage of the systems which like customers to the semantic web services since they hide the WSMO description complexity. (Lopez-Cobo et al. 2006)

## ACKNOWLEDGEMENT

This work is carried out under EU Project INFRAWEBs - IST FP62003/IST/2.3.2.3 Research Project No. 511723.

## REFERENCES

- Arroyo, S., Lara, R., Gomez, J. M., Bereka, D., Ding, Y., Fensel, D. 2004. Semantic Aspects of Web Services. In: *Practical Handbook of Internet Computing*. Munindar P. (Ed.) Chapman Hall and CRC Press, Baton Rouge
- Bussler C., A. Duke, D. Roman, M. Stollberg. 2005. *Proceedings of the First International Workshop on Web Service Choreography and Orchestration for Business Process Management*. In conjunction with the Third International Conference on Business Process Management (BPM 2005), Nancy, France, September 2005.
- De Bruijn J., C. Bussler, J. Domingue, D. Fensel, M. Hepp, U. Keller, M. Kifer, B. König-Ries et al. 2005. Web Service Modeling Ontology (WSMO). *W3C Member Submission 3 June 2005*. <http://www.w3.org/Submission/WSMO/>
- Feier C., D. Roman, A. Polleres, J. Domingue, M. Stollberg, and D. Fensel. 2005. Towards Intelligent Web Services: Web Service Modeling Ontology (WSMO). In *Proc. of the Int'l Conf on Intelligent Computing (ICIC) 2005*, Hefei, China, August 23-26, 2005.
- Lopez-Cobo J.-M., A. Lopez-Perez, J. Scicluna. 2006. A semantic choreography-driven Frequent Flyer Program. *Accepted for the EASST-EU Workshop*, 2006.
- Martin D., M. Paolucci, S. McIlraith, M. Burstein, D. McDermott, D. McGuinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, N. Srinivasan, K. Sycara. 2004. "Bringing Semantics to Web Services: The OWL-S Approach", *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, July 6-9, 2004, San Diego, California, USA.
- Roman D., H. Lausen, U. Keller. 2005. Web Service Modeling Ontology (WSMO). *WSMO Final Draft 10 February 2005*. <http://www.wsmo.org/TR/d2/v1.1/20050210/>.
- Scicluna J., A. Polleres, D. Roman. 2006. Ontology-based Choreography and Orchestration of WSMO Services. *WSMO Working Draft 3rd February 2006*. <http://www.wsmo.org/TR/d14/v0.3/>.
- Sirin E., B. Parsia, J. Hendler. 2005. Template-based Composition of Semantic Web Services. In *AAAI Fall Symposium on Agents and the Semantic Web*, Arlington, Virginia, USA, November 2005.
- Stollberg M. 2005. Reasoning Tasks and Mediation on Choreography and Orchestration in WSMO. In: *Proceedings of the 2nd International WSMO Implementation Workshop (WIW 2005)*, Innsbruck, Austria, 2005.
- Stollberg, M., C. Feier, D. Roman, D. Fensel. 2006. Semantic Web Services - Concepts and Technology. In N. Ide, D. Cristea, D. Tufis (eds.): *Language Technology, Ontologies, and the Semantic Web*. Kluwer Publishers, 2006 (to appear).

# **AUTHOR LISTING**



## AUTHOR LISTING

Agre G. ....	130	Lambert P. ....	56
Asai K. ....	63	Lin Q. ....	105
Asai Y. ....	15		
Atanasova T. ....	115/122	Mituyama Y. ....	15
Boucouvalas A. ....	23	Nait-Sidi-Moh A. ....	98
		Nakamura Y. ....	15
Chalhoub G. ....	10	Neo N. ....	105
Chan W. ....	33	Notebaert S. ....	56
Chbeir R. ....	10		
		Onoye T. ....	15
Daskalova H. ....	115/122		
Datcu D. ....	33	Rothkrantz L.J.M. ....	28/33
Davison A. ....	5		
De Cock J. ....	56	Satbhai S. ....	69
		Satoshi M. ....	46
Eberhardt G. ....	98		
		Takase N. ....	63
Fitrianie S. ....	28	Tanemura Y. ....	15
		Triantafillou V.D. ....	53
Garofalakis J.D. ....	53		
Goulas D.S. ....	53	Van de Walle R. ....	56
Grebenstein K. ....	92	van der Mast C.A.P.G. ....	69/77
Grigorova V. ....	138		
		Wack M. ....	98
Hadziliass E.A. ....	87	Wong W.S. ....	33
Haßlinger G. ....	41		
Hideki S. ....	46	Xu Z. ....	23
Hooplot F.S. ....	77	Yetongnon K. ....	10
John D. ....	23	Zhang L. ....	105
Kassel S. ....	92		
Kazutoshi F. ....	46		
Kobayashi H. ....	63		
Kondo T. ....	63		
Kotani A. ....	15		
Kusuma I. ....	105		