Proceedings of DSIE' 09

Doctoral Symposium on Informatics Engineering

4th edition dsie '09

5 | 6 Feb

Foreword

DSIE'09 - *Doctoral Symposium in Informatics Engineering 2009*, follows the three previous editions of similar events and aims to be a forum for the discussion and application of good practices of scientific research, namely in what Computer Science and Informatics Engineering is concerned. DSIE'09 is organized in the context of the Doctoral Program on Informatics Engineering (ProDEI) at the Faculty of Engineering of the University of Porto.

DSIE'09 implicitly displays what the PhD students have learnt at an early stage, during their first semester course on Scientific Research Methodologies (MIC), including appropriate methods to deal with their own research, as well as students' capabilities to produce well written scientific texts. This symposium is also a good opportunity for students to be jointly involved in all the aspects of a scientific meeting organization and participation, although lightly supervised by the course's responsible professors.

Being mainly devoted to the students of the Doctoral Program in Informatics Engineering current edition, DSIE'09 also accepted submissions from other students of the same program and from different, though related, post-graduate programs.

Current contributions already reveal that many of the students have interesting ideas (although in an embryonic state) about what specific research topic they wish to tackle. Also, papers can be found in which the state of the art description still is the major contribution. Once again, participants still are in the beginning of their own scientific work, so this can be perfectly understood and accommodated in the Meeting program.

This DSIE'09 Proceedings volume includes twenty three papers that have been revised and selected according to guidelines informed by the aforementioned principles and are assembled according to the different research topics that will be presented in eight different technical sessions: Robotics (4 papers), Computing Architectures (3 papers), Programming Modeling (2 papers), Specification and Testing (2 papers) Information Extraction and Processing (3 papers), Networks and Multimedia (3 papers), Pedagogical and Competence Issues Management (2 papers), Miscellaneous theories and applications (3 papers).

On the top of those sessions, distributed among a two-day duration time, DSIE'09 also includes two invited talks.

We, the professors responsible for the MIC course, would like to acknowledge all those who were deeply involved in the success of this event that, we hope, will contribute for a better understanding of both the themes that have been addressed during the course, the best scientific research methods and the good practices for writing scientific papers.

Eugénio Oliveira and A. Augusto Sousa
(in charge of the MIC course – Scientific Research Methodologies)

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Program

5 Feb

8:30	Registration - Room B032	
Opening Session - Room B032		
9:00	Prof. Eugénio Oliveira (ProDEI Director) Prof. Alvaro Cunha (FEUP Board) Doutora Lígia Ribeiro (U. Porto Pro-Rector)	
Special Session	**Chair: Prof. A. Augusto Sousa** **Co-chair: Tito Vieira**	
9:30	A Path for Performance Improvement: the Personal Software Process (PSP) and Team Software Process (TSP) - Invited speaker - João Pascoal Faria (FEUP)	
First Session	**Chair: Prof. A. Augusto Sousa** **Co-chair: Cristovão Sousa**	
11:00	Automatic Extraction of Quotes and Topics from News Feeds - Luís Sarmento, Sérgio Nunes	
11:20	Virtual Tourism Business Networks in Developing Countries - Luís Santos Barradas, João José Pinto Ferreira	
11:40	A New Paradigm for Automated Schematic Maps - João Manuel Mourinho	
Second Session	**Chair: Prof. Luís Paulo Reis** **Co-chair: Cláudio Barradas**	
14:00	Robot Dance based on Online Automatic Rhythmic Perception - João Lobato	
14:20	Biped locomotion methodologies applied to humanoid robotics - Hugo Rafael Picado	
14:40	Multi-Agent Coordination through Strategy - João Certo	
15:00	Cyber-Mouse: A Deliberative Implementation - João Certo, João Oliveira	
Third Session	**Chair: Prof. J. Magalhães Cruz** **Co-chair: João Certo**	
15:50	Implementing a Multiprocessor Linux Scheduler for Real-Time Sporadic Tasks - Paulo Baltarejo Sousa	
16:10	Design and Validation of Real-Time Applications - Carlos Jorge Costa	
16:30	Interoperable Geographic Information Services from Crisis Management Perspective - Marco Oliveira	
Fourth Session	**Chair: Prof. Jorge Barbosa** **Co-chair: João Oliveira**	
16:50	Modelling the Job-Shop Scheduling problem in Linear Programming and Constraint Programming - Pedro Abreu	
17:10	Propositional Based Inductive Logic Programming: Reduced Encoding of Hypotheses and Knowledge Base - Hugo Ferreira	

6 Feb

Fifth Session	**Chair: Prof. Pascoal Faria** **Co-chair: Luís Certo**
9:50	Inspections on Testing Aspect-Oriented Programs - Rodrigo Moreira
10:10	A prototype tool for supporting joint-design collaboration in requirements specification - Cristovão Sousa
Sixth Session	**Chair: Prof. Ana Paula Rocha** **Co-chair: Teresa Mota**
11:00	WordNet as a Symbolic Free Text Classifier - Gustavo Laboreiro
11:20	Temporal Analysis of Terms in Blogs - Filipe Coelho
11:40	Integration of Events Information: A robust system based on simple techniques - Luís Certo
Special Session	**Chair: Tito Vieira**
14:00	RCTS Advanced Services and the FEDERICA Project - Invited speaker - Eng. João Nuno Ferreira (FCCN)
Seventh Session	**Chair: Prof. J. Ruela** **Co-chair: Gustavo Laboreiro**
15:00	Improving the Performance of IEEE802.11s Networks using Directional Antennas over Multi-Radio/Multi-Channel Implementation - s The Research Challenges - Saravanan Kandasamy, Ricardo Morla, Manuel Ricardo
15:20	Towards the Optimization of Video P2P Streaming over Wireless Mesh Networks - Nuno Salta, Ricardo Morla, Manuel Ricardo
15:40	VoIP as a tool for an effective voice communication cost reduction - Tito Vieira
Eigthth Session	**Chair: Prof. Miguel Pimenta Monteiro** **Co-chair: Carlos Costa**
16:30	A Pedagogical Scenario based on the ILEM Model: A case study - Dulce Mota
16:50	Competence gap analysis in the Skills Recognition process using Treemaps - Teresa Mota
Closing Session	
17:10	Prof. Augusto Sousa (ProDEI Scientific Committee) Prof. F. Nunes Ferreira (ProDEI Scientific Committee) Prof. Eugénio Oliveira (ProDEI Director) Prof. Raul Vidal (DEI Director)

Automatic Extraction of Quotes and Topics from News Feeds

Luís Sarmento and Sérgio Nunes

Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
{las,ssn}@fe.up.pt

Abstract. The explosive growth in information production poses increasing challenges to consumers, confronted with problems often described as "information overflow". We present `verbatim`, a software system that can be used as a personal information butler to help structure and filter information. We address a small part of the information landscape, namely quotes extraction from portuguese news. This problem includes several challenges, specifically in the areas of information extraction and topic distillation. We present a full description of the problems and our adopted approach. `verbatim` is available online at `http://irlab.fe.up.pt/p/verbatim`.

Key words: Quotes Extraction, News Parsing, Topic Distillation, Named Entity Recognition, Information Extraction, Online Media.

1 Introduction

The current growth in information production poses increasing challenges to consumers, confronted with problems often described as "information overflow". How to deal with such a growing number of information sources? One possible answer to this problem are automatic software tools for information structuring and filtering, bringing some order to an unstructured information landscape. Tools that work as "personal information butlers". We present a contribution to this area with the development of `verbatim`, a system for quotes extraction and classification. `verbatim` acquires information from live news feeds, extracts quotes and topics from news and presents this information in a web-based interface. We choose to investigate the extraction of quotes because it combines several different views over all news topics. Both named entities (e.g. personalities, organizations) and current topics are present when extracting and distilling quotes. Also, we think that such a tool could work as an automatic "watchdog" by confronting quotes by the same entities on the same topics over time. Our main contribution is the deployment of a fully functional prototype, available online at `http://irlab.fe.up.pt/p/verbatim`.

2 Related Work

Quotes extraction from mainstream media sources has been addressed before in academic literature. Pouliquen et al. [6] present a system, named NewsExplorer[1], that detects quotations in multilingual news. The system is able to extract quotes, the name of the entity making the quote and also entities mentioned in the quote. Our work is different since it is focused on a single language (Portuguese) and also addresses the problem of topic extraction and distillation, while most of the related work assumes that news topics have been previously identified. Krestel et al. [5] describe the development of a reported speech extension to the GATE framework. This work is focused on the English language and, contrary to ours, does not address the problem of topic distillation and classification. In [4], the authors propose the TF*PDF algorithm for automatically extracting terms that can be use as descriptive tag. This approach generates a very large set of semantically related terms, but most of them are quite uninformative and innappropriate for being used as a high-level topic tag. Google has recently launched a tool that automatically extracts quotes and topics from online news — In Quotes[2]. The web-based interface is structured in issues (i.e. topics) and displays side-by-side quotes from two actors at a time. However, no implementation details are published about this system.

3 System Overview

The challenge of extracting quotes from live news feeds can be structured in the following generic problems: data acquisition and parsing, quotes extraction, removal of duplicates, topic distillation and classification, and interface design for presentation and navigation. Each problem is going to be addressed in the following sections.

3.1 Data Acquisition and Parsing

We are using a fixed number of data feeds from major portuguese mainstream media sources for news gathering. We opted to only include generic mainstream sources in this initial selection. This allowed us to avoid the major challenges faced in web crawling — link discovery and content encoding. Since the feeds are known in advance, we customized content decoding routines for each individual source. Also, since all sources publish web feeds in XML, the content extraction was straightforward. The fetching is scheduled to be performed periodically every hour on all sources. Thus, we need to account for news items that are downloaded multiple times. We use the URL (after handling redirects) of the news item to detect duplicate content. All content in stored in a UTF-8 encoded format on the server.

[1] `http://press.jrc.it/NewsExplorer`
[2] `http://labs.google.com/inquotes`

3.2 Quote Extraction

There is a variety of ways in which quotes can be expressed. Quotes can be found either in the title or throughout the body of the news feed. Some quotes are expressed *directly* using (supposedly) the speaker's exact words, while other are expressed *indirectly*, after being rephrased by the journalist. In some cases, the speaker is identified in a position that is adjacent to the quote (e.g. in the same sentence) while in other cases anaphoric resolution has to be made to find the correct speaker. This can involve, for example, matching an existing ergonym with the speaker's name (e.g. "The [Prime-Minister] said...."). Table 1 shows examples of some of the several different possible situations. In the

#	**Position**	**Direct?**	**Source**
1	title	yes	Costa Pinto: É óbvio que PS pedirá maioria absoluta
2	body	yes	"É indispensável uma ruptura com esta política de direita e é indispensável construir neste país uma alternativa à esquerda a estas políticas", afirmou Carlos Gonçalves, da Comissão Politica do PCP.
3	body	no	Hernâni Gonçalves também não acredita que os casos de corrupção que agora aparecem nas divisões distritais sejam a consequência do que se passa no futebol português ao mais alto nível.
4	body	mix	O vice-presidente do PSD, Aguiar-Branco, considerou que o primeiro-ministro perdeu uma oportunidade de "falar a verdade ao país".
5	body	yes	"Penso que um presidente ou presidente eleito e a sua equipa devem saber fazer várias coisas ao mesmo tempo. A propósito da situação em Gaza, sou colocado a par todos os dias", indicou aos jornalistas.

Table 1. Examples of several different types of quotes found in news objects.

current version of `verbatim` we mainly addressed quotes that explicitly mention the speaker in order to avoid anaphoric resolution. More specifically, we look for sentences in the body of the news feed that match the following pattern: *[Optional Ergonym], [Name of Speaker], [Speech Act] [Direct or Indirect Quote]* Using this pattern (and some minor variations of it), we are able to extract structures such as example (3) and (4) shown in Table 1. Because these structures are quite standard, the identification of each of the elements (ergonym, name of the speaker, speech act and quote) does not require extensive semantic analysis capabilities (such as noun-phrase identification and named-entity recognition), and can be achieved by using regular expressions and lists of words. Currently, extraction is made using 19 patterns (most small variations of the one previously shown), and a list with 35 speech acts (e.g. "afirmou", "acusou", "disse", etc.). In practice, about 5% of the news feeds match these patterns. Nevertheless, there

are still many other quotes, with different structures, that we are not able to extract in the current version of `verbatim`.

3.3 Removal of Duplicates

Since `verbatim` processes news feeds from several sources, it is quite usual to extract duplicate or near duplicates news from which duplicate quotes will be extracted. Such duplicate quotes do not provide any additional information to the user. On the contrary, they pollute presentation of data and damage statistics (e.g.: who are the most active speakers?). Therefore, after extracting quotes from news feeds, we proceed by trying to aggregate the most similar quotes in *quote groups*, Q_1, Q_2,... Q_{last}, with one or more quotes. The longest quote in the group is considered the *head* of the group. Every time a new quote is found, q_{new} it is compared with the head quote of each of the k most recent quote groups: Q_{last}, Q_{last-1}, Q_{last-2}... $Q_{last-k+1}$. If the similarity between q_{new} and the head quote of any of such groups is higher that a given threshold, s_{min}, then q_{new} is added to the most similar group. Otherwise, a new group, Q_{new} is created, containing q_{new} only, which becomes the new head of the group (although possibly only temporarily). This procedure is a very simplified approach of the *streaming clustering* algorithm [1].

Comparison between the new quotes q_{new} and the head quote for group k, q^k_{head} is made in two steps. First, we check if the speaker of each quote is the same. Currently, we only check for exact lexical matches (which may be problematic when there are small lexical variations such as "Ehud Barak" and "Ehud Barack"). If the name of the speakers is the same, then similarity between quotes is measured by comparing the actual content of the quote. We obtain the vector representation of each quote using a binary *bag-of-words* approach (stop words are removed). Vectors are compared using the Jaccard Coefficient. When similarity is higher that $s_{min} = 0.25$, then quotes are considered near-duplicates. Table 2 shows some statistics about the sizes of 730 groups of quotes that result from the aggregation of a set of 1,359 quotes extracted from the news sources we are consulting (several topics are covered). Only 427 out of 1,359 quotes are found to be unique. For most of the cases, the number of duplicate and near-duplicate is considerable.

# Quotes in Group	# Groups	# Quotes in Group	# Groups
1	427	5	20
2	125	6	6
3	98	7	4
4	47	≥ 8	3

Table 2. Number of quotes in the groups for 1,359 extracted quotes.

3.4 Topic Classification

`verbatim` tries to a assign a *topic tag* to each quote. However, because there is a very wide variety of topics in the news, topic classification becomes a quite complex problem. In fact, the set of possible topics in news is open: unseen topics can be added as more news are collected. Thus, efficient topic classification of news requires that the system is able to (i) dynamically identify new topics tags as they appear in the news, (ii) automatically generate a training set using that includes examples for the new topic tags, and (iii) re-train the topic classification procedure accordingly.

Identification of Topic Tags and Generation of Training Set The identification of topic tags is made by mining a very common structure found in news titles: "*topic tag*: *title headline*". For example: "*Operação Furacão*: Acusações estão atrasadas, admite PGR", "*Música*: Morreu Ron Ashton, guitarrista dos Stooges", "*Telecom*: Acordo sobre Redes de Nova Geração assinado 4ª feira..." or "*Sócrates/Entrevista*: Primeiro-ministro esteve 'longe' da verdade...". From a set of about 26,000 news items, we were able to find 783 different topic tags (occuring in at least two titles). For illustration purposes, the top 20 most common topic tags up to the first week of January 2009 are presented in Table 3 (the distribution of topic tags changes with time). We are thus able to generate a training set that matches each topic tag t_i from the set of topic tags found, $\mathcal{T}$, with a set of news items for that topic, $I_i = (i_{1i}, i_{2i}... i_{Ni})$ (i.e those items that contained the topic tag in the title). We will denote the complete training set as the $\mathcal{T} \rightarrow \mathcal{I}$, mapping from $\mathcal{T} = (t_1, t_2... t_k)$ to $\mathcal{I} = (I_i, I_2,... I_k)$.

#	Topic Tag	# Titles Found	#	Topic Tag	# Titles Found
1	Médio Oriente	172	11	Guiné-Bissau	95
2	Futebol	164	12	BPN	90
3	Música	150	13	Tailândia	87
4	Crise	137	14	Casa Pia	83
5	Lisboa	133	15	Brasil	72
6	Índia	132	16	Manoel de Oliveira	69
7	EUA	124	17	Literatura	67
8	Educação	113	18	Espanha	62
9	Saúde	108	19	PSD	55
10	Cinema	104	20	Cultura	53

Table 3. Top 20 most common topic tags.

Training the Topic Classifiers We use two different text classification approaches to assign topic tags to quotes: *Rocchio classification* [7] and *Support*

Vector Machines (SVM) [2]. Both involve representing news feeds as vectors of features. We use a bag-of-words approach for vectorizing news feed items. Let i_k be a news item, composed of title string, i_k^{title}, and a body string i_k^{body}. Let $[(w_1^t, f_1^t), (w_2^t, f_2^t), \ldots (w_m^t, f_m^t)]$ and $[(w_1^b, f_1^b), (w_2^b, f_2^b), \ldots (w_n^b, f_n^b)]$ be the bag-of-words vector representation of i_k^{title} and i_k^{body} respectively. The vector representation of i_k is obtained by adding the bag-of-word descriptions of i_k^{title} and i_k^{body} while concatenating the prefix "t_" or "b_" to each of the features depending on whether they come from the title or the body:

$$[i_k] = [(t\_w_1^t, f_1^t), (t\_w_2^t, f_2^t), \ldots (t\_wt_m^t, f_m^t), (b\_w_1^b, f_1^b), (b\_w_2^b, f_2^b), \ldots (b\_w_n^b, f_n^b)] \tag{1}$$

This vector representation allows us to keep information about the source of each word since such information may be important in topic classification.

Rocchio classification provides a straight-forward way to classify items using a nearest-neighbour assignment. For each class c_i, of a set of $|C|$ classes, we obtain $[c_i]$, the vector representation of the class. $[c_i]$ is computed from the vector representation of items from that class, $[i_{ij}]$, taken from the training set. Usually, $[c_i]$ is the arithmetic average of vectors $[i_{ij}]$, i.e. the *vector centroid*, although other item aggregation and feature weighting schemes are possible. Having the vector representation for each of the $|C|$ classes, classification is made by comparing the vector representation of a new item, $[i^{new}]$, against all the vector descriptions of the $|C|$ classes, $[c_1]$, $[c_2]$ ... $[c_{|C|}]$. Usually, the cosine metric is used to compare $[i^{new}]$ with $[c_i]$, i.e. $cos([i^{new}], [c_i])$. Item $[i^{new}]$ is then classified as belonging to the class corresponding to the closest $[c_i]$ vector. We used a variation of TF-IDF weighting for producing class descriptions. For each topic tag t_j in the training set $\mathcal{T} \rightarrow \mathcal{I}$ we start by adding the vector representation (Equation 1) of the corresponding news items $i_{ij} \in I_j$:

$$[c_k^*] = \sum_i [i_{ik}] = \left[(t\_w_1^t, F_1^t), \ldots (t\_wt_{|T|}^t, F_{|T|}^t), (b\_w_1^b, F_1^b), \ldots (b\_w_{|B|}^b, F_{|B|}^b)\right] \tag{2}$$

with F_1^t, F_i^b being the summation of individual frequencies found in each $[i_{ik}]$ for the $|T|$ and $|B|$ distinct word features extracted from the title and the body of the news items respectively. Let $f_{tpc}(w)$ be the *topic frequency* of word w, i.e. the number of vectors $[c_k^*]$ in which we find a non-nil component for the feature word w. Then, vector representation of class c_j, corresponding to topic tag t_j is given by:

$$[c_k] = [\left(t\_w_1^t, \frac{F_1^t}{f_{tpc}(t\_w_1^t)}\right), \ldots \left(t\_w_{|T|}^t, \frac{F_{|T|}^t}{f_{tpc}(t\_w_{|T|}^t)}\right), \left(b\_w_1^b, \frac{F_1^b}{f_{tpc}(b\_w_1^b)}\right), \ldots \left(b\_w_{|B|}^b, \frac{F_{|B|}^b}{f_{tpc}(b\_w_{|B|}^b)}\right)] \tag{3}$$

Support Vector Machines (SVM) provide an alternative approach to classification. SVMs have proven to be quite effective in classification tasks where

items are described by vectors in high-dimensional spaces, as is the case of text classification. A SVM efficiently finds the hyperplane that allows to separate items of two classes with *maximum* margin. SVMs are *binary classifiers*, and therefore require training multiple classifiers to perform classification under a multi-class scenario. Thus, for each topic tag t_k in the training set $\mathcal{T} \rightarrow \mathcal{I}$, we will need to train one SVM classifiers, using $I^+(k) = I_k$, the *positive examples*, and $I^-(k) = \mathcal{I} - I_k$, the *negative examples*:

$$svm_k = train_{svm}(I^+(k), I^-(k)) = train_{svm}(I_k, \mathcal{I} - I_k) \tag{4}$$

The training procedure converts each news item to its vector description $[i_k j]$ using the procedure explained before (Equation 1). After training for a given topic tag t_k, the corresponding svm_k, will produce a value from 1 to -1 when used to classify a given news item, i^{news}:

- $svm_k([i^{news}]) > 0$ if i^{news} belongs to class corresponding to topic t_k
- $svm_k([i^{news}]) < 0$ if i^{news} does not belong to class corresponding to topic t_k

We use the SVM-light [3] package to train an SVM for each topic, svm_k. Training is made using the package default parameters, and resulting SVM descriptions are saved in files for being used later, during the topic classification procedure.

Topic Classification Procedure The topic classification of a given quote is achieved by classifying all news item from which the quote was extracted. The intuition here is that using information from the entire news item should help us obtaining more evidence about which topic tag should be assigned to the quote. Thus, topic classification of a quote, is equivalent to topic classification of the news item i^{qt} from which the quote was extracted. Using the Rocchio classifier, classification is achieved by performing vector comparison between $[i^{qt}]$ (Equation 1) and the vector representation of all $|\mathcal{T}|$ classes, $[c_k]$ (Equation 3), and then choosing the tag that corresponds to the most similar $[c_k]$, as measured by the cosine metric, $cos([c_k], [i^{qt}])$. Likewise, classification using SVM's is made by feeding $[i^{qt}]$ to each of the $|\mathcal{T}|$ SVM classifiers, svm_k, and choosing the topic tag t_k that corresponds to $max(svm_k([i^{qt}]))$.

The two classifiers do not operate in the same way and, thus, do not always agree. Since we wish to obtain the most out of both classifiers, we developed a procedure for combining classification results. Let $\mathcal{T} = (t_1, t_2 ... t_k)$ be the set of topic tags over which the SVM and the Rocchio classifiers were trained, let i^{qt} be news items to classify, and let $[i^{qt}]$ be its vector representation. Then:

- compute svm_{max}, the maximum value given by $svm_k([i^{qt}])$, corresponding to $k = k^{svm}_{max}$.
- compute roc_{max} be the maximum value given by $cos([c_k], [i^{qt}])$, corresponding to $k = k^{roc}_{max}$.
- if $svm_{max} \geq min_{svm}$, then topic for i^{qt} will be $t_{k^{svm}_{max}}$
- else if $roc_{max} \geq min_{roc}$, then topic for i^{qt} will be $t_{k^{roc}_{max}}$
- else do not classify news item i^{qt}

This way, we give preference to classification made by the Support Vector Machines. We only refer to results provided by the Rocchio classifier when results from SVMs have low degree of confidence. In practice, we found that such combination procedure achieves reasonable results with $min_{svm} = -0.2$, and $min_{roc} = 0.2$. The negative value for min_{svm} arises from the fact that the number of positive examples used while training the SVMs, $\#(I^+(k))$, is always much less than the number of negative examples, $\#I^-(k) = \#(\mathcal{I} - I_k)$, leading to rather conservative classifiers, i.e. classifiers biased to produce false negatives. Still, for about 23% of the quotes we are not able to find any topic for which $svm_{max} \geq min_{svm}$, or $roc_{max} \geq min_{roc}$, These quotes remain unclassified and, therefore, are not shown to the user.

3.5 Web Interface

`verbatim`'s end-user web interface was developed over a standard LAMP technology stack. All data is stored on a MySQL database and published using CGI Perl scripts. Two interface screenshots are presented in figures 1 and 2. The homepage is shown in Figure 1 where several features are highlighted: (a) AJAX powered search over named entities (i.e. speakers); (b) last quotes found, ordered by number of supporting news; (c) most active topics, with the number of quotes in parentheses; (d) most active named entities, also with the number of quotes in parentheses.

Each quote has an individual page where the news items used to support the quote are listed. Figure 2 shows the interface for the topic "Médio Oriente" (Middle East). For each topic, the following features are available for direct inspection: (a) last quotes extracted for this topic; (b) navigational links to filter by named entity within this topic; (c) named entity search box, available in all pages. There is also a similar page for each single named entity where additional filters by topic are available.

`verbatim` includes user-friendly URLs for search, topics and named entities. For instance, the see all of Barack Obama's quotes the URL is — `http://irlab.fe.up.pt/p/verbatim/?w=Barack+Obama`. Finally, we have also included a data access API based on Atom web feeds. We publish web feeds for the last extracted quotes, last quotes by a given named entity and last quotes for a given topic.

3.6 Overall Update Routine

Because news are constantly being produced, `verbatim` needs to be updated cyclically. There are two update routines: the *quote extraction* routine, which runs every hour, and the *classifier re-training* routine, which runs every 24 hours. The quote extraction routine is the following:

- Read web feeds made available from the chosen set of news providers and store parsed news items in local database;
- For each unprocessed item in the news database, run the quote extraction procedure. Store the extracted information (id of source news item, name

Fig. 1. `verbatim` homepage.

Fig. 2. `verbatim` example topic page.

of speaker, optional ergonym, speech act, direct or indirect quote) in the database;
- Run the quote duplicate detection routine to group duplicate quotes together;
- For each unclassified group of quotes (which can be a singleton), run the classification procedure. Store classification in the database. Unclassified quotes are kept in the database (but will not be visible to the user) since it might be possible to classify them when classifiers are re-trained.

The classifier re-training routine updates SVMs descriptions and Rocchio class vector descriptions. This step is required for including new topics that might be found in the most recent news items, and for refining the descriptions of already existing topics with more recent information. SVMs require a complete training from scratch (i.e. training is not incremental), which, given the number of topics (several hundreds) and the number of news items at stake (several thousands), may require significant CPU resources. Therefore, we opted for performing such retrain every 24 hours. The classifier update routine is the following:

- Take all news items from database and run the topic detection procedure in order to build the training set $\mathcal{T} \rightarrow \mathcal{I}$, for topics $\mathcal{T} = (t_1, t_2 ... t_k)$ and new items $\mathcal{I} = (I_i, I_2, ... I_k)$;
- Vectorize all news items for all I_i in $\mathcal{I}$;
- Train Rocchio:
  - Compute topic frequency for all feature words;
  - For each t_k in $\mathcal{T}$ generate $[c*_k]$ by summing vectors associated with I_k, and obtain $[c_k]$ by weighing $[c*_k]$ with information about topic frequency for each feature word.
- Train SVM:
  - For each t_k in $\mathcal{T}$ generate $I^+(k) = I_k$, the set of *positive examples*, and $I^-(k) = \mathcal{I} - I_k$, the set of *negative examples* and make $svm_k = train_{svm}(I_k, \mathcal{I} - I_k)$.

The descriptions of Rocchio classes and SVMs are stored in file, and loaded when the topic classification routine is executed.

4 Results and Error Analysis

The database currently stores a total of 26,266 news items (as of early January 2009) after 47 days of regular fetching. During this period `verbatim` extracted a total of 570 quotes from 337 distinct named entities over 197 different topics. A ratio of roughly 1 distinct quote for every 46 news items. We found a small number of duplicate quotes being extracted as independent quotes (5 quotes). Figure 3 presents a plot of the number of news items together with the number of quotes over time.

We conducted a manual inspection over all quotes, named entities and topics to identify extraction errors. From a total of 570 quotes extracted, 68 were errors

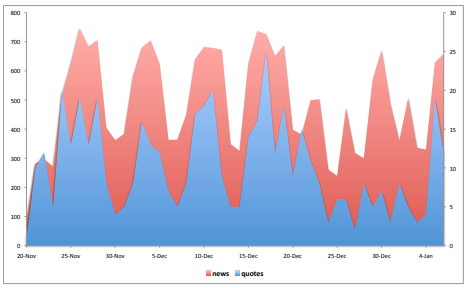


Fig. 3. News (left vertical axis) and extracted quotes (right) over time.

(11.9%). Also, in 337 named entities extracted, 6 (1.8%) were found to be errors. Finally, from the 197 topics identified, only 1 (0.5%) was an error. It is important to note that most of the errors found are minor and have little impact on the readability of the quote. Finally, to evaluate the algorithm for matching quotes to topics we also conducted a manual inspection over all quote/topic pairs. We found a total of 42 topics misattributed to quotes (7.4%).

5 Conclusion and Future Work

`verbatim` has proved to be a solid, fully functional online service working over live data from the portuguese mainstream media. Since the public launch of `verbatim` in mid November 2008, we have observed an average of 7.7 visits/day and a total of nearly 3,700 pageviews. The overall feedback, both online and offline, as been positive, and we plan to continue this line of research to improve `verbatim` in four main ways:

- **Increase the number of news sources**: after the initial proof of concept based on a small number of news feeds (8 in total), we plan to add a significant number of new sources to the database;
- **Improve quotations extraction**: as noted in the previous section, the ratio of quotations extracted is currently at 1 (distinct) quote for every 46 news items. Direct inspection of news items shows that this number can be easily improved by fine tuning additional matching rules, and creating new rules for other common pattern (both in news body and title). Using the information that we are able to gather about the mapping between names and ergonyms should also help increase the recall of extraction, since in many quotation there is no direct reference to the name but rather to the ergonym. Conflating variations of names for the same speaker, should also help to correctly group quotations of the same speaker;
- **Improving topic extraction and classification**: we found two problematic issues in topic extraction which require more attention. The first arises from the fact that news sources are not consistent about the words used to describe the topics in the headers (from which we extract the topic tag). This

leads extracting different topic tags from news, which in fact, refer to the same true topic (e.g. "Crise Finaceira", "Crise Económica"). In other cases, words used in news header refer to topics more generic than the true topic (e.g. "Desporto" instead of "Futebol"). Additionally, we would like to experiment using a different features to represent vectors, such as for example bigrams or part-of-speech tagged information.
- **Upgrade the end-user interface** As the number of quotes available in the database increase, the limitations of the current web interface become evident. For instance, as the archive of quotes by speaker becomes larger, the interface becomes cluttered. To overcome this problem we plan to add an additional navigational axis based on temporal information (e.g. filter by time feature).

Finally, we also plan to conduct a more extensive evaluation of `verbatim`'s performance by collecting a reference collection and computing traditional evaluation measures (i.e. Precision, Recall).

Acknowledgments Luís Sarmento and Sérgio Nunes were supported by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grants SFRH/BD/23590/2005 and SFRH/BD/31043/2006.

References

1. Sudipto Guha, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan. Clustering data streams. In *IEEE Symposium on Foundations of Computer Science*, pages 359–366, 2000.
2. Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
3. Thorsten Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999. software available at `http://svmlight.joachims.org/`.
4. Khoo Khyou and Bun Mitsuru Ishizuka. Topic extraction from news archive using tf*pdf algorithm. In *Proceedings of 3rd Int'l Conference on Web Informtion Systems Engineering (WISE 2002) ,IEEE Computer Soc*, pages 73–82. WISE, 2002.
5. Ralf Krestel, Sabine Bergler, and René Witte. Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), May 2008.
6. Bruno Pouliquen, Ralf Steinberger, and Clive Best. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing 2007*, Borovets, Bulgaria, 2007.
7. J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System*, pages 313–323, Englewood, Cliffs, New Jersey, 1971. Prentice Hall.

Virtual Tourism Business Networks in Developing Countries

Luís C.S. Barradas[1], J.J. Pinto-Ferreira[2]

[1] Escola Superior de Gestão e Tecnologia de Santarém,
Complexo Andaluz, Apartado 295, 2001-904 Santarém, Portugal,
claudio.barradas@esg.ipsantarem.pt

[2] Faculdade de Engenharia da Universidade do Porto/INESC Porto
Rua Roberto Frias, s/n, 4200-465, Porto, Portugal,
jjpf@fe.up.pt

Abstract. Since the rising of Computer and Information Technologies, that enterprises and organizations explore them to run their businesses, in order to explore new business opportunities, increase profits and decrease costs. Due several reasons, most of micro, small and medium enterprises located in remote communities, especially those located in developing countries, face a number of obstacles and difficulties due to their business structure and limited access to information and communication infrastructures. This paper presents the conclusive remarks of a master thesis that aims the development of a model for a virtual business network, bringing together tourism industry players located in remote communities, in order to provide them the opportunity to, at a low cost, join to the information society, explore new business opportunities and acquire global visibility. This virtual business network is fostered and relies on a JXTA P2P distributed e-marketplace model that seems to be an effective and alternative way to overtake the presented problem.

Keywords: e-Business, Business Network, e-Inclusion, Tourism.

1 Introduction

The yearning of travel and the desire to meet different people and to make relations with other civilizations has been a constant in man's history. Tourism is seen today as travel for recreation and became one of most dynamic, international and multimillionaire industries [1]. However, it is not a well defined industry due the fragmented nature of its product [2]. It requires services of leisure, lodging, transportations, hospitality, etc., that are provided by a wide range of other industry sectors. As other industries, the tourism has impact on the economy of the areas where it takes place (regions, countries or continents). These are known as touristic destinies, and most of them are totally dependent of tourism influx, for their economy support. This is a particular fact in the third world and developing countries [3]. This dependency isn't only relative to capital transference of the producing areas to the destination areas, through tourist spends [4]. Tourism business opportunities are

affected by factors such as country economy, geographic location, enterprise dimension and technological capacity [3]. Thus, a small tourism producer located in a remote community does not have the same facilities to run his business, as one located at North America or Europe.

This paper presents the work developed in the context of a master thesis [5] as well as further developments and testing realized after its conclusion. The project addresses three tourism sub-sectors, namely crafts, eco-agro-tourism and cultural heritage, which are interdependent and complementary for a number of activities and practices and strongly linked to declining rural areas. The main goal of this project is therefore to eliminate the digital divide barriers, creating equal opportunities and access to a global tourism virtual business network, where each player of target sectors can emerge globally, and everyone benefits being connected to each other. This presented virtual business network is fostered by a free and distributed P2P e-marketplace, aiming to enable local communities, the access to business collaboration services and give to their businesses a global visibility.

Section 2 presents an overview of tourism industry value chain and the Information Technologies tools that support it. In Section 3 is presented the model adopted in order to create the virtual business network, as well the core technology. Section 4 describes the experimental developments, where are described the core services developed as well the main software components that supports the network. At the end are presented the experimental tests and results. Finally at Section 5, some concluding remarks are made.

2 Tourism Industry

As in the distribution of physical products, the touristic distribution is a process composed by stages, through which flowing touristic products, since their production stage to their delivery to the consumer. This track is more or less long depending on the number of involved actors [1]. The tourism industry value chain includes four main actors:

- **Producers** – entities that produce the touristic products (e.g. agro-tourism farm, craftsman, etc.);
- **Wholesale Dealers** – typically known as tour operators, they combine goods and services that directly buy from producers.
- **Retailers** – sell the touristic packages from wholesale dealers to consumers.
- **Consumers** – the tourists.

Although not mentioned, there are other players involved in the tourism distribution chains, namely the regional tourism organizations, official organisms and governmental entities, whose activity is related to the coordination and promotion of the touristic destinies, providing also information to the wholesale dealers and retailers.

2.1. The Tourism Industry Support Network

The dimension of the worldwide tourism industry and the wide set of relations that it involves, suggest the existence of huge amounts of information being processed. The tourism industry is supported today by a large information network, which interconnects all players on its value chain [3]. This network is extremely important in the distribution, marketing and coordination of the activities, and includes: Computer Reservation Systems (CRS's) [6], Global Distribution Systems (GDS's) [7], [8], Internet based applications, and Digital Interactive Television Applications (DITA's).

CRS's are basically databases, that allow tourism producers and operators to manage their catalogs, making them simultaneously available to their business partners (e.g. room reservations and ticket emission by transportation companies).

GDS's are informatics systems that enable the availability checking, making reservations and ticket emission by tourism producers of any type, at a global scale [8]. There are currently four main GDS's available for travel agencies: Amadeus, Galileo, Sabre and Worldspan, all supported by consortiums of aerial transportation companies.

The Internet based applications, allowed to tourism consumers the direct access to touristic information and to make reservations, avoiding intermediaries. Beside this, the structure of interconnections created trough the Web, allowed to organizations the access to information about products and services, at a global scale, and simultaneously, the development of marketing actions, creating thus a bridge between the offer side and the demand/ buyer side, with great flexibility and interactivity, overtaking the electronic intermediaries, like the GDS's.

The Digital Interactive Television has been also adopted as a business channel by tourism operators. The DITA's take advantage of their ability of using the Internet, to sell or advertise touristic products and services, using a common TV set.

The market for e-Tourism has been growing quickly, but dominated by large tourism organizations, offering normalized products and supported by powerful marketing actions, communication and access to systems providing value added services, such as the ability to remotely make reservations in real time [7]. Most of the micro, small and medium enterprises of Alternative Tourism, specialty those located in remote communities, stay out of this restrict circle [1], [5], [9]. In order to eliminate this digital division, these small and medium enterprises need tools that enable them to:

- Have the opportunity to communicate and globally publish the diversity of their touristic resources, through a clear explanatory and consistent way.
- Be present on an information network that enables them to share their touristic offers at a world wide level.
- Promote an integrated and locally grounded economy chain for remote communities based on a strong local identity, leveraging on local natural, human and cultural resources.

The global reach provided by virtual business networking, gives increasing opportunities to expanding a business almost every day [10]. With new contacts, affiliations, referrals and growing customer awareness, a virtual business network lays the path for the success of a business right from the moment the business begins this venture [10].

3 The Tourism Virtual Business Network Model

With the arrival of internet, information technologies and advancements in the field of e-commerce, most of the traditional limitations and barriers are no longer a concern. Small or medium size businesses can compete nowadays in global markets. These small businesses can form groups also known as «business networks» to further improve their capabilities and reach [10]. In a business network, actors are autonomous and linked to each other through relationships, which are flexible and may change accordingly to fast changes in the environment. The stickiness that keeps the relationships is based on technical, economic, legal and especially on personal ties [11]. Organizations are moving, or must move, from today's relatively stable and slow-moving business networks, to an open digital platform where business is conducted across a quickly formed network with everyone, anywhere, anytime despite different business processes and computer systems [12]. Table 1 presents an overview of the characteristics of New Business Network Approaches, may have.

Table 1 – Characteristics of New Business Network Approaches[1]

Characteristics	Description
Products and services	Relative complex, bundled, and fast delivered products and services
Value creation	Demand networks with quick connect and disconnect relationships
Coordination and control	Network orchestration with distributed control and decision making
Information sharing	Information sharing over and with network partners
Infrastructure	Network platform with networked business operating system

3.1 Proposed Virtual Business Network Business Model

To reach the new Business Network Approaches characteristics, the business model for the proposed Tourism Virtual Business Network is based on the Peer-to-Peer distributed e-marketplace Model presented in [9]. E-marketplaces are an optimal solution, as a start point to buyers and sellers of target sectors meet each other, where suppliers try to sell their products and buyers try to satisfy their buying needs [9],[13].

P2P architectures provide far more decentralized infrastructures, while allowing a much wider range of business patterns to take place. By one hand, the interaction over a P2P network resembles the way real-world enterprises perform business with each other. On the other hand, a small set of simple services is enough to support complex business processes over a P2P infrastructure. Beside this, when compared to the traditional e-marketplaces, a P2P e-marketplace overcomes all their disadvantages [9], [13], providing more dynamic and complex relationships [9].

The proposed distributed P2P based e-marketplace model builds on the infrastructure presented by Fig. 1, that suggests a direct mapping of target tourism

[1] Adapted from [12]

sectors players, including producers, travel agencies and even consumers, on a heterogeneous P2P network.

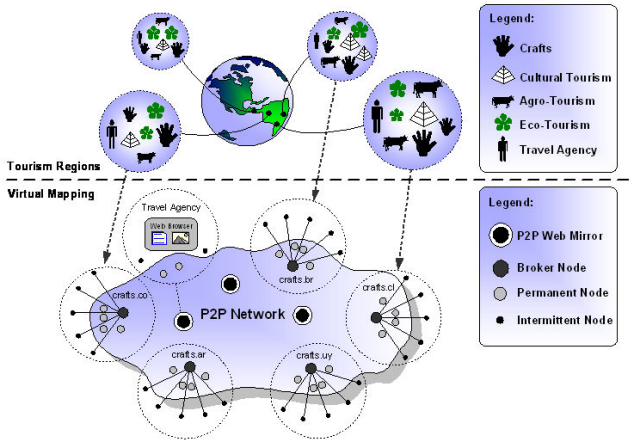


Fig. 1 - The P2P Network Infrastructure [9]

The model comprises four peer types: Web Mirror nodes, Broker nodes, Permanent nodes and Intermittent nodes. Web mirror nodes provide an entry point for all players on the Business Network. They are mainly targeted for consumers and allow them to search tourism offers on the P2P network, using a simple web browser. Broker nodes are collaborative nodes. They are mainly targeted to local tourism official organisms (or Tourism Regions Entities), and their role is to help small producers to keep their offers constantly available on the network. The remaining nodes could both be targeted to producers or travel agencies. They are different from the previous nodes because they may not have a permanent connection to Internet.

The ability to provide free services [5], [9], [13], the network scalability capacity, the speed of grow capacity and the P2P based services as Virtual Presence, Instant Messaging, Share and Collaboration, seems to be the key for a free e-marketplace that leverages a Virtual Business Network that allow to create a network effect, where everyone can benefits to be connected with everyone.

3.2. P2P Related Technology

There are today a wide range of P2P development technologies as Gnutella [14], Freenet [15], Jabber [16] or JXTA [17]. The first three solutions have been created for specific purposes as file sharing and Instant Messaging. JXTA has been created to develop heterogeneous P2P networks for general purposes. It is an open source project that defines a set of XML based protocols that establish an ubiquitous, secure and pervasive virtual network on top of IP and non-IP networks, allowing peers to directly interact and be organized independently of their locations on the network, that can be behind or not a firewall or NAT (Network address translation). The JXTA 2.0 specification [18] builds up on six independent language protocols:

- *Peer Discovery Protocol* (PDP) – Resources search and advertisement;
- *Peer Resolver Protocol* (PRP) – Generic query service;
- *Peer Information Protocol* (PIP) – Net and peer monitoring;
- *Pipe Binding Protocol* (PBP) – Addressable messaging;
- *Rendezvous Protocol* (RVP) – Propagation services;
- *Endpoint Routing Protocol* (ERP) – Message routing service;

All these protocols are available as services and can be used for the development of new richer high level services, as Instant Messaging, File Sharing, etc. JXTA has been also designed to be ubiquitous. Any device including mobile phones, PDAs, personal computers or servers are able to host a JXTA application. The information about peers, pipes and any shared resources as services, contents or files is represented by advertisements – a piece of XML structured information that describes a peer a pipe or any other resource. The communication between two JXTA peers is basically a trade of XML based messages.

The security aspects are tackled by existing and well matured technologies, as transport layer services, digital certificates and certificate authorities. Thus, JXTA provides the ability to integrate heterogeneous information sources in a decentralized, self-organized and secure way [17].

4 Experimental Developments

The developed prototype focuses mainly in the tasks of assembling and publishing tourism offers by tourism producers, as well the respective tourism offer searches by possible buyers. Another emphasized aspect is the possibility of the establishment of real time business interactions between trading parts through instant messaging. This functionality, besides allowing an easier communication between trading and business partners, induces to the development of business relationships and partnerships, through the creation and management of business partner lists, and status monitoring in the business network. The P2P tourism e-marketplace comprises three core software applications: Web mirror, Enterprise Application and Search Application. All these applications build on an infrastructure based on JXTA P2P Services, as shown in Fig. 2.

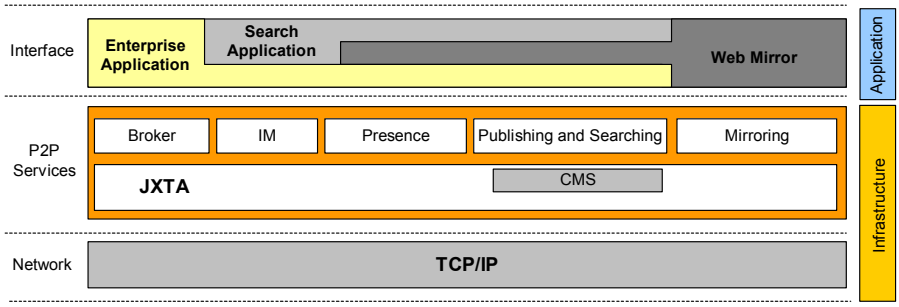


Fig. 2 Global multi-layer architecture of the tourism e-marketplace

4.1 Core P2P Services

The core P2P services provide most of the core functionalities of the e-marketplace, and are known as: Broker Service; Instant Messenger Service; Presence Service; Publishing and Searching Service; and Mirroring Service. Although all these services were vital for the support P2P network, each application implements only the services needed to fulfill its functional requirements.

Broker Service. The broker service provides the required mechanisms so that a peer can act as broker, ensuring that the owners of intermittent nodes can make their tourism offers always available on the network. The permanent availability is made by transferring the tourism offer from the intermittent node of a given Producer to a pre-selected peer running the Broker Service. This service builds on three JXTA services: *Discovery Service*, *Pipe Service* and *Content Management Service*. The first is used to find a Broker Peer; the second is used to open the communication channel between both peers, which is used for message exchanges. Messages define the type of operation to be performed by the broker: transfer, remove, share, unshare (_TRANSFER, _REMOVE, _SHARE, _UNSHARE). At last, the role of the *Content Management Service* is to transfer the offer files from their source to the node running the Broker Service. Once transferred, the broker node publishes the offer and sends back to offers owner the new JXTA Offer Advertisement, which includes the new location address of the offer file. This address is then used to state the real location of the offer file, when the Web Mirror databases are updated with the new record related to the uploaded offer.

Instant Messenger Service. This service allows the establishment of a complex web of business interactions between sellers and costumers or business partners. These interactions range from simple information requests about products our services, to the definition of contractual terms, payment conditions etc. The service architecture builds on two core JXTA services – the *Resolver Service* and the *Pipe Service*, used for chat requests and communication channel binding, respectively. The negotiation for starting a chat session, involves two different messages: the first is used for asking for a IM session (*InitiateIMRequest*) and includes the requester name and his e-mail address; the second message is the answer for the IM request (*InitiateIMResponse*) and includes the name and the e-mail address of the answerer, as well the *Pipe Advertisement* that the requester peer must use to state the communication channel. These information exchanges allow the control of the received IM requests by a user, accepting only those that he desires.

Presence Service. The presence service provides the indispensable mechanisms so that a user manages his presence status on the network, and also monitors the presence status of his business partners. The status information of a participant is represented by a Presence Advertisement, a XML based piece of information, whose structure is illustrated on Fig. 3. The presence information includes the peer identification (PeerID) the name and e-mail address of the user (E-mailAddress and Name) and finally the presence status of the user (PresenceStatus). The presence status can assume six different values: offline, on-line, busy, away, be right back, on

the phone and out to lunch, respectively. The service's architecture builds on the JXTA *Discovery service*, following the model proposed by Wilson [19]. Thus, the presence service relies on the *Discovery Service* [18], [19] capabilities to publish Presence Advertisements, as well as to discovery and get Presence Advertisements of other participants. This model allows that the user presence information can be obtained through his e-mail address.

```
<?xml version="1.0" encoding="UTF-8"?>
<PresenceAdvertisement>
    <PeerID>urn:jxta:uuid9615461646162614A78746150325033F3BC76FF
    13C2414CBC0AB993666DA53021</PeerID>
    <E-mailAddress>empresaa@mail.pt</E-mailAddress>
    <PresenceStatus>1</PresenceStatus>
    <Name>Empresa A</Name>
</PresenceAdvertisement>
```

Fig. 3. A presence advertisement

Publishing and Searching Service. This service provides the necessary mechanisms for publishing and searching tourism offers. The service builds on a user service called *Content Management Service* (CMS), whose purpose is to provide the share, search and transfer of files within a peer group [18], [19]. Relying on the CMS capabilities, this service provides the publishing, searching and transferring tasks of tourism offers on the P2P network.

Mirroring Service. The mirroring service provides the necessary mechanisms so that the summary information about an offer published by given producer can be stored on the Web Mirrors databases [5], [9]. Its architecture builds on two JXTA services: the *Discovery Service* and the *Pipe Service*. The first is used to find the existing Web Mirrors, while the second is used to the establishment of the communication channel between the peer hosting the Enterprise application and the existing Web Mirrors.

4.2 Distributed e-Marketplace Software Components

Enterprise Application. This application is targeted to tourism producers or Official Organisms [1]. Provides producers with the functionalities to create, manage and publish their offers, search offers, manage their presence and monitor the presence of their partners on the network, and manage their contact lists. All of these functionalities are provided by a set of modules:

- *Offer management module* – Allows the management of tourism offers including functions as publish offers (locally or remotely using a broker), unshare offers and remove offers (Fig. 4).
- *Offer creator module* – Provides a tool for tourism offers creation and edition. It supports plug-ins, in way to add some flexibility for different tourism sectors offers support.
- *Offer search module* – Allows the search of tourism offers on the P2P network, using offers' keywords as search key.

- *Instant Message module* – Provides a tool for users initiate IM sessions with their business partners.
- *Presence management module* – Provides a tool for user presence management on the network as well as the presence monitoring of the users available in the contacts list. It also allows the search of users on the network having their email address as search key.

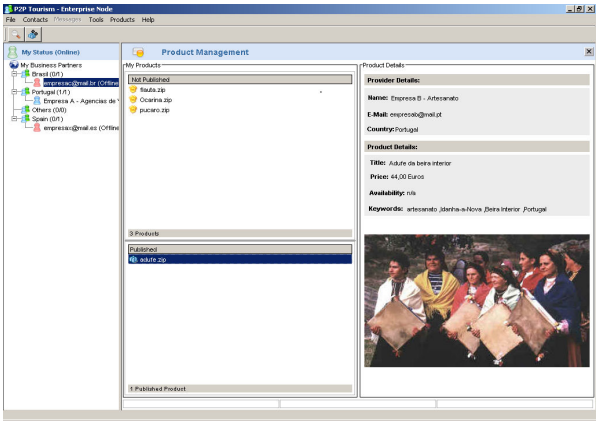


Fig. 4. Enterprise Application interface: Offer management tool

All of these modules are supported by a set of JXTA based services, namely the Presence service, the IM service, the Broker service and the Publishing and Searching service. This application can run Broker Services to help other peers.

Search Application. Once e-marketplace members, intermediaries such as tour operators and travel agencies need functionalities that enable them to search tourism offers, find and lookup business partners in an easy, fast and straight way. The Search application is targeted to these players (Fig. 5). It may run on a permanent or intermittent node. This application is in fact a simplified version of the Enterprise Application, providing the same modules and functionalities, except those for offer management and creation.

Web Mirror. The Web Mirror Application is a web based application that provides centralized points on the network, and works as the entry point for a user to join to the distributed tourism e-marketplace (Fig. 6). The services provided by the Web Mirror Application can be from different types: general users or visitors services, and e-marketplace support services. For visitors, the Web Mirror provides a service for offer searching, either on the local databases or in the P2P network through a web browser. The support services include the member's only services such as: Member account services; Software and updates services; Mirroring service for local databases update. The Web Mirror operation is supported by three JXTA based services namely the Presence Service for presence monitoring; the Publish and Searching Service for P2P searches and offers retrieving; and the Mirroring Service for local databases update.

The core technology is Java based, namely Java Beans and Java Server Pages.

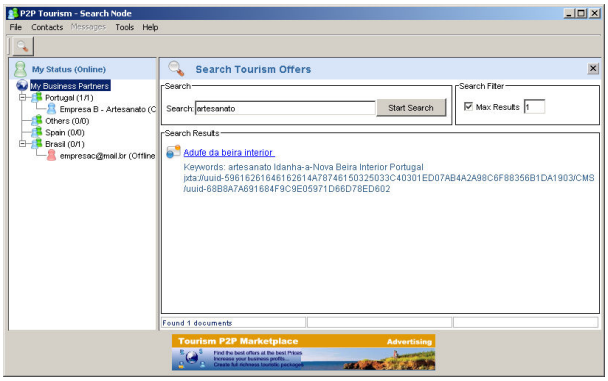


Fig. 5. Search Application interface: Offer search tool

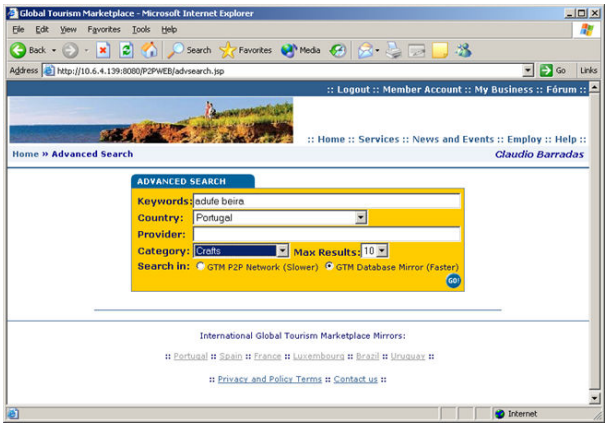


Fig. 6. - Web Mirror interface

Tourism Offers. The tourism offers consists on a combination of a set of documents including media files, as audio, video, pictures and text based contents, all arranged and structured by a XML structure [5], [9]. These media files joined together represent an offer that can travel on the JXTA P2P network [20] and presented in a web browser that supports XML and XLS transformations. By the one hand, this architecture allows conferring some tangibility to the tourism offers, a desirable characteristic for any digital tourism offer [1]. On other hand also provides the flexibility needed for integrating offers from different tourism sectors.

4.3. Security

At the current status of the prototype the security aspects are quite basic, once the main focus was centered on the aspects related to the publication and search of

tourism offers and to the establishment of contacts between business partners. Apart the JXTA platform basic security mechanisms, which requires that any JXTA peer must automatically log in the default JXTA peer group - "NetPeerGroup" [17], the tourism e-marketplace has its own peer group denominated as "*JXTA Tourism Group*". Having its own peer group, the e-marketplace peers are bounded by a logic segmentation of the JXTA network, living on a protected network space with well defined limits. All the communications and resources shares are made on the group scope, which makes them invisibles behind the group boundaries. Although these mechanisms can provide some security, even though minimum, it can be strengthened using digital certificates, encrypting all communications.

4.4. Experiments and Results

The experiments and final tests were done in a semi-closed laboratorial environment, and were focused in the tasks of sharing; searching and transferring tourism offers as well the start of basic trading interactions. Table 1 summarizes the peer types, locations as well the running services, of the peers that constitute the experimental P2P based e-marketplace.

Table 2. Summary of involved Peers in experimental tests.

Application Type	Number of running instances	Location	Running Broker Service
Web Mirror	1	Inside Firewall	N/A
Enterprise App	6	Inside Firewall	Yes
Search App	13	Inside Firewall	N/A
Enterprise App	1	Internet	Yes
Search App	2	Internet	N/A

All the services and functionalities provided by the applications were tested simultaneously, in order to simulate a real situation of the business network operation. The business network worked as expected in all tests executed inside the firewall. The peers located in the internet had some difficulties to communicate with the peers inside the firewall, taking a few seconds to complete the connection. It was also observed a considerable raise in the amount of network traffic, a typical and undesirable feature of the P2P based networks.

5 Conclusion

Typical e-marketplaces supporting technologies require some expertise, powerful hardware resources and infrastructures, which carry on large supporting costs. Consequently, most of these costs are passed to e-marketplaces' members, causing an increase of their participation costs. Promoting a virtual business network supported by a P2P based tourism e-marketplace seems to be the best strategy to achieve the low

costs and create a network effect that will ensure the success in reducing the "digital divide"[9]. The free services provided by the e-marketplace, and the low cost of operation may foster the rapid achievement of critical mass [9], [13], leading to a huge number of buyers and sellers trading and exploring new business opportunities. From a technical perspective, the performance and flexibility provided by the proposed distributed e-marketplace support P2P network ensures the support of the sustained growth of peer nodes in the network.

References

1. Cunha, Licínio: Introdução ao Turismo, 2ª Edição, Editorial Verbo, ISBN: 972-22-2085-3, (2003)
2. Rodrigues, Maria P.: Princípios Gerais de Turismo, Universidade do Algarve, (1998)
3. Holloway, Christopher J.: The Business of Tourism, 4th edition, Longman, ISBN: 0-582-29042-2, (1996)
4. Travel & Tourism's Economic Impact, http://www.wttc.org
5. Barradas, Luís CS.: Integração Inter-empresarial do Negócio em Rede, Dissertation on Electrical and Computers Engineering master degree, Faculty of Engineering of University of Porto, (2005)
6. Buhalis, D.: La empresa turística virtual. Conceptos, praticas e lecciones, Papers de TuriPME, Generalitat Valenciana (1998)
7. Costa, Jorge; Rita, Paulo; Águas, Paulo: Tendências Internacionais em Turismo, Lidel Edições Técnicas, ISBN: 972-757-145-X, (2001)
8. Vialle, O.: Les SGD dans l' industrie Touristique, Étoude réalisée pour l'OMT, (1994)
9. Barradas, Luís CS; Pinto-Ferreira, J.J.: P2P Infrastructure for Tourism Electronic Marketplace, in proceedings of 5th IFIP Working Conference on Virtual Enterprises, Toulouse, France (2004)
10. Business Networking Tutorials - Creating Virtual Business Network, http://www.exforsys.com/career-center/business-networking/creating-virtual-business-network.html
11. Hollensen, Svend: Global Marketing: A Decision-Oriented Approach, 4/E, Financial Times Press, ISBN: 9780273706786, (2007)
12. Heck, E. V., Vervest, P.: Smart Business Networks: how the network wins, Communications of ACM, Vol. 50, No. 6, June (2007)
13. Ferreira, Diogo; Pinto-Ferreira, J.J.: Building an E-marketplace on a P2P Infrastructure, in proceedings of 18th International Conference on CAD/CAM, Robotics and Factories of the Future, Porto (2002)
14. M. Ripeanu, Foster, I., Iamnitchi, A.: Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design, IEEE Internet Computing, 6, February (2002)
15. The free network project", http://freenet.sourceforge.net
16. Jabber Software Foundation, http://www.jabber.org
17. Traversat, Bernard: Project JXTA 2.0 Super-Peer Virtual Network, May 25, (2003)
18. JXTA v2.0 Protocols Specification, https://jxta-spec.dev.java.net/net (2005)
19. Wilson, Brendon: "JXTA", New Riders Publishing, 1st edition, ISBN:0735712344, (2002)
20. Contreras, P; Murtagh, F; Englmeier, K.: Distributed Multimedia Content with P2P JXTA Technology, HCI-2003, Crete, (2003)

A New Paradigm for Automated Schematic Maps

João Manuel Curralo Mourinho
Teresa Galvão Dias

Faculdade de Engenharia da Universidade do Porto, Portugal
joaomourinho@gmail.com, tgalvao@fe.up.pt

Abstract. The present paper aims to contribute to the research of enhanced solutions for the design of automated schematic maps, and is the result of a research work in progress. Schematic maps are now widely used for depicting the transportation system networks and are the result of the technical, economic and social changes that took place the last centuries. The industrial revolution brought changes to the transportation systems which had to support new economic and social models. Those transportation systems had become increasingly complex and demanded for a new kind of maps which could be easily understood by the users, the schematic maps. Nevertheless, the progress in the information and mobile computing technology area, the increasing complexity of the transport networks and the developments on the field of cognitive psychology demand a new approach for the creation of automated schematic maps in order to increase their effectiveness. The research in progress aspires to conceptualize a new paradigm for the automated schematic maps which could have high levels of user satisfaction. In order to achieve this goal, the new paradigm is based in a real-time computing model, being able to answer to on-demand user queries.

Keywords: Schematic maps, maps on-demand, design, transport system, information integration

1 Introduction

Since the most remote times, man has tried to dominate the world by capturing the two dimensions of his existence: time and space. As he made watches to represent time, he also made maps to represent space. The making of maps was only possible through the use of symbols and abstraction, as maps are mainly intended to communicate space information. They use a language, a conceptualization of the reality. However, as Harley and Woodward state in [1],

> *"(...) we must accept, although our general position is founded in semiology, that precise scientific analogies to the structure of language may be impossible to sustain."*

Therefore, maps require a different kind of conceptual structures to effectively communicate the spatial information.

At the late 18th century, the industrial revolution brought a wide set of scientific, economic and social changes which pushed new developments in the transportation field. The geographical world has already been discovered but the transportation systems (railways, roads, airways, underground systems, high-speed trains) have been growing till today, and they are expected to continue to grow. Large urban areas appeared and needed complex transportation systems, combining different transportation types. The need of highly efficient, easily understandable transportation maps pushed the evolution of the traditional maps, and new forms of cartographic representation have emerged.

Among the new forms of cartographic representation that have emerged are the schematic maps. The schematic maps were a new kind of map which appeared in response to the need of better and simpler maps to describe complex transport networks. One famous schematic map applied to a transportation network was the Harry Beck's London Underground diagram, depicted in figure 1.

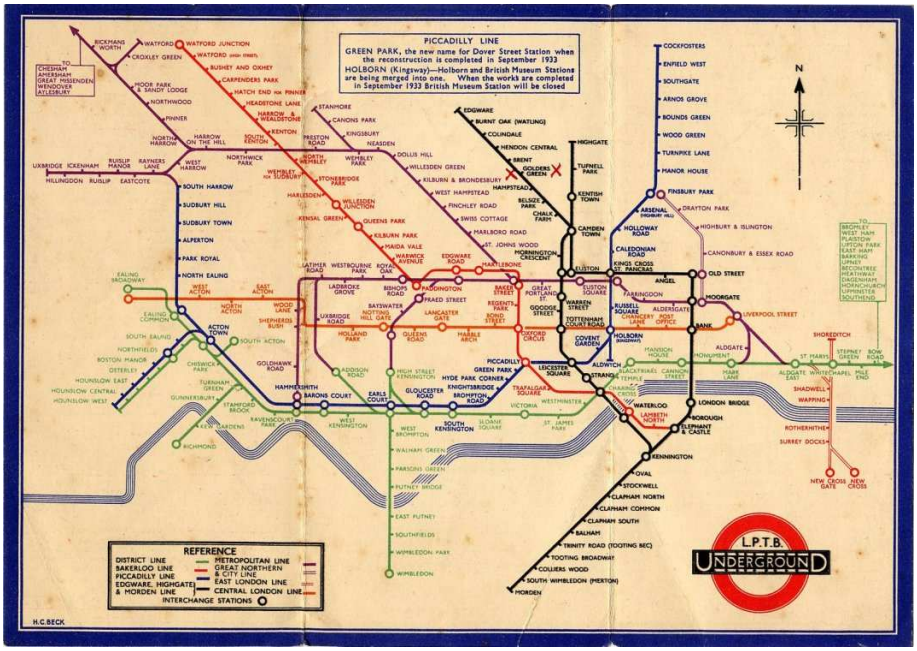


Fig. 1. London Underground Diagram, published in 1933

This diagram was considered an innovation, as for the first time lines were drawn either horizontally, vertically or diagonally at 45^{o}. This map also uses a non-linear scale, so the central area of the diagram is shown at a larger scale than the extremities. It shall be noted that although it doesn't mimic the geography of London, this map give the traveler some clues about the transportation and

schematic maps in the area and his/her location (one particular example is the river location).

Since Beck's map was published, the schematic maps produced are still based its principles, but the transportation networks are currently more complex than in 1933. Nowadays, there are huge cities which have several kinds of interconnected transportation infrastructures. In the globalization era, people need to travel more often and quicker. The information era brings a enormous amount of information to the people but they just want to know what they need to know. Time is becoming a precious resource and so people don't have time to dig the information they need. So a new kind of maps is required, in order to reduce the wasted time spent by people watching and trying to understand maps. A new paradigm in geographic representation is essential to put together the latest advances in computing technology, human perception and transportation system information. This new paradigm needs to be centered in the individual user and not in the "common user", as each person has its individual preferences. Maps should be tailored to each user's demands. Nowadays there is a predefined set of maps which are directed to large groups of people, thus not respecting their personal dimension, capabilities and experience.

The last decade has seen the democratization of the mobile devices which have been increasingly powerful and portable, and its use has become inherent to people's life. Making a call, checking the email box, listening to music, playing games or even knowing the exact geographic positions of our current position are now tasks at the distance of a finger. The information society is evolving and changing its habits, is becoming more agile and demands for new paradigms. In the transportation and schematic maps area, the implementation of the new paradigm is realistic and pertinent as it would bring tangible benefits to the society and economy, and would bring higher satisfaction levels to its users. Although there are some steps in that direction, there is the lack of a truly integrated vision which could drive the next generation of schematic maps.

This paper is organized as follows. Section 2 gives an overview of the related work in the area, regarding the concepts in which this study is based. Section 3 describes the procedures for the implementation of the new paradigm. It finishes with the conclusions on section 4 and some suggestions for further research.

2 Related Work

Although schematic maps are widely used, the few related studies available are recent. The researchers usually investigate isolated areas of the problem, not having in consideration the schematic maps as a whole.

2.1 The concept of Schematic Maps

There were some researchers who developed the schematic maps by systematizing the concepts who were used previously in an ad-hoc way. Neyer studies [2] proposed the 4 direction line schema (vertical, horizontal and both diagonals).

Nevertheless this study focuses on just one path (in the case of metro maps, they usually have more than one path), not analyzing the interaction between paths.

Barkowsky *et al* presented a systematic algorithm [3] to schematize maps. Although this study used discrete curve evolution, it didn't take in consideration the study of any perception factor nor scale normalization in crowded areas. This issue was studied by Avelar and Muller [4] which improved the schematic map generation preserving its topology by using only simple geometric operations. Nevertheless it caused some design conflicts as labels weren't considered in the study.

Cabello *et al* [5] implemented an algorithm to displace stations to achieve a better readability, and a more recent study [6] considers an alternative algorithm which proposes the creation of cells (which may be circled/squared/Voronoi) around each point for better alignment. Once again, no consideration was made on the human perception factors regarding this alternative algorithm.

2.2 The automated drawing of Schematic Maps

The only two extensive attempts to research the automated drawing of schematic maps were carried on by Martin Nöllenburg [7] and by Silvania Avelar [8]. Martin Nöllenburg studies make an extensive research on the discrete mathematical foundations which are the basis of the algorithms used in the drawing of schematic maps and makes some brief considerations about their implementation. Nevertheless, his studies doesn't cover the human perception factors nor a a concrete computer framework for drawing schematic maps. Silvana Avelar presents a wider study, by including some human perception factors and makes a complete research on the "schematic maps on demand", one of the components to be integrated in the new paradigm. She goes further on by presenting a framework for electronic schematic maps which can answer user queries and studies the generation of automated generation of schematic maps. Nevertheless, the study of the human perception factors is limited to what she calls the "aesthetic factors", and the framework lacks the complexity demanded in the real world.

2.3 User Centered Map Design

Recent studies [9] had been made on the exocentric/egocentric perspective of the human vision. They propose using 3D map visualization and GPS positioning to achieve more efficient, less erroneous and more user friendly maps regarding the traditional paper maps or even electronic north-up or head-up maps. The tests made proved that the egocentric 3D visualization is always better. Nevertheless, the studies don't mention how to achieve real time on demand maps by the use of a computer framework and don't make feasibility studies.

2.4 Customer satisfaction

As users makes use of the services provided by someone (for example a transportation company), they may be considered a customer. Although customer satisfaction is a subjective concept (and inherently difficult to measure precisely), there are studies [10] which consider customer satisfaction a result of the combination of the following sets of quality factors:

Basic Factors: Requirements that if not met will generate dissatisfaction, but if they are met they will not cause customer satisfaction. (Also known as Dissatisfiers).

Excitement Factors The requirements that if not met, will not generate consumer dissatisfaction but if met they will increase customer satisfaction. Those factors are considered to be the ones that provide differentiation and competitive advantage.

Performance Factors These factors are related to objective customer demands: they increase satisfaction if their performance is high, and cause dissatisfaction if it is low.

Recent studies [11] on user satisfaction investigate how different elements of information quality of mobile information services affect consumer satisfaction and, eventually, acceptance of these services. They concluded the following facts about user satisfaction:

1. Content quality, connection quality, interaction and contextual quality are all positively related to user satisfaction
2. User satisfaction is positively related to intention to use the service
3. The behavioral goals of the user affect the magnitude of the relations between the information quality dimensions and user satisfaction. The effect of content quality is much stronger for the users with hedonic goals.

The authors of those studies conclude that:

> *"All in all, our results suggest that, for a mobile service to become successful, all quality dimensions must be in top form. Continuous use of a mobile, service is determined, based on the level of user satisfaction, and although 'content is king', user satisfaction is effected by the whole package. When the service is closely linked to the use context and the content is relevant and effortlessly accessible, the service is likely to create real added-value to the users and the potential of success is increased."*

3 Procedures for the implementation of the new Paradigm

The conception of the new paradigm on the generation of schematic maps needs a multidisciplinar approach. The following areas of knowledge are fundamental:

- Information Systems

- Cognitive Psychology (pattern recognition capability, graphical mental representation)
- Geography (geographic item representation, topological information)
- Transportation System (schematic maps, analysis of user behaviours, trends in transportation networks)
- Information Integration (integration of sparse information needed to enhance the effectiveness of the whole system)

The information systems are the functional base of the new paradigm as they are transversal to the generation and visualization of the schematic maps. The architecture of the information system needs to support the on-demand schematic map philosophy.

3.1 Information System Model

The information system that shall support the new paradigm needs to function in a real-time way, as the maps shall be generated on-demand. People who demand the maps do not want to wait lots of time for the map (if it happens, people will probably not use the system anymore), as they need to travel, so time is a critical factor here. Sommerville[12] distinguishes two types of real time systems: hard real-time systems and soft real-time systems. The author defines both concepts as follows:

> *"A real-time system is a software system where the correct functioning of the system depends on the results produced by the system and the time at which these results are produced. A soft real-time system is a system whose operation is degraded if results are not produced according to the specified timing requirements. A hard real-time system is a system whose operation is incorrect if results are not produced according to the timing specification."*

Sommerville suggests a generic model of a real-time system, shown in figure 2.

This model fits perfectly the requirements of a new paradigm. Nevertheless there is some criticism about this model, most of it related with the impossibility of defining an explicit deadline – i.e., an exact, very specific, time from where the value of the service exhibits a notorious degradation. However, the new paradigm would make use of a soft real-time instead of a hard real-time architecture. Some authors [13] claim that although hard real-time systems are very predictable, they are not sufficiently flexible to adapt to dynamic situations. They are built under pessimistic assumptions to cope with worst-case scenarios, so they often waste resources. Soft real-time systems are built to reduce resource consumption, tolerate overloads and adapt to system changes. They are also more suited to novel applications of real-time technology, such as multimedia systems, monitoring apparatuses, telecommunication networks, mobile robotics, virtual reality, and interactive computer games. Therefore, soft real-time systems are more suited to this new schematic map paradigm.

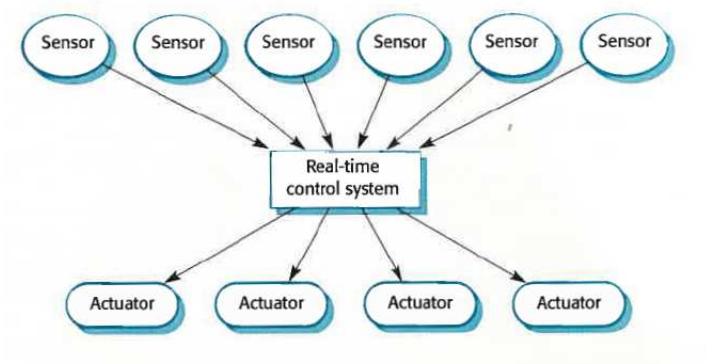


Fig. 2. General model of a real-time system

The sensors could be the schematic map hotspots located at every metro map of a metro network, for example. The hotspots would redirect the requests to the central real-time control system which could make an analysis of the actual load of the transport network and, in an intelligent way, generate the best possible on-demand schematic maps which would fit in a better way the requests and send it to the actuators. In other way, the sensors may be servers waiting for requests sent by mobile devices, as nowadays people often use them (ex. PDA). The servers would redirect the requests to the core real time system which would again send the response through the actuators.

3.2 Human perception factors and the good schematic maps

In what concerns the human perception factors, the development of schematic maps has referred them as being only "aesthetic criteria" [8]. Nevertheless, there are no scientific studies which determinate the significant set of human perception factors nor its relation with the quality of the schematic maps. So a deep investigation needs to be done, by making experiments and using quantitative (true experimental plans) and qualitative research methods (interviews, observation, document analysis, content analysis).

Recent studies [14] also claim that schematization is a necessary process as the human cognition is limited: it absorbs simplified information much faster than complex information. Therefore complex information has a negative impact upon wayfinding performance while schematized maps enhance the wayfinding performance [CITAR JACKSON 1998, P1000 -> Tá no paper].

3.3 Transportation systems and information integration

Nowadays, big cities have complex transportation systems which contain different kinds of transport networks (bus networks belonging to several companies, metro systems, railway systems, airways, waterways systems, bicycle networks).

The great challenge is to integrate all that information in order to make them working together, making it easier for the people to travel efficiently (without waste of time). The information integration plays here a fundamental role: by using the sensor information the information systems of the transportation networks should be able to have an information base which may share between them. This information may be retrieved to the users in the form of on-demand effective schematic maps and route plans.

4 Conclusions

The social and economic progress lived by the societies after the industrial revolution, together with the advances in computation technologies bring the need of a new paradigm on the schematic maps. The schematic maps, by their inherent simplicity and symbolic meaning are considered [8] the best maps for being used in the transportation area as they are far more intuitive than conventional maps. Nevertheless, the development of the schematic maps needs to integrate in a systematic way the human perception factors and the advances in the information systems. This is the motivation for the conceptualization of a new paradigm for the automated schematic maps. This new paradigm is expected to generate higher quality schematic maps. This quality may be measured through the content and contextual quality of the generated schematic maps. The content and context will be user-centered as it is the user who makes the requests, so the user-centric approach of the new paradigm increases the content and contextual quality, it is tailored to meet user's needs.

And according to Kanu *et al* [11], one of the quality factors that directly influence user satisfaction is the performance factor (also called "attractive factor"). This factor is related to objective customer demands: it increases satisfaction if its performance is high, and cause dissatisfaction if it is low. As the new paradigm will work in a real-time basis, the performance factors are also expected to be better (although response time degradation may increase with system request load). So it is possible to conclude that this new paradigm will bring higher user satisfaction, although user satisfaction metrics are often difficult to measure. It is also expected that the new paradigm will also will increase the efficiency of the transportation systems, optimizing transport system load balancing, reduction of waste (time and energy) and increase of the competitive advantages of the economic activities benefiting from the new paradigm implementation. Those expected consequences shall be object of further research, as well as more detailed analysis about user satisfaction concerning the new schematic map paradigm.

References

1. Harley, J. B., & Woodward, D. The History of Cartography. University of Chicago Press, Chicago & London, 1987
2. Neyer, G.. Line Simplification with restricted orientations. In *Proceedings of 6th Int. Workshop on Algorithms and Data Structures*, Vancouver, Canada, 1999. pp. 13-24.

3. Barkowsky, T., Latecki, L. J., & Richter, K.-F. Schematizing Maps: Simplification of geographic shape by discrete curve evolution. In *Lecture Notes in Artificial Intelligence*, 2000. pp. 41-53.
4. Avelar, S., & Müller, M. Generating topologically correct schematic maps. In *Proceedings of 9th Int. Symp. on Spatial Data Handling* , Zurich, Switzerland, 2000, 48a25-4a35.
5. Cabello, S., Dijk, M. d., Kreveld, M. v., & Strijk, T. Schematization of road networks. In *Proceedings of 17th Annual Symposium on Computational Geometry* , 2001. pp. 33-39.
6. Cabello, S., & Kreveld, M. v. Aproximation algorithms for aligning points. In *Proceedings of 19th Annual Symposium on Computational Geometry*, San Diego, Usa, 2003. pp14-15.
7. Nollenburg, M. Automated Drawing of Metro Maps. Karlsruhe: Institut fur Theoretische Informatik, Universitat Karlsruhe, 2005.
8. Avelar, S. Schematic Maps on Demand: Design, Modeling and Visualization. Zurich: Swiss Federal Institute of Technology, 2002
9. Porathe, T. User-Centered Map Design. In *Upa 2007 Conference Patterns: Blueprints for Usability.* Austin, USA, 2007.
10. Kano, N., Seraku, N., Takahashi, F., Tsuji, S. (1984), Attractive quality and must-be quality, In *Quality, The Journal of the Japanese Society for Quality Control Vol. 14 No.2*, Japan, 1984. pp.39-48.
11. Cheung, C., & Lee, M. User satisfaction with an internet-based portal: An asymmetric and nonlinear approach. In *Journal of the American Society for Information Science and Technology Vol. 60 No. 1,* 2009, pp. 111-122
12. Sommerville, I. Software Engineering. Addison Wesley, 2005.
13. Buttazzo *et al*, Soft Real-Time Systems: Predictability vs. Efficiency. Springer, 2005
14. Klippel, A., Richter, K.-F., Barkowsky, T.,&Freksa, C. (2005). The Cognitive Reality of Schematic maps. In *L. Meng, Zipf & T. ReichenBacher (Eds.), Map-Based Mobile Services - Theories, Methods and Implementations.* Berlin, (pp. 57-74)
15. Jackson, P.G.(1998): In search for better route guidance instructions. Ergonomics, Vol 41(7), pp. 1000-1013

Robot Dance based on Online Automatic Rhythmic Perception

João Oliveira[1], Fabien Gouyon[2] and Luís Paulo Reis[1,3]

ee03123@fe.up.pt, fgouyon@inescporto.pt, lpreis@fe.up.pt

[1] FEUP – Faculty of Engineering of the University of Porto, Portugal
[2] INESC Porto– Systems and Computers Engineering National Institute, Porto, Portugal
[3] LIACC – Artificial Intelligence and Computer Science Lab., University of Porto, Portugal

Abstract. Music and Entertainment are intimate elements of our daily routine. Dancing is beyond the essence of entertainment which implies the relation between an entertainer and an audience, throughout an exciting and unpredictable musical experience. This paper presents the development of an entertaining robotic system where a humanoid robot, based on the Lego Mindstorms NXT, tries to simulate the human rhythmic perception, from audio signals, and its reactive behaviour in the form of dance. The proposed system was composed by three modules: Music Analysis, Human Control, and Robot Control, which are parallelly processed, through a multithreading architecture, to induce a robotic dance performance in a reactive behavioural-based approach. This methodology incited the synchronous physical embodiment of low-level aspects of rhythm, shaped in three rhythmic events, representing soft, medium and strong onsets in the music. The resultant dance is concatenated through a conjunction of movements that can be dynamically defined a priori, being then performed by the robot in a reactive manner to these rhythmic events' occurrence. These movements also depend on two kinds of sensorial events, namely the colour stepped on the floor or the proximity to some kind of obstacle. The resulting dance alternates in a seemingly autonomous manner between a diversity of motion styles coupled to the musical rhythm, and varying in consonance with the colour stepped on the dance environment, without any previous knowledge of music. The human decision beyond the system behaviour granted the dynamism required to keep an interesting relation between the robot and its audience.

1 Introduction

More and more AI researchers are trying to make robots dance to music. And as the ideas and technologies develop, it's clear that dancing robots can be serious indeed. Recent generations of robots ever more resemble humans in shape and articulatory capacities. This progress has motivated researchers to design interactive dancing robots that can mimic the complexity and style of human choreographic dancing, and that even cooperate with musicians.

Musical robots are increasingly present in multidisciplinary entertainment areas, even interacting with professional musicians. They have even inspired the creation of worldwide robotic dancing contests, as RoboDance (one of RoboCup's competitions) where school teams, formed by children aged eight to nineteen, put their robots in action, performing dance to music in a display that emphasizes creativity of costumes and movement.

These public musical robotic applications lack however in perception, presenting mainly pre-programmed deaf robots with few human-adaptive behaviours. There's where we focused our efforts by designing a flexible framework for robot dancing applications based on automatic music signal analysis.

Music is generically an event-based phenomenon for both performer and listener, formed by a succession of sounds and silence organized in time. We nod our heads or tap our feet to the rhythm of a piece; the performer's attention is focused on each successive note [13]. In dance, body movements emerge as a natural response to music rhythmic events.

To obtain these intended events we focused our analysis on the detection of the music onset times (starting time of each musical note) through an onset detection function (a function whose peaks are intended to coincide with the times of note onsets) representing the energy variation along time, on music data composed by digital polyphonic audio signals.

The use of this rhythmic perception model induced our human-like robot to reactively execute proper dance movements in a time-synchronous way, but individually spontaneous, trying to simulate the dynamic movement behaviour typical from human beings.

The robot's body movement reacts to a conjunction of stimulus formed by three rhythmic events, namely: Low, Medium or Strong Onsets; and two sensorial event groups defined by the detected colour: *Blue, Yellow, Green, Red*; and by the proximity to an obstacle: *OK, Too Close*. Based on the interchange of these inputs a user can, through a proper interface, dynamically define every intended dance movements.

Contrasting to some other approaches, every body movement, as their sequence during the dance, is in this way produced by the robot in a seemingly autonomous way, without former knowledge of the music.

The paper structure is as follows. The next section presents some recent related work on musical robots. Section 3 discusses the system architecture principles presenting an overview of the Lego Mindstorms NXT hardware and explaining the software basis on the music analysis implementation and in the application interface. Section 4 presents an overview of the given experiment and results. Finally section 5 concludes this paper presenting the main conclusions and future work.

2 Related work

Academically, "dancing robots" and "human-robot musical interaction" are common terms. In an increasing number of research labs around the world (especially in Japan),

researchers follow a quest to find the perfect solution to achieve a rhythmic perceptive dancing robot that could interact with humans.

Nakazawa, Nakaoka et al. [1, 2] presented an approach that lets a biped robot, HRP-2 imitate the spatial trajectories of complex motions of a Japanese traditional folk dance by using a motion capture system. To do that they developed the learning-from-observation (LFO) training method that enables a robot to acquire knowledge of what to do and how to do it from observating human demonstrations. Despite the flexibility of motion generation, a problem is that these robots cannot autonomously determine the appropriate timing of dancing movements while interacting with auditory environments, i.e., while listening to music.

Weinberg et al. [3, 4], developed a humanoid robot, Haile, which plays percussion instruments in synchrony with a musician (percussionist). Their robot listens to this percussionist, analyses musical cues in real-time, and uses the result of it to cooperate in a rhythmic and diversified manner. To perform that they used two Max/MSP objects, one to detect the music beats and another to collect pitch and timbre information from it, granting synchronous and sequential rhythmic performance.

Tanaka et al. from Sony, built a dancing robot, QRIO, to interact with children, presenting a posture mirroring dance mode [5, 6]. This interaction was developed using an Entrainment Ensemble Model which relies on the repetition of sympathy, between the robot and the child, and dynamism. To keep the synchronism they used a "Rough but Robust Imitation" visual system through which QRIO mimics the detected human movements.

More recently, in 2007, Aucouturier et al. [7] developed a robot designed by ZMP, called MIURO, in which they built basic dynamics through a special type of chaos (specifically, chaotic itinerancy (CI)) to let the behavior emerge in a seemingly autonomous manner. CI is a relatively common feature in high-dimensional chaotic systems where an orbit wanders through a succession of low-dimensional ordered states (or attractors), but transits from one attractor to the next by entering high-dimensional chaotic motion. The robot motor commands are generated in real time by converting the output from a neural network that processes a pulse sequence corresponding to the beats of the music.

Michalowski et al. [8, 9] investigated the role of rhythm and synchronism in human-robot interactions, considering that rhythmicity is a holistic property of social interaction. To do so they developed perceptive techniques and generated social rhythmic behaviours in non-verbal interactions through dance between Keepon, a small yellow creature-like robot, and children.

Burger and Bresin [10] also used the Lego Mindstorms NXT to design a robot, named M[ε]X, that expresses movements to display emotions embedded in the audio layer, in both live and recorded music performance. Their robot had constraints of sensors and motors, so the emotions (happiness, anger and sadness) were implemented taking into account only the main characteristics of musicians' movements.

Yoshii et al. [11] used Honda's ASIMO to develop a biped humanoid robot that stamps its feet in time with musical beats like humans. They achieved this by building a computational mechanism that duplicates the natural human ability in terms of

associating intelligent and physical functions. The former predicts the beat times in real time for polyphonic musical audio signals. The latter then synchronizes step motions with the beat times by gradually reducing the amount of errors in intervals and timing. Their robot represents a significant step in creating an intelligent robot dancer that can generate rich and appropriate dancing movements that correspond to properties (e.g., genres and moods) of musical pieces, in a human-like behaviour.

In contrast to previous approaches, in this paper we propose a framework in which users have a deterministic role, by dynamically defining the robot dance as the concatenation of selected individual dance movements.

3 System Architecture

This system architecture is composed by a robotic agent (Fig. 2a), built with the Lego Mindstorms NXT kit (Fig. 1), a dance environment which incorporates a multi-colour floor (Fig. 2b), to induce dance variations dependent on the stepped colour, and a covering wall to delimit the dancing space; and three software modules (Fig. 3), namely *Music Analysis*, *Robot Control*, and *Human Control*; each one responsible for the treatment of specific events. All these modules are processed in a multi-threading paradigm to keep the parallelism of behaviours, imposed by robot dancing in synchrony to the musical rhythm.

3.1 Robotic Agent and Lego Mindstorms NXT[1]

Lego Mindstorms NXT is a programmable robotic kit designed by Lego (see Fig. 1). It is composed by a brick-shaped computer, named NXT brick, containing a 32-bits microprocessor, flash and RAM memories, a 4 MHz 8-bit microcontroller and a 100x64 LCD monitor. This brick supports up to four sensorial inputs and can control up to three servo-motors. Lego NXT supports USB 2.0 connection to a PC, and presents a Bluetooth wireless communication system, for remote control and data exchange. It offers many sensor capabilities through its ad-hoc sensors. In the scope of this work we provided our robot with a colour sensor, to detect and distinguish visible colours, and an ultrasonic sensor, capable of obstacle detection, retrieving the robot's distance to it.

Fig. 1. Lego NXT brick and some of its sensors and servo-motors.

[1] For more information consult http://mindstorms.lego.com/eng/default.aspx

Based on this technology we built a humanoid-like robot (see Fig.2a) using two NXT bricks that control six servo motors (one for each leg and each arm, one for a rotating hip and one for the head) and two sensors, already referred. This robot design granted 16 individual dance movements defined as "*BodyPart-Movement* (to the Left-*L*, Right-*R*, or one part to each side-*Alternate*): *Legs-RRotate, Legs-LRotate, Legs-Forward, Legs-Backward, Head-RRotate, Head-LRotate, Body-RRotate, Body-LRotate, RArm-RRotate, RArm-LRotate, LArm-RRotate, LArm-LRotate, 2Arms-RRotate, 2Arms-LRotate, 2Arms-RAlternate, 2Arms-LAlternate*.

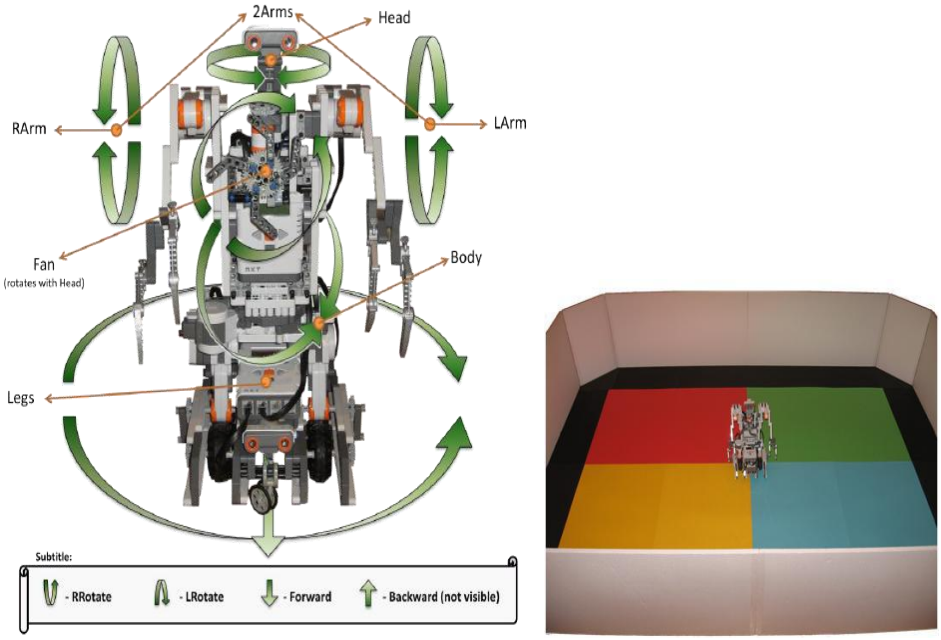


Fig.2. The robot's degrees of freedom (DOFs) to the embodiment of dance movements **(a)**; The dance environment **(b)**.

3.2 Software Modules

This system's modular architecture, and its intrinsic interconnection, was designed to achieve the primary goal of robot dancing, in synchrony to the analyzed rhythm and with flexible human control. As illustrated in Fig. 3 this implementation was composed by three interdependent modules.

The *Music Analysis* module uses a rhythm perception algorithm based on Marsyas' onset detection function, to detect rhythmic events. These events are then sent in real-time, via TCP/IP sockets, to the *Robot Control* module which remotely controls the robot via Bluetooth. In control of the former two, the *Human Control* module is composed by a user graphical interface (GUI) which grants the user interactivity with

the system, by definition of the main control parameters and composition of the resultant dance.

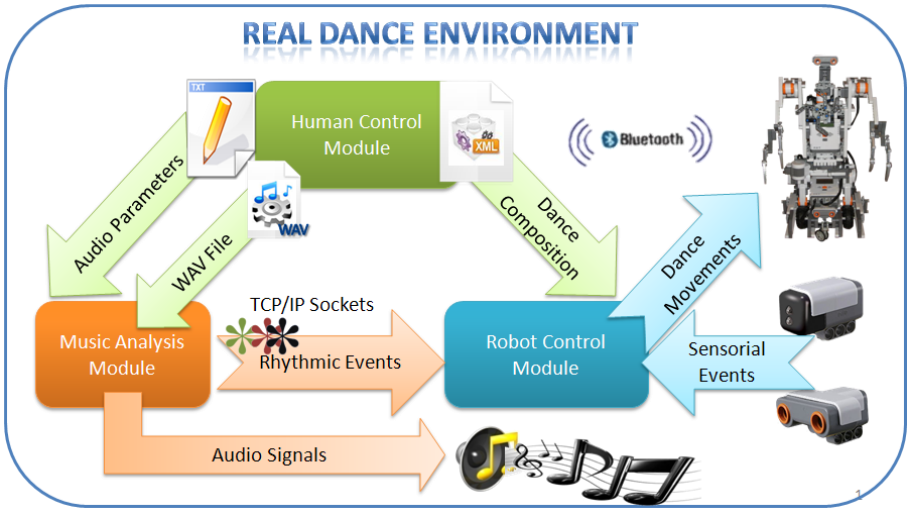


Fig.3. System architecture.

3.2.1 Music Analysis Module and Marsyas[2]

Our *Music Analysis* module was designed in Marsyas. Marsyas – Music Analysis, Retrieval and Synthesis for Audio Signals is an open source C/C++ software framework for computational audio signal analysis; designed and written by George Tzanetakis with help from students and researchers from around the world. It provides fast performance, essential for developing real-time applications, through its functional blocks, which are designed as basis on the state-of-the-art algorithms in Computer Audition.

Under Marsyas we built a MarSystem (an aggregation of functional blocks) that performs onset detection from polyphonic audio signals, in real-time, based on frame energy variations along the music – Spectral Flux detection function. This model was proposed by Dixon [12], which advocates that the Spectral Flux achieves the best results in the simplest and fastest way.

Fig. 4 illustrates this model's composition. First the stereo input audio signal is converted to mono (with *Stereo2Mono*), and then consecutive frames are overlapped (with *ShiftInput*) to grant a more stable analysis. The analysis step is called hop size, which equals the frame size minus the overlap (typically 10 ms). To the Shifted signal is applied the FFT (Fast Fourier Transform) algorithm (with *Spectrum*) using a Hamming window (in *Windowing*) to obtain the music spectrum. To the *Spectrum* output is applied a *PowerSpectrum* function that retrieves the energy variation (magnitude – in dBs) along the music.

[2] For more information consult http://marsyas.sness.net/

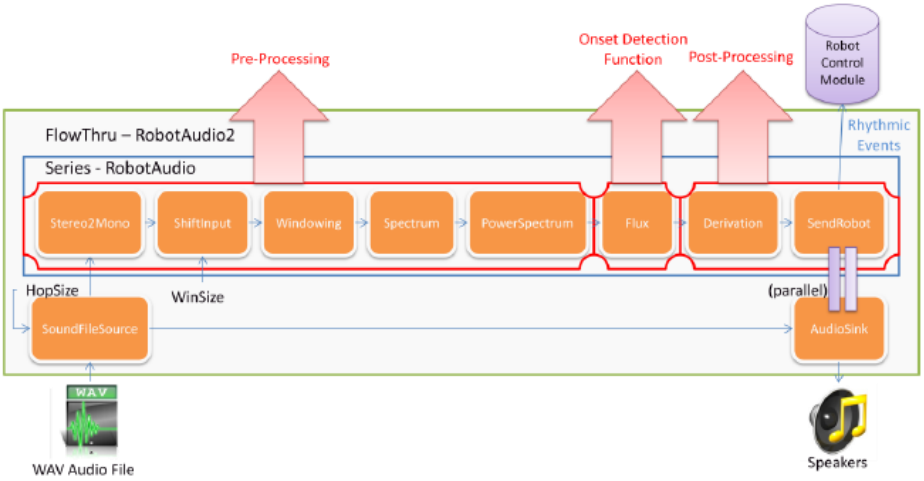


Fig. 4. MarSystem constitution with onset detection function blocks.

Then to this signal is submitted to the Spectral *Flux* function that represents the actual onset detection method. SF measures the change in magnitude in each frequency bin (*k*) of each frame (*n*), restricted to the positive changes and summed across all *k*, with the given Onset Detection Function (OF):

$$OF = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H\left(|X(n,k)| - |X(n-1,k)|\right) \quad , \tag{1}$$

where $H(x) = \frac{x + |x|}{2}$ is the half-wave rectifier function and *X(n,k)* the FFT.

The *Derivation* block retrieves only the crescent *Flux* output, by subtracting the *n* frame to the *n-1* one.

Finally the *SendRobot* block acts as our Peak Picking function and TCP client. It applies a peak adaptive thresholding algorithm to distinguish three rhythm events: *Strong*, *Medium* and *Soft* onsets; which are then sent to the *Robot Control* module via TCP sockets.

3.2.2 Robot Control Module

The Robot Control Module uses a C++ NXT Remote API, designed by Anders Søborg[3], to remotely control the robot, through the transmission of motor commands and reception of sensor data.

[3] For more information consult http://www.norgesgade14.dk/index.php.

3.2.3 Human Control Module

The Human Control Module was decomposed in two blocks (see Fig. 5): *Robot Control Panel* and *Dance Creation Menu*. The *Robot Control Panel* is a user-definable control panel where one can set the Bluetooth connection with one or two NXT bricks, depending on the design; the definition of the audio file to be analyzed and reproduced, and its correspondent parameters (possibly saved in a proper .txt file). The *Dance Creation Menu*, allows the user to dynamically define each individual dance movement in correspondence to a given rhythmic and colour; saving the resultant dance in a proper .xml file (see Fig.5 a) and Fig.5 b)).

The high-level position of this module gives (human) control to the whole process. The user has then a deterministic role in the system behaviour, by dynamically defining the robot choreography through selected individual dance movements, and by defining the polyphonic audio data (WAV file) to be analyzed, and the audio parameters which shall highly influence the resultant rhythmic perception.

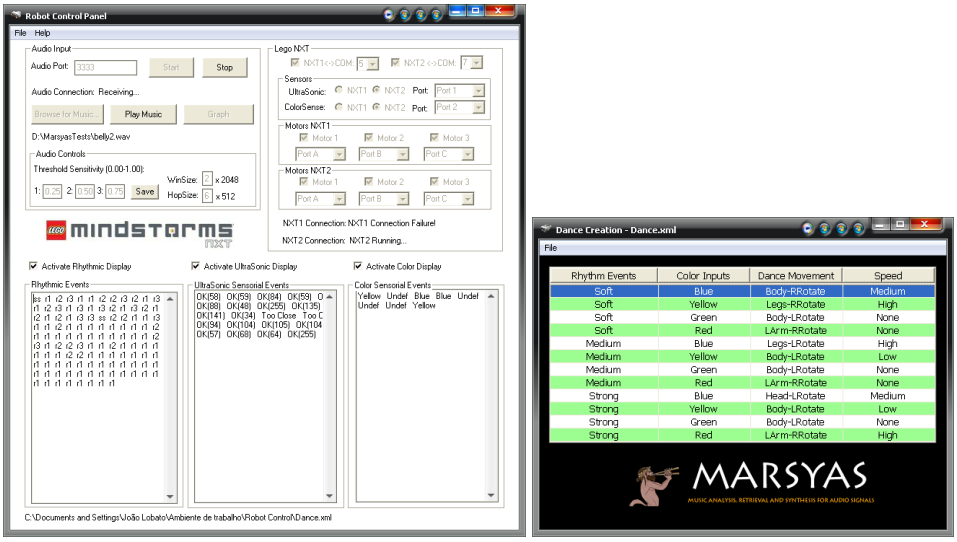


Fig. 5. Application Interface. **a)** *Robot Control Panel.* **b)** *Dance Creation.*

4 Experiments and Results

Our experiments focused on efficiency and synchronism issues related to the music onset detection and to the robot performance with proper and clear dance movements. In order to reduce the sensitivity of our onset function to the main onsets, we started to apply a Butterworth low-pass filter to the *Flux* output, using many different coefficient values. This however incited a group delay that increased with the decrease of the normalized cutoff frequency (*Wn*), promulgating a minimum delay of 10 frames (aprox.

0.7s) which is, in addition to the whole process natural delay, considerably high facing the requirements. In a way to bypass this issue we decided to slightly increase the window and hop size (to *WinSize* = 4096 and *HopSize* = 3072) which granted a lower sensitivity in onset detection focusing on the more relevant ones, with no delay imposed in the process.

In order to restrict and distinguish our three rhythmic events we designed a Peak Picking (PP) function with peak adaptive thresholding (always related to the highest onset detected so far) as follows:

$$PP(x) = \begin{cases} Strong, & if\ x > \delta_3 \\ Medium, & if\ \delta_2 < x < \delta_3 \\ Soft & if\ \delta_1 < x < \delta_2 \end{cases}, \quad \textbf{(2)} \qquad \text{where,} \quad \begin{cases} \delta_1 = thres_1 \times peak \\ \delta_2 = thres_2 \times peak \\ \delta_3 = thres_3 \times peak \end{cases}. \quad \textbf{(3)}$$

$$0 < thres_i > 1$$

The values of $thres_1$, $thres_2$, $thres_3$, as the values of window size and hop size can be dynamically assigned in the application's interface. The function waits 35 frames (equal to 2.43s due to fs_{Flux} = 14.36Hz) to initialize the onset detection, starting with peak = (1/2) * - highest onset detected until then. This acts as the function normalization due to potential inconsistency in the beginning of some music data.

To check the adequate rhythm perception parameters to a large set of music data, we embraced our application interface with a graph mode that uses the parameters inserted by the user to plot the respective output showing the three kinds of rhythm events detected along the music. This representative graph is plotted in MatLab due to the Marsyas' MatLab engine capabilities.

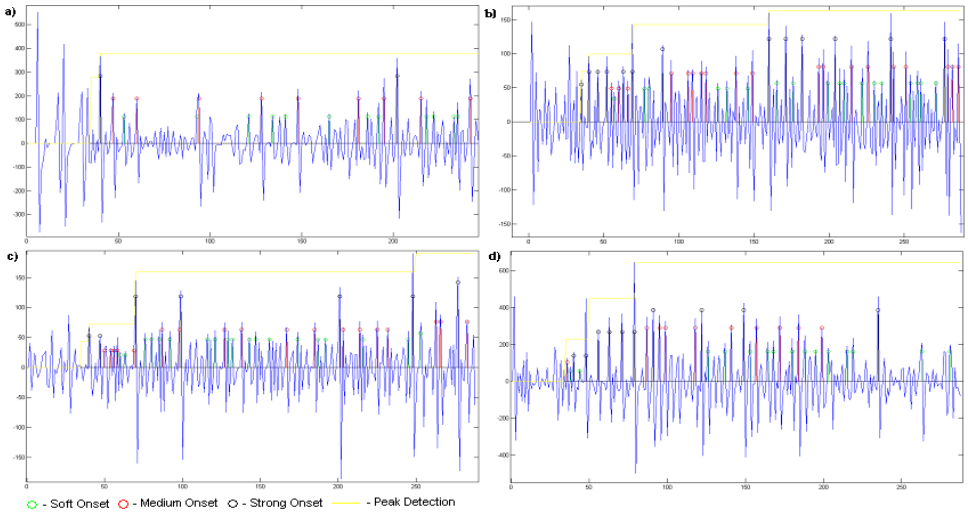


Fig. 7. Peak Picking and Onset Detection output. **a)** PN excerpt using $thres_1 = 0.30$; $thres_2 = 0.50$; $thres_3 = 0.75$. **b)** PP excerpt using $thres_1 = 0.35$; $thres_2 = 0.50$; $thres_3 = 0.75$. **c)** NP excerpt using $thres_1 = 0.30$; $thres_2 = 0.40$; $thres_3 = 0.75$. **d)** CM excerpt using $thres_1 = 0.25$; $thres_2 = 0.45$; $thres_3 = 0.60$.

The set of tests were performed on diverse music styles, consisting of 4 short excerpts (each with around 20s) from a range of instruments, classed into the following groups [13]: NP — non-pitched percussion, such as drums; PP—pitched percussion, such as guitar; PN — pitched non-percussion, in this case some violin; and CM — complex mixtures from popular and jazz music. Below we show some screenshots (Fig. 7) and a table (table 1) with the tests results.

Table 1. Resultant onset counting for the performed tests (above).

Music Style	Soft Onsets	Medium Onsets	Strong Onsets	Total
PN	12	9	2	23
PP	19	18	7	44
NP	15	10	10	35
CM	18	19	13	50

Due to inconsistency among the different music styles, as shown, we were compelled to define different parameters for each music data. To go around this issue we created a text file to each music file containing the respective parameters, from where the application imports them.

The analysis of the resultant robot dancing performance was essentially based on empirical live observation, in comparison to the meaningful data of human behaviour in a real world dance environment. We focused our analysis on synchronism, dynamism, and realism factors:

- **Synchronism:** due to the complexity of the algorithm and the processing and Bluetooth limitative capabilities we‘ve verified some lacks in synchronism. This was essentially caused by the use of multithreading on a single processor. The use of this architecture granted the required simultaneity between the modules processing, but caused some synchrony flaws due to race condition (dependence in a certain sequence of threads processing in order to complete a certain function) issues instigated by the complex (highly time-consumptive) task of dance movement decision and the robot Bluetooth communication overflow (as it can only receive/send data via BT in time-intervals of approximately 50-100ms and it takes around 30ms to transition from receive mode – motors, to transmit mode - sensors).

- **Dynamism:** the dynamism of our work is granted by the enormous variety of possible dance style definitions (in a total of 15^{12}-1), formed by 14 distinct individual dance movements (plus “None”) distributed through 12 (3 rhythmic events x 4 colour events) different event‘s conjunctions (being possible to repeat the same movement in two or more conjunctions); and enforced by the robot‘s perambulation around the dance environment. This dynamic behaviour is, so, transposed to the versatility of human decision, which has the power to adapt the robot performance to its own image through a flexible conduction. Realism: realism can be defined as the fidelity to nature or to real life through representation, in adherence to the actual facts.

- **Realism:** our robotic system keeps the interfacing to the real world through perception and action in a reactive behaviour-based dance performance, which tries to replicate the human behaviour. The robot, through its dance motion, experiences the world (dance environment) directly (*embodiment*), and this world directly influences its behaviour, through sensation – *situatedness* (colour and ultrasonic sensing). The resulting dance alternates in a seemingly autonomous manner between a diversity of motion styles coupled to the musical rhythm, and varying in consonance with the colour stepped on the dance environment.

In conclusion, despite some synchrony issues, referred above, the robot seems to react dynamically in real-time, showing a notable sense of realism. The dynamic and flexible behaviour, in compromise with synchronism, assures an interesting and entertaining relationship between an artificial agent and its human audience.

5 Conclusions and Future Work

We developed a biped humanoid robot that reacts to music in real-time, performing dance movements in synchronism to rhythm in a dynamic and seemingly autonomous way. This was achieved with a proper system architecture constituted by three modules: *Music Analysis, Robot Control*, and *Human Control*. The *Music Analysis* module is composed by a rhythm perception model based on an onset detection function, with peak picking and adaptive thresholding, constructed with Marsyas. The *Robot Control* reacts to the rhythm events sent by the former module, in real-time, and to the received sensorial events, promoting robotic dance movements, as defined in the *Dance Creation* interface from the *Human Control* module.

This way our robot enforces the significant first step in creating an intelligent robot dancer that can generate rich and appropriate dancing movements in correspondence to the rhythm of musical pieces, and supporting human-machine interaction through dynamic dance definitions.

Designing an entertainment system that exhibit such dynamic compromise between short-term synchronization and long-term autonomous behaviour was the key to maintain an interesting relationship between a human and an artificial agent.

In future work, we will apply an automatic music style definition that also addresses the issue of automatic parameter estimation, with the aim of producing a fully automatic onset detection algorithm. We will also add some beat prediction capability applying a beat tracking algorithm to complement the onset detection, and this way design a more efficient and realistic rhythm perception module.

In our robotic system we will also address the issue of multi-robot dance, implementing a swarming system that allows robot-robot interaction while dancing, allowing the creation of synchronous and dynamic choreographies. We will also improve the robots sensitivity by adding other sensorial events, such as acceleration and orientation.

Finally, and as a proof of concept we intend to evaluate this system performance, while an entertainment framework, by promoting the interaction of several subjects, as intervenients and spectators of the robot dancing performance; within different experimental conditions. In consequence we would implement a didactic software for children (and other people) to create their own robotic dances, and even to be used as a framework for creating fully functional systems for RoboDance competitions.

References

1. A. Nakazawa, S. Nakaoka, K. Ikeuchi, K. Yokoi, "Imitating Human Dance Motions through Motion Structure Analysis," IROS, pp. 2539–2544 (2002).
2. S. Nakaoka et al., "Learning from Observation Paradigm: Leg Task Models for Enabling a Biped Humanoid Robot to Imitate Human Dance," Int'l J. Robotics Research, vol. 26, no. 8, pp. 829–844 (2007).
3. G. Weinberg, S. Driscoll, M. Parry, "Musical Interactions with a Perceptual Robotic Percussionist," Proceedings of IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN) Nashville, TN (2005).
4. G. Weinberg, S. Driscoll, "The Perceptual Robotic Percussionist – New Developments in Form, Mechanics, Perception and Interaction Design," Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction (2007).
5. F. Tanaka, H. Suzuki, "Dance Interaction with QRIO: A Case Study for Non-boring Interaction by using an Entertainment Ensemble Model," Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN), pp. 419-424, Kurashiki, Japan (2004).
6. F. Tanaka, B. Fortenberry, K. Aisaka, J. Movellan, "Plans for Developing Real-time Dance Interaction between QRIO and Toddlers in a Classroom Environment," Proceedings of 2005 4th IEEE International Conference on Development and Learning (ICDL), pp. 142-147, Osaka, Japan (2005).
7. J.-J. Aucouturier, Y. Ogai, "Making a Robot Dance to Music Using Chaotic Itinerancy in a Network of FitzHugh-Nagumo Neurons," Proceedings of the 14th International Conference on Neural Information Processing (ICONIP), Kitakyushu, Japan (2007).
8. P. Marek, Michalowski, S. Sabanovic, H. Kozima, "A Dancing Robot for Rhythmic Social Interaction," 16th IEEE International Conference on Robot & Human Interactive Communication, Jeju, Korea (2007 a).
9. P. Marek, Michalowski, H. Kozima, "Methodological Issues in Facilitating Rhythmic Play with Robots," 16th IEEE International Conference on Robot & Human Interactive Communication, Jeju, Korea (2007 b)).
10. B. Burger, R. Bresin, "Displaying Expression in Musical Performance by Means of a Mobile Robot," In Paiva, A., Prada, R., & Picard, R. W. (Eds.), Affective Computing and Intelligent Interaction (pp. 753-754). Berlin / Heidelberg: Springer (2007).
11. K. Yoshii, K. Nakadai, T. Torii, Y. Hasegawa, H. Tsujino, K. Komatani, T. Ogata, H. Okuno, "A Biped Robot that Keeps Steps in Time with Musical Beats while Listening to Music with Its Own Ears," Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2007), 1743-1750, IEEE, RSJ, San Diego (2007).
12. S. Dixon, "Onset Detection Revisited," In Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06), pages 133–137, Montreal, Quebec, Canada, Sept. 18–20 (2006).
13. J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. Sandler, "A Tutorial on Onset Detection in Musical Signals." IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035–1047 (2005).

Biped locomotion methodologies applied to humanoid robotics

Hugo Picado[1,2], Nuno Lau[1,2], Luis Paulo Reis[3,4], Marcos Gestal[5]
hugopicado@ua.pt, nunolau@ua.pt, lpreis@fe.up.pt, mgestal@udc.es

[1] Institute of Electronics and Telematics Enginnering of Aveiro, Portugal
[2] Dep. of Electronics, Telecommunications and Informatics, Univ. Aveiro, Portugal
[3] Artificial Intelligence and Computer Science Lab., Univ. Porto, Portugal
[4] Faculty of Engineering of the University of Porto, Portugal
[5] Artificial Neural Network and Adaptive System Lab., Univ. Coruña, Spain

Abstract. Controlling a biped robot with several degrees of freedom is a challenging task that takes the attention of several researchers in the fields of biology, physics, electronics, computer science and mechanics. For a humanoid robot to perform in complex environments, fast, stable and adaptable behaviors are required. Developing robust behaviors requires the development of methods for joint trajectory planning and low-level control. Several methods are part of the state of the art, including trajectory-based methods, virtual model control, passive-dynamic walking and central pattern generators. This paper proposes a solution for automatic generation of a walking gait using genetic algorithms. The experimental results are shown in terms of the evolution of the ground projection of the center of mass, walking velocity and average oscillation of torso. Genetic algorithms proved to be a powerful method for automatic generation of humanoid behaviors. The robot was able to reach a walk forward velocity of 0.51m/s which is a good result considering the results of the three best teams of RoboCup 3D simulation league for the same movement.

Key words: biped, locomotion, genetic algorithms, humanoid, robotics

1 Introduction

For a long time, wheeled robots were used for research and development in the field of Artificial Intelligence and Robotics and many solutions were proposed [1]. However, wheeled robot locomotion is not adapted to many human environments [2]. This increased the interest in other types of locomotion like biped locomotion and especially in humanoid robotics. This field has been studied over the last years and many different approaches have been presented, although the ability for robots to walk in unknown terrains is still in a young stage. Several approaches to biped locomotion have been developed. This section explains the most common methods which may be broadly divided in four main categories: trajectory-based approaches, virtual model control, passive-dynamic walking and central pattern generators.

1.1 Trajectory-based approaches

Trajectory-based methods consist of finding a set of kinematics trajectories and using a stabilization criterion to ensure that the gait is stable. The most popular stabilization criteria are the Center of Mass (CoM), Center of Pressure (CoP) and Zero Moment Point (ZMP). The gait is stable when one of these criteria remains inside the support polygon (the convex hull formed by the contact points of the feet with the ground).

Center of Mass The CoM of a system of particles is the point at which the mass of the system acts as if it was concentrated [3]. In other words, CoM is defined as the location of the weighted average of the system individual mass particles, as defined by the following equation:

$$p_{CoM} = \frac{\sum_i m_i p_i}{M} \tag{1}$$

where $M = \sum_i m_i$ is the total mass of the system, m_i denotes the mass of the i^{th} particle and p_i denotes its centroid.

Center of Pressure Most humanoid robots are equipped with force-torque-sensors at the feet of the robot. The Center of Pressure (CoP) is the result of an evaluation of those sensors and is defined as the point on the ground where the resultant of the ground reaction forces acts [4]:

$$p_{CoP} = \frac{\sum_i p_i F_{N,i}}{\sum_i F_{N,i}} \tag{2}$$

where the resultant force $F_R = \sum_i F_{N,i}$ is the vector from the origin to the point of action of force $F_{N,i} = |F_{N,i}|$.

Zero Moment Point The Zero Moment Point (ZMP) is perhaps the most popular stability criterion and was originally proposed by Vukobratovic [5] in 1972. It can be defined as the point on the ground about which the sum of the moments of all the active forces equals zero [5]. An alternative, but equivalent, interpretation was given by Arakawa and Fukuda [6] (See Figure 1). They define ZMP as the point p, where $T_x = 0$ and $T_y = 0$, where T_x and T_y represent the moments around the x and y axis generated by the reaction force R and reaction torque M, respectively.

Static vs. dynamic stability The static stability criterion prevents the robot from falling down by keeping the CoM inside the support polygon by adjusting the body posture very slowly thus minimizing the dynamic effects [7] and allowing the robot to pause at any moment of the gait without falling down. Using this criterion will generally lead to more power consumption since the robot has to adjust its posture so that the CoM is always inside the support polygon. On

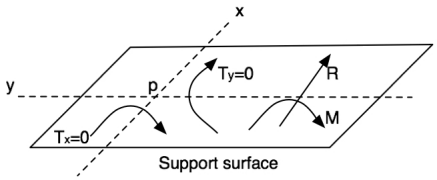


Fig. 1: Arakawa and Fukuda interpretation of ZMP concept [6].

the other hand, humans move in a dynamic fashion, a state of constant falling [7], where the CoM is not always inside the support polygon. While walking, humans fall forward and catch themselves using the swinging foot while continuing to walk forward, which makes the CoM moves forward without expending energy to adjust the CoM trajectory. Dynamic stability relies on keeping the ZMP or CoP inside the support polygon and this is a necessary and sufficient condition to achieve stability. Dynamic balance is particularly relevant during the single support phase, which means that the robot is standing in only one foot. This generally leads to more fast and reliable walking gaits.

1.2 Virtual model control

The most important drawback of ZMP is the use of complex dynamic equations to compute the robot's dynamics. This complexity can be crucial when designing humanoid robots, specially when the programmer wants to minimize the power and memory consumption of the biped. Developed by Jerry Pratt [8], Virtual Model Control (VMC) is a framework based on heuristics that uses virtual components such as springs, dampers or masses to generate the joint torques that control the biped's stability and velocity. The generated joint torques create the same effect that the virtual components would create if they were in fact connected to the real robot. This heuristic makes the design of the controller much easier. First it is necessary to place some virtual components to maintain an upright posture and ensure stability. Using the example provided by Pratt [8], imagine that the goal of the robot is to knock a door. With VMC it is only needed to place a virtual mass with a specified kinetics energy to the robot's hand using a virtual spring and damper. The robot's hand will then move to strike out and once given the desired impact, the hand will get back due to mass resonating with the virtual component attached to the hand.

1.3 Passive-dynamic walking

In the Passive-Dynamic Walking (PDW) approach, the biped walks down a slope without using any actuator [9]. The mechanical charecteristics of the legs (e.g. length, mass, foot shape) determine the stability of the generated walking motion.

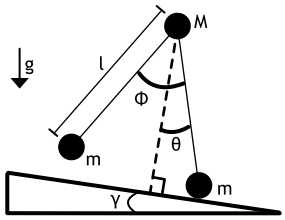


Fig. 2: Passive dynamic walking model. ϕ is the supporting surface, θ is the angle of the support leg, m is the point mass at the respective foot, M is the mass at the hip, γ is slope inclination and g is the gravity. Both legs have length l. Adapted from: [10].

PDW is based on the inverted pendulum model [9], which assumes that, in the single support phase, human walking can be modeled as an inverted pendulum. Inverted pendulum has been applied for years in several situations [11]. The swinging leg (assuming there is just the hip and the ankles and no knees) is represented by a regular pendulum, while the support leg is represented by an inverted pendulum. The support leg is then controlled by the hip joint's torque. However, in the specific case of PDW the only actuating force is the gravity. Figure 2 represents the PDW model. Tad McGeer, in 1990, was the first to apply this idea to humanoid robotics by developing a 2D bipedal robot with knee joints and curved feet [9]. The developed robot was able to walk down a three degree slope. This work demonstrated that the morphology of the robot might be more important than the control system itself. This method is known for the low power consumption.

1.4 Central pattern generators

It is assumed, by the fields of biology, that vertebrate locomotion is controlled by a spinal central pattern generator (CPG) [12]. A CPG is a set of circuits which aims to produce rhythmic trajectories without the need for any rhythmic input. In legged animals, the CPG often contains several centers that control different limbs. There has been a growing interest in these CPG models in robotics. This trajectory planning method does not need, necessarily, any sensory feedback information to generate oscillatory output for the motor neurons [12]. However, it is possible to integrate the sensory feedback information such as force resistors and gyroscopes to produce motion correction and compensation [13].

By coupling the neural oscillators signals when they are stimulated by some input, they are able to synchronize their frequencies. In the field of artificial intelligence and robotics, it is possible to build structures that are similar to the neural oscillators found in animals by the definition of a mathematical model. However, most of the times these CPGs are designed for a specific application and there are very few methodologies to modulate the shape of the rhythmic signals, in particular for online trajectory generation [14], which lead the researchers to use other methods as the ones presented previously. Sven Behnke [15] proved

that it is possible to apply CPGs to generate a omnidirectional walking gait, where the input is simply the forward, lateral and rotational walking speed.

The solution presented in this paper for joint trajectory generation has the smooth properties of the central pattern generators since it is based on oscillators but on the other hand can be more easily defined.

2 Genetic algorithms

Proposed by the mathematician John Holland in 1975 [16], A Genetic Algorithm (GA) is an optimization method inspired by the evolution of biological systems and based on global search heuristics. GA belongs to the class of evolutionary algorithms. In spite of being different, evolutionary algorithms share common properties since they are all based on the biological process of evolution. Given an initial population of individuals (also called chromosomes), the environmental pressure causes the best fitted individuals to survive and reproduce more. Each individual (chromosome) is a set of parameters (genes) and represents a possible solution to the optimization problem. The algorithm starts by creating a new population of individuals. Typically, this population is created randomly but any other creation function should be acceptable. The genes of each individual should be inside a range of acceptable values that is defined for each gene. The algorithm then starts the evolution which consists of creating a sequence of new populations. At each step, the algorithm uses the individuals in the current population to create the next population by applying several operators. These genetic operators are described as follows [17]:

- **Selection:** Chooses some parents for crossover according to predefined rules (cost function or fitness);
- **Elitism:** Defines the number of chromosomes in the current generation that are guaranteed to survive in the next generation;
- **Crossover:** Generates offspring by exchanging some genes of different parent chromosomes;
- **Mutation:** Generates a mutant gene by changing one or several genes of an individual.

In the end of the optimization process, the individual in the current population that have the best fitness value is chosen as the best individual. Algorithm 1 shows the pseudo code of a generic GA.

There are many alternatives to the use of GAs for gait optimization. Some of these alteratives might be Hill Climbing (HC) [18], Simulated Annealing (SA) [19], Tabu Search ([20], or even machine learning methods such as Reinforcement Learning (RL) [21]. The option for GA in the present work is due to the good results obtained in previous experiments in the same research field [22].

Algorithm 1 Generic Genetic Algorithm

Population ← *CreateInitialPopulation*()
Evaluate(*Population*)
while TerminationConditionNotMet() **do**
  [Selection] *Parents* ← Selection(Population)
  [Elistism] *Elite* ← Elitism(Population)
  [Crossover] *Children* ← Crossover(*Parents*, p_c)
  [Mutation] *Mutants* ← Mutation(*Children*, p_m)
  Population ← *Elite* + *Mutants*
  Evaluate(*Population*)
end while
return Best(*Population*)

3 Walking gait definition

A walking gait was developed and the tests were performed with the simulated humanoid NAO in the scope of the RoboCup 3D Soccer competition using the Simspark Simulation Environment [23]. Figure 3 shows the humanoid structure and the referential axis considered.

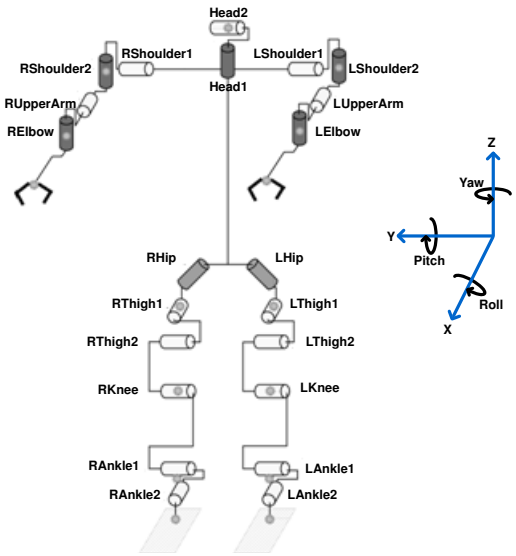


Fig. 3: Humanoid structure and global referential. The arrows around the axes represent the positive direction of the pitch, roll and yaw rotations. Adapted from [24].

3.1 Joint trajectory planning

The method used for joint trajectory planning is based on Partial Fourier Series (PFS). Some human-like movements are inherently periodic and repeat the same set of steps several times (e.g. walk, turn, etc). The principle of PFS consists of the decomposition of a periodic function into a sum of simple oscillators as represented by the following expression:

$$f(t) = C + \sum_{n=1}^{N} A_n \sin\left(n\frac{2\pi}{T}t + \phi_n\right), \forall t \in \Re_0^+ \tag{3}$$

where N is the number of frequencies, C is the offset, $A_{n=1..N}$ are amplitudes, T is the period and $\phi_{n=1..N}$ are phases.

The main idea behind the definition of the walking gait is to place an oscillator on each joint we pretend to move in order to define its trajectory. The oscillators are placed on the following joints: LShoulder1, RShoulder1, LThigh1, RThigh1, LThigh2, RThigh2, LKnee, RKnee, LAnkle1, RAnkle1, LAnkle2 and RAnkle2. Hence, 12 single-frequency oscillators are used. Since each single-frequency oscillator will have 4 parameters to define, 48 parameters are needed to completely define the gait. It is common to assume a walk sagittal symmetry, which determines the same movements for corresponding left and right sided joints with a half-period phase shift. Hence, it is possible to reduce the number of parameters by half of the original size, resulting on 24 parameters. Additionally, the period of all oscillators should be the same to keep all the joints synchronized by a single frequency clock. This consideration reduces the number of parameters to 19. A set of equations can be obtained for the left-sided joints:

$$f_{LShoulder1}(t) = C_1 + A_1 \sin\left(2\pi t/T + \phi_1\right) \tag{4}$$
$$f_{LThigh1}(t) = C_2 + A_2 \sin\left(2\pi t/T + \phi_2\right) \tag{5}$$
$$f_{LThigh2}(t) = C_3 + A_3 \sin\left(2\pi t/T + \phi_3\right) \tag{6}$$
$$f_{LKnee}(t) = C_4 + A_4 \sin\left(2\pi t/T + \phi_4\right) \tag{7}$$
$$f_{LAnkle1}(t) = C_5 + A_5 \sin\left(2\pi t/T + \phi_5\right) \tag{8}$$
$$f_{LAnkle2}(t) = C_6 + A_6 \sin\left(2\pi t/T + \phi_6\right) \tag{9}$$

where $f_X(t)$ is the trajectory equation for the joint X, $A_{i=1..6}$ are amplitudes, T is the period, $\phi_{i=1..6}$ are phases and $C_{i=1..6}$ are offsets. The right-sided joints can be obtained with no additional parameters: For roll joints the left and the right side perform the same trajectories over the time. For pitch joints, the right side can be obtained by adding a phase, π, on the corresponding oscillator. The unknown parameters together form the genome that will be used by the genetic algorithm to generate the gait.

3.2 Automatic generation of parameters

The parameters described in the previous section were defined by a GA. The algorithm was configured using an initial population of 100 chromosomes initialized randomly. The roulette method used for selection consists of simulating a roulette-wheel where the parents are selected with a probability that is proportional to their fitness. The mutation follows an uniform distribution with a probability defined by $p_m = 0.5$. For crossover the scattered method was used. Scattered creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent. The fraction of the population that is created by crossover is defined by the parameter $p_c = 0.8$. For the elitism, 10 chromosomes are selected to survive for the next generation.

The fitness function has to be chosen carefully in order to achieve good results. In the case of the forward walking, a simple but effective fitness function to minimize can be the distance to some point in the forward direction (let's call it target), assuming that the robot is initially placed far enough from it. Additionally, the torso average oscillation [25] is also used in order to obtain more stable gaits. The final version of the fitness function is stated as follows:

$$fitness = d_{target} + \overline{\theta} \tag{10}$$

where d_{target} is the distance to the target point (in meters) and $\overline{\theta}$ is the average oscillation of the torso (in radians per second). The generation process took five entire days to complete using a Core 2 Duo 2.4Gz CPU with 1GB of physical memory. Figure 4 shows the evolution of the fitness during the optimization process.

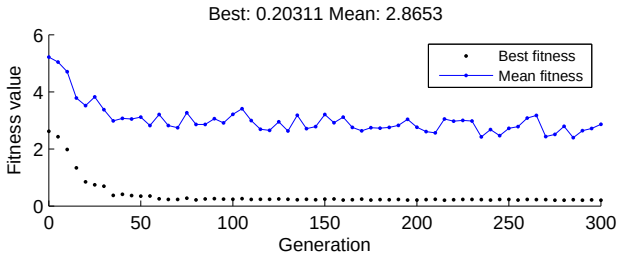


Fig. 4: Evolution of the fitness.

The fitness decreases fast and stabilizes in a hundred of generations. The minimum fitness is 0.20311, which is a very good result. Table 1 shows the values of the best individual for $A_{1..6}$, $\phi_{1..6}$ and $C_{1..6}$. For the period, T, the optimal generated value was 0.3711.

A_1	57.1842	ϕ_1	2.9594	C_1	-88.4624
A_2	5.6445	ϕ_2	-2.2855	C_2	3.6390
A_3	57.1211	ϕ_3	0.0887	C_3	35.9536
A_4	39.6205	ϕ_4	-1.8292	C_4	-39.9481
A_5	46.6315	ϕ_5	1.7640	C_5	28.5095
A_6	3.7947	ϕ_6	-1.2067	C_6	-2.9360

Table 1: Best generated individual

Figure 5 shows the evolution of the CoM and the placement of feet in the XY plane. It is possible to note that the robot tends to shift the CoM to the support foot while walking. Another characteristic shown by the same graphic is the large size of the steps.

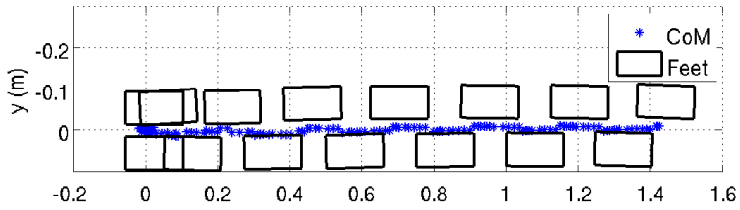


Fig. 5: The CoM and the feet in the XY plane

The average velocity (Figure 6a) shows very good results. More than 50 centimeters per second were achieved. This is a good velocity taking into account the torso average oscillation, that is represented in the Figure 6b.

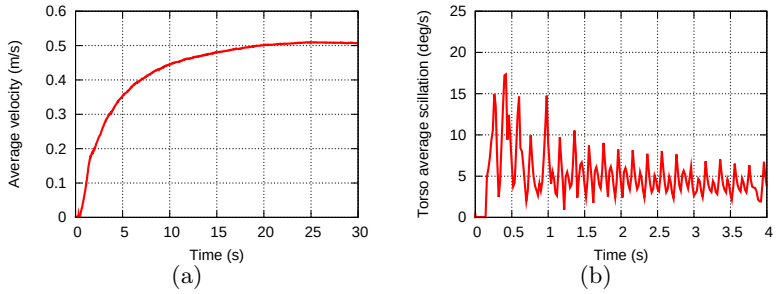


Fig. 6: (a) Average velocity over time (b) Torso average oscillation over time

The obtained results can be considered good results, comparing with the three best teams of the RoboCup 3D simulation league competition of 2008 (Suzhou, China). They were able to reach forward velocities of 1.20m/s (SEU-RedSun [26]), 0.67m/s (WrightEagle [27]) and 0.43m/s (LittleGreenBats [28])[6]. Figure 7 shows NAO walking forward using the proposed solution. At t=1.44s the biped already covered a great distance. It is also possible to note the large steps and the height of the steps.

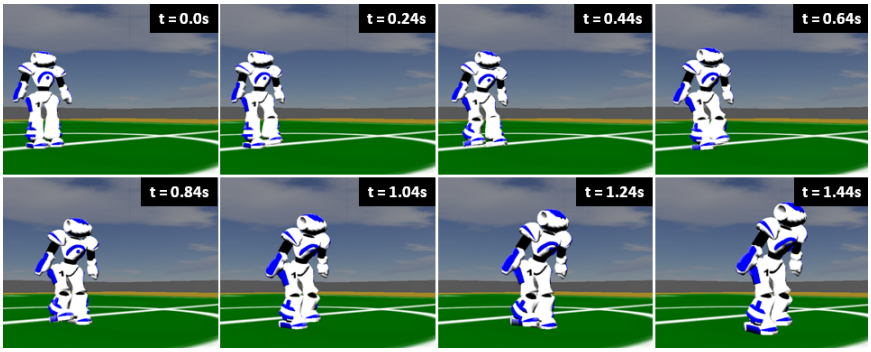


Fig. 7: Walking gait screenshots

4 Conclusion and future work

This paper presented the common approaches to biped locomotion. These approaches were broadly divided in trajectory-based approaches which consist of keeping some stablization criteria inside the support polygon of the humanoid, the virtual model control which is based on heuristics aiming at reducing the computational effort, the passive-dynamic walking where a robot walks down a slope using only the gravitational force and, finally, the central pattern generators that pretend to imitate the locomotion mechanism that is assumed to exist in vertebrates. It was presented a solution for automatic generation of a forward walking gait using a genetic algorithm. GAs are based on the biological evolution of species by applying selection, mutation and crossover operators to a set of chromosomes that are formed by genes represeting the parameters to be optimized. GA proved to be very good on achieving results. The generated walking gait is fast, stable and resistant to environment disturbances. These results were sucessfully compared to the results of the three top teams of 3D simulation league of RoboCup in 2008.

[6] This results were retrieved from the logfiles of the RoboCup 2008 competition, which may be found in http://www.robocup-cn.org/.

Several improvements are possible and needed. The generated walk was mainly based on the trajectory of CoM to monitor the quality of the gait. However, the calculation and monitoring of the ZMP trajectory is essential for achieving dynamic stability. Additionally, the use of inverse kinematics to compute the trajectories of end effectors instead of controlling the joints directly is very useful to have more flexibility among the generation of behaviors. Moreover, motion capturing, which consists of monitoring the human behaviors, provides a great way to define the humanoid behaviors, due to the anthropomorphic characteristics between the both. As future work, it is predicted to invest not only in genetic algorithms, but in machine learning methods such as reinforcement learning. These methods provide great advantages for the automatic generation of behaviors which reduce, and possibly eliminate, the human intervention during the optimization or learning process.

Acknowledgements

This research was partially supported by FCT-PTDC/EIA/70695/2006 Project – "ACORD - Adaptive Coordination of Robotic Teams".

References

1. Robin Murphy. *Introduction to AI Robotics*. MIT Press, 2000.
2. Maria Prado, Antonio Simón, Ana Pérez, and Franscisco Ezquerro. Effects of terrain irregularities on wheeled mobile robot. *Robotica*, 21:143–152, 2003.
3. Raymond Serway. *Physics for Scientists and Engineers with Modern Physics*. Brooks Cole Publishing Company, 6^{th} edition, 2003.
4. Ambarish Goswami. Postural stability of biped robots and the foot-rotation indicator (fri) point. *The International Journal of Robotics Research*, 18(6):523–533, 1999.
5. Miomir Vukobratovic and Yury Stepanenko. On the stability of anthropomorphic systems. *Mathematical Biosciences*, 15 i1:1–37, 1972.
6. Takemasa Arakawa and Toshio Fukuda. Natural motion generation of biped locomotion robot using hierarchical trajectory generation method consisting of GA, EP layers. In *Proceedings of 1997 IEEE International Conference on Robotics and Automation, ICRA'97*, volume 1, pages 211–216, 1997.
7. Karl Muecke, Patrick Cox, and Dennis Hong. *DARwIn Part 1: Concept and General Overview*, pages 40–43. SERVO Magazine, 12 2006.
8. Jerry Pratt, Chee-Meng Chew, Ann Torres, Peter Dilworth, and Gill Pratt. Virtual model control: An intuitive approach for bipedal locomotion. *The International Journal of Robotics Research*, 20:129–143, 2001.
9. Tad McGeer. Passive dynamic walking. *The International Journal of Robotics Research*, 9(2):62–82, 1990.
10. Max Kurz and Nicholas Stergiou. Do horizontal propulsive forces influence the nonlinear structure of locomotion? *Journal of NeuroEngineering and Rehabilitation*, 4:30+, 2007.
11. José Lima, José Gonçalves, Paulo Costa, and António Moreira. Inverted pendulum virtual control laboratory. In *Proceedings of the 7th Portuguese Conference on Automatic Control*, pages 11–13, 2006.

12. Eric Kandel, James Schwartz, and Thomas Jessell. *Principles of Neural Science*. McGraw-Hill Medical, 4^{th} edition, 2000.
13. Felix Faber and Sven Behnke. Stochastic optimization of bipedal walking using gyro feedback and phase resetting. In *Proceedings of 7th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Pittsburg, USA*, 2007.
14. Huashan Feng and Runxiao Wang. Construction of central pattern generator for quadruped locomotion control. In *Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Xi'an, China*, 2008.
15. Sven Behnke. Online trajectory generation for omnidirectional biped walking. *In Proceedings of 2006 IEEE International Conference on Robotics and Automation (ICRA' 06)*, pages 1597–1603, 2006.
16. John Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975.
17. Melanie Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, 1998, 1998.
18. Shahid Bokhari. On the mapping problem. *IEEE Trans. Computers*, 30(3):207–214, 1981.
19. Scott Kirkpatrick, Daniel Gelatt Jr., and Mario Vecchi. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.
20. Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, 13(5):533–549, 1986.
21. Amit Konar. *Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain*. CRC Press, Inc., Boca Raton, FL, USA, 2000.
22. Hugo Picado. Development of behaviors for a simulated humanoid robot. Master's thesis, University of Aveiro, 2008.
23. Oliver Obst and Markus Rollmann. Spark - a generic simulator for physical multiagent simulations. In Gabriela Lindemann, Jörg Denzinger, Ingo J. Timm, and Rainer Unland, editors, *MATES*, volume 3187 of *Lecture Notes in Computer Science*, pages 243–257. Springer, 2004.
24. David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jerome Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. The NAO humanoid: a combination of performance and affordability. *CoRR*, abs/0807.3223, 2008.
25. Milton Heinen and Fernando Osório. Applying genetic algorithms to control gait of physically based simulated robots. In *Proceedings of 2006 IEEE Congress on Evolutionary Computation*, pages 500–505, 2006.
26. Xu Yuan, Shen Hui, Qian Cheng, Chen Si, and Tan Yingzi. SEU-RedSun 2008 soccer simulation team description. In *Proceedings CD of RoboCup 2008*, 2008.
27. Xue Feng, Tai Yunfang, Xie Jiongkun, Zhou Weimin, Ji Dinghuang, and Xiaoping Chen Zhang Zhiqiang. Wright Eagle 2008 3D team description paper. *Proceedings CD of RoboCup 2008*, 2008.
28. Sander van Dijk, Martin Klomp, Herman Kloosterman, Bram Neijt, Matthijs Platje, Mart van de Sanden, and Erwin Scholtens. Little Green Bats humanoid 3D simulation team description paper. In *Proceedings CD of RoboCup 2008*, 2008.

Multi-Agent Coordination through Strategy

João Certo[2,3], Nuno Lau[1,4], Luís Paulo Reis[2,3]

joao.certo@fe.up.pt, lau@det.ua.pt, lpreis@fe.up.pt
[1] IEETA – Institute of Electronics and Telematics Engineering of Aveiro
[2] LIACC-NIAD&R– Artificial Intelligence and Computer Science Lab, University of Porto,
[3] FEUP – Faculty of Engineering of the University of Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[4]DETI –Informatics, Electronics and Telecommunications Dep., University of Aveiro,
Campus de Santiago, 3810-193 Aveiro, Portugal

Abstract. Coordinating heterogeneous multi-agent systems can be a difficult task. This paper presents an approach to this problem based on a developed meta-model for a multi-purpose, adaptable, strategic coordination layer. Based on previous work developed for the RoboCup Soccer simulation, small-size, middle-size and legged leagues, a generic coordination model was built that allows the management of heterogeneous teams, for both centralized and decentralized environments, with reduced use of communication. The proposed meta-model is hierarchical using concepts such as strategy, tactics, formations, sub-tactics and roles, from high to low level. Hybrid methods are used to switch formations and tactics. In order to test the model, two strategy instances, for RoboCup Rescue Simulation and RoboCup Soccer, were developed. Strategies are designed with the help of a graphical tool. Results achieved by the team in RoboCup Rescue and Soccer Simulation competitions demonstrate the usefulness of this approach.

Keywords: Multi-Agent Systems; Artificial Intelligence; Heterogeneous Agent Simulation Systems; Agent Coordination; RoboCup.

1. Introduction

RoboCup was created as an international research and education initiative, aiming to foster Artificial Intelligence (AI) and Robotics research, by providing standard problems. RoboCup has two main league types: simulation and robotics. Simulation leagues enable research on AI and multi-agent coordination while waiting for the availability of hardware to enable the same type of research.

In RoboCup Soccer leagues two opposing teams play a soccer match, thus creating a dynamic environment. A soccer match provides important scientific challenges, both at an individual level (perception, moving, dribbling, shooting) and at a collective level (strategy, collective play, formations, passing, etc.).

Proposed by Kitano [2], RoboCup Rescue simulated environment consists of a virtual city, immediately after a big catastrophe, in which heterogeneous, intelligent agents, acting in a dynamic environment, coordinate efforts to save people and

property. The agents are of six different types: Fire Brigades, Police Forces, Ambulance Teams and the three respective center agents. Fire Brigades are responsible for extinguishing fires, Police Forces open up blocked routes and Ambulance Teams unbury Civilians. In order to obtain a good score, all these agents work together communicating through supervising center agents.

FC Portugal's research focus is on the development of new coordination methodologies. After successfully developing such methodologies for soccer simulation leagues[1], the team is working on adapting these methodologies to the Rescue Simulation League.

Although a partial adaptation in rescue led to limited success[2], due to rescue simulation specificity, some of the methodologies could not be implemented and other methodologies developed specifically for the rescue team did not fit into the existing soccer strategic level. Members of the team are also involved in different robotic soccer teams (simulation 2D, simulation 3D, small-size, middle-size and legged) that, in order to collaborate between themselves, need a common strategic layer. Furthermore, a strategy developed for one soccer league, has many similarities with strategies in other soccer leagues. Also in some of the leagues our participation includes collaboration with other universities and thus the need of a common strategy enabling cooperation from robots developed by different universities.

One of the expectations of RoboCup is to stimulate technology development in the hope that it can be applied to other areas. The model and tools developed aim to simplify the portability of research between RoboCup leagues and expand its usefulness to areas outside this domain.

This paper describes the specification and application of a multi-purpose, multi-domain, adaptable, strategical layer on multi-agent systems. This layer allows the management of homogeneous and heterogeneous agents, the centralized or decentralized management of the strategy. The paper also presents a graphical strategy building tool, compliant with the defined model.

The rest of this paper is organized as follows. The next section presents related work. In section 3 the strategic layer is described. Section 4 presents an implementation on the rescue team. Section 5 presents the graphical tool, showing a strategy for the soccer team. Section 6 concludes this paper and points out to future work.

2. Related Work

Stone et al. [7, 8] previously defined periodic team synchronization (PST) domains as domains with the following characteristics: "There is a team of autonomous agents that collaborate towards the achievement of a joint long-term goal". Then they decomposed the task at hand, into multiple rigid roles, assigning one agent to each role. Thus each component of the task was accomplished and there were no conflicts

[1] FC Portugal won several World and European championships in different RoboCup soccer leagues in the past seven years.

[2] FC Portugal rescue simulation team achieved very good results in RoboCup, including winning a rescue European champion using these coordination methodologies.

among agents in terms of how they should accomplish the team goal. As it was defined, a role consisted of a specification of an agent's internal and external behaviors. The conditions and arguments of any behavior could depend on the agent's current role, which was a function of its internal state.

Due to inflexibility to short-term changes (e.g. one robot is non-operational), inflexibility to long-term changes (e.g. a route is blocked), and a lack of facility for reassigning roles, a formation was introduced as a teamwork structure within the team member agent architecture. A formation decomposes the task space defining a set of roles with associated behaviors. In a general scenario with heterogeneous agents, subsets of homogeneous agents could flexibly switch roles within formations, and agents could change formations dynamically. Formations included as many roles as there were agents in the team, so that each role is filled by one agent.

Much of FC Portugal's related research was done for soccer simulation leagues. The rest of this section explains the concepts and mechanisms developed for those leagues. The work here presented either serves as a basis for the construction of the strategical layer or is directly usable in conjunction with this layer for the specific case of RoboCup Soccer.

FC Portugal's team strategy definition extends the concepts introduced by Stone [7, 8] and is based on a set of player types (roles) and a set of tactics that include several formations for different game situations (defense, attack, etc) [5]. Formations assign each player a positioning (that determines the strategic behavior) and each positioning a player type (that determines the active behavior).

When Stone defined a situation, the concept was bound to set-plays. A situation was a set of world state conditions that triggered a series of predefined behaviors within the roles. FC Portugal's members have expanded on this concept and defined situations [4] as a group of easily identifiable logic conditions set for high-level, world state, parameters. These situations were defined so that they would not suffer a considerable, temporal, variation. The situations were then associated with formations, however not every situation had to have its own formation using, in this case, a set of replacement situations.

Situation Based Strategic Positioning(SBSP) mechanism [3, 4] is used for strategic situations (in which the agent believes that it is not going to enter in active behavior soon). For active situations, the agent position on the field is calculated using ball possession and recovery or playoff decision mechanisms. To calculate its strategic positioning, the agent analyzes which is the game situation. Then the agent calculates its base strategic position in the field in that formation, adjusting it according to the ball position and velocity, situation and player type strategic information. This behavior enables the team to move similarly to a real soccer team, covering the ball while the team remains distributed along the field.

The DPRE , Dynamic Positioning and Role Exchange (and Dynamic Covering) [5], was based on previous work from Peter Stone et al. [7, 8] which suggested the use of flexible agent roles with protocols for switching among them. The concept was extended and players may exchange their positionings and player types in the current formation if the utility of that exchange is positive for the team. Positioning exchange utilities are calculated using the distances from the player's present positions to their strategic positions and the importance of their positionings in the formation on that situation.

In the case of communication in single channel, low bandwidth, and unreliable domains the challenge is deciding what and when to communicate. In ADVCOM [5] (Intelligent Communication Mechanism), agents use communication in order to maintain world states updated by sharing individual world states, and to increase team coordination by communicating useful events (e.g. a positioning swap). The main innovation of this communication strategy is that agents communicate when they believe that the utility of their communication is higher than those of their teammates, using mutual modeling to estimate these utilities.

3. Model for the Strategic Layer

The model here depicted provides a structured method of representing, building and managing a strategy in a scenario where a team of agents is used. The terms *scenario* and *agent* should be considered as broader terms. *Scenario* can be a simulation, a game, or any other kind of set where there is an environment, with *agents* who have one or more objectives. Likewise *agents*, besides being software computational entities, can be any kind of independent units like robots or even persons.

This model handles static, dynamic, reactive or nonreactive environments and is designed to manage team strategy and cooperation. A team is an aggregation of *agents* with common goals. When *agents* in a team work together cooperatively they do *teamwork* [1, 9]. In this model, homogeneous and heterogeneous *agents* can be used. In heterogeneous environments the term *agent type* is used for differentiation.

3.1. Structure

In order to better explain the model, a top-down approach will be followed. Figure 1 represents the proposed model and depicts the interconnections between the concepts presented in this model. The figure only expands one branch for each concept.

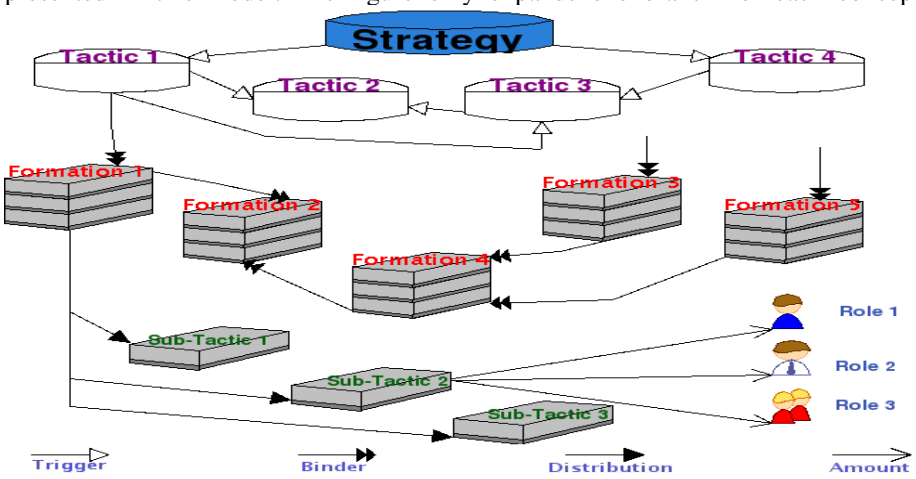


Fig. 1. Schematic of strategic concepts.

3.1.1 Strategy

Informally, a strategy is the combining and employment of means in large-scale, long-range planning and the act of directing operations for obtaining a specific goal or result.

Formally, a *strategy* is a combination of *tactics* used to face the scenario and the *triggers* to change between tactics:

$$\text{Strategy} = \{\text{Tactics, Triggers}\} \quad (\mathbf{1})$$

where a *strategy* can have several available *tactics*:

$$\text{Tactics} = \{\text{Tactic 1, Tactic 2, ... , Tactic t }\} , \forall\, t \in \infty \quad (\mathbf{2})$$

and triggers that set the conditions to change between tactics:

$$\text{Triggers} = \{\text{Trigger 1, Trigger 2, ... , Trigger tg }\} , \forall\, tg \in \infty \quad (\mathbf{3})$$

3.1.2. Tactic

A tactic is an approach to face the scenario in order to achieve a goal. Tactics deal with the identification of different situations and the correspondent use and deployment of agents in the scenario for those situations.

Formally, a *tactic* defines *agents*' *formations* as the arrangement of *agents, situations* as the combination of scenario conditions that can be seen as a more particular problems and *binder* as the association between a *formation* and a *situation* or between several *situations* and a *formation*. *Tactics* can optionally also set *tactical parameters,* the default thresholds on which *agents* base their decisions.

A *tactic* should be self-sufficient, i.e., it does not need other tactics to function through all the simulation. There can be only one *tactic* active at one given time.

$$\text{Tactic} = \{\text{Formations, Situations, Binders, [Tactical Parameters]}\} \quad (\mathbf{4})$$

where a *tactic* has several *formations* that can be used:

$$\text{Formations} = \{\text{Formation1, Formation2, ... , Formation f }\} , \forall\, f \in \infty \quad (\mathbf{5})$$

and defines different, useful, *situations*:

$$\text{Situations} = \{\text{Situation 1, Situation 2, ... , Situation s}\} , \forall\, s \in \infty \quad (\mathbf{6})$$

and the *binders* that associate *formations* with *situations*:

$$\text{Binders} = \{\text{Binder 1, Binder 2, ... , Binder b}\} , \forall\, b \in \infty \quad (\mathbf{7})$$

and the optional, *tactical parameters*:

$$\text{Tactical Parameters} = \{\text{Tactical Parameter1,..., Tactical Parameter tp}\} , tp \in \infty \quad (\mathbf{8})$$

In a *situation*, the conditions that make it unique are defined:

$$\text{Situation} = \{\text{Condition 1, Condition 2, ... , Condition cd }\} , \forall\, cd \in \infty \quad (\mathbf{9})$$

A *binder* sets the *situations* that lead to a *formation*. Optionally, a *binder* can set the connection between several origin *formations* and a terminus *formation* through *situations*:

$$\text{Binder} = \{[\text{Origin Formations}], \text{Situations}, \text{Terminus Formation}\}, \quad [\text{Origin Formations}], \text{Terminus Formation} \in \text{Formations} \tag{10}$$

4.1.3. Formation

A formation is a high-level structure that aggregates all the agents with the intent of assigning them to specific sub-tactics. The aggregation is either wrought by using agents that belong to the same type, have the same more immediate goals, or both.

Formally, a *formation* is a specific association of *sub-tactics* with a defined *distribution* that may specify an *agent type*. Only one *formation* can be active at any given time. As such, the *formation* must include *sub-tactics* for all *agents*.

$$\text{Formation} = \{ \text{Distribution}, \text{Sub-Tactics}, [\text{Agent Types}] \} \tag{11}$$

The same *sub-tactic* can be used more than once in a *formation*. This allows an implicit definition of *Group*. Let *sub-tactics* be a multiset [6]:

$$\text{Sub-Tactics} = \{(\text{Sub-Tactic1}, m(\text{Sub-Tactic1})), (\text{Sub-Tactic2}, m(\text{Sub-Tactic2}), \ldots, (\text{Sub-Tactic st}, m(\text{Sub-Tactic st}))\}, \forall\, st \in \infty \tag{12}$$

where m(Sub-Tactic st) defines the multiplicity of a *sub-tactic*.

For each element in *sub-tactics* there is correspondent value in a *distribution*:

$$\text{Distribution} = \{\text{Value1}, \text{Value2}, \ldots, \text{Value v}\}, \; v = \sum m(\textit{Sub-Tactics st}) \tag{13}$$

A distribution specifies either absolute or percentage distribution values for each sub-tactic in the formation. Distribution *values* always refer to *agent types* when applicable. In this manner, the total of *values* can surpass 100%, but not for a specific *agent type.*

The association with *agent type* is implicit when a *sub-tactic* can only be applied to one *agent type*. Otherwise, when more than one *agent type* can be used (see section 4.1.5), an *agent type* must be specified for that *sub-tactic*:

$$[\text{Agent Types}] = \{\text{Type1}, \text{Type2}, \ldots, \text{Type ty}\}, \forall\, ty \in \infty \tag{14}$$

3.1.4. Sub-Tactic

A sub-tactic reflects the approach to face the scenario of a limited set of agents either partially for a number of situations or during the whole scenario.

Formally, a *sub-tactic* is an association of *roles* with one default *amount* of *agents* assigned to those *roles*. Additionally a *sub-tactic* may also have *sub-tactical parameters* to reflect specific thresholds, *agent* parameters, coordination options or other values that are needed to configure the *roles* used on the *sub-tactic*.

$$\text{Sub-Tactic} = \{\text{Amounts}, \text{Roles}, [\text{Sub-Tactical Parameters}]\} \tag{15}$$

A *sub-tactic* can have one or more *roles*:

$$\text{Roles} = \{\text{Role 1}, \text{Role 2}, \ldots, \text{Role r}\}, \forall\, r \in \infty \tag{16}$$

For each *role* in *sub-tactic* there is an *amount* in *amounts*:

$$\text{Amounts} = \{ \text{Amount 1, Amount 2, ... , Amount a} \} , \text{a} = \sum role \text{ r} \quad (17)$$

Like in a distribution, an *amount* specifies either absolute or percentage values for each *role* in the sub-*tactic*. Percentage *amounts* in a given *sub-tactic* must total 100%.

Sub-tactics *can be divided into Typed Sub-Tactic*s and *Generic Sub-Tactics*. In a *typed sub-tactic* at least one of the *roles* is associated with an *agent type*, which becomes the *sub-tactic's* type. In order to ease the handling of different *agent types*, it is not possible to use *roles* of different *agent types* in the same *sub-tactic.* As such, *typed sub-tactic* can only use *roles* for one *agent type* together with *generic roles*. As a consequence, to build a *formation* with different *agent types,* there should be at least one *sub-tactic* for each *agent type*.

A *generic sub-tactic* is a particular kind of *sub-tactic* without any association with an *agent type*. Thus, in a generic *sub-tactic,* only *generic roles* can be used. As it was previously stated, if a *generic sub-tactic* is used in a *formation* that contains *sub-tactics* for more than one *agent type*, an *agent type* must be specified. This type is specified together with a *distribution value* when *agents* are assigned to a *generic sub-tactic*.

In the event that there are no *agent types,* or there is only one type of agent in the *tactic*, all *sub-tactic* kinds are generic and can be refereed simply as *sub-tactic*.

3.1.5 Role

A role is a normal or customary activity of an agent in a particular environment.

Formally, a *role* is a set of *algorithms* in a defined sequence that describes an *agent*'s behavior. The behavior description is expected to include, when relevant, the specification on how the *agent,* should, "in field", coordinate with *agents* in the same *role* or in other *roles*. The coordination can be of three different kinds:

- All *agents* with the same assigned *role* form one *group;*
- All *agents* with the same assigned *role* form several smaller *groups* (with a splitting rule specified inside the *role*);
- All *agents* with the same assigned *role* act independently.

The *role* also defines partial objectives accordingly to the coordination method used. Although *roles* can describe the behavior for an entire scenario, they can also describe the behavior for only a given time frame or *situation*. *Teams* form their *roles* by combining different motion and action mechanisms with partial objectives.

The *role* level is the lowest in the proposed model. For teams who use sequenced task/objective/state based *agents*, a conversion to *role* based *agent* is discussed in section 5.

Similarly to the *sub-tactics*, *roles* can be divided into *Typed Role* or *Generic Role*. A *typed role* is a particular kind of *role* that can only be assumed by one *agent type*. Using heterogeneous *agents* does not necessarily means that *typed roles* or *agent types* will be used in the *strategy*. *Typed roles* are use when, in heterogeneous *agents*, there is a need to use the different *agent*'s properties or capabilities.

A *generic role* is a kind of *role* that can be assumed by any of the *agent types* used in a *tactic*. Analogously to a *generic sub-tactic*, in the event that there are no *agent types,* or there is only one type of *agent* in the *tactic*, all *role* kinds are generic.

3.2 Decision, Supervising and Communication

The decision maker depends on the *agents*' organization and types set by the scenario. In teams where there is only a supervisor and all the *agents* are "dummy", the strategical layer will obviously only be applied to the supervisor.

In multi-agent systems, the first rule is that all *agents* have full knowledge of the strategical layer being used. Then if all *agents* have a good, shared, world state knowledge using the layer can be done with no extra communication. This is accomplished because all the *agents* switch their *tactics*, *situations* and *formations* based on the same conditions and at almost the same time. When a team uses a mechanisms like ADVCOM (section 2), the no-communication version of the strategical layer can be applied to scenarios were the normal communications are limited and unreliable.

If *agents* have more limited computational resources but still have good world state knowledge synchronization, the layer can be computed only by a supervising *agent*. This supervising *agent* would only have to communicate a new *formation* whenever declared by the strategical layer. The supervising *agent* is chosen taking into account the *agent* who normally has more computational resources. Some scenarios specifically have supervising *agents*. In environments where the world state sharing is unreliable, the layer must be computed by a supervising *agent* choosing, typically, the best informed *agent* or communication may be used for team synchronization.

3.3 Agent Assignment

The strategical layer defines both absolute and percentage forms for *distribution values* and *role amounts.* This possibility is given so that strategies can be built independently from the *agent* number used in the scenario.

Another possibility of the model is to use both absolute and percentage forms simultaneous. In this model, for both *distribution values* and *role amounts*, absolute forms for values take priority over percentage value forms. This means that *agents* are assigned first to *roles* in a *sub-tactic* specified with absolute *distribution values* and with absolute *role amounts* in the referred *role.* Next *agents* are assigned to *sub-tactics* with only absolute forms of *distribution values.* The succeeding priority is assigning *agents* to *roles* specified by absolute *role amounts*, in a sub-tactic with a percentage *distribution values.*

Finally, for the remainder *agents* that use percentage forms in the mixed method, or when the percentage form is the only assignment method used, the assignment priorities follow. When converting to absolute numbers, the values are truncated and assigned. If there are any *agents* left, one *agent* is assigned to each *sub-tactics* and *roles* that did not received any *agents* in the decreasing order of their respective percentages. If there are any available *agents* left they are assigned sequentially to the *sub-tactics* and *roles* with the highest remainder values.

If *agent types* are in use, the previously defined assignment method is applied separately to each *agent type.* As it is easily concluded the mixed method allows the definition of priority *roles* in environment where the total *agent* number is unknown.

The *agent* assignment methods defined what *roles* needed to be used, particularly for environments where the total *agent* number is unknown. Next, the assignment of a specific *agent* to a specific *role* is discussed.

In order to assign *roles*, each *agent* must be capable of differentiate himself from others. Generally, this differentiation consists of attributing a different number or a id to each *agent*. There are a number of methods used to get an unique id namely it can be hard coded, attributed by a simulator or a referee, or even defined based on a relative position rule.

In its simpler form, the *role* assignment can be done by sequentially assigning one id to a *role*. Optimal *role* assignment depends on scenario conditions like proximity to objectives, relative *agents*' positions, etc.. Based on this fact, the model does not specify a method. In fact, a method like DPRE (section 2) that uses dynamic *role* exchanges is strongly advisable. To be noted that the strategical layer is still compatible with dynamic, situation based positioning like SBSP (section 2). This is accomplished because the positioning systems are coded inside the *role*, and are not assigned with the *role*.

4. Rescue Implementation

In order to adopt the strategic layer, our rescue team needed to use *role* concept. The previous code was based on a sequential selection of algorithms based on world state conditions. To reach the *role* level the following classifications were used:

- Action: a simple deed performed by an agent. E.g.: Action: Refill; Description: filling a Fire Brigade tank in a refuge.
- Task: set of actions performed by an agent that leads to a goal. E.g.: Task: Rescue civilian; Actions: Move to civilian; Unbury Civilian; Load Civilian; Move to refuge; Unload Civilian.
- Algorithm: set of tasks performed by one or more agents used to solve a particular field problem in a specific manner. E.g.: Algorithm: Clear main roads by prioritizing the main roads; Tasks: Each chosen road or set of roads is assigned to a specific Police Force, and then Police force agents clear the roads.

After identifying the algorithms, they were associated into *roles*. If there were two relevant algorithms with the same function but with different manners of solving the problem, they would be associated with two different *roles*. Some partial, *generic roles* like finding civilians were also created. Although these *roles* only included algorithms related to search and dislocation and do not have algorithms to act after all civilians are found, they are extremely useful.

The following figures depict a simplified rescue strategy. Some additional knowledge of the rescue simulation league is advisable to fully perceive the strategy. In Figure 2 the strategy is only expanded in one tactic and one formation

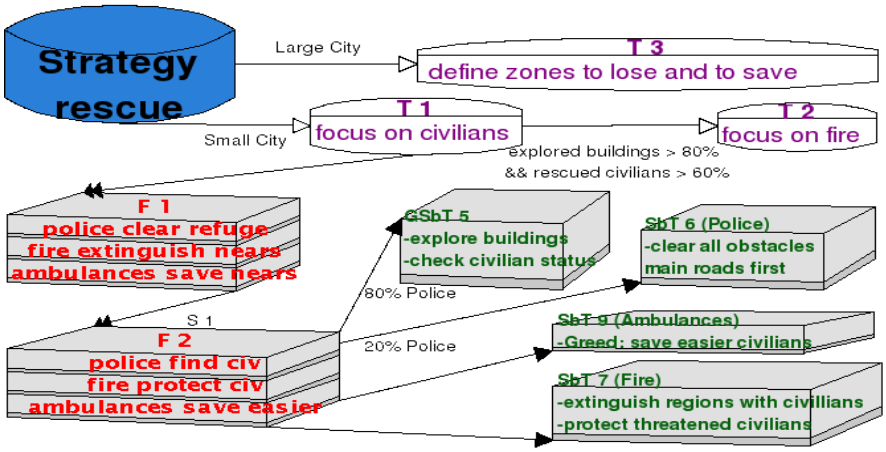


Fig. 2. Partial rescue strategy

In Figure 3 the situation to switch formation, S1 is defined.

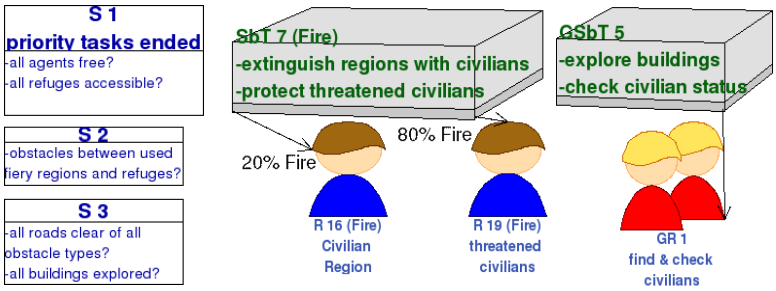


Fig. 3. a) Some rescue situations; b) two rescue sub-tactics.

5. Graphical Tool for Building Strategies

The graphical tool provides a visual interface for building strategies. By using graphical reorientations of the strategic layer components, it is possible to interconnect them. The tool already exports the built strategy to an XML file which can be used to implement the layer in *agents*. It also features a C++ code generator, still in its early stages and with some efficiency flaws. The graphical tool's GUI is provided by Kivio, a flowcharting and diagramming application for the KOffice[3] application suite.

Using soccer as an example, for a simple strategy, the same sheet can be used to represent the entire layer as seen in Figure 4. In this figure, on the left, is possible to see a developed, customized, installable, stencil set with the layer objects.

[3] KOffice is an office suite for the K Desktop Environment released under free software/open source licenses. Available at http://koffice.org/.

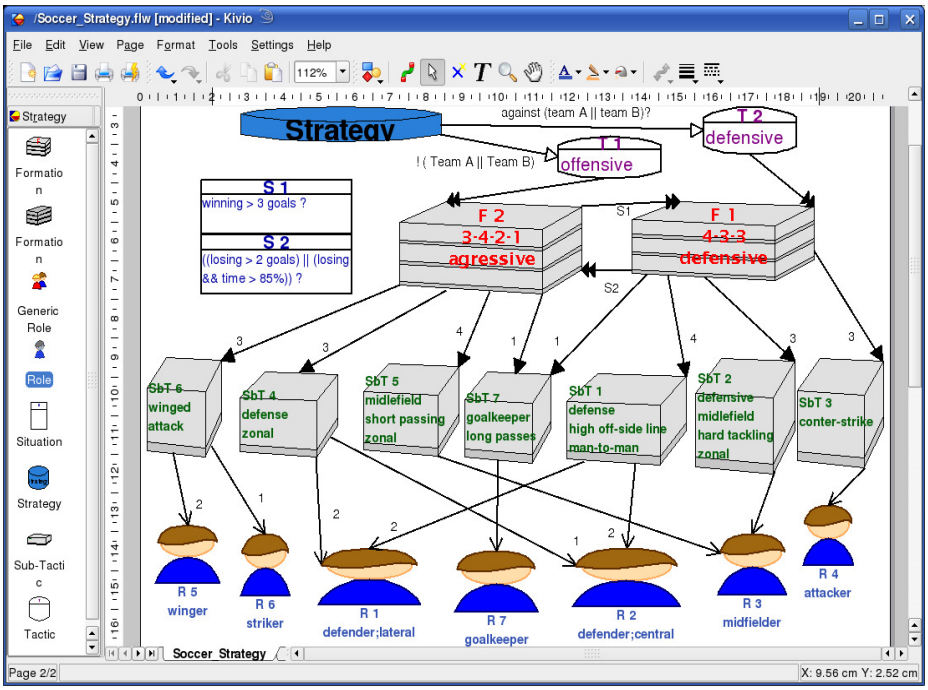


Fig. 4. Screenshot of the graphical tool featuring a soccer strategy

As shown, a trigger can originate in the *strategy* thus defining the initial *tactic*. Likewise a *binder* can originate in a tactic thus defining the initial *formation*. For a more complex *strategy* a multi-sheet is recommended separating *strategy*, each *tactic*, *formations*, *situations* and *sub-tactics*.

In fact, when defining several *tactics* which use at least one *binder* with no precedence (*origin formation)* a separate sheet for each tactic is mandatory. Figure 1 although only expands on one strategy branch (not possible under the layer), it uses formation 3 and formation 5 with no precedences.

6. Conclusion and Future Work

The proposed strategic layer is now fully integrated with our soccer and rescue teams and is successfully being used in our rescue team. The specified meta-model maintains full compatibility with all our RoboCup soccer teams, as the soccer model is a particular case of the specified generic layer.

The model flexibility enables using it both in an environment where a single program manages all homogeneous “dummy” robots, as in heterogeneous, multi-agent systems. When domains have similar nature like in soccer simulation and soccer robotic leagues, the strategies defined in one, can easily be adapted to the others. This is achieved by only modifying the roles in the existing sub-tactics. Innovative concepts like sub-tactics allow localized coordination and problem segmentation. The

developed priority and hierarchical agent's assignment methods, in the meta-model, allowed a great flexibility in strategy development for an unknown agent number. Results in international competitions for both the domains tested, proved the success of the layer.

Strategies are easily developed through the use of a very user-friendly graphical tool. By using a frequently improved, open source, editor as its base, the developed graphical tool can take advantages of its innovations.

In the future, further development of the graphical tool is expected, mostly on the source code generator. The graphical tool should also be able to generate efficient language independent code for the built strategy. These developments will enable a more generalized use of the strategic layer in the context of RoboCup and in other cooperative domains. Thus, we plan to use the strategical layer, with different instantiations, built using the graphical tool, in all our teams (simulation 2D, simulation 3D, small-size, middle-size, legged, simulation rescue and mixed-reality) participating in European and world RoboCup competitions in 2009.

References

1. Cohen, P. R. and Levesque, H. J., "Teamwork", *Noûs, Vol. 25, No. 4, Special Issue on Cognitive Science and Artificial Intelligence* Sep., 1991, pp. 487-512.
2. Kitano, H., Tadokoro, S., Noda, I., Matsubara, H., Takahashi, T., Shinjou, A., and Shimada, S., "RoboCup Rescue: search and rescue in large-scale disasters as a domain for autonomous agents research", in Proceedings of Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, 1999, pp. 739-743.
3. Lau, N. and Reis, L. P., "FC Portugal 2001 Team Description: Configurable Strategy and Flexible Teamwork", in *RoboCup-2001: Robot Soccer World Cup V, Lecture Notes in Artificial Intelligence*, Birk, A., Coradeschi, S., and Tadokoro, S., Eds. Berlin: Springer-Verlag, 2002.
4. Reis, L. P., Lau, Nuno, and Oliveira, E. C., "Situation Based Strategic Positioning for Coordinating a Team of Homogeneous Agents", in *Balancing Reactivity and Social Deliberation in Multiagent Sytems: From RoboCup to Real Word Applications*, Hannenbauer, M., Wendler, J., and Pagello, E., Eds., Berlin Springer-Verlag LNAI 2103, 2001.
5. Reis, L. P. and Lau, N., "FC Portugal Team Description: RoboCup 2000 Simulation League Champion", in *RoboCup-2000: Robot Soccer World Cup IV*, Stone, P., Balch, T., and Kraetzschmar, G., Eds., Berlin LNAI Springer-Verlag, 2001, pp. 29-40.
6. Stanley, R. P., "Enumerative Combinatorics ", vol. 1 and 2: Cambridge University Press., 1997, 1999.
7. Stone, P., *Layered Learning in Multiagent Systems: A Winning Approach to Robotic Soccer* MIT Press, 2000, ISBN: 0262194384.
8. Stone, P. and Veloso, M., "Task Decomposition, Dynamic Role Assignment, and Low-Bandwidth Communication for Real-Time Strategic Teamwork", *Artificial Intelligence*, vol. 110(2), June 1999, pp. 241–273.
9. Tambe, M., "Towards Flexible Teamwork", *Journal of Artificial Intelligence Research*, vol. 7, 1997, pp. 83-124.

Cyber-Mouse: A Deliberative Implementation

João Certo[1], João Oliveira[1] and Luis Paulo Reis[1]

[1] Faculdade de Engenharia da Universidade do Porto
{joao.certo, ee03123, lpreis}@fe.up.pt

Abstract. This paper presents an agent based maze solving problem facing two distinct situations: a known and unknown map. Different methodologies were implemented, namely an odometry based localization system, target localization through adaptive triangulation, a regressive obstacle distance function for mapping, a quad-tree strategy for map representation, and the A* algorithm for path planning. The integration of these methods granted the design of efficient deliberative agents for both experiments. A simulation environment called ciber-rato was used to test the different agent solutions through different complexity mazes to prove the original hypothesis. After evaluating the experiments, with different parameters, the robot's deliberative performance was compared to a previous reactive agent, for proof of concept.

Keywords: Software Agents, Robotics, Artificial Intelligence.

1 Introduction

Cyber-Mouse (Ciber-Rato) is a modality included in the "Micro-Rato" competition, directed to teams interested in the algorithmic issues and software control of mobile autonomous robots [1]. This modality is supported by a software environment, which simulates both robots and a labyrinth [2]. The Cyber-Mouse has seven sensors but only two, user selectable, are available at any given time. The final purpose is to reach the cheese, identified by a ground sensor and detectable through a direction providing beacon sensor visible trough low walls. Mouse's performance is evaluated through success on reaching the cheese, the time it took and the number of collisions.

The cyber-mouse competition has been used, amongst other applications as a testbed for long-term planning [3], as a scenario for the detection and avoidance of dangerously shaped obstacle [4], or even as a tool for the teaching of Artificial Intelligence and Robotics [5]. In this paper we evaluate the problems of mapping, localization and path planning by building a deliberative agent that can find its way from the starting position to the target without prior knowledge of the maze. Ultimately architecture performance is compared to a previous reactive approach.

This section introduced the Cyber-Mouse environment. Next the problems of mapping and self-localization, navigation and path planning are introduced together with some related state of the art algorithms leading to the chosen approach. Section 3 presents the implemented architecture based on several deliberative functionalities. Section 4 contains a description of the testing environments and the respective results. Finally in section 5 we conclude this paper and point to future work.

2 Robotic Mapping and Planning Overview

Mapping is the process of building an internal estimate of the metric map of the environment [6].The mapping problem is generally regarded of most importance in the pursuit of building truly autonomous mobile robots, but still mapping unstructured, dynamic, or large-scale environments remains largely an open research problem.

Planning is the process of deciding which route to take based on and expressed in terms of the current internal representation of the terrain. Typically this process calculates the cost of each motion decision towards the target, based on a given heuristics, and chooses the "cheapest" one.

In this section we present an overview on robotic mapping and planning and we introduce some state of the art algorithms in these fields. Based on this study, in the next section we explain our approach.

2.1 Mapping and Localization Problem

Robotic mapping addresses the problem of acquiring spatial models of physical environments through mobile robots, which are then used for robot navigation. To acquire a map, robots must possess sensors that enable it to perceive the outside world. Sensors commonly brought to bear for this task include cameras, range finders using sonar, laser, and infrared technology, radar, tactile sensors, compasses, and GPS. However, all these sensors are subject to errors, often referred to as measurement noise, and to strict range limitations.

So, considering these issues several different challenges can arise to robotic mapping: statistically dependent sensors measurement noise; high dimensionality of the entities that are being mapped; data association problem – problem of determining if sensor measurements taken at different points in time correspond to the same physical object in the world; environments change over time; robot exploration.

The motion commands issued during environment exploration carry important information for building maps, since they convey information about the locations at which different sensor measurements were taken. Robot motion is also subject to errors, and the controls alone are therefore insufficient to determine a robot's pose (location and orientation) relative to its environment. If the robot's pose was known all along, building a map would be quite simple. Conversely, if we already had a map of the environment, there are computationally elegant and efficient algorithms for determining the robot's pose at any point in time. In combination, however, the problem is much harder.

Considering the map representation problem, which has a significant impact on robot control [7], we can account for three main methods: Free space maps (road mapping), as spatial graphs, including Voronoi diagrams, and generalised Voronoi diagrams; object maps; and composite maps (cell decomposition) as point grids, area grids and quad trees.

Virtually all state-of-the-art algorithms for robotic mapping in the literature are probabilistic. They all employ probabilistic models of the robot and its environment,

and they all rely on probabilistic inference for turning sensor measurements into maps. In Fig. 1 a performance of the 8 major mapping algorithms is presented [6].

	Kalman	Lu/Milios	EM	Incremental ML	Hybrid	Occupancy Grids	Multi-Planar Maps	Dogma
Representation	landmark locations	point obstacles	point obstacles	landmark locations or grid maps	point obstacles	occupancy grids	objects and polygons	occupancy grids
Uncertainty	posterior poses and map	posterior poses and map	maximum likelihood map	(local) maximum likelihood map	maximum likelihood map	posterior map	maximum likelihood map	posterior map
Convergence	strong	no	weak?	no	no	strong	weak	weak
Local Minima	no	yes	yes	yes	yes	no	yes	yes
Incremental	yes	no	no	yes*	yes	yes	no	no
Requires Poses	no	no	no	no	no	yes	yes	yes
Sensor Noise	Gaussian	Gaussian	any	any	any	any	Gaussian	any
Can map cycles	yes	no	yes	no	yes, but not nested	n/a	n/a	n/a
Map dimensionality	$\sim 10^3$	unlimited	unlimited	unlimited	unlimited	unlimited	unlimited	unlimited
Correspondence	no	yes	yes	yes	yes	yes	yes	yes
Handles raw data	no	yes	yes	yes	yes	yes	yes	yes
Dynamic env's	limited	no	no	no	no	limited	no	yes

Fig. 1. Mapping algorithms comparison, [6].

Our mapping algorithm is based on a discrete model representing the distance to an obstacle given the obstacle sensor value. The map is represented using quad-tree decomposition since this algorithm grants good performance for this application, with low processing cost. A more detailed explanation of our strategy is given in Section 3.

2.2 Navigation and Path Planning

In artificial intelligence, planning originally meant a search for a sequence of logical operators or actions that transform an initial world state into a desired goal state [9]. Robot motion planning focuses primarily on the translations and rotations required to navigate, considering dynamic aspects, such as uncertainties, differential constraints, modelling errors, and optimality. Trajectory planning usually refers to the problem of taking the solution from a robot motion planning algorithm and determining how to move along the solution in a way that respects the mechanical limitations of the robot. The classic path planning problem is then finding a collision-free path from a start configuration to a goal configuration, in a reasonable amount of time, given the map representation, retrieved in the mapping process, and the robot's body constitution.

In an unknown environment the mapping and motion planning must be processed in parallel through exploration and dynamic navigation decisions. This structure requires plans updating. A natural way of updating plans is to first select a path based on the present knowledge, then move along that path for a short time while collecting new information, and re-planning the path based on new findings.

Considering the application many algorithms have been proposed for path planning: A and A Star (A*), Dijkstra, Best-First, Wavefront Expansion, Depth-First Search, Breadth-First Search. Our strategy uses the A* algorithm with a quad-tree representation of the map, as it will be explained in the next section. The decision was made by balancing implementation cost with a guarantee of a solution.

3 Architecture

Our architecture is presented in four independent modules, concerning the self-localization, target (goal) localization, mapping and navigation, and path planning problem. Previously we integrate these modules to solve various mazes facing different conditions: in a known map, with knowledge of the start and target position; and in an unknown environment without any previous knowledge.

3.1 Self-Localization

The self-localization is based on the robots' odometry which is defined by a dynamic movement model [8]. Eq. 1 represents the power of each motor considering the robot's inertia; Eq. 2 models the linear velocity, and Eq. 3 the rotation, which represents the angle with the North and initially assumes the compass direction. Finally, Eq. 4 and Eq. 5 results in the X and Y axis value of the robot, assuming an initial value for the robot's starting position:

$$\begin{cases} lOutPow_t = (lOutPow_{t-1} + lInPow_t)/2 \\ rOutPow_t = (rOutPow_{t-1} + rInPow_t)/2 \end{cases} (Um) \quad (1)$$

$$lin_t = \frac{|lOutPow_t + rOutPow_t|}{2} \quad (Um/cycle) \quad (2)$$

$$\begin{cases} rot_t = compass\ direction, & if\ Time = 0 \\ rot_t = rot_{t-1} + (rOutPow_t - lOutPow_t), & if\ Time > 0 \end{cases} (rad) \quad (3)$$

$$X_t = X_{t-1} + lin_t * \cos(rot_t) \quad (Um) \quad (4)$$

$$Y_t = Y_{t-1} + lin_t * \text{sen}(rot_t) \quad (Um) \quad (5)$$

Due to Gaussian noise this model induces a linear motion maximum error of:

$$\delta \leq \frac{Max(MotorPow) * NoiseDeviation + MotorResolution/2}{Max(MotorPow)} \ (\%) \quad (6)$$

As such, for each position estimate there is a maximum deviation for the Cartesian coordinates and 2* for the rotation angle. The simulator defines *Max(MotorPow)=0.15*, *NoiseDeviation=1.5%* and *MotorResolution=0.001* which infers ≈ 1.83% and a rotation error of 3.66%, acceptable for this application.

3.2 Target Localization

For the target localization we based our strategy in triangulation with recursive adjustments. This strategy is then divided in two subsequent steps.

A first strategy, as shown in Fig. 2(a), triangulates the beacon position by intersecting two lines given by two reference points (two different robot positions). The target is visible when the beacon is within the angular sensing range of the mouse

and there are no high obstacles between the mouse location and the beacon. In this situation, while exploring the map, a first point (*A*) is traced by memorizing the robot's position, the beacon direction (□) and the current rotation angle (□). Then within a Euclidean distance of 3um and an angle difference of at least 30º, to minimize the error, a new point (*B*) is memorized along with the beacon and rotation angle. Given the beacon sensor latency (4*ut*) and its Gaussian noise, in each point the mouse stops for 20*ut* and the beacon final angle is given by the average of the values retrieved from the 5th to the 20th cycle. Given these parameters the target position (X_C, Y_C) is achieved, in Eq.9, by the following deduction:

$$\begin{cases} Y_A = m_A * X_A + b_A \\ Y_B = m_B * X_B + b_B \end{cases} (Um)\ ,\ m = \tan\theta \tag{7}$$

$$\begin{cases} \theta = \alpha + \beta + 2\pi, & if\ \theta < -\pi \\ \theta = \alpha + \beta - 2\pi, & if\ \theta > \pi \\ \theta = \alpha + \beta, & else \end{cases} \quad (rad) \tag{8}$$

$$\begin{cases} X_C = (b_B - b_A)/(\tan\theta_A - \tan\theta_B) \\ Y_C = \tan\theta_A * X_C + b_A \end{cases} (Um) \tag{9}$$

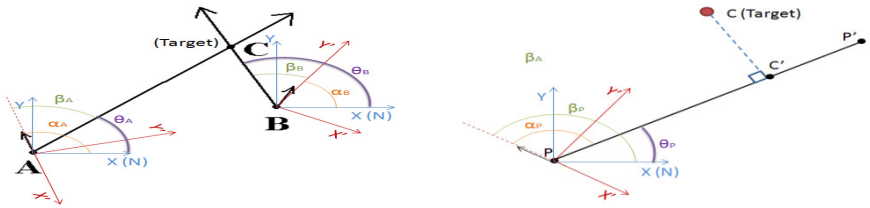


Fig. 2. a) Triangulation scheme; **b)** Target position adjustment.

The second step consists on adjusting the target point with recursive (25 in 25 cycles) new measures. Given each new position (*P*) and the correspondent beacon and rotation angles, it's traced a new line (linear equation) and the former beacon estimate is compared with its closest point on this new line (see Fig. 2(b)). This closest point ($X_{C'}$, $Y_{C'}$) is calculated as follows (Eq. 10 to14):

$$\begin{cases} Y_{P'} = any \\ X_{P'} = (Y_P - b_P)/\tan(\theta_P) \end{cases} \tag{10}$$

$$new\_line = P + u(P' - P) \tag{11}$$

$$(C - new_{line}).(P' - P) = 0 \tag{12}$$

$$u = \frac{(X_{C'} - X_P)(X_{P'} - X_P) + (Y_{C'} - Y_P)(Y_{P'} - Y_P)}{||P' - P||^2} \tag{13}$$

$$\begin{cases} X_{C'} = X_P + u(X_{P'} - X_P) \\ Y_{C'} = Y_P + u(Y_{P'} - Y_P) \end{cases} \tag{14}$$

The adjustment is then weighted considering the confidence of the current target estimate position. So depending on the number of previous adjustments, l, the new target location ($X_C(t)$, $Y_C(t)$) is given in Eq. 15:

$$\begin{cases} X_C(t) = X_C(t-1) * \left(1 - \frac{1}{n}\right) + X_{C'}(t) * \left(\frac{1}{n}\right) \\ Y_C(t) = Y_C(t-1) * \left(1 - \frac{1}{n}\right) + Y_{C'}(t) * \left(\frac{1}{n}\right) \end{cases}, \quad n = 3 + l, \ AND \ n < 6 \tag{15}$$

3.3 Mapping

The navigation and consequent mapping is based on the obstacles disposition along the map. To calculate the robot's distance to an obstacle we studied its relation to sensor values through successive empirical measurements.

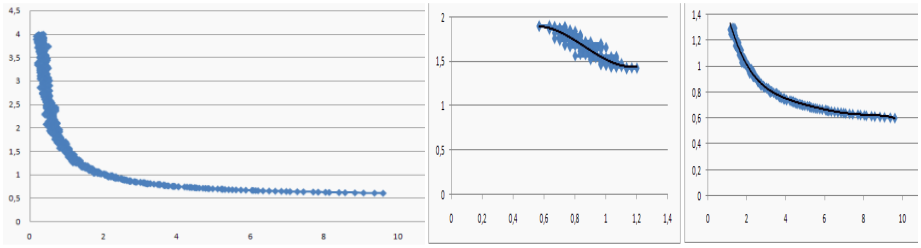


Fig. 3. Obstacle distance distribution: obstacle sensor values in horizontal axis (units); obstacle distance in vertical axis (mouse units). **a)** the full distribution; **b)** distribution for sensor values ranging from 0.9 to 1.0; **c)** distribution for sensor values ranging from 1.1 to 4.5.

In the experiments in Fig. 3 the distance, d, is in function of the given sensor values, x, and through linear regression obtained the following equation (Eq. 16):

$$\begin{cases} d = 4{,}1981x^3 - 10{,}84x^2 + 8{,}1978x - 0{,}0403, & if\ 0.9 < x < 1.0 \\ d = -0{,}0001x^5 + 0{,}0046x^4 - 0{,}0561x^3 + 0{,}3435x^2 - 1{,}095x + 2{,}2027, & if\ 1.1 < x < 4.5 \end{cases} \tag{16}$$

These functions estimate the obstacle frontal distance with a low error, δ, to a maximum of 0.213um for distances in the 0.9-1.0 sensor value range, and 0.189 in the 1.1-4.5 range.

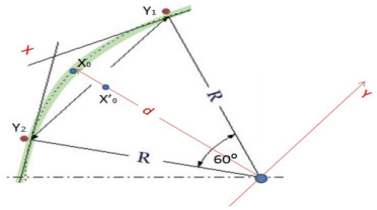


Fig. 4. Obstacle sensor coverage.

Considering the sensor aperture angle (60º) depicted in Fig. 4, the total sensor coverage is mapped in Eq.19 and 20.

$$|Y_2| = |Y_1| = X'_0 * sen(30) \tag{19}$$

$$X'_0 = Xo - Y_1 * sen(30) \tag{20}$$

3.3.1 Quad-Tree Map Representation

In the presence of an obstacle we used a Quad-Tree gridding to subdivide each of the obstacle cells. Our quad-tree strategy uses an adaptive division depth to a minimum maximum cell size (granularity) of 0.1*um*,. Fig. 5 illustrates different granularities for different known maps.

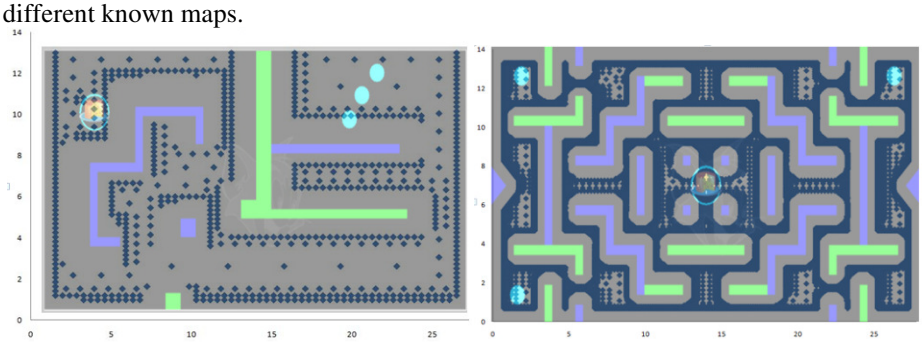


Fig. 5. Real map overlapped with Quad-Tree map representation, using different granularities: RTSS06Final with 0.7*um* depth (left); 2005Final with 0.1*um* depth(right).

Considering the minimum spacing between obstacles of 1.5um we implemented a method to grant the robot passage. This method consists on, after dividing each obstacle to the minimum defined cell size, also dividing each adjacent cell (see Fig. 6c)) to this minimum. Ultimately, the map outer walls were also subdivided to account for their presence, when calculating the passable cells (see Fig. 6b)).

The grid granularity (Quad-Tree depth) is user definable for testing different values in order to optimize the maze-solving performance, verifiable in the experiments (section *4*).

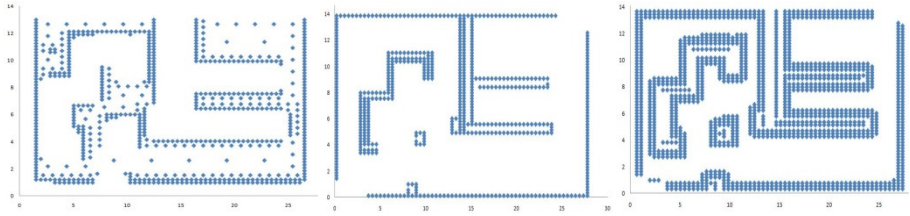


Fig. 6. Quad-Tree cells representation, for RTSS06Final map, with 0.7*um* depth (from left to right): passable cells (a); obstacle cells (b); adjacent cells (c).

3.4 Path Planning

To find a path between a previously computed objective and the agent, the A-Star (A*) algorithm is used. The following function represents the cost between the source point and the target point, which passes through node n:

$$f(n) = g(n) + h(n) \tag{21}$$

Here *g(n)* is the real cost from the source to node *n*, and *h(n)* is the estimated cost between node n and the target. f(n) is the total cost of the path that passes through node *n*. The used heuristic function is the Euclidean distance between the source and the target position. This function is implemented over the Quad-Tree map representation, by defining the shortest path towards the target by marking waypoints in the correspondent map cell centres.

As example, bellow, in Fig. 7, is represented the A* path planning around a wall.

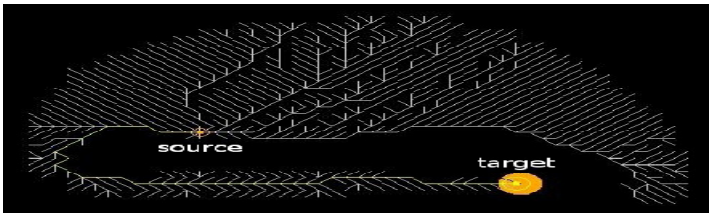


Fig. 7. Points visited around a wall, with lines to parent nodes in A*. The path is searched from the source to the target, [7].

3.5 Navigation and Control

The robot's is controlled to navigate towards the target, by following each waypoint centre given by the A* algorithm and its successive heuristic estimates. The waypoint centres are reached within a certain error margin which depends on the localization method in use (odometry-based or GPS). The agent navigation speeds are adjustable and dependent on several factors, namely the angle difference considering the target direction which makes the agent adjust its rotation angle as necessary; the distance to the next waypoint which, between other parameters, depend on the Quad-Tree granularity (chosen depth); and the slope between two consecutive waypoints.

4 Experiments and Results

This section comprises two maze solving experiments: a known map, using a specific map representation and path planning; an unknown map using real-time mapping and adaptive path planning. After explaining the implemented approach in each experiment, the correspondent results are evaluated and compared to the previously developed reactive architecture, "Smart Follower" [10].

4.1 Maze solving in a known map

In this experiment each tested map is parsed from its XML file to retrieve its exact representation along with the target and the robot's starting position. Then the quad-tree method is applied by subdividing each obstacle cell to the minimum required

size, delimiting passable areas, as previously showed in Fig. 5 and Fig. 6, for different maps and granularities (Quad-Tree depths). Using these areas the robot is controlled, as explained in section 3.3, to follow each waypoint given by the A* algorithm, considering the self-localization method (see 3.1) or the GPS for positioning. This procedure is done recursively until reaching the target area.

4.2 Maze solving in an unknown map

Presently the map solving isn't working due to lack of exploration. The robot does triangulate and adjust beacon position. The robot also uses correctly the previously described distance to obstacle polynomial function. However, mapping the obstacles with the previous simple model, using a sensor, is proving to be difficult.

4.3 Evaluation Scenarios

In order to evaluate each experiment the following, gradually increasing difficult scenarios, were chosen: *Basic* with a small wall between mouse and beacon (Fig. 8a); *MicRato98,* an easy map with only low walls (Fig. 8b); *2001Final,* a medium difficulty map with only low walls (Fig. 8c); *RTSS06Final,* a hard map with low and high walls (Fig. 8d);*2005Final*, a very hard map with low and high walls (Fig. 8e).

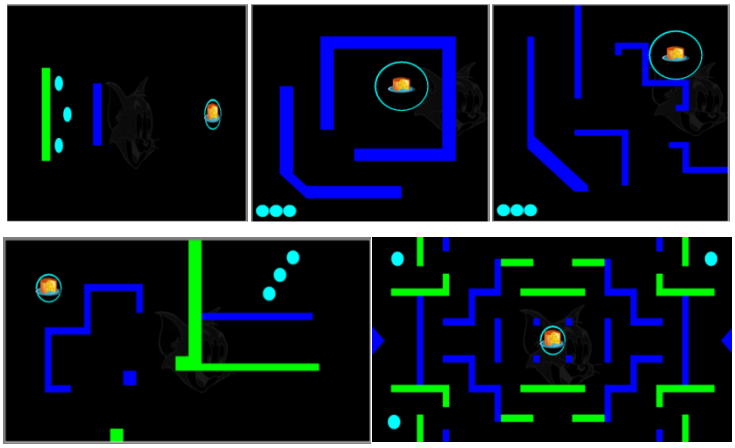


Fig. 8. Evaluation maps (from top-left to bottom-right): **a)** Basic; **b)** MicRato98; **c)** 2001Final; **d)** RTSS06Final; **e)** 2005Final.

4.4 Evaluation method

The evaluation is done by observing if the mouse reaches the cheese or not and the time it takes to do it. Since collisions impose errors in the self-localization procedure which would make the robot fail the waypoints (given by A*) towards the target, the number of collisions weren't considered. Additionally, when relevant, an

observational description of the mouse' behaviour during the experiment may be included to further evaluate and compare the approaches.

4.5 Results

Each map was tested within the two maze solving experiments, although the second experiment only tests the reactive agent, since the deliberative implementation in an unknown environment is still under implementation. For results comparison, in both experiments the tests were made with two self-localization systems: through odometry measurement (see 3.1) and with GPS; and for two minimum cell maximum sizes (considered in the quad-tree decomposition), with a 0.1 resolution: 0.1*um* and the one granting the best performance. Since the simulator adds some noise in the sensors and actuators, three different runs for each map and agent were performed. As such, conclusions can be made from the averaging of the results overcoming the stochastic nature of the simulator.

4.5.1 Maze solving in a known map

In this experiment we used the implementation described in 4.1. The achieved results, for each localization method and for the two considered minimum cell sizes, are found in the followings Table 1 and Table 2.

Table 1. Maze solving in a known map results for the minimum cell maximum size of 0.1*um*.

Known Map	Experiments						
	Minimum Cell Max. Size	1 Time	2 Time	3 Time	Average Time	Success. Exp.	Observations
Map1 - Basic-GPS	0.1	1378	1198	1215	1264	3	NA
Map1 - Basic-Odom.	0.1	958	958	NA	958	2	Miss Target-Loc. Errors
Map2 - MicRato98-GPS	0.1	2110	2131	2121	2121	3	NA
Map2 - MicRato98-Odom.	0.1	1506	1506	1506	1506	3	NA
Map3 - 2001Final-GPS	0.1	2899	2845	2205	2650	3	NA
Map3 - 2001Final-Odom.	0.1	2522	NA	2522	2522	2	Collision
Map4 - RTSS06Final-GPS	0.1	6674	6573	6691	6646	3	NA
Map4 - RTSS06Final-Odom.	0.1	5442	NA	5442	5442	2	Collision
Map5 - 2005Final-GPS	0.1	2843	3027	3049	2973	3	NA
Map5 - 2005Final-Odom.	0.1	2087	2087	NA	2087	2	Collision

Table 2. Maze solving in a known map results for the minimum cell maximum size for the best performance.

Known Map	Experiments						
	Minimum Cell Max. Size	1 Time	2 Time	3 Time	Average Time	Success. Exp.	Observations
Map1 - Basic-GPS	0.5	684	629	853	722	3	NA
Map1 - Basic-Odom.	0.5	597	NA	597	597	2	Miss Target-Loc. Errors
Map2 - MicRato98-GPS	0.2	1660	1678	1624	1654	3	NA
Map2 - MicRato98-Odom.	0.2	1843	1843	NA	1843	2	Miss Target-Loc. Errors
Map3 - 2001Final-GPS	0.2	2403	2205	2613	2407	3	NA
Map3 - 2001Final-Odom.	0.2	NA	2090	2090	2090	2	Collision
Map4 - RTSS06Final-GPS	0.5	4268	4433	3700	4134	3	NA
Map4 - RTSS06Final-Odom.	0.5	3120	3120	NA	3120	2	Miss Target-Loc. Errors
Map5 - 2005Final-GPS	0.2	2380	2306	2447	2378	3	NA
Map5 - 2005Final-Odom.	0.2	NA	NA	2367	2367	1	Collision

As observable the unsuccessful tests were provoked by collisions or target missing when using the odometry self-localization method, due to the consequent positioning errors. Each map has a correspondent minimum cell maximum size value for best performance, depending essentially on the spacing between walls and the target area.

4.5.2 Maze solving in an unknown map (reactive agent evaluation)

In this experiment we tested our best (quasi-)reactive agent, *Smart-Follower*, for paradigm comparison. The results are shown in Table 3. The experiment and correspondent results related to our deliberative implementation in unknown environments will be evaluated as future work, as exposed in *4.2* and in section *4*.

Table 3. Experimental results for the *Smart-Follower* agent.

Smart-Follower	Experiments									Observations
	1		2		3		Average			
	Time	Collisions	Time	Collisions	Time	Collisions	Successful Exp.	Time	Collisions	
Map1 - Basic	1260	0	228	0	1244	0	3	911	0	NA
Map2 - 2001Final	638	0	790	0	NA	NA	2	714	0	Closure Conflict
Map3 - 2002Lab1	772	0	762	1	902	1	3	812	1	NA
Map4 - RTSS06Final	NA	NA	1366	0	NA	NA	1	1366	0	Wall-Beacon Conflict
Map5 - 2005Final	NA	NA	NA	NA	NA	NA	0	NA	NA	Wall-Beacon Conflict

Here, the *(en)closure* conflict happens when the mouse is surrounded by walls on both side sensor and front, and an obstacle on the back. Although the side sensors detect an obstacle there was enough room for the mouse to pass. The *wall-beacon* conflict noted on the observation row happens when the target area is impossible to reach due to entrapment between walls, caused by the conflict of following the beacon and avoiding obstacles at the same time.

4 Conclusions and Future Work

As observed, in the known map experiment the use of our self-localization method granted better results than the GPS, since the exactitude of the GPS positioning makes the robot constantly adjust its position towards each waypoint centre. This effect is also evident through the discrepancy between GPS runs with the same conditions. In contrast, tests using self-localization, always achieves the same time results, as it discredits errors imposed by the motor's noise and consequently accounts for the motion inaccuracy. Yet this localization error only makes the robot collide or miss the target in about one third of the runs, with an exception to the *2005Final* map where the exclusive presence of 1.5um spacing demands great positioning accuracy (with a maximum error of about 1.5%), only granted by GPS. Still the majority of successful tests validate the method.

The minimum cell maximum size can be adjusted to improve the performance. For each map there is an optimum value which is dependent of the minimum spacing between obstacles and by the target area. This way the obstacle cells must be small

enough to grant the robot's passage between walls and to grant a waypoint in the target area centre, and big enough to keep a good performance.

As observable in the second experiment, quasi-reactive approaches, featuring some deliberations, can quite effectively resolve most of the maps and situations with simple algorithms. Only the *2005Final* map couldn't be resolved. Comparing its results with the ones achieved by our deliberative agent, within known maps, we can infer that the reactive implementation simplicity granted better timing performances. Nevertheless the deliberative agent proved to be goal-effective with an accuracy of approximately 75%, only failing the target due to an error-prone odometry self-localization method. When using GPS the target was always reached, independently on the map. As a final remark one might refer that the *Smart-Follower* senses space and time, by sensing the world with its multidisciplinary sensors, while our deliberative agent is deaf, blind and mute, founding its planning and correspondent navigation on an internal map representation.

In the future, in known map environments, the proximity sensors can be used in order to avoid obstacles collisions or even to correct odometry errors ensuring a success rate of 100% in odometry based systems. A greedy type approach can also be used for determining the maximum cell size for solving a known map. This approach should improve the time taken to reach the beacon.

Regarding unknown maps, a probabilistic model should be used when determining if a cell as an obstacle. Successive readings would thus improve the mapping process.

References

1. Almeida, L., Fonseca, P., Azevedo, J.L.: The Micro-Rato Contest: a popular approach to improve self-study in electronics and computer science. SMC'2000, IEEE Int. Conference on Systems, Man and Cybernetics, Vol. 1, Nashville, USA (2000) 701 - 705
2. Lau, N., Pereira, A., Melo, A., Neves, A., Figueiredo, J.: Ciber-Rato: Um Ambiente de Simulação de Robots Móveis e Autónomos. Revista do DETUA **3** (2002) 647 - 650
3. Ribeiro, P.: YAM (Yet Another Mouse) - Um Robot Virtual com Planeamento de Caminho a Longo Prazo. Revista do DETUA **3** (2002) 672-674
4. Luís, P., Martins, B., Almeida, P., Silva, V.: Detecção de Configurações de Obstáculos Perigosas: Aplicação no Robô EnCuRRalado. Revista do DETUA **3** (2002) 659-661
5. Reis, L.P.: Ciber-FEUP - Um Agente para Utilizar o Simulador Ciber-Rato no Ensino da Inteligência Artificial e Robótica Inteligente. Revista do DETUA **3** (2002) 655-658
6. Thrun, S. – Robotic Mapping: A Survey CMU-CS-02-111 (2002).
7. Reis, L.P., Lau, N., Mapping and Navigation, support slides presentation for Intelligent Robotics course, online at: http://paginas.fe.up.pt/~lpreis/robo2008/docs/5IR0809-Mapping_Navigation.pdf, accessed 15 December, 2008.
8. Lau, N., CiberRato 2008 Rules and Technical Specifications, online at: http://microrato.ua.pt/main/Docs/RegrasMicroRato2008_EN.pdf accessed 15 December 2008.
9. LaValle, S.: Planning Algorithms. Cambridge University Press, 2006
10. Certo, J., Oliveira, J., Gonçalves, R.: Cyber-Mouse: Analysis on a Quasi Reactive Approach. Robótica Inteligente, FEUP, 2008, pp.6.

Implementing a Multiprocessor Linux Scheduler for Real-Time Sporadic Tasks

Paulo Baltarejo Sousa

Faculty of Engineering, University of Porto,
Rua Dr. Roberto Frias,s/n, 4200-465 Porto Portugal
pro08009@fe.up.pt

Abstract. The advent of multicore systems have renewed the interest of research community on real-time scheduling on multiprocessor systems. Real-time scheduling theory for uniprocessors is considered mature, but real-time scheduling theory for multiprocessors is an emerging research field. Being part of this research community we have decided to implement a multiprocessor Linux scheduler for a new real-time scheduling algorithm that was designed to schedule real-time sporadic tasks on multiprocessor systems. Testing and debugging operations have provided a better understanding of the algorithm and also of the platform used for implementation (Linux 2.6.24). Additionally, we have proposed some changes to the algorithm, in order to increase the performance and robustness.

Keywords: Real-time systems, Multiprocessor scheduling, Linux

1 Introduction

Multicore platforms are being commercialized coming with an increasing number of cores, expecting to reach hundreds of processors per chip in the future [1]. These new platforms have renewed the interest of the research community for real-time scheduling on multiprocessor systems, since this research field has started a long time ago [2, 3].

Real-time scheduling theory is well-developed for uniprocessor systems, but for multicore systems it is emerging. Usually these theoretical scheduling algorithms assume a set of assumption that do not have correspondence in a real system. For instance, they usually consider negligible the time for context switching and the execution time of the scheduler. We have chosen to study real-time multicore scheduling theory with practice.

In this paper we present an embryonic implementation of the Sporadic Multiprocessor Linux Scheduler (SMLS), which implements the recently published Sporadic Multiprocessor Scheduling (SMS) algorithm [4] that was designed to schedule real-time sporadic tasks on multiprocessor systems. According to our best knowledge there are not any previous implementations of this algorithm, so this is the first attempt to deeply study this algorithm in practice.

The SMLS has been implemented by modifying the general purpose Linux 2.6.24 kernel version. Our choice did not fall with any real-time Linux version, because (i) Linux 2.6.24 kernel version is provided with items that we believe to be relevant for implementing this algorithm, such as high resolution timers, preemption and dynamic ticks, and (ii) there are too many versions of real-time Linux that shows there are no consensus on what constitutes a real-time Linux [5].

The rest of paper is structured as follows. In Section 2 we begin an overview of the most relevant concepts of real-time multiprocessor scheduling and we also introduce the Linux modular scheduling framework. Based mainly on material from [4], we describe the main concepts of the Sporadic Multiprocessor Scheduling (SMS) algorithm. Next, we proceed with in a brief description of the SMLS in Section 4. In Section 5 we outline some issues related to the algorithm and to the platform used to implement the algorithm. In this section we also present our contributions to improve the performance of the SMS algorithm. We conclude the paper and discuss our plans for improving the SMS algorithm implementation in Section 6.

2 Background

The purpose of this section is to give to the reader the necessary background to understand the content of this document.

2.1 Real-Time Scheduling Algorithms on Multiprocessors

The most common definition of real-time systems is: *Real-Time Systems* are defined as those systems in which the correctness of the system depends on the logical result of computation, but also on the time at which the results are produced [6]. This time is usually referred to as *deadline*.

Real-time applications are usually composed by multiple tasks. Depending on the criticality level, tasks can be classified as: i) *hard real-time*, when missing a deadline produces undesirable or fatal results and (ii) *soft real-time*, where missing a deadline is not desirable but the system can still work correctly.

Another characteristic of real-time tasks is the periodicity, which defines the frequency in which they are activated or appear in the system. They can be classified as: (i) *periodic*, which appear regularly with at some known rate, (ii) *aperiodic*, which appear irregularly with at some unknown rate and (iii) *sporadic*, which appear irregularly with at some known rate.

The real-time scheduling algorithm should schedule tasks according to their demands such that their deadlines are met. One of the most used for uniprocessor systems is the *Earliest-Deadline-First* (EDF) [2]. The EDF scheduling algorithm is a dynamic priority driven algorithm in which higher priority is assigned to the task that has earlier deadline.

Advances in process technology allow integration of multiple processors on a single chip, called *multicores*. Multicore processors are now mainstream, with

the number of cores increasing, expecting to reach hundreds of processors per chip in the future [1, 7].

Unfortunately, real-time scheduling on multiprocessors did not enjoy such a success as it did on a uniprocessor. As early as in the 1960s, it was observed by the inventor of EDF that [3]: "*Few of the results obtained for a single processor generalize directly to the multiple processor ...*".

The research community has focused it interests on developing new scheduling algorithms [4, 8] for multiprocessors systems, which in many cases use concepts and principles used in the scheduling algorithms for uniprocessor.

The multiprocessor scheduling algorithms have traditionally been categorized as *global* or *partitioned*. Global scheduling algorithms store tasks in one global queue, shared by all processors. At any moment, the m (assuming that the system is composed by m processors) highest-priority tasks among those are selected for execution on the m processors. Tasks can migrate from one processor to another during the execution, that is, a execution of a task can be preempted in one processor and resume its execution on another processor. In contrast, partitioned scheduling algorithms part the task set such that all tasks in a partition are assigned to the same processor. Tasks may not migrate from one processor to another. An important issue is, in both systems, one task can be executed by only one processor at any given time instant.

2.2 Linux Modular Scheduling Framework

The introduction of scheduling classes, in the Linux 2.6.23 kernel version, made the core scheduler quite extensible. The scheduling classes encapsulate scheduling policies and are implemented as modules [9]. Then, the kernel consists of a scheduler core and various modules. These modules are hierarchically organized by priority and the scheduler dispatcher will look for runnable task of each module in a decreasing order priority.

Currently, Linux has three native scheduler modules: *RT* (Real-Time), *CFS* (Completely Fair Scheduling) and *Idle*. Thus, in this system the dispatcher will always look in the run queue of *RT* for a runnable task. If the run queue is empty, only then it moves to the *CFS* run queue and so on. The *Idle* module is used for idle task, every processor has an idle task in its run queue that is executed when there is no other runnable task.

3 Sporadic Multiprocessor Scheduling Algorithm

The Sporadic Multiprocessor Scheduling (SMS) algorithm [4], was designed to schedule real-time sporadic tasks. The SMS algorithm tries to clamp down on the number of preemptions (which involve operating system overheads) and can be configured to achieve different levels of utilization bound and migration costs. This algorithm can be categorized as semi-partitioned, since it assigns $m - 1$ tasks (assuming that there are m processors in the systems) to two processors and the rest to only one processor. Next, the details of the algorithm will be discussed.

3.1 System Model

The SMS algorithm consider the problem of preemptively scheduling n sporadic tasks on m identical processors. A task τ_i is uniquely indexed in the range $1..n$ and a processor in the range $1..m$. Each task τ_i is characterized by worst-case execution time C_i and minimum inter-arrival time T_i and by the time that the execution must be completed, the deadline (D_i). In this algorithm it is assumed that T_i and C_i are real numbers and $0 \leq C_i \leq T_i$ and also $D_i = T_i$.

A processor p executes at most one task at a time and no task may execute on multiple processors simultaneously. The system utilization is defined as $U_s = \frac{1}{m} \cdot \Sigma_{i=1}^{n} \frac{C_i}{T_i}$. The SMS algorithm divides time into slot of length $S = \frac{TMIN}{\delta}$. Where $TMIN$ is the minimal inter-arrival time of all tasks ($TMIN = min(T_1, T_2, \cdots, T_n)$) and δ is a parameter assigned by the designer to control the frequency of migration of tasks assigned to two processors.

α is an inflation parameter and is computed as follows: $\alpha = \frac{1}{2} - \sqrt{\delta \cdot (\delta + 1)} + \delta$. Later in this document the purpose of α will be explained. SEP is the utilization bound of SMS algorithm and is computed as follows: $SEP = 1 - (4 \cdot \alpha)$.

The SMS algorithm can be divided into two algorithms. An offline algorithm for task assignment and an online dispatching algorithm. These algorithms will be detail in the next sections.

3.2 Tasks Assigning Algorithm

The first step of the algorithm is to sort the task set by task utilization ($U_i = \frac{C_i}{T_i}$) in descending order, such that τ_1 is the heaviest and τ_n is the lightest tasks, respectively. Tasks whose utilization exceed SEP (henceforth called *heavy tasks*) are each assigned to a dedicated processor. Then, the remaining tasks are assigned to the remaining processors in a manner similar to next-fit bin packing [10]. Assignment is done in such a manner that the utilization of processors is exactly SEP. Task splitting is performed whenever a task causes the utilization of the processor to exceed SEP. In this case, this task (henceforth called a *split task*) is split by the current processor p and by the next one $p + 1$. Then, in these processors there are time window (called *reserves*) where this split task has priority over other tasks (henceforth called *non-split tasks*) assigned to these processors. The length of the reserves are chosen such that no overlap occurs, the split task can be scheduled, and also all non-split tasks can meet deadlines. The non-split tasks are scheduled under EDF.

Consider a system with four processors ($m = 4$) and seven tasks ($n = 7$). Table 1 shows the worst case execution time (C_i), minimum inter-arrival time (T_i) and utilization ($U_i = \frac{C_i}{T_i}$) of each task τ_i. As a matter of simplicity, the task set is already sorted in descending order by U_i. Assume also that $\delta = 4$ and consequently the utilization of the SMS algorithm is 88.85 % ($SEP = 0.8885$). The time units are intentionally omitted in Table 1, because they are not important for understanding the algorithm.

The task assignment algorithm works as follows: since τ_1 is a *heavy task* it is assigned to a dedicated processor (P_1). τ_2 is assigned to processor (P_2), but

Table 1: Task set

Task	C	T	U
τ_1	9	10	0.9000
τ_2	7	12	0.5833
τ_3	7	13	0.5385
τ_4	8	16	0.5000
τ_5	6	14	0.4286
τ_6	6	16	0.3750
τ_7	3	17	0.1765

assigning task τ_3 to processor P_2 would cause the utilization of processor P_2 to exceed SEP ($0.5833 + 0.5385 > 0.8885$). Therefore, task τ_3 is split between processor P_2 and processor P_3. A portion of task τ_3 is assigned to processor P_2, just enough to make the utilization of processor P_2 equal to SEP, that is 0.3052. This part is referred as $hi\_split[P_2]$ and the remaining portion (0.2332) of task τ_3 is assigned to processor P_3, which is referred as $lo\_split[P_3]$. The task set assignment to processors is shown on Table 2. Note that, the U_p of each processor is shown in the last column.

Table 2: Task set assignment

Processor	**Tasks and Utilization**			U
	lo_split		*hi_split*	
P_1		τ_1: 0.9000		0.9000
P_2		τ_2: 0.5833	τ_3: 0.3052	0.8885
P_3	τ_3: 0.2332	τ_4: 0.5000	τ_5: 0.1533	0.8885
P_4	τ_5: 0.2733	τ_6: 0.3750 and τ_7: 0.1765		0.8247

3.3 Dispatching Algorithm

On a dedicated processor, the dispatching algorithm is very simple, whenever there is one task ready to be executed, the processor executes this task. Recall that the time is divided into timeslot of length $S = \frac{TMIN}{\delta}$ and non-dedicated processors usually execute split and non-split tasks. For that, the timeslot might be divided in three parts. The first time units x are reserved for executing the $lo\_split[p]$ and the last time units y are reserved for executing the $hi\_split[p]$. The remainder is reserved for executing non-split tasks. However, it is important to note that one split task executes one portion on processor p and the remaining portion on another processor $p + 1$. This means that a split task τ_i will execute on both processors but not simultaneously (Fig. 1).

Reserves x and y for each split task must be sized such that $\frac{x+y}{S} = \frac{C_i}{T_i}$. Depending on the phasing of the arrival and deadline of τ_i relative to timeslot boundaries, the fraction of time available for τ_i between its arrival and deadline

may differ from $\frac{x+y}{S}$, since a split task only executes during the reserves. Consequently, it is necessary to inflate reserves by α in order to always meet deadlines: $x = S \cdot (\alpha + lo\_split[p+1])$ and $y = S \cdot (\alpha + hi\_split[p])$. Note that, the timeslot composition is usually different for every processors. The online dispatching

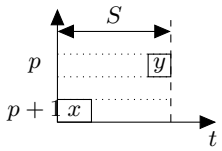


Fig. 1: x and y reserves for one task

algorithm works over timeslot of each processor and whenever the dispatcher is running, it checks to find the time elapsed in the current timeslot: (i) if the current time falls within a reserve (x or y) and if the assigned split task is ready to be executed, then the split task is scheduled to run on processor. Otherwise, the non-split task with the earliest deadline is scheduled to execute on processor. (ii) If the current time does not fall within a reserve, the non-split task with the earliest deadline is scheduled to run on processor.

Table 3 presents the timeslot composition for every processor for the system model presented on Section 3.2 and Fig. 2 shows a simplified execution timeline. The timeslot length is $S = 2.5000$ and the inflation factor is $\alpha = 0.2786$. In execution timeline presented on Fig. 2 only one activation of each task is assumed and also that the release time of all tasks is at the same instant. The execution of the tasks is represented by rectangles labeled with the task's name. A black circle states the end of execution of a task. As one can see, the split tasks execute only within reserves. For instance, task τ_3 on processor P_2 executes only on reserves. Outside its reserves it does not use the processor, even if it is free. In contrast, the non-split tasks execute mainly outside the reserves but potentially also within the reserves, namely, when there is no split task ready to be executed. There are two clear situations in the Fig. 2 that illustrate this. First (a), task τ_7 executes at the beginning of the timeslot, which begins at 12.50, because the split task τ_5 has finished its execution on the previous timeslot. Second (b), split task τ_5 ends its execution a little bit before the end of timeslot that finishes at 12.50 and there is some available time on the reserve, which is used by non-split task τ_4.

4 Sporadic Multiprocessor Linux Scheduler

In this section, the implementation of the Sporadic Multiprocessor Linux Scheduler (SMLS) will briefly described. Here, only the important details necessary for understanding the content of this document are referred. A detailed description

Table 3: Timeslot composition

Processor	x	**non-reserve**	y
P_1	0.0000	2.5000	0.0000
P_2	0.0000	1.6673	0.8327
P_3	0.6528	1.3893	0.4579
P_4	0.7529	1.7471	0.0000

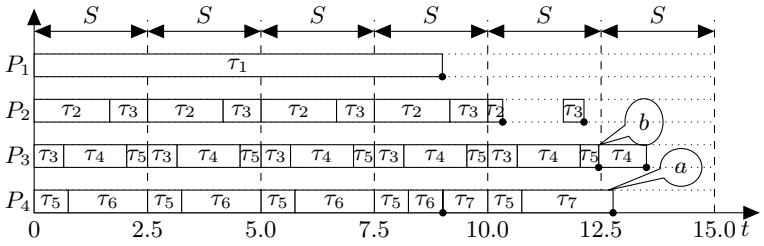


Fig. 2: Execution timeline

of this implementation can be found in [11]. The SMS algorithm was implemented using the Linux 2.6.24 kernel version.

4.1 Data Structures

A Linux *process*/task is an instance of a program in execution [12]. To manage processes, the kernel maintains information about each process in a *process descriptor*. The information stored in each process descriptor (`struct task_struct`) concerns with run-state of process, address space, list of open files, process priority, just to mention some. All process descriptors are stored in a circular doubly-linked list.

To support the SMS algorithm some fields must be added to this data structure. Listing 1 shows the most important fields added. Fields `processor1` and `processor2` are used to set the logical identifier of processor(s) in which the task will be executed. Note that, according to the SMS algorithm each non-split task executes only on one processor, and each split task executes on two processors. In the former, these fields are set with the same identifier, in the latter, the $lo\_split[\tau_i]$ and $hi\_split[\tau_i]$ are executed on processors which identifiers are set on `processor1` and `processor2`, respectively. The relative deadline of each task is set on the `deadline` field.

```
struct task_struct {
  ...
  int processor1;
  int processor2;
  unsigned long long deadline;
};
```

Listing 1: Fields added to the `struct task_struct` data structure

Each processor holds a run queue of all runnable processes assigned to it. The scheduling algorithm uses this run queue to select the "best" process to be executed. The information for these processes is stored in a per-processor data structure called `struct rq`. Listing 2 shows new data structures added to the `struct rq` data structure.

The information about each SMS task is stored using the `struct sms_task` data structure. Thus, `task` field is a pointer to the process descriptor. The absolute deadline is stored on the `deadline` field. A data type `struct rb_node` field is required for using SMS tasks on a red-black tree (`node_edf`). The linux kernel has already implemented red-black tree. Basically, red-black trees are balanced binary trees whose external nodes are sorted by a key, the most operations are done in $O(log_{(n)})$ time.

All SMS tasks assigned to one processor are managed using the `struct sms_rq` data structure (Listing 2). The root of the red-black tree is the field `rb_edf`. All non-split tasks are organized in a red-black tree by the absolute deadline. To manage the reserves and the split tasks, this data structure has two fields (`x` and `y`) to specify the length of reserves and also two pointers for process descriptor of the split tasks (`lo_split` and `hi_split`).

```
struct sms_task {
  struct task_struct *task;
  unsigned long long deadline;
  struct rb_node node_edf;
  ...
};
struct sms_rq {
  struct rb_root rb_edf;
  struct split_task {
    ...
    struct task_struct *lo_split;
    struct task_struct *hi_split;
  }split_task;
  unsigned long long x;
  unsigned long long y;
  ...
};
```

Listing 2: New data structures

4.2 New Scheduling Policy

To add a new scheduling policy to the Linux kernel it is necessary to create a new module. In this implementation, the *SMS* module was added on the top of the modules hierarchy, thus it is the highest priority module. Our system is hierarchically organized as it is shown in the Fig. 3.

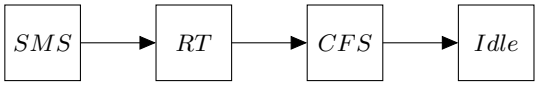


Fig. 3: Priority hierarchy of scheduler modules

Note that, each scheduler module is coded in a file. The *RT*, *CFS* and *Idle* are coded in the `/kernel/sched_rt.c`, `/kernel/sched_fair.c` and `/kernel/`

`sched_idletask.c` files, respectively. Then, to implement the *SMS* module we have created the `/kernel/sched_sms.c` file.

According to the modular scheduling framework rules each module must implement the set of functions specified in the `sched_class` structure. Listing 3 shows the definition of `sms_sched_class`, which implements the *SMS* module. The first field (`next`) of this structure is a pointer to `sched_class` which is pointing to the `rt_sched_class` that implements the *RT* module.

The other fields are functions that act as callbacks to specific events. The `enqueue_task_sms` is called whenever a SMS task enters in a runnable state. This function must check if it is a split task or a non-split task. In the former case it must update the pointers to split tasks, in the latter it must insert a node on the red-black tree. When a SMS task is no longer runnable, then the `dequeue_task_sms` function is called that undoes the work of the `enqueue_task_sms` function. As the name suggests, `check_preempt_curr_sms` function, checks whether the currently running task must be preempted. This function is called following the enqueueing or dequeueing of a task and also following any interruption. This function only sets a flag that indicates to the scheduler core that the currently running task must be preempted. `pick_next_task_sms` function selects the task to be executed by the current processor. This function is called by the scheduler core whenever the currently running task is marked to be preempted. `task_tick_sms` function is mostly called from time tick functions. In the current implementation this function calls the `check_preempt_curr_sms` function, to check, if the current task must be preempted. The last three functions work according to the dispatching algorithm.

```
const struct sched_class sms_sched_class = {
  .next = &rt_sched_class,
  .enqueue_task = enqueue_task_sms,
  .dequeue_task = dequeue_task_sms,
  .check_preempt_curr = check_preempt_curr_sms,
  .pick_next_task = pick_next_task_sms,
  .task_tick  = task_tick_sms,
  ...
};
```

Listing 3: `sms_sched_class` definition

5 Tests and Results

Since SMS algorithm implementation is at an embryonic state, the results presented here come especially from testing and debugging operations. Usually, from theory to practice there are some obstacles and barriers that depend on the platform used to. On the other hand, the implementation brings a set of details that were not considered in the theoretical work. In this section, these details that the scheduler development and the underlying platform became aware of, are described.

5.1 New Timeslot Selection

The algorithm defines at most two reserves x and y in the timeslot of each processor and one split task τ_i executes one portion on reserve y of the processor p and the other portion on reserve x of the processor $p+1$. Nevertheless, looking for two consecutive timeslots we realize that whenever a split task finishes the execution on processor p, the task has to immediately resume execution on its reserve on processor $p+1$. Actually, this situation can imply more overhead and more preemptions. This will be explained using the Listing 4, which shows part of the `check_preempt_curr_sms` function code and assuming the task set presented in Section 3.2, more specifically the situation illustrated in Fig. 4.

Let us assume that the current processor is processor P_3.One of the arguments of the `check_preempt_curr_sms` function is a pointer (`struct rq * rq`) to the run queue of the processor P_3 where all runnable SMS tasks assigned to it are stored as well as the other important data necessary for SMS scheduling algorithm, such that x and y reserves. The relative time instant within in the current timeslot is given by invocation of the `get_timeslice_reserve(rq)` function. Assuming that, `get_timeslice_reserve` invocation returns `RESERVE_LO_SPLIT`, which means that the current time instant falls in the x reserves. Then, the next step is to get the pointer to the split task τ_3 (`get_lo_split_task(&rq->sms_rq)`). Then, there is the need to check if task τ_3 is currently running on processor P_3. If it is, nothing is done. Otherwise, `resched_task(rq->curr)` function is invoked to preempt the currently running task and it also checks if task τ_3 is running on the processor P_2 (`if((cpu_curr(lo_split->processor1))==lo_split)`. If it is, processor P_3 sends an interprocessor interrupt to force rescheduling on processor P_2 (`resched_cpu(lo_split->processor1)`) to stop the execution of the task τ_3.

```
static void check_preempt_curr_sms(struct rq *rq, struct task_struct *p)
{
  ...
  r=get_timeslice_reserve(rq);
  switch(r){
    case RESERVE_LO_SPLIT:
      lo_split=get_lo_split_task(&rq->sms_rq);
      if(lo_split!=NULL){
        if(lo_split!=rq->curr){
          resched_task(rq->curr);
          if((cpu_curr(lo_split->processor1))==lo_split){
            resched_cpu(lo_split->processor1);
          }
          return;
        }
      }
  ...
}
```

Listing 4: Part of the `check_preempt_curr_sms` function code

This causes additional overhead, that could be avoided if the x reserve was available some time units later. So, the authors have changed the timeslot composition in such manner that x reserve is available some $M < 2 \cdot \alpha \cdot S$ time units later as shown in the (Fig. 5). Note that, the scheduling analysis presented in [4] is still valid, thus every system that is schedulable under the previously published SMS algorithm is also schedudable under this changed algorithm.

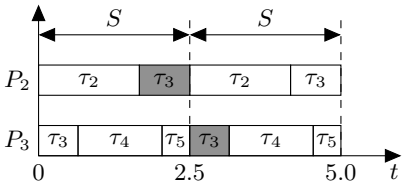


Fig. 4: Timeline execution of the split task τ_3

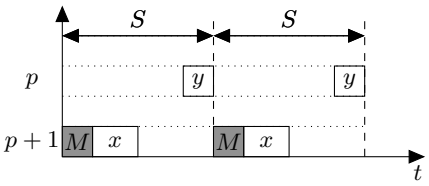


Fig. 5: New timeslot composition

5.2 Improving Performance

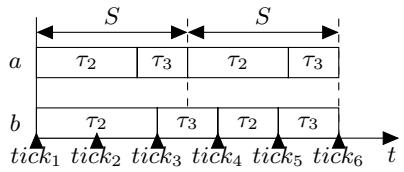
According to the original SMS algorithm $TMIN$ is computed as the minimal arrival time of all tasks ($TMIN = min(T_1, T_2, \cdots, T_n)$) and the timeslot length $S = \frac{TMIN}{\delta}$. However, we realize that the T_i of the heavy tasks could be excluded. Thus the S could be potentially larger, if T_i of these tasks were the smaller. Note that, larger timeslot imply fewer preemptions and each task executes more time in each timeslot, consequently, the performance increases.

5.3 Timing Behavior

The *SMS* module is not provided by any additional timer interrupt mechanism, thus the SMLS is a tick-based scheduler. So, since the `HZ` macro on our system is set to 1000, the frequency of timer interrupt is approximately of 1 ms, more precisely $999,848$ ns. This timer interval is called a *tick*. This means that the granularity of our system is 1 ms. We did not experiment higher timer frequencies, but according to [5] experimentation with higher timer frequencies resulted in an unstable system.

To get a better grasp of the SMLS timing behavior, Fig. 6 shows a timeline execution of task of processor P_2 according to the task set presented in Section 3.2, in which we now assume the time unit in milliseconds. As one can see, the timing behavior of the SMS algorithm and the SMLS are different. The timeslot length is equal to 2.5 ms and the tick occurs every 1 ms, therefore the timeslot length is not a multiple of the tick. This could lead to undesirable behavior of the scheduler. As one can see in the Fig. 6, tasks do not execute the same amount of time in every timeslot. On the other hand, the context switch does not occur at the correct time on the SMLS.

To solve the identified problem we need a mechanism by which timer interrupts are allowed to occur with nanosecond precision, which we think that is the ideal granularity for SMLS, but not necessarily on every nanosecond. For that, it is our intention to adapt the dynticks mechanism [13, 14] to the SMLS. In the current Linux version, the dynticks mechanism eliminates the ticks when the processor is idle. Instead, we will try to change the tick-based scheduler for event-based scheduler, in such manner that the events occurred according to the time events of the SMS algorithm.



a) SMS algorithm b)SMLS

Fig. 6: Timing behavior of the SMS algorithm and of the SMLS

Another, problem with *SMS* module implementation concerns the processors. According to the SMS algorithm the processors must be aligned, but we realize that they present some drift in terms of time.

6 Conclusion and Future Work

Nowadays, real-time scheduling for uniprocessor is considered mature, but real-time scheduling theory for multiprocessors has been triggered by the advent of the multicore systems. Multiprocessor systems are much more complex than the uniprocessor systems. Usually, real-time scheduling analysis is based on a set of assumptions that in a real implementation are not possible. For instance, they do not consider some source of overheads like the context switch and the execution of the scheduler, just to mention some. So, we have decided to study emerging real-time scheduling theory for multiprocessors with practice, that is, using real implementations of the algorithms. Motivated by this idea, we have implemented a Sporadic Multiprocessor Linux Scheduler (SMLS) based on a recently published real-time scheduling algorithm, the Sporadic Multiprocessor Scheduling (SMS) algorithm. The SMS algorithm schedules sporadic tasks for multiprocessor systems in order to met their deadlines. This algorithm tries to clamp down on the number of preemptions (which involve operating system overheads) and can be configured to achieve different levels of utilization bound and migration costs.

This first implementation gives us a better understanding of the algorithm as well as the platform used to (Linux 2.6.24 kernel version), in a such way that we have proposed a set of improvements to the algorithm in order to get better performance and robustness. Further, we are now aware of the specific features and details related to the implementation of the algorithm in this specific platform.

Future work will include investigations into timing behavior of the SMLS to get near of the SMS algorithm. The source of overheads will be explored in order to reduce some latencies. Finally, performance comparisons will be done with other algorithms to evaluate the performance of the SMS algorithm and also of the SMLS.

7 Acknowledgments

We would like to thank Björn Andersson and Konstantinos Bletsas for useful discussions and also to thank Paulo Santos Matos for his help in LaTeX.

References

1. J. Held, J. Bautista, and S. Koehl. From a few cores to many: A tera-scale computing research overview. White paper, Intel Corporation, 2006.
2. C. L. Liu. Scheduling algorithms for hard-real-time multiprogramming of a single processor. *JPL Space Programs Summary*, II(1):37–60, 1969.
3. C. L. Liu and J. W. Layland. Scheduling algorithms for multiprogramming in a hard-real-time environment. *J. ACM*, 20(1):46–61, 1973.
4. B. Andersson and K. Bletsas. Sporadic multiprocessor scheduling with few preemptions. In *ECRTS '08: Proceedings of the 2008 Euromicro Conference on Real-Time Systems*, pages 243–252, Washington, DC, USA, 2008. IEEE Computer Society.
5. B. Brandenburg, A. Block, J. Calandrino, U. Devi, H. Leontyev, and J. H. Anderson. LITMUS$^{\mathrm{RT}}$: A status reports. In *Proceedings of the 9th Real-Time Linux Workshop*, pages 107–123. Real-Time Linux Foundation, 2007.
6. J. A. Stankovic. Misconceptions about real-time computing: A serious problem for next-generation systems. *Computer*, 21(10):10–19, 1988.
7. D. Geer. Industry trends: Chip makers turn to multicore processors. *Computer*, 38(5):11–13, 2005.
8. J. H. Anderson and A. Srinivasan. Mixed pfair/erfair scheduling of asynchronous periodic tasks. *Journal Computer and System Sciences*, 68(1):157–204, 2004.
9. A. Kumar. Multiprocessing with the completely fair scheduler. Technical report, IBM, 2008.
10. Jr. E. G. Coffman, M. R. Garey, and D. S. Johnson. Approximation algorithms for bin packing: a survey. In *Approximation algorithms for NP-hard problems*, pages 46–93. PWS Publishing Co., Boston, MA, USA, 1997.
11. P. B. Sousa. Sporadic multiprocessor linux scheduler. Hurray-tr-090102, IPP-HURRAY!, Polytechnic Institute of Porto, 2009.
12. D. Bovet and M. Cesati. *Understanding The Linux Kernel*. O Reilly & Associates Inc, 2005.
13. S. Siddha, V. Pallipadi, and A. Van De Ven. Getting maximum mileage out of tickless. In *Proceedings of the Linux Symposium*, pages 201–207, Ottawa, ON, Canada, 2007. Intel Open Source Technology Center, Linux Symposium.
14. V. Srinivasan, G. R. Shenoy, S. Vaddagiri, D. Sarma, and V. Pallipadi. Energy-aware task and interrupt management in linux. In *Proceedings of the Linux Symposium*, pages 187–197, Ottawa, ON, Canada, 2008. Linux Symposium.

Design and Validation of Real-Time Applications

Carlos J. A. Costa

Centro de Estudos em Educação, Tecnologias e Saúde, ESTGL, Instituto Politécnico de Viseu, Av. Visconde Guedes Teixeira, 5100-074 Lamego, Portugal
ccosta@estgl.ipv.pt

Abstract. Real-time scheduling models tend to ignore the impact of system tasks (supporting platform overhead) upon the timeliness of application tasks. However, this system overload is not negligible. Accurate scheduling models from where control engineers can confidently design their real-time applications are thus required. In this paper such models are proposed by augmenting static scheduling and fixed priority models. This is achieved by inclusion on models with time constraints imposed by the system tasks to run application tasks in real-time kernel schedulers. Other contribution is presenting a set of procedures from where time parameters can be driven. These procedures are based on a set of real-time benchmarks. This solution is validated experimentally doing a comparative between the previsions of system performance computed, and results obtained by an experimental system developed. We achieved make accurate prevision of system performance.

Keywords: Computer-aided Control System Design, Real-Time Systems, Scheduling Algorithms, Architectures, Real-Time Benchmarking, Overload, Overhead Estimation.

1 Introduction

Real-time systems must perform correctly in both the value and the time domain [1]. Computer controllers are probably the most known real-time systems. This is because a controller must conform to some time constraints in sensing the environment and performing control actions accordingly [2]. Such time constrains are usually provided in the form of deadlines [3].

Real-time scheduling aims to devise conditions for guarantying processes deadlines in concurrent computer control applications. Current literature is rich in techniques and algorithms for real-time scheduling [4], [5], [6], and [7]. Yet, most of the authors depart from load models that only cover application processes characteristics – e.g. maximum execution time, minimum time interval between execution requests, and deadlines. Furthermore, until recent research, they used methods based on these same tasks characteristics, to address performance issues with exact response time analysis (RTA) for fixed priority preemptive systems and achieve schedulability [8], or try to improve the schedulability bound, for instance rate monotonic bound, like consider the relative period ratios in a system, by reducing the difference between the smallest and largest virtual period values [9]. The

computational overhead introduced by the operating system for interrupt handling, task scheduling management and context switching are usually not taken into account. As a consequence, a real-time schedule declared feasible according to classical or improved models may fail in practice to guarantee processes deadlines. This is particularly true when high processor utilization is achieved and deadlines tend to be met by a small margin, if any.

The motivation for this paper is thus the notion that the mathematical frameworks provided by scheduling algorithms do not perfectly handle the complexity of real-time applications. Therefore, its first aim is to use augment scheduling models from where control engineers can confidently analyze a schedule and deduce if it is feasible or not. But since it is impossible to cope in here all the real-time scheduling algorithms, this paper is restricted to static and fixed priority preemptive scheduling schemes implemented on general single processor systems. These are the most common scheduling schemes. This study is based on a research [10] which presents useful conclusions under cyclic scheduling implementations (e.g., round robin scheduling scheme), and complements some of the results obtained here.

The measurement of the impact of the system software execution upon the overall system performance requires the quantification of a few timing parameters. Some operating system suppliers provide the timing values required for such quantification. However, the conditions under which such parameters were defined are not always revealed or are different from those of interest. Sets of experimental procedures that can bridge this gap are presented. They result from an eclectic view on several fine-grained [11], [12], [13], [14], [15] and application oriented [16], [17], [18], [19] real-time benchmarks that can be easily tailored to any real-time single processor system. There is another research for implementation of real-time solutions [20], also based on the quantification of timing parameters of supporting platform. However, it relies on special hardware that increases performance by reducing the overhead of supporting system. This reduction is achieved by transferring to hardware some functionality or using some tricks. This was not the approach followed in this study.

The remainder of the paper is organized as follows. Section 2 presents augmented real-time scheduling models through the inclusion of the overhead introduced by the supporting software system. Preemptive and non-preemptive, as well as static (cyclical) and dynamic fixed-priority scheduling strategies are covered. Section 3 provides a set of experimental procedures that enable the quantification of the overheads discussed in the previous section. Section 4 investigates how the devised scheduling models differ from classical models in a quantitative form. The results show that the computational overhead ignored in classical models can easily make real-time systems to miss their deadlines when high processor utilization and fast response is a major concern, as it is usually the case. Section 5 ends the paper summarizing the most important conclusions.

2 Augmented Scheduling Models

This section presents augmented scheduling models that express the impact of kernel tasks upon the timeliness of application tasks. Two kernel costs are considered:

overhead and *blocking*. *Overhead* is the time taken by the kernel in performing a service on behalf of a specific task – such as invoking, resuming or terminating it. *Blocking* is the amount of time that a task is prevented from running due to the execution of a kernel operation that cannot be avoided – e.g. handling an interruption that does not cause any task switching.

Although two scheduling approaches – *event-driven* and *time-driven* – are founded in most real-time kernels here it is considered only *time-driven*.

In *time-driven* implementations a cyclic timer interrupt activates the scheduler to perform its duty.

As usual, it is assumed that the system kernel includes three data structures that store the information required for scheduling support. The *run queue* stores task control blocks (TCB) of tasks ready to run. The *start queue* includes TCB of (periodic) tasks that already executed for completion in the present period and are waiting for their next period to start again. The *active queue* stores the TCB of the task that is currently running on the single available processor. It is assumed that the *start queue* is ordered by task's release time and the *run queue* is ordered by task's priorities. It is also assumed that:

- Every time an interrupt occurs, the handler saves same registers, performs some operations and then calls the scheduler. Later, if the scheduler doesn't preempt the active task, the saved registers are restored and the active task is resumed. If a task switch is performed, the TCB of the active task is stored in the *run queue* according to priority. The TCB in head of the *run queue* is passed to the *active queue*, and the new active task executes.
- When a task ends execution, it traps the scheduler. The trap handling routine selects the new active task from the head of the *run queue*.
- The TCB of a periodic task is stored on the *start queue* on a *time-driven* basis. This is performed by a mechanism internal or external to the kernel. TCBs are passed from the start to the *run queue* by kernel or the interrupt handling routine.

From these assumptions, the following atomic parameters define the *overheads* and the *blocking* times that a predictable scheduling model must include. These apply to non-preemptive and preemptive scheduling:

- C_{int} – the time to handle an interrupt request.
- C_{sched} – the time to execute the scheduling code to determine the next task to run. It includes comparing the head of the *run queue* with the active task and moving TCB from the *start* to the *run queue*.
- C_{resume} – the time to resume the active task when task switch does not occur. It includes the time to restore the registers saved by the *interrupt handler*.
- C_{store} – the time to save the active task to a TCB, and sort it into the *run queue*.
- C_{load} – the time to load the new active task state from the *run queue*.
- C_{trap} – the time to handle the trap generated by normal completion of a task. This includes, storing the TCB of the completing (periodic) task into the *start queue* and selecting the head of the *run queue* to be the next active task.

2.1 Static scheduling

We begin by the simplest case – static scheduling. In practice is assumed it states to a model that includes *n* independent tasks. Each task, τ_i, is characterised by its period, T_i, and worst-case execution time (if never interrupted), C_i. The dispatcher is based on an interrupt driven real-time clock and must follow a task execution order established off-line. No idle time exists between the executions of two consecutive tasks – Fig. 1.

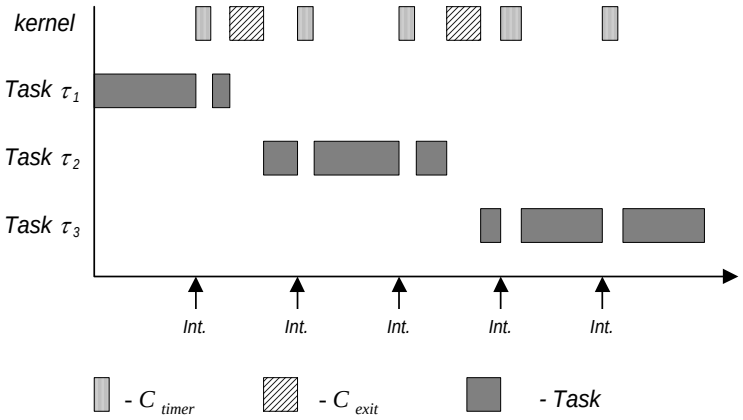


Fig. 1. Static scheduling

In this case C_{exit} (the *overhead* introduced by the completion of a task) is given by:

$$C_{exit} = C_{trap} + C_{load} \ . \tag{1}$$

A blocking time also exists due to periodic timer interrupts. In the worst case such blocking takes the value:

$$C_{timer} = C_{int} + C_{sched} + C_{resume} \ . \tag{2}$$

The total blocking time for a task τ_i depends on the number of interrupts occurred during its execution. Denoting the time between two consecutive interrupts by T_{tic}, one finds that the total blocking time of a task τ_i takes the form:

$$\left\lceil \frac{C_i}{T_{tic}} \right\rceil \times C_{timer} \ . \tag{3}$$

When *overhead* and *blocking* terms are ignored, the cyclic execution of the *n* tasks requires a time interval given by:

$$C_{total} = \sum_{i=1}^{n} C_i \ . \tag{4}$$

When *overheads* and the *blocking* terms are considered this time interval reverts to:

$$C'_{total} = \sum_{i=1}^{n} \left(C_i + \left\lceil \frac{C_i}{T_{tic}} \right\rceil \times C_{timer} \right) + nC_{exit} \quad . \tag{5}$$

2.2 Fixed Priority Scheduling

Preemptive fixed-priority scheduling has gained an increasing practical importance since the publication of Liu and Layland's work [21]. They showed that, for synchronous tasks (that share a common release time) that comply with a restrictive system model and that have deadlines equal to their periods $D_i=T_i$, then rate monotonic priority ordering (RMPO) is optimal. RMPO assigns priorities in order of task periods such that the task with the shortest period is given the highest priority. Yet, only in 1989 Lehoczky *et al* found an analytical condition from where the feasibility of a fixed-priority schedule can be assessed [22].

Whist the work developed by Lehoczky *et al* is very important, it is not complete in the sense that it ignores the impact of the time required to perform system tasks. And there are reasons to believe that such *overhead* in not negligible, since interrupt handling, task switching and preemption are vital to fixed priority scheduling and may occur frequently.

Two implementations are possible for a fixed priority scheduler [20]: *event-driven* and *time-driven*. Both scenarios are deeply analyzed in [23].

Time-driven scheduling relies on a timer that periodically interrupts active application task and invokes the scheduler – Fig 2.

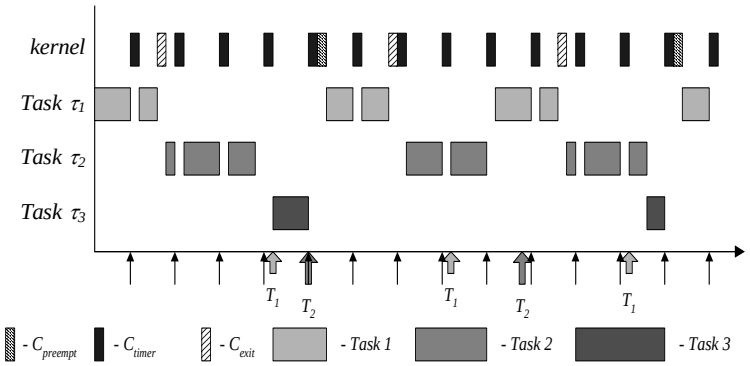


Fig. 2. Fixed-Priority Time-Driven Scheduling.

In this case the $C_{preempt}$ parameter (the *overhead* by low priority task preemption) is given by:

$$C_{preempt} = C_{store} + C_{load} \quad . \tag{6}$$

In this case Costa's work [23] proves that the following expression holds:

$$W_i(t) = \sum_{j=1}^{i} \left(C_j + C_{preempt}\right) \times \left\lceil \frac{t}{T_j} \right\rceil + \left\lceil \frac{t}{T_{tic}} \right\rceil \times C_{timer} + T_{tic} + \sum_{j=1}^{i-1} \left\lfloor \frac{t}{T_j} \right\rfloor \times C_{exit} \quad (7)$$

The blocking time C_{timer} is the same given by expression (2). Since the scheduler is invoked every timer interrupt, a task of high priority can be blocked by a task of low priority until the next interruption. Thus, is considered a blocking factor T_{tic}, the time between two consecutive timer interrupts witch guarantees worst case blocking time.

As it can be easily noticed, if both *overhead* and *blocking* conditions are ignored, expressions (7) revert to the classical workload model stated in [22]:

$$W_i(t) = \sum_{j=1}^{i} \left(C_j\right) \times \left\lceil \frac{t}{T_j} \right\rceil \quad (8)$$

3 Experimental Quantification

The quantification of the parameters introduced in the last section is vital to the exact characterisation of a scheduling model. Therefore, it is very convenient to define a set of procedures from where *overhead* and *blocking* times can be driven. These procedures – based on a set of *fine-grained* and application oriented real-time benchmarks – are briefly justified in this section. The illustrating values are for a general hardware, without any evolutions to beneficiates best effort performance, like *caches* (an 80386 @25MHz based system), running a freeware kernel: *Ctask* [24]. To carry out the time measures, it was used an I/O card – the PCL-711b-Multilab PC Card – which makes the measure process independent from test bed system and accurate. More details on the procedures can be found in [23], where pseudo-code is also provided.

3.1 Non-periodic parameters

The quantification of the C_{exit} parameter is somehow trivial. It suffices to consider n non-preemptive empty tasks and measure the time interval between the moment that the first task starts and the last task ends execution. Such time, divided by the number of tasks minus one, gives us C_{exit}. In our experiences we found $C_{exit} \approx 185\ \mu s$.

The second parameter to determine is C_{timer}. Here, we must consider a task whose nominal execution time is well known and greater than the T_{tic}. When the task executes under the kernel control, it is periodically interrupted – Fig. 3. Thus, the difference between the execution time of the task under the kernel control and its nominal execution time divided by the number of interrupts occurred during task execution provides C_{timer}.

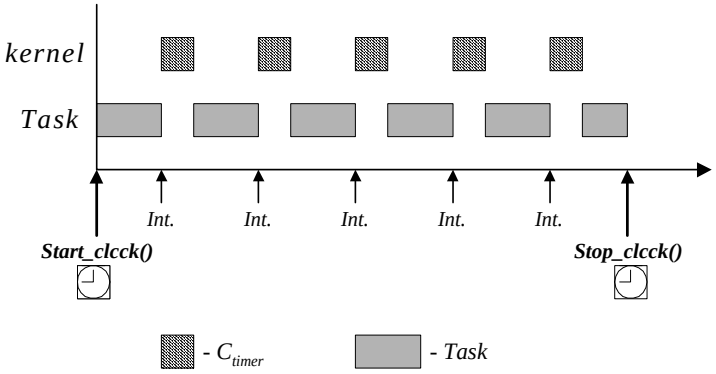


Fig. 3. Measuring C_{timer}

The number of interrupts occurred during task execution is given by:

$$\text{Number of Interrupts} = \left\lceil \frac{\text{Nominal Execution Time}}{T_{tic}} \right\rceil . \quad (9)$$

In our experiences we found $C_{timer} \approx 136\ \mu s$.

3.2 Periodic parameters

To obtain $C_{preempt}$ it suffices to consider two tasks with different priorities. Both execute for a sufficiently long time in a loop. The low priority task also includes an inner loop whose execution time is nearly one timer tick long, and where it waits preemption. The high priority task suspends by a timer tick, allowing the kernel to execute the low priority task. When the suspension period ends, the high priority task preempts the second task. Once again, tasks are firstly executed without kernel control. Then, they run under the control of the kernel. $C_{preempt}$ is obtained by the difference between the two execution times divided by the number of iterations performed. In our experiences we found $C_{preempt} \approx 280\ \mu s$.

The last parameter to define in the scope of fixed priority scheduling is C_{exit}, This parameter is greater than or equal to the C_{exit} considered on expression (1). In fixed priority scheduling, there is an extra time to order insertion of the periodic task's TCB into the *start queue*. In this case, we consider two tasks whose priorities were defined according to the rate monotonic scheduling algorithm. The higher priority task executes for completion and reads the elapsing time before completing and low priority task reads the elapsed time and executes. In this scenario, C_{exit} is the difference between these both times. In our experiences it was found $C_{exit} \approx 300\ \mu s$.

3.3 Experimental System

It is worth stating that a considerable set of experiences for both static and fixed priority models was performed in order to validate the proposed models. Such experiences departed from *application workloads* similar to those considered for some case studies presented in real-time literature. The resulting schedules were first evaluated according to the scheduling models presented above. Later, the task sets were executed on the experimental system, and the most relevant parameters were measured. In both cases – static and fixed priority scheduling – expected and measured values were found to closely match.

4 Overload analyses

The *overhead* and *blocking* terms can have a considerable impact upon the application tasks. This is because system services are performed very often, or can take a considerable time comparably to task execution times. Moreover, since each *overhead* and *blocking* term wields a particular influence upon the execution of application tasks, some workloads and scheduling schemes are expected to be more susceptible to a given parameter than to another. We make a comparative between the prevision of system performance from the application of the proposed models and the results obtained experimentally. The present section discusses this subject.

We first consider static scheduling. In this case, and according to expression (5), application tasks can have their execution delayed due to both C_{exit} and C_{timer}.

Consider a system that executes an 8-Task set which workload is defined by expression (4). Impact of system tasks is given by difference between (5) and (4). Has it can be seen at Table 1, the accuracy of prevision depends on greatness relationship between tasks execution times and kernel overhead.

Table 1. System Performance at (1/55*ms*) Interrupts Frequency

System Workload	Expected Performance	Experimental Results	Relative Difference
1,554.9	1,560.4	1,573.9	-0.85%
3,325.0	3,335.2	3,345.4	-0.30%
6,030.0	6,046.6	6,051.8	-0.09%
11,723.7	11,754.3	11,765.8	-0.10%
45,728.4	45,843.5	45,813.8	0.06%
91,204.7	91,432.5	91,372.1	0.07%
157,379.4	157,771.0	157,678.1	0.06%
180,049.1	180,496.9	180,397.2	0.06%
182,161.1	182,614.1	182,491.0	0.07%
364,114.5	365,018.2	364,789.7	0.06%

The impact of C_{timer} upon the execution of application tasks can be easily testified. It suffices to trace the execution time of a given and well-known workload for different timer interrupt frequencies. The experience result is provided in Fig. 4. As

suspected, the overhead increases as the timer period decreases. This means that when time resolution is a major concern, a considerable overhead is inevitable. The same conclusion was found in [10], under round robin scheduling implementation.

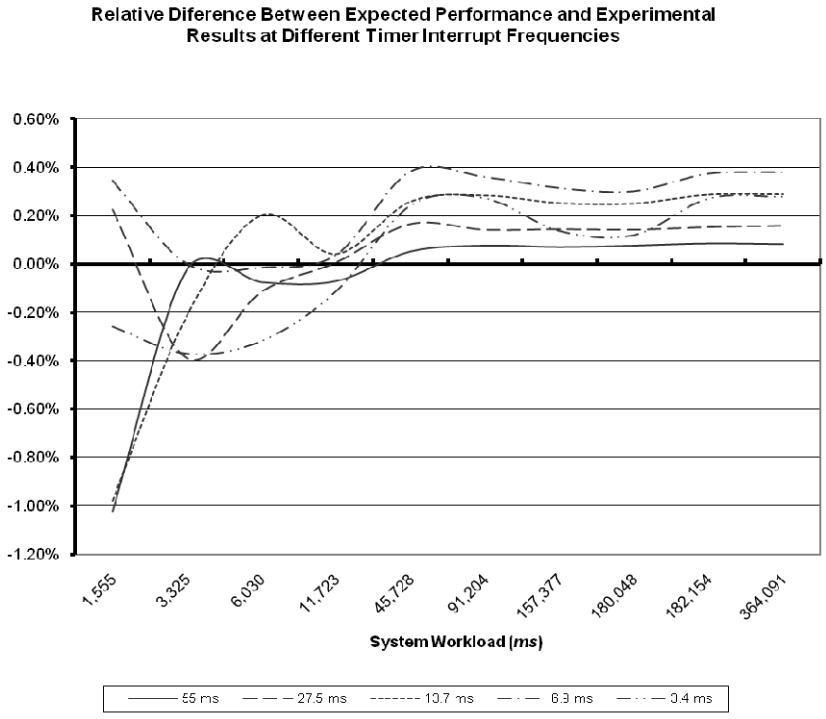


Fig. 4. *Expected Performance vs. Experimental Results at Different Timer Resolutions*

Let now see the case of *rate monotonic policy*, implemented upon a *time-driven* fixed priority preemptive scheduler. It is consider the following 3-Task set – Table 2.

Table 2. Characteristics for a 3-Task set

Task	Execution Time (ms)	Period (ms)	U_i
1	201.6	549.5	0.367
2	234.5	879.1	0.267
3	549.0	2,087.9	0.263
		U=	0.897

This task set is schedulable by condition (8), because when $i = 3$ and $t = 1650$ *ms*, $W_3(1650) = 1622.9 < 1650$ *ms*. If it is considered the system overhead, it continues to be schedulable by expression (7) – $W_3(1650) = 1639.1 < 1650$ *ms*.

When the blocking factor T_{tic} is considered, the system no longer is schedulable, in terms of Expected Performance. W_3 (1650) = 1694 > 1650. It is considered, that this system may not be schedulable. If, a low priority task blocks a high priority task until next interrupt timer and low priority task executes near last interrupt, this Task set it is no more schedulable. However, experimentally system always met all deadlines. Fig. 5 presents the quality of our performance prevision.

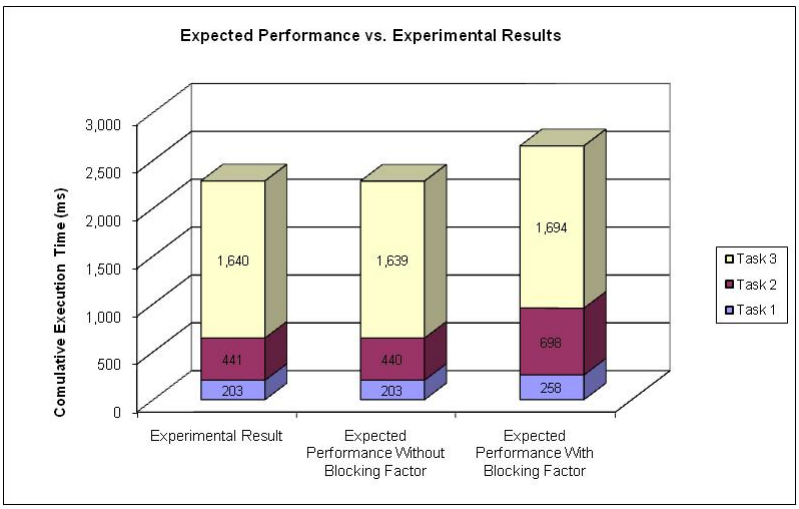


Fig. 5. Comparative Expected Performance (with and without blocking factor) vs. Experimental Results

Now let us see what happens, if we increase the Liu and Layland [21] upper bound over 90%. This is achieved by increasing the workload of first task to 209.4 *ms* (U_1 = 0.381). All other parameters of 3-Task set in Table 2, remain unchanged, and U = 0.911. The high priority task increases 7.8 *ms* and makes the timeliness of high priority task tighter.

From expression (8) the set is schedulable, because also here, $W_3(1650) = 1646.2 < 1650$ *ms*. As we can see in Table 3, the set is no more schedulable, when the system overhead is considered, because the time we expect to end last task is 2 110.5, which is greater than period (2 087.9). This is confirmed experimentally.

Table 3. Expected Performance and Experimental Results for a Non-Schedulable 3-Task set.

Task	Experimental Result (ms)	Expected Performance Without Blocking Factor		Expected Performance With Blocking Factor	
		Execution Time (ms)	Relative Difference	Execution Time (ms)	Relative Difference
1	211.0	211.2	0.05%	266.1	26.07%
2	448.6	447.6	-0.22%	713.8	59.11%
3	2,113.7	2,110.5	-0.15%	2,165.4	2.45%

The notion that system overload can take a considerable magnitude in time-driven fixed priority scheduling is an important conclusion. Another important conclusion is that C_{exit} has a considerable impact upon fixed priority application tasks that execute for a short time [10], since C_{exit} is greater in fixed priority scheduling than in static scheduling. Therefore, it is not surprising to find an overhead greater than 15% in a fixed priority schedule.

It is worth noting than in fixed priority scheduling – and according to the rate monotonic algorithm – high priority tasks tend to have a short execution time and are thus very prone to system overhead. Yet, high priority application tasks tend to meet their deadlines by a comfortable margin (see execution times, booth expected and experimental devised on Table 3). As a consequence, system overhead does not usually make a high priority application task to miss its deadline.

However, the system overhead that impacts the execution of high priority application tasks also makes low priority application tasks to delay their executions. Therefore low priority tasks are also very prone to system overhead. Moreover, since low priority tasks tend to meet their deadlines by a short margin, if any, one finds that these tasks are much more prone to system overhead than high priority tasks. A set of experiments that illustrate this important conclusion can be found in [23].

5 Conclusions

The paper has shown that the overhead introduced by system tasks upon the timeliness of application tasks is not negligible even for very simple real-time systems. Therefore, augment scheduling models from where control engineers can confidently analyse a real-time schedule are required. These models were presented in this paper for the most important single processor scheduling schemes used in practice. Techniques for quantifying the overhead of system tasks upon the application tasks were also presented. Consequently, the author of the paper feels that it has contributed to improve the practice of real-time systems in control applications. Also felt is that a similar research must be carried out in the context of actual computer systems, eventuality distributed, and over TCP/IP networks. This is because these computer systems are general, present in most of environments, and optimized to best effort case. Thus, *caches* and TCP/IP suite introduce a larger set of system parameters – *overheads* and *blocking* conditions upon the application tasks – that must be maximized. More, dynamic preemptive scheduling over multicomputer concurrent system is a major problem and must be addressed.

References

1. Stankovic, J. and Ramamritham, K.: Advances in Real-Time Systems. IEEE Computer Society Press (1993).
2. Middleton, R. and Goodwin, G.: Digital Control and Estimation. Prentice-Hall Int. (1990).
3. Magalhães, A. P.: A Survey on Estimating the Timing Constraints of Hard Real-time Systems. Design Automation for Embedded Systems, vol. 1, no. 3, pp. 213-231, (1996)

4. Stankovic, J., Spuri, M., Di Natale, M. and Buttazzo, G. C.: Implications of Classical Scheduling Results for Real-Time Systems, IEEE Computer (1995)
5. Ghosh, K., Mukherjee, B. and Schwan, K.: A Survey of Real-Time Operating Systems – Draft, Technical report, GIT-CC-93/18, College of Computing, Georgia Institute of Technology, Atlanta, Georgia, February (1994)
6. Ramamritham, K. and Stankovic, J.: Scheduling Algorithms and Operating Systems support for Real-Time Systems. Proceedings of the IEEE, vol. 82, no. 1, January (1994)
7. Burns, A.: Scheduling hard real-time systems: a review. Software Engineering Journal, pp. 116-128, May (1991)
8. Davis, R., Zabos, A., and Burns, A.: Efficient exact schedulability tests for fixed priority real-time systems. IEEE Transactions on Computers, vol. 57, no. 9, pp. 1261-1276, Setember (2008)
9. Lu, W., Wei, H, and Lin, K.: Rate monotonic schedulability conditions using relative period ratios. Proc. of 12th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, RTCSA 2006, pp. 3-9, IEEE RTCSA 2006 (2006)
10. Costa, C. J. and Magalhães, A. J.: Experimental Deduction of Real-Time Scheduling Models. In: Conference Proceedings, Controlo'2000: 4th Portuguese Conference on Automatic Control, pp. 191-196, APCA, Portugal (2000)
11. Kar, P. and Porter, K.: RHEALSTONE A Real-Time Benchmarking Proposal. Dr. Dobb's Journal, pp.14-24, February (1989)
12. Kar, P.: Implementing the Rhealstone Real-Time Benchmark. Dr. Dobb's Journal, pp.46-55, April (1990)
13. McRae, E.: Benchmarking Real-Time Operating Systems. Dr. Dobb's Journal, pp.48-58, May (1996)
14. Martínez, A. G., Conde, J. F. and Viña, A.: A Comprehensive Approach in Performance Evaluation for Modern Real-Time Operating Systems. In: Proceedings of EUROMICRO-22, pp. 61-68, IEEE (1996)
15. Sacha, K. M.: Measuring the Real-Time Operating System Performance. In: Proceedings of the 2th EUROMICRO Workshop on Real-Time Systems, IEEE (1995)
16. Curnow, H. J. and Witchmann, B. A.: A Synthetic Benchmark. The Computer Journal 19(1), pp. 43-49, February (1976)
17. Weicker, R. P.: DHRYSTONE: A Synthetic Systems Programming Benchmark. Communications of the ACM, vol. 27, no. 10, October (1984)
18. Donohoe, P.: A Survey of Real-Time Performance Benchmarks for the Ada Programming Language. Technical Report, CMU/SEI-87-TR-28, ESD-TR-87-191, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, December (1987)
19. Kamenoff, N. and Weiderman, N.: Hartstone Distributed Benchmark: Requirements and Definitions. In: Proceedings of the Real-Time Systems Symposium, pp. 199-208, IEEE Computer Society Press, December (1991)
20. Katcher, D. I., Arakawa, H. and Strosnider, J. K.: Engineering and Analysis of Fixed Priority Schedulers. IEEE Transactions on Software Engineering, vol. 19, no. 9, pp. 920-934, September (1993)
21. Liu, C. L. and Layland, J. W.: Scheduling Algorithms for Multiprogramming in Hard-Real-Time Environments. Journal of ACM, vol. 20, no. 1, pp. 46-61, January (1973)
22. Lehoczky, J., Sha, L. and Ding, Y.: The Rate Monotonic Scheduling Algorithm: Exact Characterization and Average Case Behavior. Proc. IEEE Real-Time Symp., CS Press, Los Alamitos, Calif., pp. 166-171 (1989)
23. Costa, C. J.: Análise e Concepção de Sistemas de Tempo-Real. Msc Dissertation, FEUP, October (1998) (in Portuguese)
24. CTask - A Multitasking Kernel for C, Version 2.2d, Released 93-06-08, http://www.ifi.uzh.ch/ailab/embedded/multitaskers/ctask.readme

Interoperable Geographic Information Services from Crisis Management Perspective

Marco Amaro Oliveira

INESC Porto
mao@inescporto.pt

Abstract. Nowadays, we face a considerable increase in the complexity of the living environment of the western world. This trend is particularly evident in the domain of critical infrastructures. One of its negative consequences is manifested by the fact that the society has become more vulnerable and thus the catastrophic potential that may arise from the failure of a critical infrastructure has been identified.
In this paper we propose a conceptual high-level architecture with focus on interoperable geographic information (GI) services from the crisis management perspective. Based on Open Geospatial Consortium standards and initiatives, we present the building blocks of the interoperable solution for supporting crisis management that is proposed as a result of the EU sponsored project MEDSI.
In particular we focus on the application and operation of several OGC standards, some adopted and some still under discussion, as well as their integration and cooperation within a single framework of several OGC standards.
We identified that such an architecture can be feasibly implemented to support a distributed and collaborative approach to Crisis Management, and that it has proved a strong asset concerning the aspect of solving interoperability issues that necessarily arise when using distributed heterogeneous systems and data sources of geographic information.

Key words: distributed geographic information systems, geographic information systems interoperability, geographic information, critical infrastructures

1 Introduction

Nowadays, we face a considerable increase in the complexity of the living environment of the western world. This trend is particularly evident in the domain of critical infrastructures. One of its negative consequences is manifested by the fact that the society has become more vulnerable and thus the catastrophic potential that may arise from the failure of a critical infrastructure has been identified. For that reason, advancing the field of crisis management for protecting critical infrastructures has been recognized as one of the top priorities in European countries [1], and several programs have been launched [2,3,4] both in Europe and the rest of the world.

One of the main concerns when responding to a crisis is how to support system's interoperability and how to integrate, in a meaningful way, information provided by various and heterogeneous sources and through different media. Thus, the crucial task is to define strategies to obtain timely and accurate geospatial information to quickly visualize and understand the context of emergency situations.

From the fact that each of these sources may be using different Geographic Information Systems (GIS) or Computer-aided Design (CAD) software, running on different operating system platforms and having it's own data storage formats, four main issues can be identified: data exchange, data access, data visualization and remote data processing. Data exchange issue concerns to the fact that implementing support for all the different formats is not feasible and for proprietary formats may not be legal, thus to tackle this issue a open and standardized format for storage and exchange of geographic information is required. Data access and remote data processing issues regard to system's access and exchange of data, to solve this issue it is required to have standardized interfaces for exchanging, querying and requesting processing tasks on remote data. Data visualization issue regards to the need of having a consistent and uniform visual representation of geographic information provided by remote and heterogeneous sources, thus a standardized format for allowing user-defined symbolization and coloring is required.

To effectively handle this issues, Open Geospatial Consortium [1] (OGC) established standards for Geographic Information (GI) sharing and processing. Among others, they promote standards like Web processing Service [5] (WPS) to define how a client can request the execution of a process, Sensor Observation Service [6] for retrieving real-time sensor data through the use of Sensor Model Language [7] (SensorML). Web Map Service [8] (WMS) to retrieve geographic information in image format and Web Feature Server [9] (WFS) to retrieve GI in vector format through the use of Geography Markup Language [10] (GML). Furthermore, OGC also proposes standards to allow the usage of user-defined symbolization and coloring, like Styled Layer Descriptor [11], and standards that allow to storage, in a portable and platform-independent format, the description of a map composed from requests made to multiple, and distributed, map servers - Web Map Context Documents (WMC) [12]. Although the adoption of the OGC standards by GIS software providers has been a slow process, these standards are widely accepted by the GI community and represent a firm foundation for constructing distributed GI-based software systems.

The incorporation of state of the art information technology advances in the field of supporting and enhancing decision-making capabilities in crisis management represents one of the key aspects of EU sponsored project MEDSI [13]. Since good technology is always built on the foundation of good technology, we made a commitment to apply and operationalize the standards and initiatives of OGC to propose a modern interoperable infrastructure for supporting crisis management within the project. By enforcing the OGC standards, MEDSI in-

[1] OGC website: `http://www.opengeospatial.org`

tends to exploit in particular the interoperability issues raised by the multiplicity of formats and heterogeneous ways to access various sorts of data.

To promote the protection of critical infrastructures as one of its key areas, OGC has established a specific line of action for the Critical Infrastructure Protection Initiative [14]. Adopting this experience, by incorporating many of the premises and concepts of this initiative, and by building upon their work, we expect to be able to contribute for accelerating the launch of the assets of the European management decision support for critical infrastructures.

The resulting framework will be tested on realistic user scenarios, like, for example, fire and explosion of hazardous materials in an industrial area, and river flooding. Those scenarios will allow to validate not only the suitability of the proposed conceptual architecture but also the application of the instantiated system in a real user environment.

Potential users of the proposed software framework are private and governmental organizations, including city management, regional management, central institutions and agencies, international crisis management organizations, public safety and security forces, and intelligence services.

In the paper we first present the proposed conceptual high-level architecture for supporting crisis management decisions. Then, we give an overview of technological platform stating to relevant functionality in supporting collaborative crisis management and symbology. We conclude by calling attention to the most important points presented in this paper.

2 Crisis Management Support

Given the broad scope of the concept "Critical Infrastructure" [15] and the wide variety of potential users for such a system (local, regional, national or trans-national level), it was obvious from the feasibility and scalability point of view that a distributed network of self-contained cells (named MEDSI Cells) would have to be put in place. Moreover, these cells, which can vary in size and geographic distribution, need to be self-sufficient in the protection of the infrastructure they are aimed to protect, while they also should be able to cooperate in case of broader emergency situations (Figure 1).

Also, the central role of GI together with the ability to swiftly exchange the most updated representations of geographical data (maps) among crisis management actors was a premise in the inception of the project. This assumption remains true not only for exchanges between MEDSI Cells, but also inside the scope of a single cell, instantiated in a crisis management center.

Originating from these premises, the use cases retrieved from the specific application scenarios and being aware of the interoperability issues that arise when trying to use together several heterogeneous GIS software and data sources, we have decided to enforce the use of OGC Web Services (OWS) for all sorts of GI access and exchange.

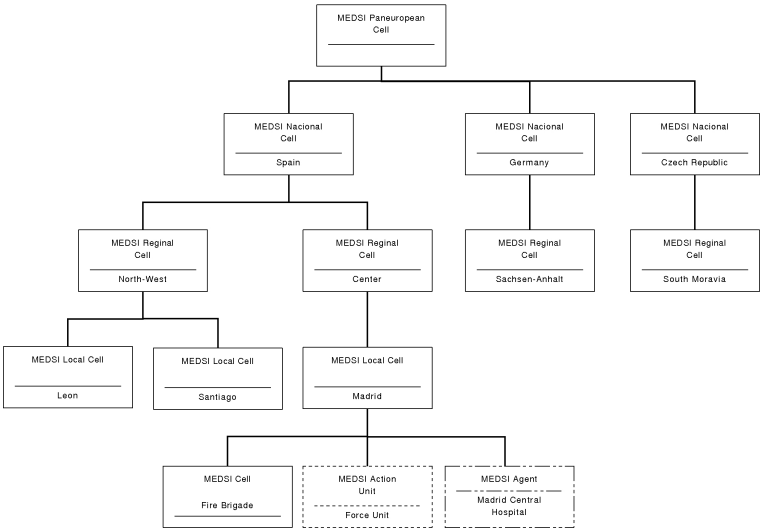


Fig. 1. Integrated network of cells.

Users of the system have the capability of finding the appropriate GI sources by means of a catalogue service where the Web Services providing the relevant GI have been previously registered.

It should be stressed that the services providing GI may be available not only inside the crisis center, where typically resides information like local aerial photos, streets, buildings, available resources, etc., but also from external providers like up to date satellite photos, sensor/weather data or even the most updated "map" of an endangered facility, made available upon request as can be inferred by the conceptual architecture presented in figure 2.

The need to request the execution of specific spatial information processing to legacy and/or complementary systems, such as the ones specific for calculating hazardous materials area of impact, was overcome with the use of XML Web Services. The use of the WPS was not taken into account due to the fact that the discussion paper for the definition of an implementation specification of a Web Processing Service was not available at the time, being issued afterwards.

Although this paper will not focus on the specific use cases , it's worth mentioning that the complete framework instantiation consists of a rich client application capable of providing support to domain-specific tasks such as risk analysis, crisis plans, standard operation procedures, etc. In other words, the system handles a set of specific domain objects, some of them comprising associated geometries, and it tries to do so maintaining the lowest possible coupling with the data-source level. For the GI realm, this means that the domain objects can be accessed through a WFS interface. As such, it is possible that another web service makes a "request" to the WFS for its schema in order to be able to inter-operate with it.

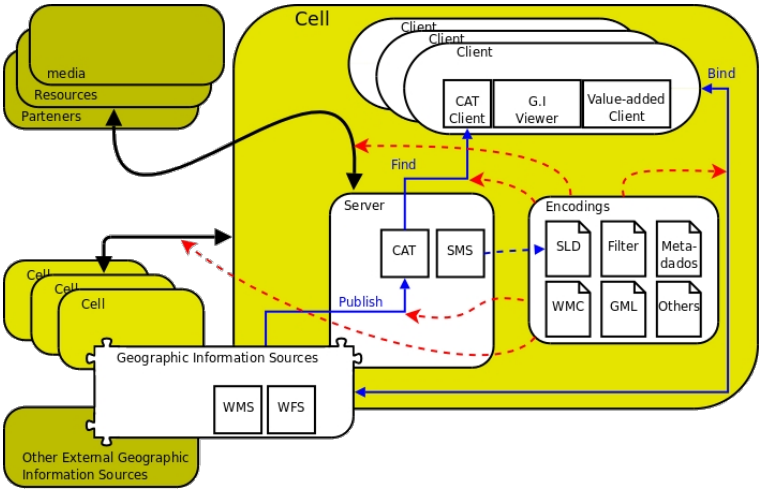


Fig. 2. Conceptual Architecture.

As outputs from the system, the most recent "situation maps" are produced to generate reports and updates to the units operating in the crisis scene, as well as for leading updated information to the several actor responsible for taking decisions in a crisis situation.

Moreover, it's worth mentioning that due to the ability of exchanging "Context Documents" amongst crisis management actors, these can interact in a much more efficient way as they are able to see exactly the same "map representation" (as if using paper based maps), but with the capability to obtain continuous updates and the possibility of continuing to browse afterwards like in any on-line map.

3 Conceptual Architecture Realization

The system comprises several modules supporting crisis management specific tasks ranging from Analysis and Planning to Simulation, Decision Support, Resources Management and also some horizontal types of functionalities like messaging and reporting.

These modules which implement the main specific functionalities will not be detailed under the scope of this paper, however they have the particularity of being connected to the GI infrastructure in one way or another. First because objects addressed in the crisis management domain correspond directly to a geographical feature. And second, because although sometimes no direct mapping is feasible, the geographical representation of involved areas provides a meaningful context to understand and respond to a crisis situation or help revealing a potential threat.

A special attention has also been paid to symbology, given its high expressiveness which helps the user to quickly absorb a significant level of information, provided that the used symbols are commonly understandable.

In this section we describe the technological platform being built for MEDSI project focusing mainly on the GI infrastructure.

As previously said, MEDSI has chosen to use in its prototype WMS as a portrayal service for displaying maps in image format and WFS for data services returning GML, which is then rendered in vector format. Because aerial photographs provide a high level of understandability even for users not familiarized with cartography, a Web coverage Service (WCS) [16] was also used in MEDSI prototype realization.

Although other services such as Web Terrain Server (WTS) [17] could also have been used for enhanced terrain visualization, the aforementioned ones have been deemed adequate for establishing a proof of concept.

As the first step in the prototype implementation, the consortium has configured several geographical data sources, in different technologies capable to output GI under the form of WMS and WFS. We have successfully configured and tested interoperable access to several platforms providing GI by means of web services. From then the consortium has been seamlessly using both open source geographic information sources such as Deegree [18], UMN map server [19] and Geoserver [20], but also popular proprietary solutions such as Geomedia Web Server [21] and ESRI ArcIMS [22] through their respective WMS and WFS connectors.

A catalogue service has also been deployed [23] and enhanced with the capability of classify these GI services according to a proposed ontology to facilitate finding of the appropriate data sources.

On client side, we used an Open Source GI viewer [24] able to view WMS which we extended to support WFS accesses as well as other functionalities required for proper integration with other implemented modules.

A catalogue browser is also used in the client for finding the needed data sources. Some editing and annotation functionalities, along with some basic spatial analysis have also been used to provide added value to MEDSI client.

Although the resulting prototype has proved itself as a reliable, highly customizable and interoperable framework for accessing distributed geographical information, other functionalities have been identified as a prerequisite for supporting a collaborative approach to decision support in crisis management.

One of such functionalities is the ability to store and further reproduce the status of a GI view, i.e. the set of map requests that originated that same view. Saving and reproducing map context information is essential to support any underlying workflow mechanism and also to create personal views for each user profile or for each type of crisis.

A context is some sort of "memento" for maps, comprising the description in a portable, platform-independent format of the grouping of one or more map requests from one or more map services for storage in a repository or for transmission between clients [12].

As such, a specific framework component for loading and saving WMC documents (XML) was developed, as well as a repository for storing and retrieving them which also provides an ad-hoc workflow mechanism allowing the exchange of contexts between users and/or groups of users.

Furthermore, a gazetteer service [25] can be used to find a geographic feature by its name (e.g. a street) returning a new geographical extent to update the present context.

Other important factor is the ability of crisis management actors within a cell to be able of visualizing geographic information from other cells, e.g. points of interest, according to their own used symbols, leading to the requirement of changing symbology in runtime.

A collaborative approach to crisis management requires establishing a common "language" for team communication. Symbology takes an essential role for quick visual identification of the most important spots within the crisis geographical extent. Within MEDSI project we have established a framework for the generation of symbols from a structured definition containing the symbol description.

Styled Layer Descriptor[11] is a language that can be used to customize the output of WMS and WFS on the client side, as it defines styles for presenting different map layers. A symbol for denoting a specific phenomenon on the map is first dynamically constructed from an icon selected from the symbology repository and augmented with dynamic information from the data base. Then, the necessary SLD file is built and placed in a location accessible from the web.

4 Results

MEDSI decided to follow the path of OGC Web Services for its own GI infrastructure, while aiming to support a distributed and collaborative approach to Crisis Management. This has proved a strong asset concerning the aspect of solving interoperability issues that necessarily arise when using distributed heterogeneous systems and data sources.

By enforcing the use of standards like WMS and WFS to help solving issues resulting from the definition of a distributed architecture, MEDSI has aligned itself with the European and International tendencies on building common Spatial Data Infrastructures, which can help solving relevant problems such as the ones involved in the protection of critical infrastructures.

The definition of a solid application layer for a specific domain, over a set of distributed geographic information sources in the form of web services has brought MEDSI before the need of complementing the standard interfaces for accessing data with many other abstractions and mechanisms to provide the necessary functionalities while keeping independence from the information sources.

An example of these abstractions was a kind of context information that could be used to save, exchange and restore the status of a composed GI view (map). The Web Map Context documents (a position paper by the time we

started) was identified as the vehicle to store and transport this information across the network, thus enabling collaboration.

Symbols can be seen as lingua franca for interpreting emergency maps, however different communities may use different symbols to represent the same concepts. Dynamic binding of Styled Layered Descriptions (SLD) to map services (WMS) has helped overcome this issue.

On the other hand, the low-coupling between maps services and symbols and the fact that they can be loaded in run-time has helped taking the dynamic aspects of symbols a step further into augmenting symbols with database information for better and faster map interpretation.

The conceptual architecture can now be used as a high-level, dynamic and upgradable architecture that can support logical construction methods in order to achieve a sustainable development strategy for Distributed Geographic Information Systems. Such an architecture will benefit of the inclusion of Web Processing and Sensor Observation Services support.

5 Acknowledgement

The work reported in this article is a result of the research project "MEDSI IST-2002-506991" funded by the Commission of European Communities. We are grateful to all the MEDSI Consortium members for their contributions to the project.

References

1. J. Solana. A european route to security. The International Herald Tribune, December 12 2003. Also published in other international newspapers.
2. Committee-of-the-Regions. Opinion of the committee of the regions on the communications from the commission to the council and the european parliament prevention, preparedness and response to terrorist attacks prevention of and the fight against terrorist financing through measures to improve the exchange of information, to strengthen transparency and enhance the traceability of financial transactions preparedness and consequence management in the fight against terrorism critical infrastructure protection in the fight against terrorism. Official Journal of the European Union, April 04 2006. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/c_081/c_08120060404en00010005.pdf.
3. European Commission. Green paper on a european programme for critical infrastructure protection. Online, November 17 2005. http://eur-lex.europa.eu/LexUriServ/site/en/com/2005/com2005_0576en01.pdf.
4. European Commission. Pilot project 2005-2006 european programme for critical infrastructure protection (epcip). Official Journal, January 21 2006. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/c_016/c_01620060121en00240024.pdf.
5. Open Geospatial Consortium. Web processing service specification, version 1.0. Available in World Wide Web, June, 08 2007. http://portal.opengeospatial.org/files/?artifact_id=24151.

6. Open Geospatial Consortium. Sensor observation service specification, version 1.0. Available in World Wide Web, October, 26 2007. http://portal.opengeospatial.org/files/?artifact_id=26667.
7. Open Geospatial Consortium. Sensor model language encoding specification, version 1.0. Available in World Wide Web, July, 17 2007. http://portal.opengeospatial.org/files/?artifact_id=21273.
8. Open Geospatial Consortium. Web map service specification, version 1.3. Available in World Wide Web, August 2004. http://portal.opengeospatial.org/files/?artifact_id=5316.
9. Open Geospatial Consortium. Web feature service specification, version 1.1. Available in World Wide Web, May, 03 2005. https://portal.opengeospatial.org/files/?artifact_id=8339.
10. Open Geospatial Consortium. Geography markup language, version 3.0. Available in World Wide Web, January, 29 2003. https://portal.opengeospatial.org/files/?artifact_id=7174.
11. Open Geospatial Consortium. Styled layer descriptor implementation specification, version 1.1. Available in World Wide Web, June, 29 2006. http://portal.opengeospatial.org/files/?artifact_id=22364.
12. Open Geospatial Consortium. Web map context documents, version 1.1. Available in World Wide Web, May, 03 2005. https://portal.opengeospatial.org/files/?artifact_id=8618.
13. Bojan Cestnik, Artur Rocha, and Martin Endig. Emergency response through collaborative crisis management. In *Proceedings of Eastern European eGov days 2005*, Budapest, Hungary, March 17-18 2005.
14. Open Geospatial Consortium. Critical infrastructure protection initiative 2. Available in World Wide Web, 2002. http://www.opengeospatial.org/projects/initiatives/cipi2.
15. European Commission. Communication from the commission to the council and the european parliament - critical infrastructure protection in the fight against terrorism. Online, October 20 2004. http://eur-lex.europa.eu/LexUriServ/site/en/com/2004/com2004_0702en01.pdf.
16. Open Geospatial Consortium. Web coverage service, version 1.0. Available in World Wide Web, October, 16 2003. https://portal.opengeospatial.org/files/?artifact_id=3837.
17. Open Geospatial Consortium. Web terrain server, version 0.3.2. Available in World Wide Web, August, 24 2001. http://portal.opengeospatial.org/files/?artifact_id=1072.
18. Deegree. Available in World Wide Web, 2005. http://deegree.sourceforge.net/.
19. University of Minnesota MapServer. Available in World Wide Web, 2005. http://mapserver.gis.umn.edu/.
20. The GeoServer Project. Available in World Wide Web, 2005. http://geoserver.sourceforge.net/.
21. Intergraph. Geomedia. Available in World Wide Web, 2005. http://www.intergraph.com/geomedia/.
22. ESRI. Gis standards and it interoperability. Available in World Wiede Web, 2005. http://www.esri.com/software/standards/.
23. Open Geospatial Consortium. Catalog service specification, version 2.0. Available in World Wide Web, August, 02 2004. http://portal.opengeospatial.org/files/?artifact_id=5929&version=1.

24. The JUMP Project. Available in World Wide Web, 2005. http://www.jump-project.org/.
25. Open Geospatial Consortium. Gazetteer service profile for a wfs, version 0.0.9. Available in World Wide Web, September, 03 2002. https://portal.opengeospatial.org/files/?artifact_id=7175.

Modelling the Job-Shop Scheduling problem in Linear Programming and Constraint Programming

Pedro Abreu

Faculdade de Engenharia da Universidade do Porto

Abstract. The Job-Shop Scheduling problem is a well-know combinatorial optimization problem. In this problem, we have to schedule several tasks minimizing the total time to process all the tasks. There are several constraints like the order of processing, resource capacity limitation and others. There are several algorithms to search the best solution. We will be looking here at approaches according to the Linear Programming and Constraint Programming paradigms. These algorithms search in a set of possible solution, within defined constraints, with the objective of minimize a function that gives a value to solutions. The solutions can be represented in several ways with different constraints and evaluation functions. In this work, we present different representation for the solution and the respective evaluation function and constraints for Linear Programming and Constraint Programming. We compare the differents methods applyed to different type of instances of the problem. The instances differ in the dimensions, the durations and the method to generate the instances. The results show that the Linear Programming Model with starting time for solution representation dominates in terms of the performance average the others in almost all instances and classes of the instances.

Key words: modelling, constraint programming, linear programming, job-shop, combinatorial optimization

1 Introduction

The Job-Shop Scheduling (JSS) problem is a well-know combinatorial optimization problem. In the JSS, the data of the problem are K tasks ($\{O_j m : j \in J, m \in M\}$) where $J = \{J_1, \ldots, J_n\}$ is a set of jobs and $M = \{M_1, \ldots, M_n\}$ is a set of machines. It is assumed that a task (O_{jm}) belongs to a job (J_j) and is processed in a specific machine (M_m). The tasks of the same job have an order to be processed. Each task has a processing duration. There are some restrictions such as: in each machine, can only be processed one task at each time and other restriction is that the processing of a task cannot be interrupted. The solution to the problem is the schedule of the beginning time of processing the tasks with the objective to minimize the value of the total time to process all the tasks (makespan). So, the solution with the less makespan is called the solution.

There are several algorithms to search optimal solution for optimization problems. The Linear Programming (LP) and Constraint Programming (CP) are two well-know optimization techniques. The two techniques have the objective to minimize/maximize a function subject to constraints that limit the function domain. The major differences between the two, is that LP is a technique only for optimize linear functions subject to linear constraints. The LP algorithm is called Mixed Integer Programming (MIP) if the domain is a set of integer variables. The CP algorithm is called Constraint Programming for Finite Domains (CLPFD) if the domain is a set of integer variables. So, having a optimization problem, we can define, for LP and CLP, the domain as the solutions for the problem. The constraints are the restrictions of the problem. For example, for the JSS problem, we can define as the domain the starting time of each task and the function to minimize is the total time to process all the tasks.

However, there are several possible representation for the solution of the problem. Above we give a possible representation of the domain, but we can also define the solution for the JSS, for example, as the order of processing the tasks [1]. So, a question can be ask: what algorithm or/and representation we should use for a specific problem of JSS? There are theorems such the No-Free Lunch theorems [10] and benchmark works [3] that shows there aren't any best algorithm for all classes of the problem. In this work, we present 4 methods: 2 different solution representation for the CP algorithm and 2 different solution representation for the LP algorithm. We compare the performance of the 4 methods in several different instances of JSS. The methods are:

- MIP algorithm using the domain as an integer variable for each task that is the starting time of the tasks processing;
- MIP algorithm using the domain as a set of binary variables. Each variable of the domain represents if a task begins in a instance of the time;
- CP using the domain as a integer variable for each task that is the starting time of the tasks processing.
- CP using the domain as an integer variable for each task that is the starting time of the tasks processing. The difference between this method and the previous are some programming issues. This method uses a built-in constraint provided by the software used, unlike in the previous one that uses a simple constraint.

In each method above, we have some characteristic in common thus we can group the methods in different classes. For example, the first, third and the fourth methods have the same representation of the solutions. The first and the second methods are techniques from LP.

These methods are applied to several different instances of JSS problem. The instances are generated randomly given the number of jobs, number of machines, the maximum value of the tasks duration and the method to generate the durations of the tasks. Given different values to the characteristic of the generation parameters, we can group the instances in classes. For example, an instance with 3 jobs and 4 machines and an instance with 3 jobs and 2 machines belong to the class of instances with 3 jobs. As we said above, on of the objectives of the

work it's to compare the performance of the methods proposed. So, the results of the experiments effectuated, we have not only compare the performance of the methods individually in all instances, but also we have analyze the performance of the classes of methods within the classes of instances describe above. We want to find out if there are some characteristic of the methods or/and of the instances that influences more the performance.

The Job-Shop Scheduling problem and details about the generation of instances are described in Section 2. The methods used in this work are describe in detail and explained in Section 3 the methods of LP and in Section 4 the methods of the CP. The experiments and results are in Section 5 that are discussed in Section 6. We present the conclusions about this work in Section 7.

2 Basic Job-Shop Scheduling

2.1 Description

The deterministic job-shop scheduling problem can be seen as the most general of the classical scheduling problems. Formally, this problem can be described as follows. A finite set J of n jobs $\{J_1, J_2, \ldots, J_n\}$ has to be processed on a finite set M of m machines $\{M_1, M_2, \ldots, M_m\}$. Each job J_i must be processed once on every machine M_j, so each job consists of a chain of m tasks. Let O_{ij} represent the operation of job J_i on machine M_j, and let p_{ij} be the processing time required by task O_{ij}.

The operations of each job J_i have to be scheduled in a predetermined given order, i.e. there are precedence constraints between the operations of each job J_i. Let $\prec$ be used to denote a precedence constraint, so that $O_{ik} \prec O_{il}$ means that job J_i has to be completely processed on machine M_k prior to being processed on machine M_l. Each job has its own flow pattern through the machines, so the precedence constraints between operations can be different for each job. Other additional constraints also have to be satisfied. Each machine can only process one job at a time (capacity constraints). Also, preemption is not allowed, so operations cannot be interrupted and must be fully processed once started. Let t_{ij} denote the starting time of operation O_{ij}. The objective is to determine starting times t_{ij} for all operations, in order to optimize some objective function, while satisfying the precedence, capacity and no-preemption constraints. The duration in which all operations for all jobs are completed is denoted as the makespan $C_{\max}$. In this paper, we consider as objective function the minimization of the makespan:

$$\begin{aligned} C^*_{\max} &= \min\left(C_{\max}\right) \\ &= \min_{\text{feasibleschedules}}\left(\max\left(t_{ij} + p_{ij}\right)\right), \\ &\quad \forall J_i \in J, M_j \in M. \end{aligned}$$

In following, we present an example of a JSS instance. The instance has 3 jobs and 3 machines. In Table 1, we describe the data of the instance that includes the processing duration and the order within the job of all the tasks. In the Table 1, we can obtain the information of the tasks, for example, in the first line

of the data we conclude that the tasks from the job 1 (first column) processed in machine 1 (second column) is processed in 1 unit of the time (third column) and is the second (last column) to be processed in the job that belongs.

Table 1. Information of a JSS instance

Job	Machine	Duration	Order
1	1	1	2
1	2	3	1
1	3	4	3
2	1	4	2
2	2	2	1
2	3	4	3
3	1	4	3
3	2	4	2
3	3	2	1

Above, we defined that the objective is to determine the starting time t_{ij} for all tasks, the set of all starting times is called a schedule. In Figure 1, we show two schedules of two possible solutions for the example above. The x-axis is represented the time and in y-axis the job identification. The color rectangles are the tasks and those tasks with the same color are processed in same machine with the identification number at the top left.

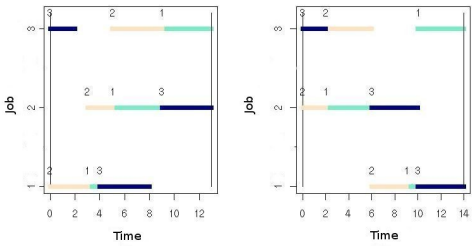


Fig. 1. Two possible schedules

However, it isn't mandatory that the search of the solution be in the starting time solution-space. The solution-space is the set of all solutions and starting time is a way to represent the solution for the problem. There are others possible representations. One example is a list of priority list between the tasks for each machine [1].

The job-shop scheduling problem is NP-hard [2, 6], and notoriously difficult to solve. Many papers have been published on the job-shop scheduling problem. A comprehensive survey of job shop scheduling techniques can be found in [4].

2.2 Instance Generation

In this section, we describe the method for generate different instances of JSS. We need to define the number of jobs, number of machines, the maximum duration and the method for generate the duration of the tasks. There are several methods for generate the duration of the tasks concerning the probability distribution type and the type of the tasks target. In this work, we use the method called No-Correlated for generate the duration of the tasks. This method uses one probability distribution to generate different values for each tasks that represents the duration. The value must be between 1 and the maximum duration defined. The probability distribution can be an Uniform Distribution [8] or a Gaussian Distribution [9], with average the half value of the maximum duration value. For generate an instance, we have to define the processing order, for each job, of the tasks. For this purpose, we generate N permutations randomly from a sequence of integers from 1 to M. The order of each job is the sequence obtained. For example, an instance with 3 jobs and 4 machines and a sequence 4 2 3 1, for the job 1, obtained with randomly permutation of the sequence 1 to 4. So, the first task to be processed in job 1 is the task to be processed in machine 4, the second task to be processed in job 1 is the task to be processed in machine 2, and so on.

3 Linear Programming

A linear programming problem may be defined as the problem of maximizing or minimizing a linear function subject to linear constraints. The constraints may be equalities or inequalities. Formally, a linear programming problem is to find an n-vector, $X = (X_1, \ldots, X_n)^T$, to maximize/minimize

$$c^T x = c_1 X_1 + \cdots + c_n X_n$$

subject to the constraints

$$\begin{gathered} a_{11} X_1 + a_{12} X_2 + \cdots + a_{1n} X_n \leq b_1 \\ a_{21} X_1 + a_{22} X_2 + \cdots + a_{2n} X_n \leq b_2 \\ \cdots \\ a_{m1} X_1 + a_{m2} X_2 + \cdots + a_{mn} X_n \leq b_m \end{gathered}$$

and

$$X_1 \geq 0, X_2 \geq 0, \ldots, X_n \geq 0$$

The following models uses the parameter:

$nrJobs$ is the number of jobs;
$nrMachines$ is the number of machines;
p_{jm} is the duration of the task from job j processed in machine m;
maq_{jo} is the machine of the task from job j that is processed in the order o

3.1 Model 1

In this section, we describe one model used. In this model, the solution for the problem is represented as binary variables for each instant of time and each task. Each variable is 1 if in that instant the task began case contrary the value is 0. Let's formally describe the model. The variable that represents the domain of the method are:

$$X_{jmt} = \begin{cases} 1 \text{ if the task from job j processed in machine m} \\ \quad \text{starts in instant t} \\ 0 \text{ case contrary} \end{cases}$$

$C_{\max}$ is the duration in which all operations for all jobs are completed;

The model (M_1) is the following:

$$v(M_1) = minimize C_{\max} \tag{1}$$

$$\forall j \in J, o \in \{2, \ldots, |O|\}, \sum_{t \in T} t \times X_{j,maq_{jo},t} \geq p_j m + \sum_{t \in T} t \times X_{j,maq_{j(o-1)},t} \tag{2}$$

$$\sum_{j \in J} \sum_{k \in \{max(0,t-d_{jm}) \ldots t\}} X_{jmk} \leq 1 \tag{3}$$

$$p_{j,maq_{j|M|}} + \sum_{t \in T} t \times X_{j,maq_{j,nrMachines},t} \leq C_{\max} \tag{4}$$

$$\forall j \in J, \forall m \in M, \sum_{t \in T} X_{j,m,t} = 1 \tag{5}$$

The goal is defined in (Equation (1)), i.e., to minimize the makespan ($C_{\max}$). The makespan $C_{\max}$ is the time that the processing of all tasks are finish as in Equation (4).

The processing order of the tasks, for each job, is imposed in Equation (2) and the capacity of one task per time in each machine is in Equation (3). In Equation (2), we have that the start time ($\sum_{t \in T} t \times X_{j,maq_{jo},t}$) of the task from job j and processed in machine of the order o in job j (maq_{jo}) is higher than the end time of the previous one (in order $o-1$) in the same job j. In Equation (3), the second sum part equal to one if in instant t, the task from job j processed in machine m is being processed. So, for each machine in each instance, the sum in all jobs must be 0 or 1. Finally, the Equation (5) impose that a task can only begin one time.

3.2 Model 2

In this section, we describe one model used. In this model, the solution for the problem is represented as the starting times for each task. Let's formally describe the model. The variable that represents the domain of the method are:

$Y_{mj_1j_2}$ $\begin{cases} 1 \text{ if, in machine m, the task from job } j_1 \text{ is processed} \\ \quad \text{first than the task from job } j_2 \\ 0 \text{ if, in machine m, the task from job } j_2 \text{ is processed} \\ \quad \text{first than the task from job } j_1 \end{cases}$

X_{jm} is the starting of processing time of the task from job j processed in machine m;

$C_{\max}$ is the duration in which all operations for all jobs are completed;

The model (M_2) is the following:

$$v(M_2) = minimize C_{\max} \tag{6}$$

$$\forall j \in J, \forall m \in M : X_{jm} \leq K \tag{7}$$

$$\forall j \in J, \forall m \in M : X_{jm} \geq 0 \tag{8}$$

$$\forall j \in J, \forall o \in \{2, \ldots, nrMachines\} : X_{jmaq_{j(o-1)}} + p_{jmaq_{j(o-1)}} \leq X_{j,maq_{jo}} \tag{9}$$

$$\forall j_1 \in J, \forall j_2 \in J, \forall m \in M : Xj_1m \geq X_{j_2m} + p_{j_2m} - K \times Y_{mj_1j_2}, j_1 \neq j_2 \tag{10}$$

$$\forall j_1 \in J, \forall j_2 \in J, \forall m \in M : X_{j_2m} \geq X_{j_1m} + pj_1m - K \times (1 - Y_{mj_1j_2}) \tag{11}$$

$$\forall j \in J : p_{j,maq_{j,nrMachines}} + X_{j,maq_{j,nrMachines}} \leq C_{\max} \tag{12}$$

The goal is define in (Equation (1)), i.e., to minimize the makespan ($C_{\max}$). The makespan $C_{\max}$ is the time that the processing of all tasks are finish as in Equation (12). The bounds of the variable X_{jm} are defined in Equation (7) and (8).

The processing order of the tasks, for each job, is imposed in Equation (9) and the capacity of one task per time in each machine is in Equation (10) and Equation (11). The Equation (9), we have that the start time ($X_{j,maq_{jo},t}$) of the task from job j and processed in machine of the order o in job j (maq_{jo}) is higher than the end time of the previous one (in order $o-1$) in the same job j. Using the representation of this model (X_{jm}) as the starting time of the task j in machine m, it's enough to impose the restriction if two tasks of same machines cannot be processed at same time. We can use the expression $X_{j_1m} + p_{j_1m} < X_{j_2m} \vee X_{j_2m} + p_{j_2m} < X_{j_1m}$. However, this expression is non-linear. So, we use an auxiliary variable $Y_{mj_1j_2}$.

4 Constraint Programming

The Constraint Programming is a programming language that is oriented to relationships or constraints among entities.

The Constraint Programming can be used for tackling combinatorial problems such as the JSS. For combinatorial problems, we deal with finite domains and the CP class of techniques used for finite domains are called Constraint Languages Programming for Finite Domains (CLPFD). For CLPFD, it must be define the search variables, the domain of all variables used in the search and the constraints. The domain of a variable is the set of possible values that

the variable can take, but not all points (all variables with a value) are feasible solutions for the problem. These feasible solution are validated by constraints. Given a domain for all variables and the constraints between the variables, the search is done using backtracking search [7].

In this work, we use for CP the Sicstus [1] software that it's a Prolog engine. The domain of the variables is defined using the term $domain(+Variables, +MinValue, +MaxValue)$ where $+Variables$ is a list of variables, $+MinValue$ is the minimum value of the domain and $+MaxValue$ is the maximum value of the domain. For the backtracking search is use the built-in $labeling(: Options, +Variables)$ predicate. The : $Options$ argument controls the order in which variables are selected for assignment (variable choice heuristic), the way in which choices are made for the selected variable (value choice heuristic), and whether all solutions or a single, optimal solution should be found [5]. In this work, we only define the objective of the search that is to minimize the makespan, so $Options = [minimize(C_{\max})]$. The argument $Variables$ is a list of the variable where the backtracking search is done.

4.1 Model 1

In this section, we describe a model using CP for JSS. The search variable is X_{jm} the end of processing time of the task from job j processed in machine m. The domain of the variables X_{jm} is from 0 to MAX where $MAX = \sum_{i \in J, j \in M} p_{ij}$. The objective is to $minimize(C_{\max})$ where $maximum(C_{\max}, X_{jm})$,i.e, minimize the makespan that is the maximum of the end time of processing time of the tasks.

The restrictions are imposed by the user predicates *precedenceConstraint/1* and *machineConstraint/1*. The predicate $precedenceConstraint(+AllTasksJobs)$ is the constraint of the order of processing between tasks of the same job. The parameter *AllTasksJobs* is a list of all tasks of the problem represented with the term $task(O_i, D_i, E_i, H_i, Id_i)$ where O_i is the start time, D_i the non-negative duration, E_i the end time, H_i the resource consumption (if positive) or production (if negative), and Id_i is the task identifier [5]. The predicate *precedenceConstraint/1* run the list of lists where each is grouped the tasks belonging to same job. For each list, that are ordered the tasks by the same order impose by the problem, the predicate run two by two tasks and impose that

$$E_1 \# =< S_2$$

where E_1 is the final time of a task 1 that is previous to the task 2 with start time of processing S_2. The second constraint $machineConstraint(+AllTasksMachines)$ impose that only one task be processed by time in each machine. The parameter *AllTasksMachines* is a list with terms $task(O_i, D_i, E_i, H_i, M_i)$ where O_i is the start time, D_i the non-negative duration, E_i the end time, H_i the resource consumption (if positive) or production (if negative), and M_i a machine identifier[5].

[1] `http://www.sics.se/sicstus.html`

This predicate *machineConstraint/1* run the list and compares all pairs of tasks and if the two are processed in same machine then

$$|med_1 - med_2|\# \geqslant \frac{p_1 + p_2}{2} \tag{13}$$

where $med_i = end_i - \frac{duration_i}{2}$ is the middle point between the start and end time of the task O_i, p_i is the duration of taskO_i and end_i is the end time of processing the task O_i. The Equation (13) means that the distance between the two middle points of the tasks ($|med_1 - med_2|$) must be higher than the half value of the duration of the two tasks. We show that this Inequation (13) imply the desired constraint of the problem, that is,

$$end_1 < end_2 \Rightarrow end_1 < start_2 \vee end_2 < end_1 \Rightarrow end_2 < start_1 \tag{14}$$

The demonstration is following one:

$$|med_2 - med_1| > \tfrac{d_1+d_2}{2} \tag{15}$$

$$\Leftrightarrow med_2 - med_1 > \tfrac{d_1+d_2}{2} \vee med_2 - med_1 < -\tfrac{d_1+d_2}{2} \tag{16}$$

$$\Leftrightarrow \quad start_2 > end_1 \vee end_2 < start_1 \tag{17}$$

If $end_1 < end_2$ then $start_1 < end_2$, by definition and with the result from (17), $start_2 > end_1$ is true. The same with $end_1 > end_2$ then $start_2 < end_1$, by definition, and with the result from (17), $end_2 < start_1$ is true.

4.2 Model 2

As in model from Section 4.1, the solution variable is the processing operation end time. The model is similar excepts instead of using the predicate *machineConstraint/1* for restrict the capacity of processing one task per machine at same time, we use a built-in predicate *cumulatives/3* provided by Sicstus software. The predicate $cumulatives(+AllTasksMachine, +AllMachines, +Options)$ from a set of n tasks and m machines maintains the limit of tasks to be processed by a machine at same time. The variable $AllTasksMachine$ is a list of term $task(O_i, D_i, E_i, H_i, M_i)$ where O_i is the start time, D_i the non-negative duration, E_i the end time, H_i the resource consumption (if positive) or production (if negative), and M_i a machine identifier. The predicate *cumulatives* needs information about the limits of tasks for each machine. This information is in list $AllMachines$ with the a term $machine(M_j, L_j)$, for each machine, where M_j is the identifier and L_j is the resource bound of the machine. In this problem of Job-Shop Scheduling, the resource bound is a superior limit of 1 task, so $L_j = 1, \forall j \in J$ and in variable $Options$ it must have the term $bound(upper)$. More information about the predicate $cumulatives/3$ is in the Sicstus manual [5]

5 Experiments

The experiments are applied to randomly generated instances described in Section 2. For generate the instances, we defined the dimension (number of jobs and machines), the maximum duration and method to generate the duration times. For this work, it was used the combination between all of the following parameters:

- Number of jobs: 5(only CP methods and LP-model-2),4 and 3;
- Number of machines: 5(only CP methods and LP-model-2), 4 and 3;
- Maximum duration: 5 and 10 (only CP methods and LP-model-2);
- Methods for duration: No-Correlated with Uniform Distribution and No-Correlated with Gaussian Distribution;

For each combination, it was created 20 instances of the same type. For example, there are 20 instances with 3 jobs and 3 machines with maximum duration of the operation of 5 and the were generated with the uniform distribution. The algorithms are two models of LP and two models from CP. The LP models are described in Section 3 are: the algorithm that we called LP-model-1 is described in Section 3.1 and the algorithm that we called LP-model-2 is described in Section 3.2. The CP models are described in Section 4 are: the algorithm that we called CP-model-1 is described in Section 4.1 and the algorithm that we called CP-model-2 is described in Section 4.2. For the method LP-model-1, we didn't run in instances with 5 job or 5 machines or 10 for maximum duration , because of very high running time. The software used in the LP algorithms is the "GNU Linear Programming Kit" (glpk) avaible in [2]. For the CP algorithm is used a commercial software, called Sicstus as it was mentioned before. For each method and instance, we repeated the experiment 10 times. In Table 2, we present the average results of the time needed to the algorithms or classes of algorithms (in the column 1) to solve the instances grouped by some characteristic. The first group of lines (LP and CP) is the average of time needed to solve by the algorithms of LP and CP. The second group (binaryStart and startTime), is the average of duration needed for solving the instances for the algorithms with the same domain representation. For example, the values of the column "All" is for all instances and in subcolumn "3" from column "Job" is the value for instances with 3 jobs. The fourth line of values are the average of time that algorithms with domain representation as the start time of the tasks (PL-model-2,CP-model-1 and CP-model-2) needed to solve each class of instances. In Table 3, we show the percentage of instances that each algorithm have the best performance comparing to the others. In both table, the comparation between two methods are made in the same set of instances. So, when comparing with method LP-model-1, we didn't used instance with 5 jobs or 5 machines or 10 for maximum duration, since we didn't run this type of instance with this method. In Table 3, the group A are the instances used for comparing the 4 methods and the group B are the instances used only for the methods of CP and the LP-model-2.

[2] `http://www.gnu.org/software/glpk/`

Table 2. Information of the average duration for solving the instances of each class

	All	Nr.Jobs			Nr.Machines			Distrib.	
		3	4	5	3	4	5	unif.	gauss.
LP	28.318	2.490	54.146	0.337	4.928	51.708	3.862	27.950	28.686
CP	0.089	0.086	0.0931	12.329	0.086	0.093	10.230	0.088	0.090
binaryStart	56.604	4.967	108.241	-	9.829	103.379	-	55.869	57.339
startTime	0.070	0.061	0.079	8.33	0.066	0.074	6.889	0.069	0.071
LP-model-1	56.603	4.967	108.241	-	9.829	103.379	-	55.869	57.339
LP-model-2	0.033	0.014	0.051	0.337	0.02	0.038	0.207	0.032	0.033
CP-model-1	0.090	0.086	0.094	8.77	0.087	0.093	7.050	0.089	0.091
CP-model-2	0.088	0.085	0.093	15.88	0.085	0.092	13.410	0.088	0.089

Table 3. Percentage of instance that an algorithm have the best performance

Group	LP-model-1	LP-model-2	CP-model-1	CP-model-2
A	0	95.63	1.87	2.5
B	-	62.67	19.82	17.5

6 Discussion

There are several observations that we can make from the results from Table 2. The first one, it's that ranking of methods about the performance is the same whatever is the class of instances. We can see that the method with best average performance is LP-model-2, then CP-model-1, CP-model-2 and finally LP-model-1. However, in average, the two models from CP are better than the two of LP. This is because the model 1 of LP have very poor performances. It's interesting to observe that values of the performance of the LP methods increase more than the methods of CP methods when the dimensions of the instances increases. About the distribution used for generate the duration, it have small differences between the Gaussian and Uniform distributions. However, the average performance for all method is slightly lower for instances with durations generated with Uniform distribution even when the maximum duration is very low (5) that makes the samples of the two distribution not very different. Perhaps, if we increase the maximum duration we can see more amplified the trend observed. From Table 3, we observe that the method LP-model-2 is better than the others in 96% of the instances. When we increase the dimensions of the job, machine or maximum duration the dominance of the method LP-model-2 is much lower. The percentage of winnings is only 62%.

7 Conclusion

In this work, we present several models for Linear Programming and Constraint Programming of Job-Shop Scheduling. The differences between the models are the representation of the solution and the programming language used. We compare the average of time needed for each method find the optimal solution in

different class of instances. In average, the best method is the same for all classes of the instances - a model where the representation of the solution is the starting time of the tasks using the Linear Programming techniques. There are evidences that when we increase the dimension of instances the ranking between methods can change since the performance of LP methods increase more than the CP methods and the number of instances that CP methods wins have a high increase only adding a job or a machine. Another trend noted is that the instances with durations generated with Gaussian distribution can be harder for solve, but the differences are very low for any conclusion. The model of the problem used is very important. We can see that for two different models for LP, we have a performance so different. The results between the two models from CP are surprising. We expect better results with CP-model-2 since the *cumulative* is an optimized to schedule tasks.

For future work, we should increase the number of representation and experiment all methods in more instances with higher dimensions and generate with different methods for a a higher diversification of the instances. The objective is to find some answer to the questions raised during the work.

References

1. Runwei Cheng, Mitsuo Gen, and Yasuhiro Tsujimura. A tutorial survey of job-shop scheduling problems using genetic algorithms—i: representation. *Comput. Ind. Eng.*, 30(4):983–997, 1996.
2. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, California, 1979.
3. A. Jain and S. Meeran. A state-of-the-art review of job-shop scheduling techniques. Technical report, Department of Applied Physics, Electronic and Mechanical Engineering, University of Dundee, Dundee, Scotland, 1998.
4. A. S. Jain and S. Meeran. Deterministic job-shop scheduling: Past, present and future. *European Journal of Operational Research*, 113:390–434, 1999.
5. Intelligent Systems Laboratory. *SICStus Prolog Users Manual*. Swedish Institute of Computer Science, PO Box 1263 SE-164 29 Kista, Sweden, 4.0.4 edition, June 2008.
6. J. K. Lenstra and A. H. G. Rinnooy Kan. Computational complexity of discrete pptimisation problems. *Annals of Discrete Mathematics*, 4:121–140, 1979.
7. Kim. Marriott and P. J. Stuckey. *Programming with constraints : an introduction*. MIT Press, Cambridge, Mass., 1998.
8. E. Taillard. Benchmarks for basic scheduling problems. *European Journal of Operational Research*, 64:278–285, 1993.
9. Jean-Paul Watson, Laura Barbulescu, Adele E. Howe, and Darrell Whitley. Algorithm performance and problem structure for flow-shop scheduling. In *AAAI/IAAI*, pages 688–695, 1999.
10. David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.

Propositional Based Inductive Logic Programming: Reduced Encoding of Hypotheses and Knowledge Base

Hugo Ferreira[1]

Faculty of Engeneering of the Universarity of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL,
pro07026@fe.up.pt,
WWW home page: https://www.fe.up.pt/~hmf/web/welcome.html

Abstract. Inductive Logic Programming is a machine learning technique that allows one to induce first-order logic theories. The induction of such theories is nevertheless time-consuming due to the very large search space that such hypothesis present. More specifically, the standard coverage test used by most ILP systems is based on the inefficient resolution decision procedure. Propositional logic based inference presents an efficient alternative. However, the propositionalization of first-order logic results in very large formulae. This is a preliminary report investigating how we may compactly encode and efficiently use propositional logic in ILP. Some tentative conclusions are presented based on the worst–case analysis of several possible representations.

1 Introduction

Inductive Logic Programming (ILP) [1] or more generally Multi-relational Data Mining (MRDM) [2] refers to a set of machine learning techniques that are concerned with the identification of patterns from multi-relational data. Unlike many other machine learning algorithms, ILP and MRDM deal specifically with first–order clausal logic or relations. Earlier work in this area focused on the use of Horn clause logic. However, intense research in this subject has led to the development of numerous extentions, adaptations and innovations resulting in a myriad of approaches that can be classified into two main groups: *predictive induction* and *descriptive induction*.

Predictive induction performs data analysis through hypothesis generation and testing. It includes the induction of first–order classification rules, decision trees and Bayesian classifiers. Descriptive induction on the other hand performs exploratory data analysis aimed at discovering regularities and uncovering patterns that are required in knowledge discovery. In this case symbolic clustering, association rule learning and subgroup discovery are possible.

Many of today's ILP systems execute in two phases: the *saturation* phase when the set of valid candidate hypotheses is extended with additional hypothesis and the *reduction* phase when these newly introduced candidate hypotheses

are evaluated and eventually removed if they do not meet a set of necessary and sufficient conditions. One of these necessary conditions is that a hypothesis logically entail all positive examples. This *covering* test is usually done deductively using SLDNF resolution (Selective Linear Definite clause resolution with Negation as Failure; available for example in Prolog systems) to test the hypotheses against each and every example. The reduction phase is therefore very time consuming thereby making it impractical to attempt the larger and more interesting problems. We are therefore investigating the possibility of using alternate decision procedures based on propositional logic in order verify if any efficiency gains are possible. Such a stance has already been taken in other research areas, most notably in AI (Artificial Intelligence) planning [3] and model checking [4].

In order to be able to use any of the deductive inference mechanisms that are available for propositional logic, we must first be able to translate all of the relational data into propositional data. This is referred to as *propositionalization*. The most pressing problem with propositionalization is the generation of very large amounts of information that may ultimately be irrelevant for the induction of the model. We therefore study the issues regarding propositionalization and identify several potential solutions that may effectivelly reduce the number of propositions that are generated.

This article is structured as follows: we first review research concerning propositionalization and the use of propositional logic in the implementation of ILP systems (Section 2). Next we consider several possible strategies that may be employed in the reduction of the propositional hypothesis space (Section 3). Because this is only a preliminary study we will also point out several additional tasks that still need to be performed in order to produce the necessary results (Section 4). Finally, we present some tentative conclusions based on the work done so far (Section 5).

2 Related Work

Investigations has been conducted in order to determine if the use of propositional logic may or may not be advantageous relative to its ILP counterparts [5]. It has been shown that in several applications propositional learners outperform relational learners [5, 6]. In addition to this, the use of propositionalization allows us to take advantage of existing attribute-value learners that are readily available [6] and include many interesting properties such as the ability to handle noisy data [7, 8], produce understandable and efficient rules [9]. The use of propositionalization however presents several major disadvantages, which include [6]: incorrect use of joining and aggregation of relational data, the need to exhaustively generate all propositions prior to induction and the lack of opportunity to prune the hypothesis search space.

One of the earliest research efforts in ILP studied the issues related to the upgrading of attribute-value learners for use with first–order logic [9]. This work established a direct relation between propositional learners and ILP systems based on the *learning from interpretation* setting. Even though [9] clearly stated that

the "relational representation can overcome limitations of the propositional representation" and provides a methodology to upgrade the attribute-value learners in order to produce efficient equivalents, it also established the need for *bias* when searching the hypothesisis space in first–order logic representation. Such bias is required in order to make the search tractable because the hypothesisis space in first–order logic is very large, potentially infinite, and suffers from the determinacy problem [9]. It is therefore conceivable that the bias applied to ILP in first–order logic may and attribute-value learners may also be adopted by propositional logic based learners.

The LINUS system is an example of the use of attribute-learners to induce first–order theories [7]. Its main goal was to take advantage of the noise–handling capability of the propositional learners. Such a capability was not available to existing ILP systems at that time. This work is directly concerned with the issues related to the propositionalization of relational data [7, 8]. Its main restriction is that the hypothesis language is limited to deductive hierarchical databases clauses. Further language bias was also introduced in the DINUS system, which enhanced LINUS, by limiting the number of variables used in the refined hypothesis (*i–j determinacy*) [8].

At least one effort specifically tried to answer whether or not there are any "objective and quantifiable advantages of learning in first–order logic over learning in propositional logic" [5]. The author admits that answering these questions in all settings may not be possible. He does however attempt to provide some insight into these issues by limiting his study to the case of *tree induction* and, in regards to first–order logic, learning from interpretation only. He points out that *propositionalization need not be complete*. As with first–order logic ILP systems, heuristics may be employed in order to prune the search space. Experimentation is conducted using stochastic search that employs the hill-climbing strategy. Results showed very good performance and the conclusion (which concurs with other researchers' work) is that attribute-value learners in the form of tree induction (S-CART: upgraded classification and regression trees) is to be preferred over relational learning. The author nevertheless emphasizes that efficient ILP-like techniques were used in the propositionalization phase of induction in a very specific setting, so additional systematic experimentation for other setting is required.

We have already seen that language bias may be used to reduce the search space. In the case of the LINUS and DINUS systems such bias is statically defined prior to induction. Alternatively, *language bias can also be dynamic*. In such cases ILP systems may allow for a shift in bias towards a more expressive language during the induction process itself. Similarly [10] describes an ILP based system that uses a propositional learner whose features are constructed in a goal-directed fashion. It uses a two-phased learning algorithm. In the first phase the MolFea domain specific inductive database selects a set of features according to weights assigned to the (positive and negative) examples. The propositional learner then induces a model with these features. The induced hypothesis is then used to evaluate the error rate. Each example's weight is then changed according to the

induce hypothesis error rate. This process is then repeated until either the error rate is below or above a given threshold or a maximum number of iterations is reached.

Learning unrestricted clausal theories from complete evidence (interpretation based on the complete Herbrand base) is known as *identification*. Finding non-redundant theories in such a setting is referred to as *reformulation*. Its practical implementation in first–order logic presents significant challenges. However [11] demonstrates that in propositional logic the induction consists of only two steps of purely symbolic manipulation: determining false evidence from true evidence (negation) and determining the true hypothesis from the false evidence (simplification of the dual representation of formula in CNF (Conjunctive Normal Form) as DNF (Disjunctive Normal Form)). The work cited shows us that symbolic manipulation of propositional logic formula may provide efficient alternatives to generating and manipulating features.

3 Reducing the Hypothesisis Space

ILP systems in general use two basic strategies in order to make induction more efficient. The first is the use of language (*syntactic* or *semantic*) *bias* [9]. We will limit ourselves to the analysis of syntactic language bias because semantic language restrictions depend on a specific domain and are therefore not easily adopted. In addition to language bias, the search performed on the hypothesis space may employ several *search strategies* and *heuristics*. Search strategies include complete strategies (depth–first, breadth–first and iterative deepening) and incomplete strategies such as (best–first, hill–climbing and beam search). Heuristics are used to guide the search of the hypothesis space and establish the stopping criterion. A large variety of heuristics exist and include for example: accuracy, informativity, accuracy gain, information gain, relative frequency, Laplace estimate and the m-estimate [8]. We have seen the cross–pollination of techniques between the attribute-value learner and ILP systems. This work attempts to identify uses of ILP language bias (pruning, domain encoding), search strategies and heuristics (generality ordering, ranking, symbolic logic manipulation) in order to reduce the hypothesis space expressed in propositional logic.

3.1 Generality Ordering of Propositional Logic Search-Space

The majority of ILP systems structure their search space according to the first–order logical based Θ–subsumption framework [9]. This has the important advantage that the search space may be ordered and efficiently pruned during cover testing. In the case of propositional logic the search space and cover testing is also done according to the logical generality relation and search ordering and pruning is therefore possible. However in this case cover testing is based on logical implication instead of the logically weaker relation of Θ–subsumption in first order logic ($H\Theta$–subsumes $E \Rightarrow H \models E$ but not the converse). The resulting search space is therefore more restricted (it is now anti–symmetric) thereby promoting greater efficiency.

3.2 Symbolic Boolean Logical Manipulation

The related work reviewed here shows us that propositional logic provides a means, for example, of efficiently and deterministically inducing non-redundant theories [11]. We propose the use of alterante propositional logic inference methods in order to compactly represent and efficient manipulate the set of features. Some examples follow:

- The conversion and manipulation of CNF formula may be done efficiently with BDDs (Binary Decision Diagrams) [12, 13, 4]. It also provides a very compact representation of formulae and thereby allow for the representation of very large sets of hypothesis and examples.
- The induction of theories in [11], for example, required the calculation of the prime implicates (minimization of the boolean expressions), which was done using the Quine-McCluskey method. Prime-implicates may also be obtained via other means such as: integer linear programming, SAT (satisfiability) solvers and BDDs [14].
- SAT solvers' efficiency are due to their use of clause learning and non-chronological backtracking methods. Learned clauses may be retained by SAT solvers during the repeated execution of coverage testing thereby potentially increasing one of ILP's most time-consuming phases.
- The cover testing may be performed symbolically via BDDs thereby allow for the simultaneous coverage testing of several hypothesis for a given set of examples.

We would like to point out that there is a very large body of research in automated reasoning and model checking, which may provide many interesting solutions. It is important to note that many decision procedures and their improvements have in the past been deemed useless due to the high utility cost they incur. However in the ILP setting such methods may still yield good results. Consider for example the recompilation of large sets of examples that increase the efficiency of satisfiability checking. Because in ILP such querying of the knowledge based is done countless times, the utility cost is drastically lower. Many previously discarded solutions should therefore be reconsidered.

3.3 *i*-Determinacy

In first–order logic ILP, irrespective of whether the refinement operator computes the generalization or specialization of a clause, such an operator presents two basic difficulties [9]. Consider for example the case of specialization :

Infinite chains of refinements: when adding literals and binding variables we may generate an infinite number of refinements whose equivalence class does not change. For example the clause `daughter(X,Y)` ← `parent(X,Y)`. It may be refined to `daughter(X,Y)` ← `parent(X,Y), parent(X,Z)` then further on to `daughter(X,Y)` ← `parent(X,Y), parent(X,Z), parent(X,W)` and so on.

Determinacy problem: occurs when a given refinement of a clause will not alter the coverage. Consider for example a knowledge base that describes molecules. Because all molecules have a bond refining the clause $\oplus \leftarrow$`atom(X)` to $\oplus \leftarrow$`atom(X), bond(X,Y)` will not alter the coverage.

The problems referred to above make it difficult to identify an upper or lower bounds of the hypothesis space. In addition to this they may inadvertently mislead the heuristic search of the hypothesis space. Normally such difficulties are circumvented by applying some form of language bias. The simplest of these is to set a fixed limit on the number of predicates and variables that each clause may have. The DINUS system for example employs a more sophisticated form of language bias known as i-determinacy [8]. The i-determinacy simply enforces the restriction that all literals be determinate up to a maximum variable depth of i. A variable in the head of the clause has depth 0. The depth of all other variable is one plus the *maximum* distance of any its preceding variables appearing in prior clauses. All of these techniques also apply to propositional logic learners because they can be used before the propositionalization of the relational data or during hypothesis refinement.

Within the same setting as the LINUS system (function free, non-recursive definite clauses) a simple observation shows us that the number of propositions generated under the i-determinacy restrictions may be reduced further. Table 1 shows the set of propositions generated under this restriction (predicate `p/2` stands parent, `m/1` for mother and `f/1` for father):

Table 1. Propositional form of the *daugter* relationship problem [8].

E	Variables		Features							
	X	Y	f(X)	f(Y)	m(X)	m(Y)	p(X,X)	p(X,Y)	p(Y,X)	p(Y,Y)
$\oplus$	sue	eve	true	true	false	false	false	false	true	false
$\oplus$	ann	pat	true	false	false	true	false	false	true	false
$\ominus$	tom	ann	false	true	true	false	false	false	true	false
$\ominus$	eve	ann	true	true	false	false	false	false	false	false

Consider the case where a proposition has the same valuation for all examples (for instance the last column labeled `p(Y,Y)` in Table 1). These features cannot be used to discern between positive ($\oplus$) and the negative ($\ominus$) examples and are therefore *irrelevant*. They can therefore be safely removed from the knowledge base prior to predictive induction. In the case of descriptive induction no negative examples are used. Removal of any such data would render the process incomplete. Nevertheless such a pruning strategy may still be used as a heuristic.

Next, consider the case where a given proposition has both true and false valuations for all positive examples (for example the column labeled `f(Y)` in Table 1). Such a proposition is said to be *inconsistent* for a given equivalence class. If

we know *a priori* that all positive examples represent a single concept that may be described by a single clause, then this attribute may be removed from the knowledge base before proceeding with induction. In the event this assumption is not valid, such a pruning strategy is inadmissible but may nevertheless still be used as a heuristic.

It is important to note that we cannot use the consistency or inconsistency of the negative examples as a basis for the removal of any of the features. This is because negative examples are, by definition, inconsistent (no single feature can be used as a model or counterexample).

For the knowledge base represented above (see Table 1) LINUS induces the following rule (a daughter is anyone who is a female and has a parent): `daughter(X,Y)` ← `female(X), parent(Y,X)`. If we proceed with the removal of the propositions according to the criterion described in the previous paragraphs (eliminate the columns labeled `p(X,X), p(X,Y), p(Y,Y), f(Y), m(Y)`), we can observe that none of the final hypothesis's literals have been removed. In this example we see a significant reduction in the number of propositions (62.5%) which allows for faster coverage checking and ultimately more efficient induction. Nevertheless we must be aware that the amount of reduction attained depends both on the structure of the knowledge base and the set of literals used during hypothesis search.

The criterion above were described and demonstrated solely for the removal of propositions from the hypothesis space prior to induction. They may also be used in a variety of other ways. For example, we may rank features by their consistency and refine the hypothesis according to that rank (consistency here is measured as a ratio between the number of positive examples in which the proposition has the same valuation and the total number of positive example). This can also be used, for example, to deal with noisy data by assuming that a proposition need not cover all positives in order to be used in an induced clause.

3.4 Compact Logical Axiomatization and Representation

We have already seen (Section 3.2) that propositional logic may be used in the induction of hypothesis. The attentive reader may also have noticed that the pruning criterion presented above (see Section 3.3) may also be done symbolically. In order for such logical operations to be efficient[1], we must ensure that the encoding be as compact as possible. Efficient propositional logic encoding has been extensively studied in AI planning. We have adapted and trivially extended the work presented in [15] to the case of logical induction.

Efficient encoding of a problem in propositional logic depends on the: *axiomatization* of the problem (what logical formulae are used to define the problem consistently), the *representation of propositions* (how are the basic logical propositions represented), and the use of *factored* axioms (manually minimizing the set of formula that represent the domain). In this section we will study both

[1] In general the efficiency of the of the propositional logic inference engines are proportional to the number of propositions and the size of the formula.

the axiomatization and representation of the formula (unlike planning that uses a specialized language to express the planning problem that is amenable to factoring, minimizing a set of general formula can only be done via logical simplification). The various encoding will be compared according to the expected worst–case size of the formula. The size of the formula is evaluated as the maximum number of variables used in the logical formula.

Axiomatization In the case of induction both the knowledge base and the hypothesis must be expressed as logical formula. The set of all positive examples and the set of all candidate hypothesis can be encoded using a formula in the DNF. For instance in Table 1 the positive examples may be defined by the logical formula:

$$f(X) \wedge f(Y) \wedge \overline{m(X)} \wedge \overline{m(Y)} \wedge \overline{p(X,X)} \wedge \overline{p(X,Y)} \wedge p(Y,X) \wedge \overline{p(Y,Y)} \vee$$
$$f(X) \wedge \overline{f(Y)} \wedge \overline{m(X)} \wedge m(Y) \wedge \overline{p(X,X)} \wedge \overline{p(X,Y)} \wedge p(Y,X) \wedge \overline{p(Y,Y)}$$

The resulting formula have some interesting properties. First observe that the knowledge base is now expressed only in terms of the propositions that encode the hypothesis and not the original domain. This means that the size of the knowledge base has effectively been reduced. To demonstrate this, consider for example the trivial case of the single predicate `p(Y,X)` whose domain is { `sue, eve, ann, pat, tom` } in Table 1. The original knowledge base would require two propositions (`p(sue, eve)`$\vee$`p(ann, pat)`); the converted knowledge base however only requires the single proposition `p(Y,X)`. Second, because of the inherent structure of the examples the formula are amenable to efficient[2] logic simplification. For example in order to encode both positive examples in Table 1 we can use the equivalent but simpler formula:

$$f(X) \wedge \overline{m(X)} \wedge \overline{p(X,X)} \wedge \overline{p(X,Y)} \wedge p(Y,X) \wedge \overline{p(Y,Y)}$$

Note that because the size of the knowledge base depends on: a) the inherent structure of the classes and b) the number of propositions used to encode the hypothesis (size of formula and number of variables), the actual reduction in the size of the encoding is ultimately domain specific and highly dependent of the pruning strategies used by the induction algorithm. Nevertheless this reduction is guaranteed for any knowledge base whose domain size is greater than the number of propositions used to encode the hypothesis.

Unfortunately the interesting properties described in the previous paragraphs don't all hold for the case of the negative examples. Recall that the negative examples are by definition inconsistent and it is therefore highly unlikely that they exhibit any structure. Even though simplification of such formula is more difficult, this does not pose any real issue because: a) conversion of the negative examples to the hypothesis space ensures a minimum reduction in its size and b) usually the set of negative examples used in the predictive induction of theories is small compared to the number of positive examples.

[2] The efficiency here is a direct result of encoding the knowledge base in DNF.

A final note on symbolic logic manipulation is in order here. One may be tempted to manipulate the formula representing the examples as a means of inducing a theory. However it is important to be aware that such operations may be exponential in nature thereby resulting in the very inefficient induction of theories. Consider for example negating the formula that represent the negative examples in an attempt to obtain the complete set of possible hypothesis. This is equivalent to performing a conversion from DNF to CNF in linear time. Enumerating these hypothesis in order to rank and select an appropriate theory is then equivalent to performing a conversion from CNF to DNF (trivially satisfiable), which is exponential.

Representation In order to make the encoding – and therefore the inference – more efficient, alternate propositionalization of the formula is possible. Until now each first-order logic proposition have been mapped to a single corresponding variable (for example the proposition `p(Y,X)` in predicate logic is represented by a single propositional logic variable `p(Y,X)`). This is referred to as a *regular* representation. However, other encoding exist that allow one to represent a single first-order literal via several variables. The first of these is known as the *simply split* representation and it encodes each of the first-order logic predicate's arguments as a separate variable (`p1(Y),p2(X)`) thereby reducing the total number of variables required to encode the literals (see Table 2 for an estimate on size). Notice that the same predicate and any of its arguments may appear several times in the same example or hypothesis. It is therefore possible to further reduce the number of variables by encoding each predicate and all arguments (irrespective of the predicate) as a separate variable (`p,arg1(Y),arg2(X)`).

The total number of predicates and the possible combinations of arguments is finite but large. Because a large number of elements may be compactly represented using binary encoding, we may also represent each of the predicates as a set of binary variables (in other words all predicates P may encoded with $log_2|P|$ variables). This is known as the *bitwise* representation. We may also extend the use of binary encoding to all arguments and will refer to this as the *full bitwise* representation[3]. These representations, decrease the number of variables further.

All the representations described above and their (worst-case) estimated sizes are shown in Table 2 (column number of variables). Note the estimated size refers only to the encoding of a single example or hypothesis. Expressing the size based on the number of examples used or the number of hypothesis considered will not provide any additional information on the compactness of the encoding. It is important to emphasize that because these are worst-case estimates (maximum arity, no typed parameters used), the expected compression ratios will always be less that the estimated.

Unfortunately when the representation changes so must the axiomatization change in order to guarantee the consistency of the formula. The regular repre-

[3] It is important to note that the binary encoding can be applied to the regular, simply split and overload split representations. We will nevertheless limit our analysis to the case of the overload split representation that has the best estimated reduction.

sentation requires no changes. In the case of the *simply split* representation each set of literals that defines a single first-order predicate must now be expressed as a signed conjunct. The result is that a set of disjunctions are introduced into the formula, which translates to a worst–case exponential increase in size of the equivalent formula in CNF. To demonstrate this consider the first positive example in Table 1 when it is converted to the simply split representation (we show only the case for predicate `p(Y,X)`, the others do not change):

$$\oplus_1 = \overline{(p1(X) \wedge p2(X))} \wedge \overline{(p1(X) \wedge p2(Y))} \wedge p1(Y) \wedge p2(X) \wedge \overline{(p1(Y) \wedge p2(Y))} \quad (1)$$

$$= \overline{(p1(X)} \vee \overline{p2(X)} \wedge \overline{(p1(X)} \vee \overline{p2(Y)} \wedge p1(Y) \wedge p2(X) \wedge \overline{(p1(Y)} \vee \overline{p2(Y))} \quad (2)$$

$$= \overline{p1(X)} \wedge p1(Y) \wedge p2(X) \wedge \overline{p2(Y)} \quad (3)$$

The worst–case analysis of the size of the equivalent CNF formula according to the various representations is shown in Table 2. Those expressions represent the size of the equivalent CNF formula when all predicates are assigned a negative valuation and subsequently all disjunctions are removed. We can see that the regular representation incurs no additional costs. However all other representations result in an exponential increase in size according to the number of predicates. Note however that in general because $|P| < A_p$ for the overload, bitwise and full bitwise split representations, the increase in size may be acceptable. Some additional notes are in order here. The first is that even though there may

Table 2. Size of basic logical propositions according to number of variables used in the representation of a single example or hypothesis.

p(X,Y,Z)	Representation	Number of Variables [1]	Size of Conjunction
Regular	`p(Y,X)`	$\|P\|^{(A_p+1)}A_p^{A_p}$	$\|P\|^{(A_p+1)}A_p^{A_p}$
Simply Split	`p1(Y),p2(X)`	$(\|P\|A_p)^2$	$\|P\|A_p^{(\|P\|+1)}$
Overload Split	`p,arg1(X),arg2(Y)`	$\|P\| + \|P\|A_p^2$	$\|P\| + (\|P\|A_p)^{(\|P\|+1)}$
Partial Bitwise	$\texttt{p}_i$`,arg1(X),arg2(Y)`	$log_2\|P\| + \|P\|A_p^2$	$log_2\|P\| + (\|P\|A_p)^{(log_2\|P\|+1)}$
Full Bitwise	$\texttt{p}_i$`,`$\texttt{arg}_{i,j}$`(X)`	$log_2\|P\| + log_2(\|P\|A_p^2)$	$log_2\|P\| + log_2(\|P\|A_p)^{(log_2\|P\|+1)}$

[1] $|P|$ is the number of predicates,
A_p is the maximum arity of the predicates

be a worst–case exponential growth in the sizes of the formulae, usually these formulae represent some structure and are therefore amenable to simplification (as shown in Equation 1). Second, the increase in size for the simply split and the overload split representation are greatest. However the bitwise and especially the full bitwise encoding may provide a viable solution. And third, notice how once the formulae have been simplified important information regarding the relations may be lost. For example $\overline{\texttt{p1(X)}} \wedge \texttt{p1(Y)} \wedge \texttt{p2(X)} \wedge \texttt{p2(Y)}$ represents two propositions. In the general case all possible combinations of parameters must be *reconstructed*, which may incur a worst–case exponential cost.

4 Future Work

As was previously stated this is preliminary work. Not all the literature has been covered yet, so in the immediate future several other articles will be examined. In addition to this we still have an open issue regarding representation. Consider for example the set of candidate hypothesis which can be defined as the disjunction of propositions (for example `p1(X)` $\vee$ `p1(Y)` $\vee$ `p2(X)` $\vee$ `p2(Y)`). Any combination of such variables may be induced as a valid hypothesis, however only a subset of those valuations are consistent with the first-order relations. More specifically we must include restrictions that will ensure that all valuations that represent arguments of a first-order relation are true (for example if the hypothesis has a variable `p1(X)` set then either `p2(X)` or `p2(Y)` or both, in case of multiple use of the same predicate, must also be true). We have yet to analyse what additional restrictions are required and the resulting increase in complexity due to these restrictions. Once the above has been completed additional experimentation is necessary to test the various encodings. This is because the results not only depend on the propositionalization techniques used, but also on the selected inference mechanisms and the problem domain itself. As a result we can only opt for an encoding based on empirical evidence.

Several important issues related to the selection of symbolic–logic encoding have been overlooked. More precisely we have not delved into the various inference methods that are available and how these may be used to implement the standard ILP algorithms. Some of the concerns we have in this area include: declaring bias, allowing for heuristic (incomplete) search and dealing with noisy data.

5 Conclusion

No real conclusion can be presented for lack of empirical evidence, however we may speculate on the viability of using propositional logic for ILP based on the estimated worst–case size of the formulae. The analysis presented seems to indicate that propositional symbolic logic manipulation provide a compact means of encoding (especially for the full bitwise representation) and an efficient medium for inducing theories. Nevertheless, a more detailed look at the complexity of propositional encoding of the knowledge base and the candidate hypothesis shows us that such an encoding may not be viable due to their exponential size. This is true even when restricting ourselves to domains wherein only determinate, function free, non-recursive definite clauses are used. Much of this problems stems from the fact that efficiently encoding a domain is basically a compromise between reducing the number of variables used and the complexity of the formulae required. What is more the compactness of the encoding is also determined by the inherent structure of the problem and is therefore domain dependent. It is therefore our opinion that standard ILP techniques, such as languague bias and heuristc search, are still required in order to be able to take advantage of the use of propositional symbolic logic manipulation.

References

1. Stephen Muggleton and Luc De Raedt, Inductive Logic Programming: Theory and Methods, J. Log. Program., 19/20, 629–679, (1994)
2. Džeroski, Sašo, Multi-relational data mining: an introduction, SIGKDD Explor. Newsl., 5, 1, 1–16 ACM Press, New York, NY, USA, July (2003)
3. Kautz, Henry and Selman, Bart, Pushing the Envelope: Planning, Propositional Logic, and Stochastic Search, Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference, Ed. Shrobe, Howard and Senator, Ted, 1194–1201, AAAI Press, Menlo Park, California, (1996)
4. Bwolen Yang and Randal E. Bryant and David R. O'Hallaron and Armin Biere and Olivier Coudert and Geert Janssen and Rajeev K. Ranjan and Fabio Somenzi, A Performance Study of BDD-Based Model Checking, FMCAD '98: Proceedings of the Second International Conference on Formal Methods in Computer-Aided Design, 255–289, Springer-Verlag, London, UK (1998)
5. Stefan Kramer, Relational learning vs. propositionalization: Investigations in inductive logic programming and propositional machine learning, AI Commun., 13, 4, 275–276, IOS Press, Amsterdam, The Netherlands, (2000)
6. Nicolas Lachiche, Good and Bad Practices in Propositionalisation, AI*IA, 50–61, (2005)
7. S. Dzeroski and N. Lavrac, Inductive Learning in Deductive Databases, IEEE Transactions on Knowledge and Data Engineering, 5, 6, 939–949, IEEE Computer Society, Los Alamitos, CA, USA, (1993)
8. Nada Lavrac and Saso Dzeroski, Inductive Logic Programming: Techniques and Applications, Ellis Horwood, New York, (1994)
9. Wim Van Laer and Luc De Raedt, How to Upgrade Propositional Learners to First Order Logic: a Case Study, Relational Data Mining, 235–256, Springer-Verlag New York, Inc., New York, NY, USA, (2000)
10. Stefan Kramer, Demand-Driven Construction of Structural Features in ILP, ILP '01: Proceedings of the 11th International Conference on Inductive Logic Programming, 132–141, Springer-Verlag, London, UK, (2001)
11. Peter A. Flach, Normal Forms for Inductive Logic Programming, ILP '97: Proceedings of the 7th International Workshop on Inductive Logic Programming, 149–156, Springer-Verlag, London, UK, (1997)
12. Soha Hassoun and Tsutomu Sasao, Logic Synthesis and Verification, Ordererd Binary Decision Diagrams in Electronic Design Automation, Chapter 11, 285–308, Kluwer Academic Publishers, Norwell, MA, USA, (2002)
13. Henrik Reif Andersen, An Introduction to Binary Decision Diagrams, October 1997, (minor revisions April 1998)
14. Vasco M. Manquinho and Arlindo L. Oliveira and Joo P. Marques Silva, Models and Algorithms for Computing Minimum-Size Prime Implicants, In Proc. International Workshop on Boolean Problems (IWBP'98), (1998)
15. Michael D. Ernst and Todd D. Millstein and Daniel S. Weld, Automatic SAT-compilation of planning problems, IJCAI-97, Proceedings of the Fifteenth International JointConference on Artificial Intelligence, 1169–1176, Nagoya, Aichi, Japan, August 23–29 (1997)

Inspections on Testing Aspect-Oriented Programs

Rodrigo M. L. M. Moreira

Faculdade de Engenharia da Universidade do Porto
pro08007@fe.up.pt

Abstract. Aspect-Oriented Programming (AOP) is a recent programming paradigm that aims at enhancing modularity and thus solving the problem of crosscutting concerns by capturing them into new units of modularity called aspects. With the increasing usage and acceptance of AOP, the task of assuring aspect-oriented systems' correctness has become a challenge, mainly due to its nature. Although several testing techniques have been applied and improved in Object-Oriented (OO) programs through the years, it is still required to demonstrate and verify which ones can be applied to AOP. This paper provides a synopsis regarding AOP and identifies software quality harms inducted by AOP. In addition, major issues related with testing AOP programs are also described. The latter leads to a set of suggestions with the intention of improving and assuring software quality in Aspect-Oriented Systems, which is the main goal of the underlying research work. The work here reported was strictly based on surveying existing literature on the topic.

Keywords: Aspect-Oriented Programming, AOP, Testing, Software Quality, AOP testing issues, Testing Aspect-Oriented Programs.

1 Introduction

Aspect-Oriented Programming (AOP) is an emerging methodology that best suites the concept of separation of concerns (SOC). SOC is one of the key principles in Software Engineering. It refers to the ability to identify, encapsulate, and manipulate concrete portions of software that are relevant to a particular *concern*. A *concern* is an area of interest in a system. It can be seen as a requirement that must be addressed in order to satisfy a system's goal. Concerns embody the main motivation for categorizing and decomposing software into controllable and comprehensible components. In practice, a software system corresponds to the realization of a set of concerns [1].

Despite the fact that Object-Oriented programming (OOP) is the most used methodology to manage core concerns, it doesn't allow decomposition of a problem into all of its concerns. This occurs due to OOP nature that forces *coupling* between the core and crosscutting concerns, meaning that the addition of new crosscutting features will obligate changes to occur in core modules. As result, these concerns will get *scattered* and *tangled* through the modules (as illustrated in Fig.1). These are the symptoms of typical software that is non-modularized, where AOP should be applied.

AOP is a new discipline that targets to enhance modularity, by using aspects, whose code portions are *woven* into the target application according to precise

weaving rules provided in the aspect definition [2]. The main goal for AOP is to handle crosscutting concerns. AOP is applied upon existing methodologies such as OOP and procedural programming, expanding them with concepts and constructs in order to modularize crosscutting concerns. In fact, the core concerns are implemented using the chosen base methodology. AOP is an additional technique and cannot be seen as a replacement for OOP or procedural programming.

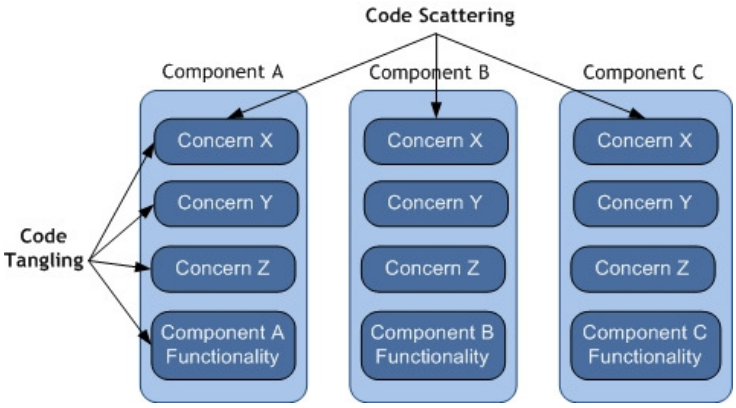


Fig. 1. Code *tangling* and code *scattering* example. Code *scattering* refers to the code for a particular concern that appears in multiple places in the design. Code *tangling* refers to the code for multiple concerns that is tangled together, forcing an object to deal with multiple concerns simultaneously (adapted from [5]).

Among the several benefits of using AOP are [1]: **improve feature localization**, since the code is better structured due to enhanced design; **higher modularity**, given that AOP provides a mechanism to address each concern separately with minimal coupling; **easier system evolution**, as AOP modularizes the individual aspects and makes core modules oblivious to the aspects. In order to add a new feature, a new aspect needs to be included and no change to the core modules is required; **improve code reuse**, since each aspect is implemented as a separate module, each module is loosely coupled. In addition, it is possible to change a system by only changing the weaving specification instead of core modules.

Due to the nature of AOP, aspects that are *woven* into the system may not serve its intended purpose. In addition, unforeseen bugs may be embedded in the software if it is not modified cautiously. Verifying the correctness of a *woven* program is not simple because behaviors, such as performance and synchronization, are influenced by aspect descriptions [3]. In addition, aspects behaviors' depend on the *woven* context and cannot be tested separately. Moreover, existing object-oriented testing techniques are not adapted to test aspect-oriented programs [4]. For this reason, testing aspect-oriented programs is a huge challenge.

To better understand the topic addressed by this paper, it is necessary to briefly introduce the most important characteristics of AOP. AOP defines standard terminology such as *join points*, *pointcuts*, *advices* and *aspects*. A ***join point*** represents a position in the execution of a program where additional behavior can be

added. In other words, it stands as the place where the crosscutting actions are *woven* in. A ***pointcut*** is a set of join points. It is able to select (query) particular *join points* and collect their context. The code to be executed at a *join point* that has been selected by a *pointcut* and thus changing the behavior of the program is known as ***advice***. Finally, an ***aspect*** represents the encapsulation of a crosscutting concern. Hence, *pointcuts* and *advices* are combined in an *aspect*.

This paper is organized as follows: section 2 identifies quality issues regarding aspect-oriented programming; section 3 introduces the key issues of testing aspect-oriented programs; section 4 presents a set of proposals for assuring software quality in aspect-oriented programming, based on the issues identified before; and finally, section 5 draws conclusions and describes possible future work directions.

2 Aspect-Oriented Programming Quality Issues

The definition of software quality can be made from several different perspectives. According to [6], there are five major perspectives, namely: *transcendental*, *user*, *manufacturing*, *product*, and *value-based*. From a *transcendental* perspective, quality is hard to define or describe in abstract terms, but it can be recognized if it is present. It is often related with a number of indefinable properties that delight users. However, in the *user* perspective, quality is fitness for purpose or meeting user's needs. In a *manufacturing* view, quality stands for conformance to process standards. Though, in a *product* perspective, the focus goes to inbuilt characteristics in the product itself. By doing so, it is expected that the controlling of these internal quality indicators will improve external product behavior. Finally, in the *value-based* view, quality stands as the customer's motivation to pay for software.

Typically, in what software quality is concerned, people have different expectations based on their roles and responsibilities [6]. People can be categorized in two main groups: *consumers* of software products and *producers* of software. Regarding consumers expectations, they look forward that the software will perform as specified, satisfying consumer's needs. In addition, consumers also expect that the software will work over a long period of time (reliability). On the *producers'* side, the most basic quality question is to accomplish their contractual obligations by producing software that conforms to its specifications. Furthermore, good quality is often associated with superior product designs that not only preserve conceptual integrity of software components but also decrease coupling across different components.

We (as users/consumers) often expect software to be trustworthy. There are situations in which the existing code should be updated to become more trustworthy [2]. It is very important to realize that the task of updating software to accomplish more trustworthiness generally has a crosscutting nature. This can be achieved by using Aspect-Oriented Programming. However, the **incorrect** and **undisciplined use** of AOP methodology, can lead into several quality issues. With AOP, when new code blocks are woven into the target application, several risks arise. Since the target application will have its behavior changed, it can cause an impact on well known software qualities, such as reliability, functionality, performance and

efficiency. In addition, software correctness might be affected, causing redundancy in the system. Further, an important goal is to obtain the architecture correct. Software architecture encloses the most important decisions that make a project successful. It contains decisions about the structure of the system, about functionality, performance, reliability and requirements. AOP also has impact on the architecture of a software system. If it is applied correctly, it can simplify the maintenance and expansibility of the system. Otherwise, it can affect system's performance, correctness and reliability.

Obliviousness is seen by some authors as a necessary property for AOP [7, 8], but not all. Obliviousness means that developers should build aspects without needing to be aware of other concerns. Although attractive, this property may sometimes represent a problem. When new aspects are added into the application, conflicts between aspects may arise due to the obliviousness property of AOP. For instance, if an aspect with certain behavior is already implemented in the system and if another aspect is added and changes the behavior of the first aspect, it will lead to the incorrect system's behavior.

From another perspective, AOP can improve quality [9]. Efficiently and disciplinarily implementation of crosscutting concerns may provide benefits to quality. One of reasons is due to enable releasing resources and thus allowing developers to focus on the core implementation. In addition, due to better modularization, AOP improves the implementation's comprehensibility and simplifies the process to integrate new requirements and to accommodate changes to current ones. Furthermore, software quality is able to improve since AOP provides:

Clear responsibilities for individual modules. Modules appear with clearly defined responsibilities due to the encapsulation of the code for crosscutting concerns. Consequently, this partitioning improves comprehensibility of each module, simplifies the initial design stage, and facilitates the implementation of requirements changes. Further, each developer can focus on building modules of their own expertise, reducing the effort required for collaboration between experts from multiple domains.

Consistent implementation. Conventional implementations of crosscutting concerns are known for their inconsistency. This inconsistency is hard to discover because the code is scattered and tangled through multiple modules.

Improved reusability. In conventional implementations, core modules most often aren't reusable. Mismatched requirements in their crosscutting concerns appear as the main cause for the lack of reusability. However, AOP allows more combination and matching due to the separation of core concerns from crosscutting. AOP facilitates reusability, such as enabling simplified ways of implementing design patterns and other reusable constructs.

Improved skill transfer. Most of the concepts learned from AOP are reusable and transferable across a diverse range of applications that may vary from enterprise to embedded, hence reducing training costs. Typically, when developers are required to learn about a new framework, they have to start from scratch, every time. This situation does not occur with AOP. Although developers might have the need to learn

several AOP languages, the core concepts and many design patterns are universally applicable.

3 Aspect-Oriented Programming Testing Issues

Testing plays an important role in software development process and cannot be perceived as a separate activity performed at the end. The definition of software testing is not consensual. Myers [10] identifies software testing as the process of executing a program or system with the intent of finding errors. Another definition is provided by Hetzel [11] claiming that software testing involves any activity aimed at evaluating an attribute or capability of a program or system and determining that it meets its required results. Testing is not a simple task to perform. In fact it is rather complex, critical and challenging [12].

Software testing is used to accomplish quality assurance, verification and validation [13]. Quality assurance refers to a well defined set of activities required to provide confidence that processes are established and continuously improved, in order to produce products that meet specifications and are fit for use. Verification ensures that the software system conforms to its specification, guaranteeing the rightness of the product. Alternatively, validation evaluates if the software system performs accordingly to the specified requirements. In summary, the main goal of software validation (verification and validation), is to demonstrate that the system conforms to its specification and that the system meets user's expectations.

3.1 AOP characteristics

AOP is an emerging programming paradigm and due to its evolution, the difficulty of testing AOP is receiving more attention. Likewise, when OO was first introduced as a new programming paradigm, it brought a unique set of both benefits and challenges. Thus, several testing techniques have been researched, improved and applied to OO through the years. It is fairly clear that AOP has different characteristics than OO. Such characteristics can pose new issues and therefore new challenges regarding testing. The following issues regarding testing Aspect-Oriented Programs have been identified in [14]:

Aspects do not have independent identity or existence. Aspects depend upon the context of some other class for their identity and execution context.

Aspect implementation can be tightly coupled to their woven context. Aspects depend on the internal representation and implementation of classes into which they are woven. Changes performed on these classes might broadcast to the aspects.

Control and data dependencies are not readily apparent from the source code of aspects or classes. When aspects are woven into the target application, neither the

resulting control flow nor the resulting data flow structure is clear. In fact, it can be unpredictable.

Emergent behavior. Faults can be located at several sources, making hard to find their root. Further, such root can lie in the implementation of a class or an aspect, or it can be a side effect of a particular weave order of multiple aspects.

Furthermore, it is reasonable to claim that it is not clear how to test effectively aspect-oriented programs due to their characteristics, and also to determine if such programs have been tested enough. From the issues identified above, such challenges cannot be addressed using traditional unit or integration testing techniques [14]. Typically, most of unit testing techniques are applicable to the core classes (OO) but not to aspects. As a result, such techniques cannot be applied to aspects essentially because aspects are not complete code units and their behavior often depends on the woven context [14]. Integration testing exercises interfaces among units with a specified scope to demonstrate that the units are collectively operable. Units are physically dependent or must cooperate to meet a requirement. Regarding AOP, the concept of integration is more fine grained and occurs with respect to the intra-method control and data flow. As such, there are no well-defined interfaces.

The unique features of AOP don't manifest in OO or even in procedural programming. Each feature can introduce new **fault types** that may lead to failures. Hence, it is not apparent how faults and failures occur in AOP.

3.2 Aspects conflicts and interferences

The aiming of AOP towards *obliviousness* has been harder to attain, as **conflicts** between aspects may arise. On the topic of OO, most of these conflicts are avoided by using encapsulation techniques and unit testing or design by contract approaches [15]. However, with AOP, such encapsulation is no longer valid. In addition, in AOP when new behavior is added into the system by an aspect, it can break previous tests. For instance, if an aspect was already tested and if another aspect is added to the system that conflicts with the first, then the first aspect can no longer prove its correctness thus affecting system's overall quality.

Another issue refers not only to detect such conflicts but also to understand their **interactions** (between aspects and base classes). It is necessary to understand how new added behavior can cause conflicts and how to detect its source.

3.3 Problems with pointcuts

Pointcuts are one of the main constructs of AOP. As explained early on, pointcuts contain specifications that match the join points of a particular type according to a signature that includes a pattern, and thus performing a specific action (advice) when triggered. When dealing with large and complex programs, developers may overlook the purpose of pointcuts. That is, they may create pointcuts that may not be correct or might fail its intended purpose. Developers might write pointcut expressions with

incorrect strength [19]. As such, one of the two scenarios will most likely occur: (i) if the pattern is too strong some required join points will not be selected (ii) if the pattern is too weak then additional join points will be selected that should be ignored. This incorrect strength causes aspects to fail.

Therefore, testing the strength of pointcuts needs to be addressed with the aim of validating the correctness in their expressions, thereby evaluate the effectiveness of the test suite.

3.4 Undisclosed type of errors / bug patterns

Due to the complexity introduced by AOP, such as behavior modification, it turns out to be difficult to detect and locate errors in aspect-oriented programs. Some bugs may be hidden on either aspects or/and in base classes. In addition, bugs may appear in a program woven by a weaver even if bugs do not exist in individual aspects or objects [17]. These kinds of bugs are caused by weaving policies such as the order weaving followed. The distinctiveness nature of AOP, leads testers to not have sufficient knowledge on how to write efficient and proper tests, in order to cover most common errors that appear in aspect-oriented programs.

3.5 Recurring / Symptomatic issues

At present AOP is being used to build generic aspects in order to monitor particular crosscutting properties for a wide range of uses, such as program tracing, runtime assertion and logging. Such diversity enhances additional difficulty on the task to identify aspects and build them for testing as **test oracles**. How to build specific testing aspect which can be identified as test oracles becomes the key problem in building aspect-oriented unit testing practical [18]. In addition, current specification-based tests generation methods are only able to test program behavior specified by terms of **invariants**. However, invariants are only addressed to functional behaviors. Xu and Yang [18] state that patterns related to non-functional properties of the program cannot be modeled, and therefore, existing specification-based test generation cannot test these special aspects of the program.

Another question relies on how to define proper *concern* coverage. The idea is to express the test adequacy criteria relative to crosscutting concerns.

Other difficulties are related with the **aspect composition order**, with the **inter-type declarations**, and with the changes in normal and exceptional **control flow**, possibly introduced by aspects [20]. In addition, the same authors identified that due to AOP requiring execution points to be intercepted by an aspect, to be specified in the aspect itself, it will most likely cause faults. For instance, when the execution to another code block is redirected, two situations may arise. The first refers to the contract breakdown between caller and callee in a method execution. The second, points to the fact that such situation might alter the state of a current object in an inappropriate way. Moreover, changing the execution order of statements might also defy some assumptions required for a given statement to work properly. The execution flow of a program can be altered by an aspect as well, possibly in an undesirable way. By using inter-type declarations in the aspects, new methods and

fields can be added to a class of the base system. Such implementation might as well lead to **inheritance relationships** to be altered and **interface implementations forced** [22].

An additional question that arises is how to identify which code blocks are directly affected (by changing its behavior) with the AOP code itself. That is, how to determine which code portions can be or should be retested when the aspects are added into the system.

Since base class tests are not necessarily valid for testing aspect-oriented programs, because aspects likely change transitions of object states, it is not clear to what extent can base class **tests be reused for testing aspects** [21]. This also leads to another challenge: when a fault occurs, is it related to the aspects or it has to do with the base classes?

Furthermore, whenever large number of unit or integration tests are generated and executed, the correctness of these test executions is still unknown, if there are no specifications for checking their correctness.

4 Improving Software Quality Assurance in AOP

A great deal of research for testing aspect-oriented programs has been conducted over the past years. In the previous section, a quite reasonable extent of issues has been identified. This section features some suggestions that might help to solve some challenges/issues identified before, and therefore aiming to improve software quality in AOP.

4.1 Fault Model

One approach that reflects the structural and behavioral characteristics of aspect-oriented programs was identified in [14], aiming that criteria and strategies for testing AOP should be developed in the terms of a **fault model**. This fault model for AOP is related the nature of faults in aspect-oriented programs and the unique characteristics of AOP. The aim of this proposal is to take an effective step towards systematic testing of aspect-oriented programs. As such, Alexander and Bieman [14] proposed a candidate fault model for AOP, as well as brief descriptions of the proposed testing criteria for each:

Incorrect strength in pointcut patterns. Such faults can lead only aspects to fail and not the core base classes' functionality. Thus, a test of the aspect is required.

Incorrect aspect precendence. These faults are related by the weave order. They will occur when multiple aspects interact. Testing all weave orders is proposed.

Failure to establish postconditions. These can cause core concern methods to fail their execution. Re-test all methods that have code weaving using the original test set is suggested.

Failure to preserve state invariants. These faults can also cause core concern methods to fail. As mentioned before, re-testing concern methods seems the solution.

Incorrect focus of control flow. These can source advice to activate at the wrong time. In order to reveal these faults, a form of condition coverage of pointcut designators can be considered.

Incorrect changes in control dependencies. These faults will affect core concern behavior, therefore re-test all core concern methods is proposed.

4.2 Bug Patterns

Bug patterns are referred as erroneous code (can be due to bad programming practices) that constantly persists failing. In [16] authors propose a **bug pattern** identification for AOP. This approach identifies what types of bugs are unique or are common related to aspect-oriented programs. In addition, the bug pattern identification can help language designers and tool developers to develop the corresponding bug finding techniques or bug detectors, which can be applied to locate syntactic bugs by program analysis. However, this approach appears in relation with with *AspectJ* [26]. *AspectJ* is the most popular and best supported extension to the Java programming language, adding to Java, aspect-oriented programming capabilities. Authors introduce six bug patterns in AspectJ:

Infinite loop. This bug pattern is due to accidental matching of unexpected join points, since the selection of join points relies on lexical-level matching.

Scope of advice. When a parameter is reassigned it will have no effect outside of the advice. The state of the parameters won't change and modifications outside of the advice won't be reflected when misunderstanding the scope of advice.

Multiple advice invocation. The execution sequence of advices might affect the result of a program. If the advice precedence declaration is missing, it will cause misunderstanding to the developer and to the *AspectJ* compiler as well. Furthermore, it does not include any type of error report.

Unmatched join point. Typically in a pointcut expression, when the wildcard (** or ..) is used for matching join points, and the type is declared in *args()* primitive pointcut does not match with the intended one. This type unmatch will cause a problem.

Misuse of *getTarget()*. This bug pattern is due to *AspectJ* reflection mechanism. For instance, the *getTarget()* method of *thisJoinPoint* object will return NULL when used inside of a static method. This result (NULL) may lead to a *NULLPointerException* or a casting failure.

Introduction interference. Changes performed in class behavior due to the introduction of new members in a class hierarchy (dynamic interference) can alter the class runtime behaviors unexpectedly.

4.3 Unit Testing

Unit testing can be used to **detect conflicts** between aspects. This challenge was addressed in [15], by means of a methodology in order to improve management of conflicts and also by a tool to support the methodology. When a new aspect is weaved into the system, it might change the behavior of previous aspects that were weaved before, and thus unit tests for that specific module, most likely have been broken. Subsequently, in [15] authors claim there is a need for aspects to be able to perform three main activities, namely: (i) to announce which aspects they are expected to break; (ii) which aspects they depend on; (iii) and which they are adding to the system. Therefore, authors propose that aspects should be able to make such announcements by means of Java annotations.

4.4 Mutation Testing

In order to **test pointcuts** for their strength, Anbalagan and Xie [19] proposed to apply mutation testing. Mutation testing is a fault-based technique that can be used to inject faults into a program. Its purpose is to help to locate and expose weaknesses in test suites. The proposed mutation testing of pointcuts is performed in two steps: create effective mutants of a pointcut expression and test these mutants using the designed test data. To reduce human efforts in mutation testing of pointcuts – due to a large set of mutants - authors developed a framework to automatically generate relevant mutants instead of arbitrary strings. In addition, this framework classifies mutants and detects equivalent ones. With the unchanged aim to **test pointcuts**, the same authors in [23] proposed APTE, an automated framework that tests pointcuts in AspectJ programs with the support of AJTE, a unit-testing framework without weaving. This approach not only verifies the correctness of pointcut expressions in existing or current versions of base code, but also identifies boundary join points.

4.5 Test Oracles

In [18], Xu and Yang, guided their effort in solving the problem that relies on how to build specific testing aspects which can be identified as **test oracles**. They present an approach to generating the unit testing framework and test oracles from aspects in AOP [18]. In addition, a new concept, *application-specific test*, is introduced. This concept stands as the separation of concerns on specific application of common aspect-oriented programs' aspects. In order to build the application-specific aspects for testing, authors discuss *Aspect-Oriented Test Description Language* (AOTDL). AOTDL is able to specify the properties for testing that can be translated into the common aspects in *AspectJ*. After weaving and compiling programs, unit testing

codes are then generated automatically, and serve as test oracles [18]. The test outcomes are decided on exceptions thrown by testing aspects.

4.6 State Based Incremental Testing

An incremental approach to test if aspect-oriented programs and their base classes conform to their respective **behavior models** is presented in [21]. Authors use an aspect-oriented extension to state models in order to capture the impact of aspects on state transitions of base class objects. In addition, they also exploit a weaving mechanism for composing aspects into their base models. Their work demonstrated for a majority of base class tests can in fact be reused for aspects, but some modifications to some of them are required. Further, authors state that their incremental testing approach is somewhat similar to traditional regression testing, with the difference that aspects are feasible to investigate systematic reuse and modification of the existing tests.

5 Conclusions

This paper discussed software quality issues introduced by aspect-oriented programming. In addition, not only issues have been identified but also a perspective on how AOP can improve quality, when applied correctly, was also given. Accordingly, a set of key testing issues that arise with AOP as well as solutions to improve quality in AOP were addressed.

Most of the proposed approaches to achieve and enhance software quality in AOP do not solve every problem, and existing ones cannot be seen as a "silver bullet" towards testing aspect-oriented programs. Furthermore, most of the existing work on testing aspect-oriented programs is being directed only towards one of the existing programming languages for AOP – *AspectJ*. In the same way that object-oriented programs can be developed and tested, for instance, in Java and .NET, the same remains valid for aspect-oriented programs. In fact, some the testing issues identified in this paper appear related with *AspectJ*. However, it is not clear if such issues also apply to other AOP tools such as, *PostSharp* [24] and *Spring.NET* [25]. In addition, these tools can also lead to other unknown issues.

There is still a lot of research to be conducted in order to discover all issues as well as solutions regarding testing aspect-oriented programs. A future work effort could be directed on the innovation on how to apply other testing methodologies such as FIT (Framework for Integrated Test) [27] to AOP. Another idea that could be explored was firstly discussed in [28] and refers to the usage of aspects for testing aspects.

References

1. Laddad, R.: AspectJ in Action: Practical Aspect-Oriented Programming. Manning Publications (2003)
2. Safonov, V.O.: Using Aspect-Oriented Programming for Trustworthy Software Development. Wiley-Interscience (2008)

3. Ubayashi, N., Tamai, T.: Aspect-oriented programming with model-checking. In: Proceedings of the 1st international conference on Aspect-oriented software development, Enschede, The Netherlands (2002)
4. Badri, M., Badri, L., Bourque-Fortin, M.: Generating unit test sequences for aspect-oriented programs: towards a formal approach using UML state diagrams. In: Enabling Technologies for the New Knowledge Society: ITI 3rd International Conference (2005)
5. Robinson, D.: Aspect-Oriented Programming with the e Verification Language: A Pragmatic Guide for Testbench Developers. Morgan Kaufmann (2007)
6. Lewis, W.E.: Software Testing and Continuous Quality Improvement. Auerbach Publications (2004)
7. Filman, R., Friedman, D.: Aspect-oriented programming is quantification and obliviousness. In: Workshop on Advanced Separation of Concerns (2000)
8. Filman, R.: What is aspect-oriented programming, revisited. (2001)
9. Laddad, R.: Aspect-Oriented Programming Will Improve Quality. In: Quality time (2003)
10. Myers, G. J., Sandler, C., Badgett, T., Thomas, T.M.: The Art of Software Testing, Second Edition. Wiley (2004)
11. Hetzel, B.: The Complete Guide to Software Testing, Wiley (1993)
12. Binder, R.V.: Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison Wiley (1999)
13. Parizi, R.M., Ghani, A.A.: A Survey on Aspect-Oriented Testing Approaches. In: Fifth International Conference on Computational Science and Applications (2007)
14. Alexander, R.T., Bieman, J. M., Andrews, A.A.: Towards the Systematic Testing of Aspect-Oriented Programs (2004)
15. Restivo, A., Aguiar, A.: Towards Detecting and Solving Aspects Conflicts and Interferences Using Unit Tests. In: Proceedings of the 5th workshop on Software engineering properties of languages and aspect technologies (2007)
16. Zhang, S., Zhao, J.: On Identifying Bug Patterns in Aspect-Oriented Programs. In: Computer Software and Applications Conference (2007)
17. Ubayashi, N., Tamai, T.: Aspect-Oriented Programming with Model Checking. In: Proceedings of the 1st international conference on Aspect-oriented software development (2002)
18. Xu, G., Yang, Z.: A Novel Approach to Unit Testing : The Aspect-Oriented Way. In: International Symposium on Future Software Technology (2004)
19. Anbalagan, P., Xie, T.: Efficient Mutant Generation for Mutation Testing of Pointcuts in Aspect-Oriented Programs. In: Proceedings of the Second Workshop on Mutation Analysis (2006)
20. Ceccato, M., Tonella, P., Ricca, F.: Is AOP code easier or harder to test than OOP code? (2005)
21. Xu, D., Xu, W.: State-Based Incremental Testing of Aspect-Oriented Programs. In: Proceedings of the 5th international conference on Aspect-oriented software development (2006)
22. Bernardi, M.: Reverse Engineering of Aspect Oriented Systems to Support their Comprehension, Evolution, Testing and Assessment. In: 12th European Conference on Software Maintenance and Reengineering (2008)
23. Anbalagan, P., Xie, T.: APTE: automated pointcut testing for AspectJ programs. In: Proceedings of the 2nd workshop on Testing aspect-oriented programs (2006)
24. PostSharp, http://www.postsharp.org/ (2008)
25. Spring.NET, http://www.springframework.net/ (2008)
26. AspectJ, http://www.eclipse.org/aspectj/ (2008)
27. FIT: Framework for Integrated Test, http://fit.c2.com/ (2008)
28. Sokenou, D., Herrmann, S.: Aspects for Testing Aspects? In: 1st Workshop on Testing Aspect-Oriented Programs (2005)

A prototype tool for supporting joint-design collaboration in requirements specification

Cristóvão Sousa [1]

[1] INESC Porto - Instituto de Engenharia de Sistemas e Computadores do Porto
Campus da FEUP, Rua Dr. Roberto Frias, 378 4200 - 465 Porto Portugal

Abstract. Ontologies are a technological key factor regarding the knowledge management domain. This paper presents a graphical-based knowledge representation approach using concept maps towards capture of requirements in work organization domain, which was translated into a Content Management System in order to manage the work design information. Some aspects related with the advantages of visual approaches for collaborative development of ontologies are discussed.

Keywords: Conceptual Modeling, Ontologies, Requirements Engineering, Content Management Systems.

1 Introduction

The more specific the requirements are greater the quality of the software product. This is a common view among the most developers nowadays. Requirements analysis is critical for the success of the development of a new software product.

Unless you are already an expert in the field, as a software engineer you need to begin to understand your client's area of expertise before you can begin making decisions. Somehow, you need to learn enough to get going. Hence, a multidisciplinary cooperation is a new challenge and a new category of methodological tools supported by modern knowledge management technologies is needed in order to make feasible such complex development processes. In this context ontologies play a key role by providing a shared conceptual model of a specific domain, that means, it provides the vocabulary which entails the type of objects and concepts that exist in a specific domain and their properties and relations. The goal of this paper is to show how graphical representation of knowledge provides a very effective means for presenting information to both users and system developers. The collaboration and sharing work design development efforts in CODEwork@vo are used as our empirical inputs. This work is organised as follow:

This first chapter is initiated with a brief introduction to the developed work, presenting its general goals.

Chapter 2 surveys the specification of information systems requirements through conceptual modelling. In this context we'll find a short description of the main knowledge representation (KR) formalisms.

Chapter 3 and 4 begins with a more specific approach to the requirements representation, concentrated on modelling information about a specific domain of discourse.

Chapter 4 discloses a practical and possible application scenario of the developed systems requirements.

2 Conceptual modelling and requirements engineering

Conceptual modelling can be the process of creating patterns interrelating concepts but not expressed in a mathematical point of view neither concerned with quantification. Conceptual Modelling is about qualitative assumptions about concepts. Diagrams, such as maps, graphs, and flowcharts, can be used in concept modelling.

According to [1] the information systems requirements have two sources:

- User-defined requirements which arise from people in the organization and reflect their goals, intentions and wishes
- Domain-imposed requirements which provide requirements reflecting real world facts and constraints on the designed system.
- Conceptual modeling is well suited for modeling information about the domain of discourse.

Conceptual modelling is the first phase of the two-phase organisation life-cycle (see Fig. 1). It aims at abstracting the specification of the required information system, i.e., the conceptual schema, from an analysis of the relevant aspects of the universe of discourse about which the user community needs information. The succeeding phase, that of system engineering, uses the conceptual schema to design and implement a working system which is verified against the conceptual schema. Conceptual modelling is situated in the broader view of information systems requirements engineering. [2]

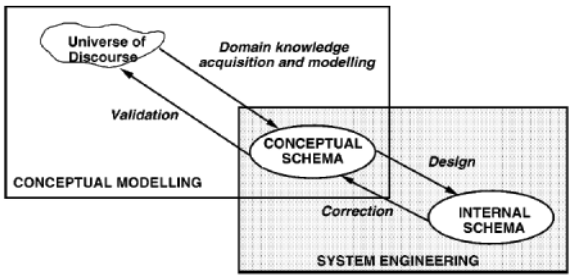


Fig. 1. Conceptual Modelling vs. System Engineering. Source: [Rolland, 2000]

In the organizational domain we may have exactly the same approach in a way that all structure of work organization can be explained trough the design of its Universe of Discourse. However, nowadays organizations are in constant mutation due the global environment and market pressures. Companies, work more often spread geographically forming a network in a complex value chain where its elements may change regularly. In this context, new forms for conceptual modelling supported by ICT in a collaborative environment are needed. This drives us to the specification of first general system requirement, which is explicit knowledge representation.

According to [2], within the scope of conceptual modeling is the requirements specification, as well as any additional knowledge not in the form of requirements.

2.1 Explicit Knowledge representation requirements

According to Patrick Hayes [3], "Knowledge Representation (KR) refers to general topic of how information appropriately encoded and utilized in computational models of cognition". KR has two main work areas which are Knowledge Representation Formalism and Knowledge Engineering which refers to the development and maintenance of Knowledge Based Systems. The capability of knowledge representation and knowledge sharing is the main challenge in all knowledge management process which involves knowledge acquisition, knowledge representation, knowledge maintenance and knowledge use.

The first step in conceptual modelling of a specific domain is the definition of its conceptual terms and structure. The "structural schema" is a set of objects, their properties and relationships between objects and properties. The representation of those terms and their relationships can be achieved by knowledge representation formalisms.

Semantic networks, conceptual graphs, rdf, topic maps and concept maps are the graphical knowledge representation formalisms briefly presented, which may also be used for ontology development. Visually, a semantic network is a set of nodes which represent concepts and instances, connected by arcs which represent relations between nodes. According to [4], semantic networks are a powerful knowledge representation system because they are easy to understand by humans and can be used in automated processing systems which means that they can also became a vehicle to archive organization knowledge. While visual tools for ontology construction, semantic networks provide the conceptual representation of a domain allowing the explicit representation of concepts, relations and instances. Semantic networks have also mechanisms for formal representation of knowledge. However, the interpretation of a semantic network may cause some confusion. No distinction is made between different types of links and the distinction between concepts and objects is not clear [5]. Due its lack of formal semantic characteristic, there are many variances of

semantic networks. Conceptual Graphs (CGs), by its turn, is a formal logic-based knowledge representation developed by John Sowa. CGs are, in fact, a variation of semantic network combined with logic! CGs are a very powerful and versatile tool for knowledge representation. They are human readable and machine process able. Nevertheless it is not possible to draw a CG without having a basic knowledge of logic and CGs itself. Resource Description Framework (RDF) is a framework for representing information in the web. RDF has the capability to formal express the data meaning allowing interoperability and provides an integration environment between different patterns of metadata. RDF language was created to represent a simple data model based on XML and using vocabulary based on Uniform Resource Locator (URL). The data model is graphically represented through triples which consist in three types of objects that describe relationships between resources regarding properties and values. In terms of ontology representation, RDF allows the explicit representation of resources (concepts), properties and statements. RDF expresses the meaning of data allowing interoperability in the web. It is therefore a knowledge representation formalism which provides the structure that is used to represent data models for objects and their relations. However it needs RDF Schema in order to provide mechanisms to declare properties and define relations between properties and resources. RDF Schema is used to describe, semantically, properties, classes of web resources and the type of data for the property values. It extends RDF with new vocabulary allowing the knowledge to be represented through ontologies. Regarding Topic Maps (TM), they have a great expressive power. In some sense they are a reformulation of semantic networks and conceptual graphs. Additionally they offer a new and standard way of encoding and exchanging knowledge. Technically a TM is formed by three concepts: topic name, association and occurrence. A Topic can be everything - an object, as person, a concept, etc. The association indicates how a topic is related with other topics. Each topic involved in an association, is said to play a rule [2]. In terms of ontology representation TM can represent facts, procedures, concepts and complex relations between concepts and real world occurrences. It is possible to represent knowledge in a formal way. However, and despite of TM's flexibility, they are very difficult to manage.

Conceptual Maps (Cmaps), are able to represent meaningful relationships between concepts linked by words to from a semantic unit [6]. The concepts are included in circles or boxes while relations between concepts are represented by links connecting the boxes. The links are labelled, describing the relation between two concepts. Propositions result from the phrases composed by the concepts and the link label (concept - verbal phrase - concept). According to [7] Cmaps are very useful in facilitating the visualization and discussion, and in providing domain experts with a tool that could be used to declare the primary elements of their knowledge. Cmaps' simplicity and explicitly make them very useful in several areas namely: Knowledge organization and creation; Collaborative learning; Domain summarization; Browsing tool.

3 Domain conceptualization with cmaps

Regarding system engineering, the customer is the driver of requirements and in this context plays the role of the domain expert. Usually there is a lack of understanding of the systems requirements between the systems engineer and the customer. This communication gap, leverage to the failure of information systems implementation. This happen because the engineer is focused on the system and as such requirements and system functions are presented to the customer from the systems point-of-view and in the language of the engineering disciplines. At the same time, the customer's emphasis is on the mission outcome, not on the tasking details to accomplish that outcome. Even though the engineer wants to understand the desired customer outcome, he queries the customer in terms of the system (i.e.: doing the tasks.) They should focus on the desired system objectives rather than the system tasks. [8]

It is absolutely imperative a functional accordance between the stakeholders.

In order to improve communications between engineers and customers (domain experts), some efforts were made. The Use Cases, for instance, are an example attempting to tackle the problem. However, the problems still remaining. A better way is needed to deal with this difficulty. In this line of thinking, we think Concept Maps do fairly the job.

Concept Maps (Cmaps) can use the most appropriate language to create, display, study, discuss and refine what the customer's needs really are and what the systems engineer will supply that will meet those needs. Afterwards, they can both test the final delivery with confidence against what was agreed upon in the Cmap. Concept Maps, however, are a deceptively simple and elegant solution to the customer / engineer requirements breach. Peers attain the elegance of this solution simply by establishing mutual understanding of the real problem. [8]

4 Using ontologies for requirements capture

An informal representation can illustrate information explicitly, but only through a formal way it is possible to share and interpret knowledge among computer systems increasing interoperability and reusability of the conceptual requirements designed. Well, this is where the Cmaps sin lies! Cmaps do not include a formal language, but just a set of recommendations for their construction; nevertheless, that lack of a

formal language in knowledge representation is what makes Cmaps so easy to use for everyone.

Ontologies play in this domain an important role. According to [9], ontologies can be applied in the software engineering lifecycle. Within the analysis and design phase, ontologies can be use in the requirements engineering task describing requirements specification and formal represent requirements knowledge. “In most cases, natural language is used to describe requirements, e.g. in the form of use cases. However, it is possible to use normative language or formal specification languages which are generally more precise and pave the way towards the formal system specification. Because the degree of expressiveness can be adapted to the actual needs, ontologies can cover semi-formal and structured as well as formal representation. Further, the “domain model” represents the understanding of the domain under consideration, i.e. in the form of concepts, their relations and business rules. In its simplest form, a glossary may serve as a basis for a domain model. However, it can be formalized using a conceptual modelling language such as the UML. Moreover, the problem domain can be described using an ontology language, with varying degrees formalization and expressiveness. ” [9]

Within the implementation phase, ontologies can be applied in the following areas:

- Integration with software modelling languages thanks to the ontology standards such as OWL and RDF.
- Coding support, here ontologies provide a globally unique identifier for concepts. While at the programming level it is convenient to have a limited set of data "types" like strings, that can be used for multiple purposes, an ontology enables developers to annotate API elements with an unambiguous concept. A potential drawback is the extra-effort for modelling the semantic layer. In the case of APIs, this is partially eased since an initial modelling effort scales well with the estimated reuse. However, the question of incentives for someone to semantically describe an API still remains. [9]
- In code documentation ontologies provide a unified representation for both problem domain and source code, thus enabling easier cross-references among both information spheres. Moreover, it is easy to create arbitrary views on the source code (e.g. concerning a variable). Reasoning is applied to create those views, e.g. to find all places where a variable is accessed either directly or indirectly. [9]

Any way, a version of CmapTools called CmapTools Ontology Editor (more detailed in the next section) has been developed that allows users to work with ontologies. CmapTools COE keep the fundamental characteristics for concept construction without a restriction to a hierarchical model, since relations between concepts are transversals and arbitrary. COE keep also the interaction with the model allowing seeing, navigating and editing. The main change was the attempt to give to COE more interoperability and formalized notations meeting ontological agreements in order to allow the producing of information in more formats such as OWL and RDF, permitting as well reusing and sharing information. In this way COE came to

cover a gap between the informal nature of conceptual maps and the formal nature of machine-readable ontology languages [10].

4.1 CmapTool COE

CmapTools Ontology Editor (COE) software provides via Concept Maps a complete collaborative and argumentative environment, based on graphical direct manipulation and representation to facilitate the continuous exchange of information among domain experts. COE was developed in order to give CmapTools more interoperability and formalized notations meeting ontological agreements, that allows users to work with ontologies.

Cmap Tools COE can be used as:

- An ontology viewer;
- An ontology editor;
- A concept search engine.

[12] COE can display any OWL or RDF ontology as a readable concept map, supports intuitive editing and construction of "ontology maps", allows uses to rapidly locate related concepts in published software ontology and outputs valid OWL/XML. The goal is to enable rapid and intuitive capture of machine interpretable knowledge by combining navigation, comprehension, selection and construction of knowledge in a single collaborative environment with an intuitive GUI. [11]

The majority of ontology-authoring tools need high level of technical skills to ontology development. In many cases it is necessary some OO (Object Oriented) sensibility. The friendly graphical interface from COE in opposition to text-based interfaces from other editors makes the mental processing of ontology, easier. The Protégé OWL, has already the capability of graphical visualization of the ontology but with a higher complexity of abstraction. COE allows the capture of knowledge structures using templates which can be dropped directly onto ontology map canvas. The templates are graphical representations for commonly used OWL structures. With these templates the knowledge construction is much more rapid and easier.

However when we want to export our ontology to OWL, is not always a peaceful process specially when we use templates for some constructors such as owl:intersectionOf (New Class All Of Definition template), owl:unionOf (New Any Of Class Definition template). Concept Maps COE tool is not specific for knowledge representation languages. Its application is much more generic to cover new developments. COE is not only focused on the conceptualization and formalization of specific domain ontology, but on all network. COE is not a tool only for software specialists; it's a tool for all domain experts.

In fact, COE it's a very useful tool, but, regarding ontology building, Protégé is, so far, the most used tool. Indeed, Protégé is very powerful and capable software, however it lights much on a technical view and "disregard" a crucial aspect that is

graphical direct manipulation of ontologies. In order to minimize this weakness, several plug-in have been emerging trying to endow Protégé for cognitive support on ontology development. All those efforts improved Protégé; however it is not possible yet to have a complete visual authoring feature, which means that Protégé do not have, so far, a direct manipulation feature over graphical representations.

5 A prototype of a joint design collaboration tool

This chapter discusses the experiences obtained in CODEWORK project through the usage of CmapTools software both in the capture of work organization requirements in form of an ontology and the way it was created in terms of concepts and knowledge representation.

At the CODEwork project, the ontology development in the work organization domain had several steps; however we did not follow any methodology in particular. The next figure shows the process that we perform for creating the ontology.

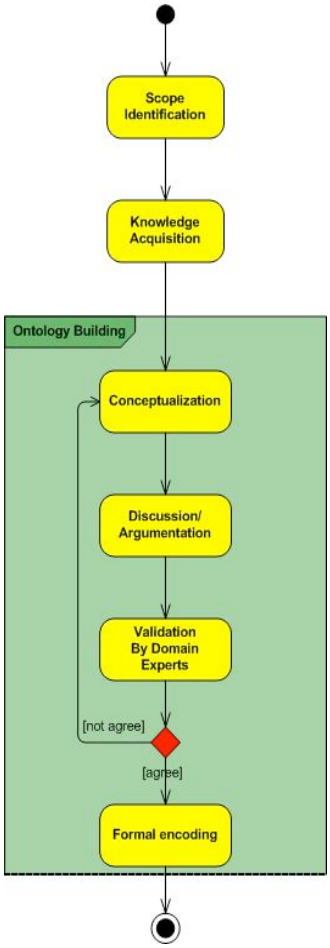


Fig. 2. Ontology development methodology

The first phase was the definition of the ontology scope. Knowledge acquisition was the next phase, where all terms about the domain within the scope range were collected. Explicit definitions and descriptions were made and the competency questions were formalized in order to see if our ontology fulfilled the main scope. Later, Cmaps were created freely, without any formal restriction. The created Cmaps represented only a particular view of the problem. The next steps were to put the map on the server for collective discussion and validation. The team argued about the conceptual map trough notes, forum or other maps, files, links…, until it is reached a common understanding. After the cmap had been validated, it was converted in a formal notation. This formal notation was achieved by using COE templates.

In this particular scenario we will describe how to set up Plone CMS content based on OWL ontologies. The objective is to reach a simple manageable hierarchical

distribution of folders and files trough a web portal, according to the concepts and relations, which compose the work organization ontology. The steps performed to achieve that goal are presented in the figure 3.

Step 1: Development of work organization ontology in COE.

Step 2: Convert owl ontology to UML trough the Protégé UML plug-in and create UML composition relations. In OWL it is not possible to specify composition relations. With Protégé UML plug-in we specified that *JobAnalysis* concept was composed by three other concepts namely: *JobSpecification*, *JobDescription* and *JobPerformane*, then we transformed the ontology into and UML project.

Step 3: This step is optional. We opened the project with ArgoUML in order to see if the UML model converted within Protégé is consistent. At this step we may apply stereotypes specifying which object types (e.g. folder, file, special folder, discussion thread, image, etc.) our classes will assume within Plone.

Step 4/5: Using ArchGenXML utility, our UML model is converted into a valid Plone product which can be installed trough Plone administrator interface.

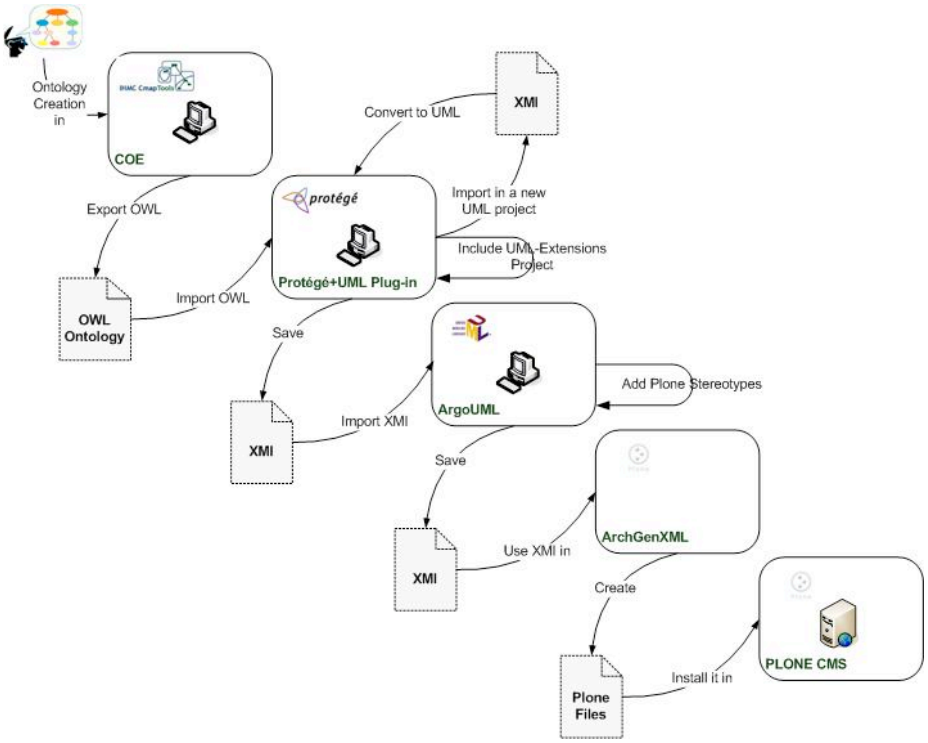


Fig. 3. Prototype Implementation Steps

6 Conclusions and future work

In this work it was described an approach to manage knowledge about work organization making use of ontologies for system requirements specification.

Through ontologies it becomes easy to involve several teams in the process of change, finding a common model, sharing knowledge and point-of-views.

Ontology is a formal specification of a domain which demands the knowledge about some technical requirements and specific representation conventions. The knowledge is not only in the ontology engineers heads, thus it is highly recommended the specification of a knowledge model supported by simple knowledge representation formalisms that allow domain specialists to discuss together about a certain domain in order to be obtained a set of concepts and its relations according with the experience, know-how and surveys of the right people without having the concern about knowledge representation technical issues.

In the first step of knowledge acquisition, Cmaps do fairly the job due its informal characteristics. However, an easy knowledge formalism is not sufficient by itself to support the construction of a knowledge model. Thence came into existence the CmapTools Software by IHMC providing a complete collaborative knowledge representation environment based on CMs. By definition ontology has formal constraints, so it is necessary to transform cmaps into a formal notation. A framework that implements a solid methodology for cmaps translation into COE formal conventions with validation mechanisms is needed. We also foresee cmaps and Ontologies as a solution for today's' new programming challenges. Nowadays, all people access to the internet to rapidly retrieve information from several different sources either for professional or personal use. This new situation asks for new ways to develop and present web content. A possible future approach could be based on concept maps as a tool to set up a CMS for implementing e.g., a web portal. At the same time, it would be interesting to have the reverse synchronous standard mechanism in which the content created in a CMS could be used to maintain and evaluate ontologies(e.g., delete/add new instances to the ontology). Interoperability Semantic Web could be the answer at a short/medium term.

References

1. Pepper, S., "The TAO of Topic Maps," presented at XML Europe 2000, Paris, France, 2000.
2. Lindland, O.I.; Sindre, G.; Solvberg, A., "Understanding quality in conceptual modeling," Software, IEEE , vol.11, no.2, pp.42-49, Mar 1994 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=268955&isnumber=6703

3. P. Hayes, T. Eskridge, R. Saavedra, T. Reichherzer, M. Mehrotra, and D. Bobrovnikoff, "Collaborative Knowledge Capture in Ontologies," in Semantic Integration Workshop. Sanibel, Island, 2003.
4. Gordon JL."Creating Knowledge Maps by Exploiting Depend Relationships", Knowledge Based Systems 2000; vol. 13: 71-79.
5. Baader F. "Logic-Based Knowledge Representation," in Artificial Intelligence Today: Recent Trends and Developments, vol. 1600/1999, Lecture Notes in Computer Science: Springer Berlin / Heidelberg, 1999 pp. 13.
6. A. Cañas, G. Hill, A. Granados, C. Pérez, and J. D. Pérez, "The Network Architecture of cmapTools," Institute for Human and Machine Cognition (IHMC) 2002.
7. Garcia A, Norena A, Betancourt,Ragan M. "Cognitive support for an argumentative structure during the ontology development process," in 9th Intl. Protégé Conference. Standford, California, 2006.
8. J. BeVier and C. Calimer , A Meeting of the Minds: A Successful Systems Engineering Experiment using Concept Maps for Effective Communications, presented at INCOSE 2005
9. Stefan S., Fzi Forschungszentrum Informatik, Universität Mannheim In 2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE 2006), held at the 5th International Semantic Web Conference (ISWC 2006)
10. P. Hayes, T. Eskridge, R. Saavedra, T. Reichherzer, M. Mehrotra, and D. Bobrovnikoff, "COE:Tools for collaborative Ontology Development and Reuse", 2006.
11. P. Mika and H. Akkermans, "Analysis of the state of the art in Ontology Based knowledge management", Vrije Universiteit, Amsterdam, Project Deliverable 14/02/2003

WordNet as a Symbolic Free Text Classifier

Gustavo Laboreiro[1] and Irene Pimenta Rodrigues[2]

[1] Faculdade de Engenharia da Universidade do Porto
gustavo.laboreiro@gmail.com
[2] Universidade de Évora
ipr@di.uevora.pt

Abstract. Text Classification is a difficult subject for small texts, where symbolic approaches can have some advantages.
This work describes a symbolic approach at solving this problem using WordNet's semantic relations for nouns, and the synergy they carry to create a new similarity measure for *synsets*. This is based on the overlapping of *synset* paths from the most general *synset* to the particular category or noun.
It is expected that this could provide good results and ample material for exploration.
The results of its application on the TRECVID 2006 data shows fair results with a good possibility of improving.

Key words: natural language processing, information extraction, document classification, document retrieval

1 Introduction

The digital universe shows no sign of slowing down, and is expected to grow tenfold within five years [1]. To cope with this, search solutions need to continue to evolve.

The two main approaches are better algorithms and augmented data. Due to several user-centric reasons [2], the latter one is rarely used outside of particular cases (like ID3 tags). Doing metainformation extraction automatically could help solve some of these problems. In the case of texts, a hint of its subject is quite helpful in search and browsing tasks. "Tiger Woods's eagle" has nothing to do with animals, but with golf.

Most text classifying systems in use appear to use supervised learning techniques. First, because the most common classification systems are SPAM filters, where naïve Bayes classifiers work very well [3,4]. Secondly, because it allows using a single implemented algorithm to work in any number of situations, changing the training set and little else. Support Vector Machines are gaining popularity mainly due to the good results they achieve with a limited number of examples [5, 6]. There are also unsupervised learning systems, like SONIA [7], that uses clustering techniques.

With the richness of any language's vocabulary, it is possible that a statistical system finds no known word present in its training data when processing a short

text. To increase the size of the training set to be near "all-encompassing" implies a much greater cost. Knowing that very small texts don't usually provide much statistically (or otherwise) significant information, their classification is a hit-and-miss affair.

Symbolic algorithms tend to run slower than their statistical counterparts, but carry with them the knowledgebase while the alternative approach needs to harvest it from a corpus. It is easy to see why symbolic approaches are more common in small text processing, like the natural language queries present in TRECVID.

TRECVID is a yearly series of conferences and workshops in video retrieval. During this event several teams are expected to present video segments that match a certain query expressed in natural language. Like "Find shots with one or more people leaving or entering a vehicle". To do so, the contending systems need to assert what to look for, matching the query to a number of predefined tags, that will help them look through the videos.

Unfortunately, text processing seems to be an afterthought at TRECVID, since it is mainly about video. Most of the time this step is described in less than half a page, if at all, and it is uncommon to publish results from just this problem. So we have only a partial idea of the direction the teams take and where they stand.

The University of Amsterdam's team used two algorithms: A statistical one called "text matching" and a symbolic "ontology querying" using WordNet [8]. The latter was done using just the most popular meaning of a word and comparing it using Resnik's similarity measure.

Helsinki University of Technology presented a cruder WordNet approach [9] that involved synonyms, but could also handle negation in queries.

As a different approach to the same problem, many of the participants seem to delegate this task — partially or completely — to outside premade generic systems. The Chinese University of Hong Kong [10] and the The University of Hong Kong [11] used the Lemur tookit[3], having the latter added some symbolic processing using WordNet, and four heuristics for the selection based on term similarity. AT&T [12] opted for LinPipe[4]. They also added reference data from published material dated from the same time as TRECVID's videos to use as reference as a way to improve their results.

K-Space [13], an inter-institution group, used a system that used Terrier[5] for its indexing and retrieval uses. KinoSearch[6] was used by COST [14], a research network of mostly european researchers, who used WordNet's synonyms for query expansion.

It is the author's position that not limiting WordNet to be used as a dictionary, but exploiting its network of relations to provide a semantic-based ontology to the fullest, it would be possible to achieve a better-than-average text

[3] http://www.lemurproject.org/
[4] http://www.alias-i.com/lingpipe/
[5] http://ir.dcs.gla.ac.uk/terrier/
[6] http://www.rectangular.com/kinosearch/

classification. To that end, a design was specified as described in Sec. 2, paying attention to how the categories were defined, how the synonyms are organized within WordNet, what can be inferred from it, and how that information can be used, leading to a new similarity function, described in Sec. 3. An approach to the selection of the best categories from the ones eligible is presented at the end.

Using the data from TRECVID 2006 a test was run, and its results are analysed in Sec. 4, and discussed in Sec. 5, as well as further directions of development to follow in search of better results.

The described approach was used as part of a contender project [15] in TRECVID 2007, and has a more detailed description in another work [16], on which this article is based.

2 Procedure, Methods and Techniques

Text classification is a very straightforward thing to explain. Given the texts and the categories, all that is needed is to create a correspondence from the former to the latter.

Before the process can start, we need to make sure we know what each category means. For instance, in a dictionary "person" is not limited to an individual, the body of a human, or the grammatical category. There are also different interpretations in sociology, law, philosophy, not to mention expressions as "in person". For this reason, short descriptions of each category are provided in TRECVID which clarify what each means.

After this step the text is analysed and only the most relevant features in it are used. Their possible meanings are semantically compared with each category, and scored. In the end, the best scores are selected as the best classification.

2.1 Category definition

Categories should be identified unambiguously. Many words have more than one meaning, depending on the context. This clarification isn't yet possible in the text (perhaps as future work), but it is possible in the category definition, as they are predefined with a short description. This way, many obscure synonyms are avoided (like using "dog" to refer to a person).

Synset[7] are used to define the categories for this reason.

It is often desirable to use more than one *synset* to define an idea than just one higher-level notion that encompasses more than it is supposed to. For instance, one category is described as

> Shots of the interior of a court-room location

and can be defined by the following *synsets*:

[7] A *synset* (synonym set) is an identification of meaning associated with a word or expression. For instance, "mad" and "crazy" share a *synset*, while "mad" and "angry" share another.

court, courtroom: a room in which a lawcourt sits "television cameras were admitted in the courtroom"

court, lawcourt, court of law, court of justice: a tribunal that is presided over by a magistrate or by one or more judges who administer justice according to the laws

Formally, we can say that we manually define a $C \times S$ relation, where C defines the categories, and S is the group of possible *synsets*.

2.2 Working with the text

The text to analyse goes trough a syntactical analyser, like VISL[8]. This provides us with the classification of each word and it's lemma.

Only nouns are of interest at this point, since they are the easyest thing to identify in a still image. This includes compound and proper nouns that may be formed by more than one word.

> 173. Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
> 194. Find shots of Condoleeza Rice

Two lists should be kept: one containing the positive, and the other *negative nouns*. The latter holds nouns preceded by a negative preposition or otherwise detected as meant to be excluded. Both will follow the same procedure until the end, but is intuitive how their results could be combined. Query 186 is a good example:

> 186. Find shots of a natural scene - with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles

Already a *stopword* can be inferred from the previous examples: "shot" can be ignored from the start of every query from now on, as it contributes nothing to solving the current problem.

2.3 Working with Relations

On one side we have a number of nouns, on the other we have the categories. We are looking for the stronger links between elements of these two groups. But it cannot be done directly, since each noun may have different meanings. As with the categories, we need to map the nouns to *synsets*.

If N is the collection of every noun, we can easily do a $N \times S$ correspondence. The hard part is to define a good function that can relate a pair of *synsets*. We will soon define a way to provide a score based on the semantic knowledge defined in WordNet.

[8] http://visl.sdu.dk/

WordNet 3.0 provides several types of $R : S \rightarrow S$ relations. The ones relevant to nouns that are useful are **hyponymy**, **meronymy** and **instance of**. They all work from a more abstract to a more concrete and defined notion. Their definition, size and an example follows.

hyponym A speciality or subordination relation (89,089 items)
A cat is a mammal.
member meronym Marks the presence in a group of elements. (12,293 items)
A fish is a member of a school.
part meronym Indicates that something is part of an object or process. (9,097 items)
Drying is part of washing.
substance meronym Explains the constitution of something. (797 items)
Bread is made of flour.
instance of Points an element as a particular case. (8,577 items)
Lucy was an Australopithecus afarensis.

Despite the differences in size, the same weight was given to all relations. Special care should be taken when mixing relations in order to keep a notion of relationship between two *synsets*.

Intuitively, two *synsets* should be more closely related the smaller the relation path that connects them. Several measures have been proposed [17] that tackle this problem. There are some problems with these approaches, which makes them ill-suited for the current challenge.

First, we find that they don't scale well. Each word has many *synsets* associated with it, and each of these has to be compared with dozens of categories. That means many searches over a big search space, done for each noun in the sentence. All this just for the first query!

Second, most of the measures only work with **hyponyms**, leading to a great deal of knowledge left unused. A simple **meronym** could show a good relation that no **hyponyms** would (and that is why we have more than one relation).

Third, *synset* similarity should not be measured in blind distance, but on how much they have in common. For instance, "sleeping" and "walking" are just 2 *synsets* away, because "sleepwalking"is a **hyponym** of both. The same goes for "recreation" and "vice" due to "gambling". "Pepper spray" combines both "spray can" and "chemical weapon" and a "programmer" and a "terrorist" are just a "hacker" away. This bad situation expands with longer paths. The problem resides in searching through "more concrete" and then "more abstract" instances. Figure 1 illustrates this situation.

Lastly, we would like to avoid the use of a corpus and statistical processing as much as possible, concentrating in using all the available knowledge to its fullest before looking for outside help.

3 The similarity measure

The relevance between two *synsets* is directly proportional to the number of nodes in the path between each *synset* and the most high-leveled term in the

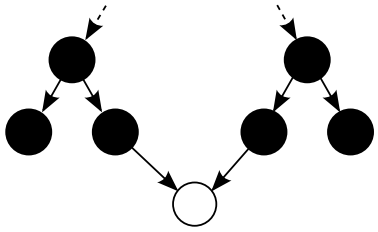


Fig. 1. The more specific hollow node bridges two more general nodes, possibly providing unexpected similarity measurements.

ontology. For example, in WordNet, everything is an "entity" through the **hyponyms**[9]. We'll call that *synset* the **root**. Given that we are working on a graph, more than one such path can exist.

A path P_i is an element of $P = \{(s_1, s_2, \ldots, s_n) : \forall i < n-1 \quad R(s_{i-1}) = s_i\}$, and we'll use the notation P_n and P_c to represent a noun's path and a category's path, respectively.

We will also say that path P_i is a prefix of P_j — which means the former is a generalisation of the latter — if it has all the *synsets* present in P_j in the same order. Symbolically: $P_i \subset P_j \Leftrightarrow \exists n \in \mathbb{N} : \forall m \leq n \quad s_{i\,m} = s_{j\,m}$ where $s_{i\,k} \in P_i, s_{j\,k} \in P_j, k \in \mathbb{N}$.

Knowing that all paths start at the root, it is trivial to compare two paths and determine their **level of convergence**, which corresponds to the number of identical elements at the beginning of both paths. We can define it as function $m : P \times P \to \mathbb{N}_0$ where $m(P_i, P_j) = \max |P_k|, P_k \in P : P_k \subset P_i \wedge P_k \subset P_j$.

There is no guarantee that *synsets* A and B, that have 5 nodes in common, are more similar than A and C that have only 4. That happens because relations don't carry the same weight. For example, observing Fig. 2 one can see that a puma is as much a carnivore as a polar bear. Since WordNet has a very tree-like topology, it is likely that B and C share a significant part of the path — that is to say that they aren't totally different things, so this assumption is bearable for now. Other factors may affect the classification later on.

This approach to pathfinding is not limited to avoiding shortcomings. There are other benefits. It is simple and fast, since it traverses the graph in one direction (opposite to the relation). Also, since it works its way towards higher level concepts in a depth first approach, branching should be a lesser problem. Loops aren't supposed to exist when following only one relation in an ontology. And it presents lots of information that is always useful. We're not comparing something with the *synset* for "good". We are comparing it with *entity/physical entity/object/unit/artifact/good* or with *entity/abstraction/attribute/quality/morality/good*.

[9] This is not entirely true. But clearly through oversight the "Underground Railroad", the "Third Crusade" and a few other *synsets* aren't integrated into the main body of the relations.

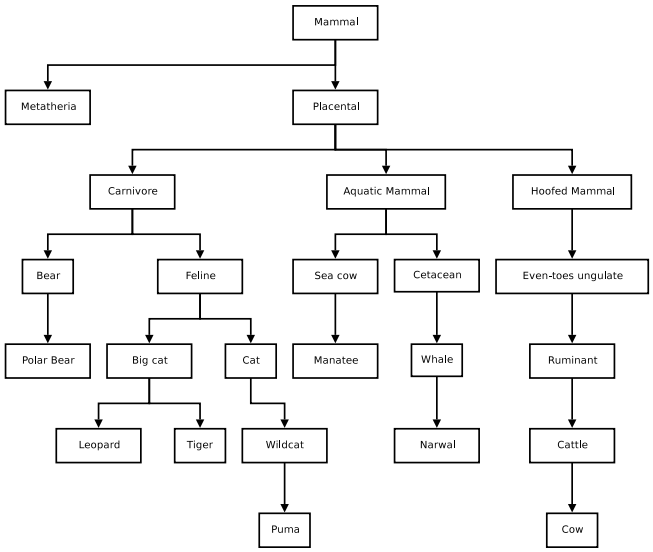


Fig. 2. Example of WordNet's hyponyms.

After taking into account the overlappig paths, what is left are the two **divergent paths**. For the moment we are interested only in their lengths, that represent how different they are. We thus define the function $\ominus : P \times P \rightarrow \mathbb{N}$ defined as $P_i \ominus P_j = |P_i| - m(P_i, P_j)$ to mean P_i's **divergent path** length relative to P_j. Table 1 shows the possible outcomes for both paths and what they represent.

Table 1. Interpreting the length of different divergent paths

Noun	Category	Relation	Meaning	Response
0	0	$P_c = P_n$	They are the same	Accept as a match
< 0	0	$P_c \subset P_n$	We have a word more specific than the category. For example, "hammer" and "tool".	Almost certainly accept it as a match.
0	> 0	$P_c \supset P_n$	The word is more general than the category. Like "animal" and "dog".	It is a weak match. May be dropped.
> 0	> 0		They specialise in different ways.	It is the worst case.

Before we quantify what is a small difference and what is a big one, let us observe the four situations depicted in Fig. 3. As can be seen from their **divergent path** lengths listed in Table 2, the length of the shorter path is more important in determining how "inline" both *synsets* are. If zero, we are dealing

with more general and concrete ideas of the same thing. If larger (and since this is the smaller path), the similarity is weakened.

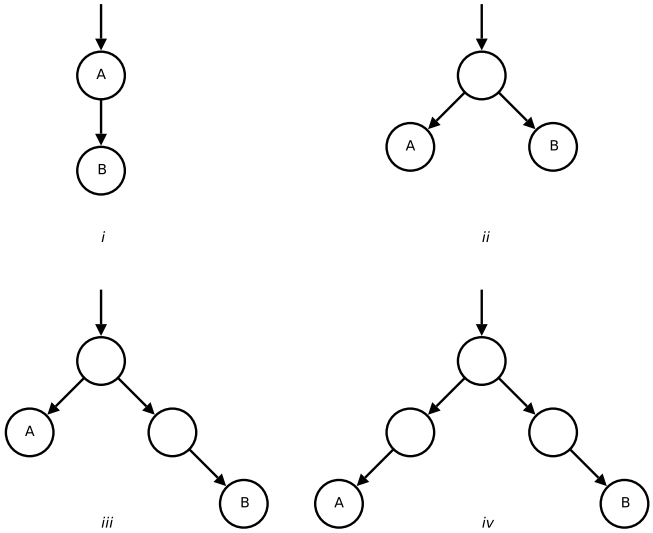


Fig. 3. Four situations ordered from better to worse.

Table 2. The four situations in Fig. 3 indicating path length

Situation	Shorter path length	Longer path length	Observation
i	0	1	The best situation when $A \neq B$
ii	1	1	Still an acceptable arrangement
iii	1	2	Can be considered a successful match in some circumstances
iv	2	2	Intolerable unless the **level of convergence** is high

Joining both measurements, we get equation (1), which presents an index of dissimilarity — that is, the bigger the index, the less alike two *synsets* are. In our tests, we found that $k = 8$ works well.

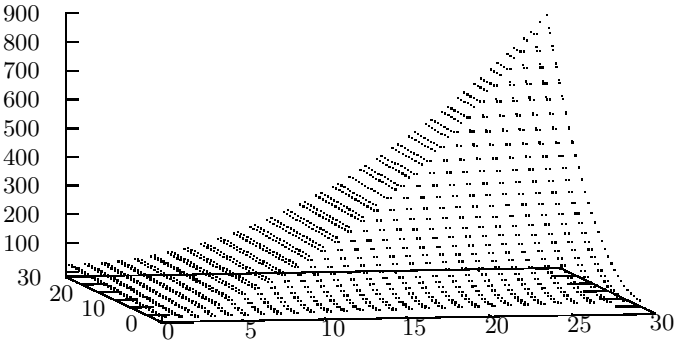


Fig. 4. Plotting of the dissimilarity function (k=1, m=0)

$$I(P_c, P_n) = \begin{cases} -100 & \text{iff} P_c = P_n \\ -50 & \text{iff} P_c \subset P_n \\ k \cdot \min(P_n \ominus P_c, P_c \ominus P_n)^2 + & \\ \quad + \max(P_n \ominus P_c, P_c \ominus P_n) + & \text{otherwise} \\ \quad -m(P_c, P_n) & \end{cases} \tag{1}$$

In Fig. 4 we can see a graph illustrating how changes in the longer divergent path have little impact in relation with the shorter one.

For each pair $C \times N$ we evaluate, only the best one is meaningful, so we keep only the best result (lower I) for each category.

3.1 Choosing the best category

Given the intended use of this work, the choice was taken to err on the side of recall. Since another selection process is to be taken afterwards (on the videos), it would be risky to do aggressive filtering at this stage. So **at least one** result is guaranteed to be returned, and in case of reasonable doubt, that result is included.

Of all the triples $(c, n, \mathbb{R})$, we select the cutting threshold T as the greatest I that is not the result of a trivial selection. I.e. $I \in]-50, +\infty[$.

- If $T > 0$, then we accept all the categories with $I < \frac{1}{2}T$. These are the good results.
- If $T = 0$, we choose the categories with $I \leq 0$. These are still considered acceptable.
- If $T < 0$, then we keep all the elements with $I < 2$. If none, then just the ones with $I = T$. This is the worst case.

In the end, we have two lists of categories: a positive and a negative. They can be combined, or passed on to a lower-level filtering system.

4 Results

To evaluate the performance of the implementation of the described algorithm, the 24 queries used in the 2006 edition of TRECVID were classified by a person, and that classification was used to rate the choices of the system. The details of the implementation and the tests, as well as all the data used are present in the previously mentioned article [16].

All the queries are identified by a number. The elements on the negative list are subtracted from the positive list, and the result is compared with the reference to provide the **recall**, **precision** and **average precision** The results can be seen on fig. 5.

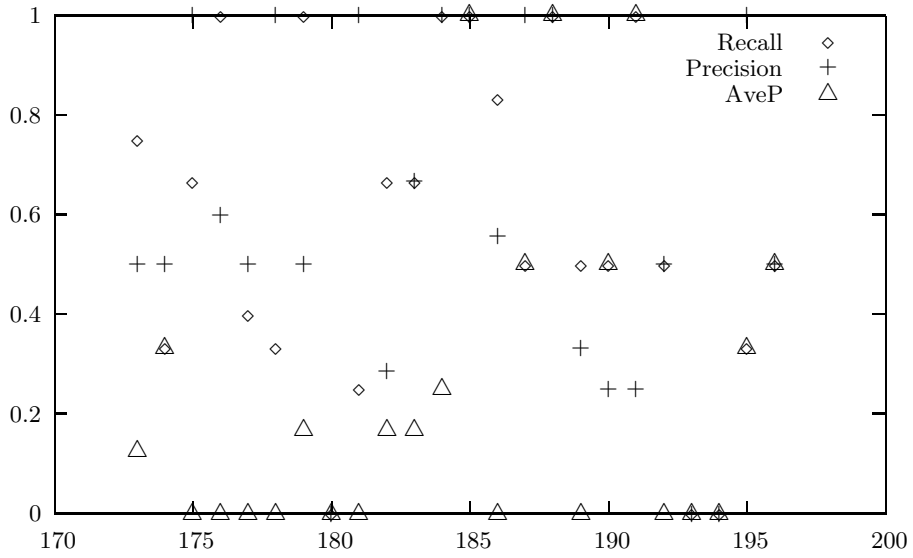


Fig. 5. The results of a testrun using the data from TRECVID 2006 (queries 173–196) against a human classification of the same sentences. Shown are the recall(mean 0.572, median 0.500), precision (mean 0.581, median 0.500) and average precision (mean 0.252, median 0.146)

As stated before, working with small amounts of information is mostly an hit-or-miss affair, and two perfect matches (185 and 188) and three complete misses (180, 193 and 194[10]) illustrate that very well.

Many of the results cannot be said to be incorrect, even if they are not the expected ones (like linking "animal" with "person"). Other errors derive from what a person expects to see on an image related with the theme (like a "US

[10] "Condoleezza Rice" was misspelled in this query, leading to it not being recognised as a proper noun, returning only the category "vegetation"

flag" when talking about "George W. Bush"), which cannot be deduced by a simple semantic analysis (but this can be fixed for some common situations).

The results for **recall** and **precision** were slightly below the expected values. Yet they seem to be workable into the 60–70% range with some work and tuning.

The **average precision** is a less useful measure in this situation. It makes more sense in document retrieval environments than in classification systems. To give an example, the following TRECVID query says:

> 184 Find shots of one or more people seated at a computer with display visible

The user expected the classification "person" and "screen". The system returned "screen" and "person", and the *average precision* of that answer is just 25%. The answer is correct, as both elements are equally important, and there is no way to rate one above the other. For this reason, the *average precision* measurement is included to allow the comparison to similar systems.

As a more general test, outside of the TRECVID area, 12 documents were picked from the first 100 Reuters-21578 documents that were classified using words present as WordNet *synsets*. This system scored a mean of 0.833 in **recall**, 0.446 in **precision** and 0.500 in **average precision**.

5 Conclusions

The variety of approaches taken by teams at TRECVID in the automatic query analysis could indicate that satisfying results weren't being met, and experimentation is still in order. Also, it is good to see that general-purpose tools exist able to provide "good enough" results to be relied on for this purpose.

The main ideas proposed here appear to be sound and worth exploring further. WordNet is suited as a base for text classification based only in symbolic principles, and the proposed similarity function is but a gentle start. The path of improvements possible seems long, even at this starting point.

Minor adjustments should raise the application from the slightly disappointing results gathered at the TRECVID test. The Reuters-21578's test was more in line with was expected, since, as stated earlier, the system was tuned towards recall and the TRECVID environment.

Several ideas can be pursued to improve the performance of the work presented:

The situation *iii* in Fig. 3 can be improved if we distinguish between A being a category or being a *noun*, where the former is better than the latter.

Mixing relations is an obvious step, as well as working with the adjectives and verbs in the query.

Some categories are hard to express as *synset* A or B. It should be possible to combine two into just one. For example "military building".

A better negation detection, capable of detecting more complicated situations, such as "Find shots of a person that is not in a crowd or wearing a jacket" could be useful.

6 Acknowledgements

Thanks go Cristina Ribeiro and Catalin Calistru for the opportunity to work on such an interesting subject.

References

1. Gantz, J.F., Chute, C., Manfrediz, A., Minton, S., Reisel, D., Schlichting, W., Toncheva, A.: The diverse and exploding digital universe. Technical report, IDC (2008)
2. Doctorow, C.: Metacrap: Putting the torch to seven straw-men of the meta-utopia (August 2001) http://www.well.com/ doctorow/metacrap.htm.
3. Pantel, P., Lin, D.: Spamcop — a spam classification & organization program. In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization, AAI Press (1998)
4. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization, AAI Press (1998)
5. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. 2nd edition edn. Prentice-Hall, Englewood Cliffs, NJ (2003)
6. Sahay, S.: Support vector machines and document classification
7. Sahami, M., Yusufali, S., Baldonado, M.Q.W.: Real-time full-text clustering of networked documents
8. Snoek, C.G.M., van Gemert, J.C., Gevers, T., Huurnink, B., Koelma, D.C., van Liempt, M., de Rooij, O., van de Sande, K.E.A., Seinstra, F.J., et al.: The MediaMill TRECVID 2006 semantic video search engine. In: Proceedings of the 4th TRECVID Workshop, Gaithersburg, USA (November 2006)
9. Sjoberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: Picsom experiments in trecvid 2006. (2006)
10. Hoi, S.C.H., Wong, L.L.S., Lyu, A.: Chinese University of Hong Kong at TRECVID 2006: Shot boundary detection and video search. (2006)
11. Jiang, Y.G., Wei, X., Ngo, C.W., Tan, H.K., Zhao, W., Wu, X.: Modeling local interest points for smantic detection and video search at trecvid 2006. (2006)
12. Liu, Z., Gibbon, D., Zavesky, E., Shahraray, B., Haffner, P.: AT&T research at TRECVID 2006. (2006)
13. Wilkins, P., Byrne, D., Jones, G.J.F., Lee, H., Keenan, G., McGuinness, K., O'Connor, N.E., O'Hare, N., Smeaton, A.F., Adamek, T., Troncy, R., et al.: K-space at trecvid 2008. Technical report, Dublin City University (October 2008)
14. Zhang, Q., Tolias, G., Mansencal, B., Saracoglu, A., Aginako, N., Alatan, A., Alexandre, L.A., Avrithis, Y., Benois-Pineau, J., Chandramouli, K., , et al.: Cost292 experimental framework for trecvid 2008. Technical report (2008)
15. Calistru, C., Ribeiro, C., David, G., Rodrigues, I., Laboreiro, G.: Inesc, porto at trecvid 2007: Automatic and interactive video search. In: TRECVID 2007 Papers. (October 2007)
16. Laboreiro, G.: Classificador para língua natural. Master's thesis, University of Évora, Largo dos Colegiais 2, 7000 Évora (October 2007)
17. Budanitsky, A.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures (2001)

Temporal Analysis of Terms in Blogs

Filipe Coelho

FEUP - Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto Portugal
filipe.coelho@fe.up.pt

Abstract. Blogs are becoming extremely popular, revealing the most relevant topics for their social communities on a daily basis. The work presented here has focused on the temporal analysis of terms usage in blogs, specifically the Portuguese *SAPO Blogs* collection, to find the most relevant terms occurred during the first half of 2008. The gathered information was stored and processed by means of a data warehouse, which facilitated the necessary calculations for terms analysis by the relevance and interestingness ranking algorithms. Term clouds were used to show the comparison between these algorithms, allowing us to quickly determine that interestingness ranking produced the best results for this collection.

Key words: information retrieval, knowledge extraction, information visualization

1 Introduction

Nowadays blogs are an outstanding communication channel for web communities to express themselves. Global, national and personal news are commonly displayed and discussed, turning blogs into social and cultural spaces about various topics. These topics and more importantly their evolution over time, represent the impact of certain facts that occur in our daily lives, like important events or marketing campaigns, for example.

Of course, not all blogs are meaningful i.e. carry useful information, and even inside a blog, not all posts contain useful topics for the majority of the community. This work has centered its study on ways to retrieve terms used in blogs and globally analyze them to find out the most important ones in a daily basis. For that matter, the *SAPO Blogs*[1] collection was used, restricted to the first semester (January to June) of 2008 and only to the 100 most active blogs. This allowed the correct preprocessing of posts contents and later on to test the chosen ranking algorithms in real-time.

1.1 Objectives

The main objectives of this work are:

[1] http://blogs.sapo.pt/

– Temporal analysis of blogs by finding the most relevant terms over time;
– Comparison of ranking algorithms used in information retrieval;
– Analysis of these algorithms and their immunity to non-relevant terms (stopwords), innate or by fine-tuning;
– Specification of proper means of storage to allow obtaining the ranks lists in real-time;
– Visualization of the obtained results through "time-aware" clouds.

The next sections will describe in greater detail the development and results achieved through this work.

2 Information Visualization through Clouds

Clouds (Figure 1)[2] are a relatively recent form of information visualization. Based on a certain criteria (frequency, relevance, similarity, among others), clouds quickly highlight existing concepts in the analyzed text.

Fig. 1. *Wordle cloud of Google News on the 9th January 2009*

Clouds can adopt different designations according to the source domains. Tag Clouds[1] show the relative importance of terms associated to resources like blog posts, photos and videos, used to discriminate their contents. Their graphical construction and the importance of some visual features as a way of rapidly compare tags is exemplified in [2–4]. When used in text mining, clouds are normally designated by term, word or (more commonly) text clouds. These are used in political speech analysis[3], discovery of dominant topics in literature books, and even in unusual tasks like preparing a PhD thesis[4], among others. However, these designations are commonly swapped since what really matters is the cloud characteristic visualization method, and not the domain of its analysis.

There are some examples of clouds with a time component, becoming "time-aware" clouds. Figure 2[5], generated through the *Tagline Generator*[6], illustrates

[2] http://wordle.net/
[3] http://politicaltagclouds.wordpress.com/
[4] http://creativitymachine.net/2007/01/25/tagcloud-of-my-phd/
[5] http://chir.ag/phernalia/preztags/
[6] http://chir.ag/tech/download/tagline/

the temporal analysis of political speeches content over the years, namely the recent use(color) and popularity(size) of certain keywords, and the political situation lived in that time. The example in Figure 2 relates to the political speeches made by USA's president John F. Kennedy when Cuba installed Soviet missiles. As expected, the words "*Soviet*", "*Missiles*", "*Cuba*", "*Economic*" and "*Strength*" were extremely relevant in those speeches.

Fig. 2. *Evolution of American presidential speeches over the years*

Another example of visualizing tags over time can be observed in Figure 3[7], about the *Yahoo Taglines*[5] project. Tags (and their sources – photos) are being replaced or changing size as the top bar (timeline) advances.

Fig. 3. *"Waterfall"* example

[7] http://research.yahoo.com/taglines/

3 Terms usage in Blogs

The temporal analysis done in the blogs collection was based on two concepts: interestingness (Sect. 3.1) of the terms, and their relevance (Sect. 3.2). Those two concepts will now be explained in further detail.

3.1 Interestingness

This concept was presented in [5], and the same notation will be used here. Every object (in this case, terms) $u \in U$ (universe of objects) is associated with a multiset[8] of timestamps (in this case, blog posts) $t \in T$, that represent their occurrences over time.

Since these occurrences are multisets, a term can occur several times in a post. Using $\gamma(u,t)$ as the number of occurrences u at time t, i.e. the term frequency in the post, $\gamma(u)=\sum_{t=0}^{T-1}\gamma(u,t)$ represents the total number of occurrences of u, that is, the total frequency of the term in the collection.

Term interestingness is defined by the following two properties:

- one term will be more interesting during a specific time interval if it occurs more frequently on that interval and less frequently outside of it;
- a rare term, that appears only in a specific time period shouldn't necessarily be the most interesting term for that period of time.

Having $I = [a,b]$ with $0 \leq a \leq b \leq T$, and using a constant C to respect Property 2, the interestingness of an object u for an interval I is given by:

$$Int(u,I) = \frac{\sum_{t \in I}\gamma(u,t)}{(C+\gamma(u))} \tag{1}$$

The most interesting terms for an interval I are the ones with the biggest values of $Int(\cdot, I)$.

3.2 *Tf-idf* ranking

Tf-idf [6] was the other chosen algorithm for term ranking. This algorithm is based on two term features: their frequency in the document and their rarity in the whole collection.

In order to study the temporal evolution of terms, it is assumed that all blog posts (and their corresponding terms) of a specific day represent one document.

The *tf-idf* algorithm is defined by the following equations:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

$$idf_i = log\frac{|D|}{d_j : t_i \in d_j} \tag{3}$$

[8] http://planetmath.org/encyclopedia/Multiset.html

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i \tag{4}$$

By combining the term frequency and its rarity, the *tf-idf* algorithm tries to retrieve relevant terms and properly rank them.

4 Development

4.1 Dataset Preprocessing

Development was started after having obtained a MySQL[9] database with *SAPO Blogs* feeds previously gathered by a crawler, containing posts from the blogs authors over the years. This initial database had over 3.5 million posts and 36.000 posts from blogs over five years (2003-2008), taking 10GB of disk storage. A subset of this database was chosen in order to extract the necessary information to test the ranking algorithms, namely the 100 most active blogs of the first semester of 2008, resulting in 500MB of information for analysis. Some cleaning was also performed after this step, removing empty blogs and orphan posts. These restrictions allowed the algorithms to be properly tested in an acceptable timeframe.

4.2 Fixing Text Encoding

One serious problem detected during the post content analysis was related to text encoding. Several posts, initially created using the *Windows-1252*[10] encoding were directly stored as *UTF-8*[11], without any attempt of conversion, probably during the crawling process. Fixing text encoding isn't a trivial process, having required some time and several "trial and error" experiments to overcome this problem.

The initial miss-stored text encoding (*Windows-1252*) was determined thanks to a tool available in the web article "*Detect Encoding for in- and outgoing text*"[12]. The next step was to determine which posts were incorrectly stored: the fix should only be applied to those posts and not to the ones originally created in UTF-8. The fixing process algorithm (in C♯) is as follows:

```
//encoders
Encoding cp1252 = Encoding.GetEncoding(1252);
Encoding utf8 = Encoding.UTF8;

//byte arrays containing converted text by the encoders
Byte[] field_text_bytes_cp1252 = cp1252.GetBytes(field_text);
Byte[] field_text_bytes_utf8 = utf8.GetBytes(field_text);

//text converted to UTF-8
string field_text_utf8 = utf8.GetString(field_text_bytes_cp1252);
```

[9] http://www.mysql.com/
[10] http://www.microsoft.com/globaldev/reference/sbcs/1252.mspx
[11] http://tools.ietf.org/html/rfc3629
[12] http://www.codeproject.com/KB/recipes/DetectEncoding.aspx

```
11
   //control variable
13 bool fix = false;

15 //cp1252 <-> UTF-8 test
   if (cp1252.GetByteCount(field_text) == utf8.GetByteCount(field_text_utf8))
17 {
       if (cp1252.GetByteCount(field_text_utf8) == utf8.GetByteCount(field_text))
19     {
           Console.WriteLine("no need to fix!");
21         fix = false;
       }
23
       else
25     {
           Console.WriteLine("needs fixing!");
27         fix = true;
       }
29 }
   else
31 {
       Console.WriteLine("already in UTF-8!");
33     fix = false;
   }
35
   if (fix)
37 {
           //update database with UTF-8 text
39         ...
   }
```

The encoding fix starts by comparing the resulting number of bytes obtained from the encoders' analysis of the original text in their corresponding array (line 16). Encoding correction is only necessary when this count is equal but the complementary count (line 18) isn't. In this case, the UTF-8 converted text (line 10) is submitted to the database, replacing the original one. This algorithm was successfully applied to the affected fields (blogs/posts titles and contents).

4.3 Terms Indexing

Based on the correct text, the terms were extracted from the posts and indexed using the software Lucene[13] v2.4.0, converted to the .NET platform[14] through the IKVM.NET[15] conversion tool.

During the indexing phase, the text was preprocessed in order to avoid HTML tags indexing and to properly handle the "&;" entities. The term vectors were uploaded to the database, which resulted in the relational model presented in Figure 4.

4.4 Warehouse

To obtain the term ranking by both algorithms in real-time, in order to "feed" the term clouds, several values need to be precalculated. This approach is related

[13] http://lucene.apache.org/
[14] http://www.microsoft.com/net/
[15] http://www.ikvm.net/

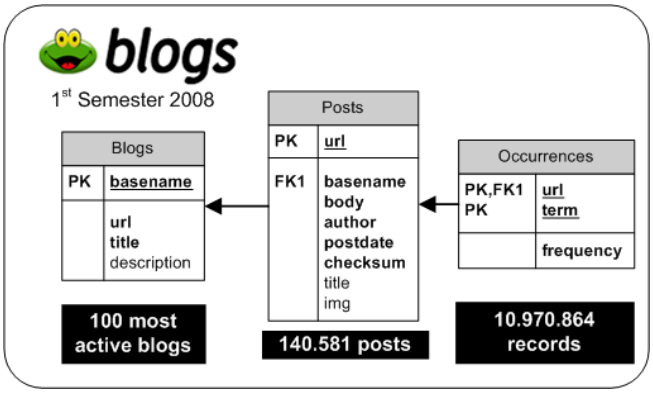


Fig. 4. *Relational model*

to data warehouses, suited for data mining and knowledge extraction operations [7]. A dimensional model was defined (Figure 5), with the necessary dimensions (red) and facts tables (blue) to store the previous relational model information, and summary tables (green) were created to contain the necessary measures for the ranking algorithms. The warehouse ETL[16] process was done in MySQL v5.1 using the MyISAM engine, particularly suitable for storing huge and "write-once-ready-many" tables.

4.5 Data Visualization

The application *Live Tag Cloud in WPF*[17] (Figure 6) was used as the visual component, with the necessary adjustments (connection to the data warehouse, specific queries to obtain the data, among other minor tweaks).

Starting at the first day of the collection (1st of January 2008), the application successively interrogates the data warehouse to obtain the 10 top (interesting or relevant, depending on the chosen algorithm) terms in each day. Those terms are added to the list or, if they already exist, their size increases and eventually their colors change, according to the following rules:

- the 15% less frequent terms are hidden;
- terms between 15% and 40% frequency become light-gray;
- terms between 40% and 70% frequency become dark-gray;
- terms between 70% and 75% frequency become black;
- terms between 75% and 95% frequency become yellow;
- the most frequent (+95%) terms become red.

The terms color and size are redetermined at each new query, in order to study their usage evolution.

[16] Extraction, Transformation and Loading
[17] http://wpfblog.info/2008/06/01/live-tag-cloud-in-wpf/

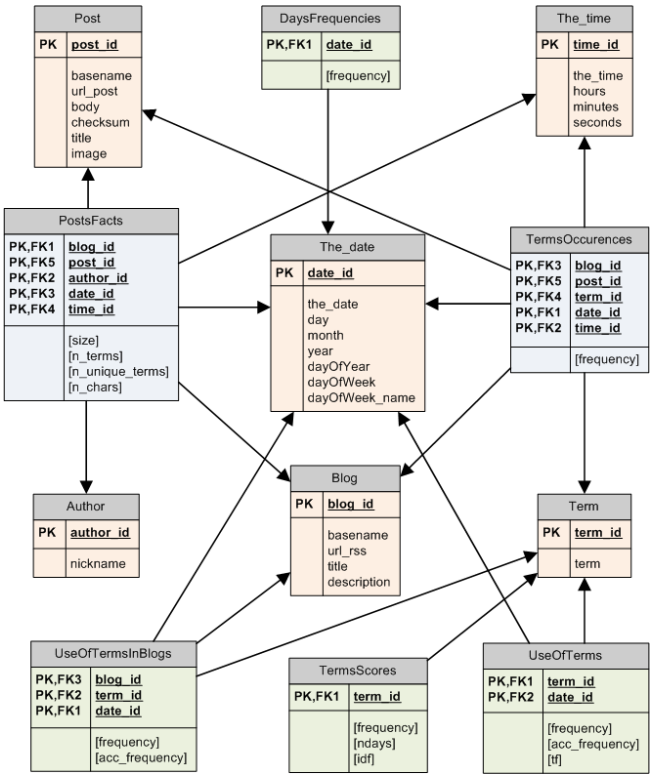


Fig. 5. *Dimensional model*

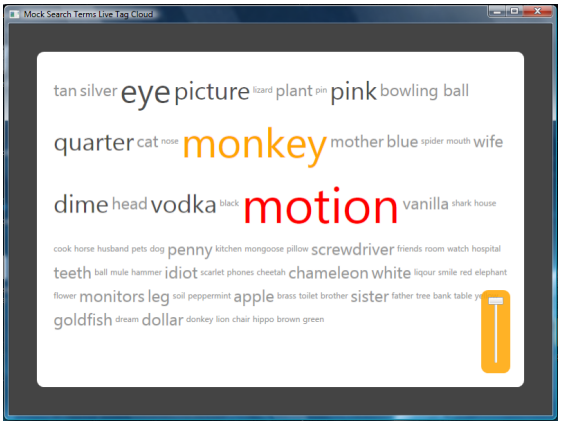


Fig. 6. *"Live Tag Cloud in WPF"*

5 Results

5.1 Comparison of Ranking Algorithms

The interestingness algorithm uses a constant C that needs to be fine-tuned to the collection, unlike the relevance algorithm, which is rather general. In order to determine the optimum value, a simple test was done: five important dates (covered by the collection timespan) were chosen, and for each, lists with the 100 top terms of each algorithm (original and variants) were obtained. These lists were manually analyzed in order to determine how many meaningful terms were present, according to the date. Table 1 shows the evaluation results.

Table 1. Comparison of ranking algorithms

	Carnival	Easter	Revolution	Children	Country	
I (C = log(max(frequency))	10	7	7	10	1	7
I (C = sqrt(max(frequency))	8	18	33	25	3	**17.4**
I (C = avg(frequency))	11	13	19	16	0	11.8
tf-idf	9	10	11	16	1	9.4
	9.5	12	17.5	16.8	1.3	

As it can be seen, when $C = sqrt(max(frequency_{term}, collection))$, the interestingness algorithm returns the biggest number of meaningful words related to the chosen date. This makes sense, as using huge values for C would focus the interestingness on the frequency alone (returning all the stopwords), and using very small values would violate Property 2 (3.1).

5.2 Ranking over time

Using the visual component initially with the interestingness algorithm and later on with the relevance one, the results shown in Figs. 7 and 8 were achieved. In both figures the clouds represent the evolution of the 10 top words in each day, over time.

Based in these figures, it can be acknowledged that:

- The interestingness cloud shows a significant number of terms, including several words related to important dates in the collection ("carnival", "Easter", among others);
- The relevance cloud doesn't contain highlighted words that could be related to importante events;
- among the interesting terms, temporal ones are definitely highlighted through their designations and abbreviations ("feb", "mon", etc);
- In both clouds, several email and websites addresses are present, including some rather meaningless numbers, which is probably related to the initial database subset, and suggest a more thorough approach during the term preprocessing and indexing stages;

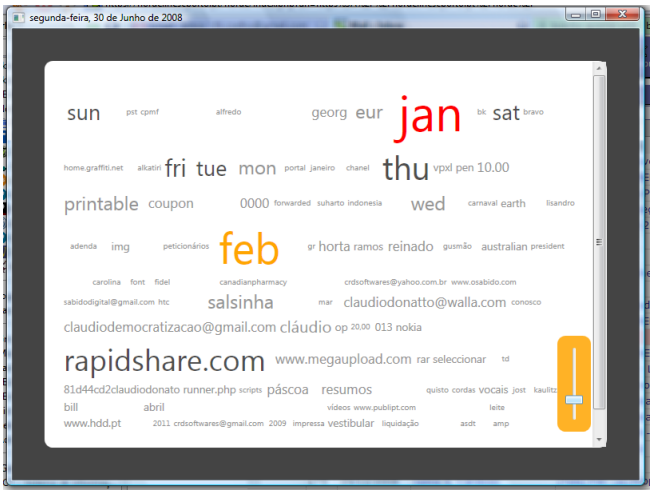


Fig. 7. *Interesting words over time*

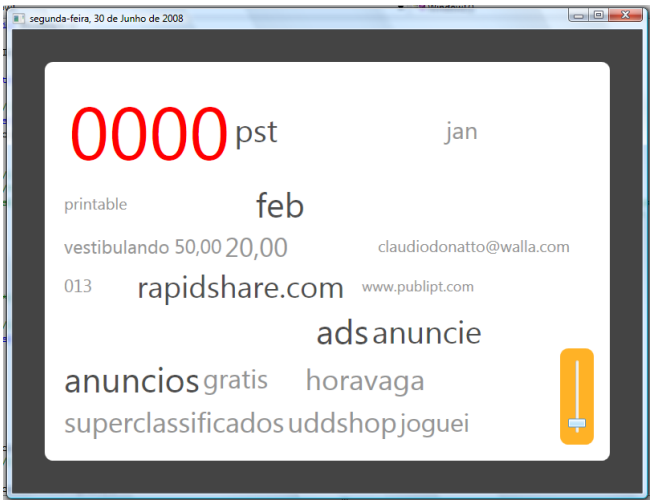


Fig. 8. *Relevant words over time*

– this "top terms over time" approach reveals that the interestingness algorithm shows greater potential in retrieving meaningful terms in the top ten results of each day, when compared to the relevance algorithm.

6 Conclusions

During this work, several issues were found and overcome, mainly the text encoding correction and the warehouse ETL (Extraction, Transformation and Loading) step, which required a significant amount of time (pre-calculation of all the necessary measures for the ranking algorithms).

Still, the initial chosen subset allowed the comparison between interestingness and relevance algorithms. It can be concluded that the former is more adequate to quickly obtain meaningful terms, specifically thanks to the fine-tuning of the C parameter to the chosen collection. The dimensional model and data warehouse used for storage have proven to be very adequate to the task at hand, allowing for quick retrieval of the ranking lists for the temporal analysis. And finally, the visual component demonstrates that clouds are a very good way of intuitively highlight the top terms in a temporal analysis.

6.1 Future Work

Based on the obtained results, some future steps can be determined to continue this work. A very important issue will be the term analysis during the preprocessing, by removing meaningless terms like numbers, incomplete dates and email addresses, which have shown not to be relevant to knowledge extraction. Next, it makes sense to proceed to a post-indexing analysis of the terms, mainly by using a dictionary that can associate important features of those terms (subjects, verbs, etc). Results have suggested a specific value for C (interestingness algorithm), based on the square root of the maximum term frequency of the collection. The manual evaluation, inherently subjective, needs to be further reviewed by using the whole collection, in order to study if the formula still returns the best value for C. Finally, the visual component could be adapted to other temporal analysis experiments, and its visual feedback can be further enhanced for a better highlighting of top terms.

7 Acknowledgments

I'd like to thank Maria Cristina Ribeiro, João Correia Lopes and Gabriel David, my professors at FEUP[18], for their guidance during this work, and SAPO[19] for their blog feeds collection database.

[18] http://www.fe.up.pt/
[19] http://www.sapo.pt/

References

[1] Fernanda B. Viégas and Martin Wattenberg. Timelines tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

[2] Scott Bateman, Carl Gutwin, and Miguel A. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In Peter Brusilovsky and Hugh C. Davis, editors, *Hypertext*, pages 193–202. ACM, 2008.

[3] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.

[4] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualization. *CoRR*, abs/cs/0703109, 2007.

[5] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM Press.

[6] G. Salton and C. Ruckley. Term weighting approaches in automatic text retrieval. *Information Processing and Managament*, 24(5):513–523, 1988.

[7] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. Wiley, April 2002.

Integration of Events Information: A robust system based on simple techniques

Luis Certo

Faculty of Engineering of the University of Porto (FEUP)

Abstract. Every day the amount of online data is increasing and with it the quantity of redundant, sparse and contradictory information. In order to reduce this problem, data extraction and information integration systems are required. Normally these systems are very complex to implement because they tend to treat data generically. While dealing with specific types of data, it is possible to reach highly structured representations of information, making the integration process to become very simple and yet very effective. This paper aims to verify that the recognition and reduction of events structure data enables the use of simple character based techniques for information integration. All the employed techniques correspond to small improvements and combinations of the well-known Smith-Waterman algorithm. Even though the system is very straightforward, tests with real data proved that the proposed concept is valid and valuable. The simplicity the system also provides strong flexibility and scalability.

Key words: information integration,data extraction, edit distance, text alignment, string comparison, regular expressions

1 Introduction

We live in an information era. The load of information is huge and this growth is being mainly powered by the advances in computer and communication technologies, like the Internet. However, as the amount of online information increases it is becoming difficult to filter the useful from the useless.
This massive insertion of information can generate what is called Information Overload, causing in many cases, redundant, sparse and contradictory information. This problem can be solved by using a system capable of integrating the information [1] from the available sources.
Integration of multiple information sources generally aims at combining multiple occurrences of the same object while eliminating redundancies, so that they form a unified new whole and give users the illusion of interacting with one single piece of information. Detecting repeated information can decrease the ammount of data which simplifies the search for important information.
From all types of content that can be found in the Internet, advertising is, obviously, one of the main causes of information overload. Throughout the last years Internet has been increasingly used to announce and advertise and every

company or even small shops have a website to promote their products. Like any other company, ticket shops, magazines, theaters, bars and clubs use the Internet to announce the events they are related to (p.e. concerts, conferences, movies in theaters). We define events as products that have a date, a title, and optionally, an hour, a price, an address and others descriptors.
Each website can use distinct forms of language to describe their events (natural language, numeric language). Since this kind of data is usually accommodated to (almost) static HTML Templates and the text patterns are not very complex, extracting this type of information is generally easy.
Multiple websites can publish different occurrences of the same event. Sometimes, people detect repeated events based only on few parts of the whole information, like the event title, the address or the hour. Occasionally, even when involved entities are unknown it is possible to detect repeated occurrences. Perceiving the structure of data can enable us to understand this effect and simplify the information integration process. Structure definition concerns the identification of primary description fields (relevant and yet more static information) and their relative importance. It is obvious that some fields are more relevant than others (title is more significant than price). But while it is possible to establish a hierarchy between some of them, others are difficult to classify. In addition, the level of importance of each field requires to be specifically described by numbers in order to be implemented. Having defined the importance of each primary field enables us to use these parameters in order to build a (1)heuristic for occurrence comparison and subsequent integration.
The inspection of some websites that provide events information revealed that some parts of the information remain almost the same. In effective, and based on this empirical rule, we assumed that the same event published by multiple websites is always described by equal dates, similar titles and similar addresses. The proposed concept combines this assumption with (1) to build an information integration system that we think can achieve good results, without demanding for a very complex implementation.
Even though there are many techniques for information integration, some of them very complex (p.e. feature based and machine learning, data reduction techniques), it is ironic to think that sometimes real data do not requires this complexity.
The structure of the remainder is as follows. In section 2 some of the most popular projects of this field are presented. Section 3 describes the problem in the detail and how it will be resolved. In section 3 the system is meticulously described, with major focus on the integration module. In section 4 the system is tested and the results are criticized. The document finishes with a conclusion and some suggestions for further system improvements.

2 Related work

One of the most popular systems in this field is GoogleNews (Figure 1) [1]. This system gathers the news of all the important newspapers of each country. Each new can be published by several newspapers. When this situation occurs, GoogleNews only states the new once, giving the user a list of all sources where the new is cited. Besides the extraction and integration capabilities, the system organizes the data according to its date and relevance (number of links pointing to the new). To access this information, the user can visit a website or subscribe a RSS Feeds service.

Fig. 1. GoogleNews.UK Screenshot

Although very popular, the internal functioning of GoogleNews is totally unknown.

Before integration data, extraction has to be performed. There are several methods and techniques to perform information extraction. Lately one of the most discussed techniques is the Semantic based [2]. These techniques try to identify, semantically, what entities are present in the text, what they mean, and what information is related to them. This kind of extraction is normally based on RDF (Semantic Metadata) [2]. These descriptors define the entities, how they relate, which properties are assigned to them and what their value is.

FOAF Project (Friend of A Friend) [2] is one of the projects that uses RDF. This system invites people to describe themselves and their homepages using the FOAF RDF, enabling the system to automatically connect (integrate) those pages forming a social network.

Other Semantic Applications try to solve the problem without any markup languages support. Instead, they use methods of natural language processing and artificial intelligence to capture the semantic of the text. OpenCalais (developed by the famous Reuters) [3] is one of these kinds of applications. This application tries to automatically capture the semantic of the text in a particular website

[1] http://news.google.com/
[2] http://www.foaf-project.org/
[3] http://www.opencalais.com/

and outputs RDF data.
A simpler approach for data extracting is proposed by the Dapper: The DataMapper (Figure 2) [4]. Using this application, the user can visually select parts of a website to be extracted. Even though this system can be very useful it does not provide a very customizable environment.

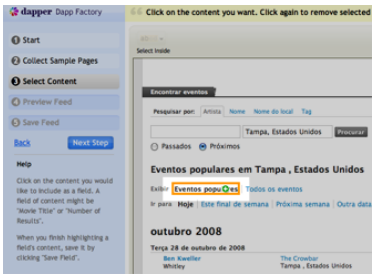


Fig. 2. Dapper: The DataMapper screenshot

This application allows the exportation of the extracted information in several formats (DapXML, RSS, etc).

3 Problem Statement

Relevant events can be referenced in multiple websites, on the other hand, minor events maybe referenced just in one website. In the first case there's a problem of duplicate and sparse information. The same event can appear on several lists and data can be repeated. Sparse information it's a common problem in these types of websites, and causes the user to consult several data sources in order to achieve a full characterization of the object. The second situation can also obligate the user to navigate through multiple websites in order to find the desired item. Both cases can be solved by integrating the data from all the sources.
While searching for an event it is very easy to find several websites that provide lists of occurrences. These websites represent the system input, where each list needs to be extracted and mapped into a predefined structure.
An event description contains several fields. Generally, every event is characterized by:

- **Time fields** - day and hour (optional) when the event occurs;
- **Entity fields**
  - **Event title** - can correspond to the Artist's name;
  - **Address**
  - **City, Country**

[4] http://www.dapper.net/

– **Other descriptors (optional)**

Fields marked as optional correspond to parts of information that are not always available. Not all websites provide information about the hour, price or the event description. Before any integration to be performed it is very important that all information is mapped into the structure, enabling a more direct comparison between data objects.
For this study the 'Other Descriptors' field was discarded, and only the first two fields were considered relevant to the event definition. When two occurrences are classified as similar, a new occurrence is created, containing the fusion between all available fields. This fusion corresponds to a simple additive operation, where each field of the new data object contains all information provided by each occurrence.

4 System Specification

The input of the system is a list of previously chosen websites containing updated and trust worth information. The system should output the list of extracted and integrated events, with no duplicate occurrences. Each occurrence should include all aggregated information, hence representing a more complete characterization of data.
The whole system can be divided into four modules. Although some of them are not the major concern of this study, they will be cited and synthetically described (Figure 3):

– **Data Extraction** - extracts all data from the selected sources and maps it to labeled variables;
– **Data Transformation** - processes extracted data and writes into homogeneous formats;
– **Data Filtering** - after all information is described by standard formats, undesired events are discarded;
– **Data Integration** - all events found are aggregated considering that different occurrences of the same event should be merged.

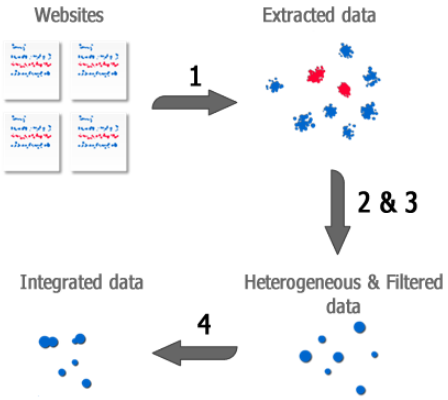


Fig. 3. Functional sequence of the four system modules

The first module extracts and outputs only relevant data. All information is labeled and organized so that the next modules can recognize each field of the structure without the need of extra processing. Each extracted occurrence should have at least a title, an address and a date. More detailed representations can also have an hour.
Some websites only publish events for a specific city, or a specific country, so this information is implicit. When this data is not present in the extracted information, the city and country fields should be manually filled in.
The second part of the system is responsible for interpreting information and rewriting it into standard formats. Data transformation is very important, as it represents the preparation of data for the integration process. In this case the date field is more complex to analyze. This field can be expressed using different sequences of numbers (12/12/2008, 12-12-2008), natural language expressions (today, tomorrow, weekend), or mixed formats (12 December, 12 Dec). Hour is a very sensitive field because there is no need for precision. It is very common to have different websites assigning slightly different values for this field. This situation should be taken into account in the last part of the system.
The next module is a simple module and it's completely optional. Some websites only publish events respecting a specific domain (country, city, etc), thus, data can be initially filtered by the sources themselves. When data is not initially restricted, or when it is necessary to use a more advanced filter, this module becomes very useful. The reason for this element to be in the third position instead of being placed directly after first module is that after the data becomes homogeneous the filter can be generic. At the bottom of the system is the data integration module which is in charge for occurrence comparison and, in case of similar occurrences, to aggregate information into a single data object.

4.1 Data Extraction and Mapping

The majority of the websites that provide events information use almost static templates that are rarely updated. Considering that the number of sources is not very large (to a maximum of some dozens), designing individual extraction applications can be a good approach to the problem. Particularly tailored applications result in precise data extraction since for each website it is possible to add particular customizations.
Using a visual extraction application like 'Dapper: The DataMapper' it is relatively easy to conceive an application to map all fields from the html structure into a well structured xml.
If the outputted result is not filtered enough a filter module can be added. This filter can use regular expressions [3] to remove html tags or other noisy data.
Although it is not directly considered for this study, sometimes the description of an event can contain data from other primary fields, like hour. This information normally matches a distinguishable pattern. Hour normally correspond to something like '11:30' pr '11h30m'. To capture these patterns regular expressions can be used. For instance, $\backslash d\{1,2\}:\backslash d\{1,2\}$ | $\backslash d\{1,2\}h\backslash d\{1,2\}m$ will capture hour data described by this pattern.
Before integration, all information should be adapted to a homogeneous structure. Date and hour are the only fields that can have recognizable morphological differences. Other fields don't need to be processed. Dates can be expressed by sequences of numbers (12/12/2008, 12-12-2008), natural language expressions (weekend, today, tomorrow) or even mixed formats (12 December, 12 Dec). To be integrated, all dates have to be interpreted and transformed into a standard structure (p.e. dd/mm/yy). The hour can have small and simply detectable variations (12:30, 12hours, 12hh30mm).
Like in the extraction phase, these situations, even the complex ones, can be easily solved by using regular expressions. For example, extracting the day, month and year values of a numeric date sequence can be achieved by $\backslash d\{1,2\}$-$\backslash d\{1,2\}$-$\backslash d\{1,4\}$.
After all variables are captured it is only needed to choose the desired format for data.

4.2 Information Integration

Distinct occurrences of the same event can have different amounts of data, can use different structures, or even distinct languages, to describe and display the 'same' information. Since the union of all data provided by multiple websites can be described by:

$$A \cup B \cup C \cup ...N = A + B + C + ...N - (A \cap B \cap C \cap ...N)$$

where A, B C and N are Websites , to integrate[4] corresponds to identify and resolve this intersection. Two occurrences are merged if the system finds that they are similar. To verify if two occurrences are similar the system uses the following similarity function[5]:

$$SIM_{TOTAL} =$$
$$[SD(eo_1, eo_2) \times SHI(eo_1, eo_2) \times SCC(eo_1, eo_2)] \times \sum_{i=0}^{n} K_i \times SIM(eo_1\ field_i, eo_2\ field_i)$$
$$\text{Where } \sum_{i=0}^{n} K_i = 1$$

The labels eo_1 and eo_2 stand for *EventOcurrence 1* and *EventOcurrence 2*. To be considered similar, SIM_{TOTAL} has to be bigger than a specific threshold θ. *SD* is the abbreviation for *SameDate*, *SHI* correspond to *SameHourInterval* and *SCC* means same country and city. Each one of these functions can return *1* or *0*. If *0* is returned by any these functions the two occurrences are considered not similar. It is obvious that, if two occurrences do not happen at the same date, they do not refer to the same event. Even if they take place at the same day but the time interval between them is bigger than α it is clear that is not the same event. $SHI(eo_1, eo_2)$ can only be used when both eo_1 and eo_2 have hour description. These fields are numeric, thus their similarity function tends to be very objective. Other fields, where text is applied, need other types of similarity functions. $SHI(eo_1, eo_2)$ can only be used when both eo_1 and eo_2 have the hour descrition. These fields are numeric and their similarity function tends to be very objective. Other fields, where text is applied, similarity functions are more complex.

Excepting for the city and the country where the text match must be exact, other similarity functions are not so clear-cut. The rest of the SIM_{TOTAL} function is composed by these other similarity functions, where each one is adapted to its specific field. In this case the comparison between fields corresponds to a name matching problem [6]. Considering the corpus of terms for both fields to be infinite unpredictable, a vector based [4] approach is not possible. Instead, these functions should implement character based [6] algorithms. Because there is no language, dictionary matching techniques [7] or phonetic approaches [8] should not be considered. The terms involved can be from any language and use any letters or symbols, so, the algorithms have to be very generic. Before applying the metrics (methods for string comparison), a couple of filters should be used in order to remove the accents, to low case or high case all characters and to remove double spaces. These filters can increase the name matching score without affecting their 'meaning'.

Although there's no specific language for names, Stop-Words [9] removal can also be considered. Having in mind that this is a name matching problem, this technique is very risky because it can damage important name terms. For example, if after removing the accents and low/high case the title of two occurrences, their values are 'TAKE THAT' and 'TAKE THIS'. If the 'THAT' word and the 'THIS' word are part of the 'Stop-Words' list, both titles will be transformed into 'TAKE' causing them to be equally matched.

Several algorithms can be used to solve the name matching problem. In this case the Smith-Waterman algorithm [10][11][12] can be used. This algorithm is designed for performing local sequence alignment. Instead of considering the whole word the algorithm tries to match all substrings, minimizing the cost (number of deletions, insertion, replacements and gaps) and consequently increasing the similarity measure. When comparing two words it should be considered that two mismatch characters is a worse situation than if one of the words has blank spaces.

A gap is a space in the middle of a string. Spaces at the beginning and at the end of a string are not considered gaps. The gap should cost less than an unmatched char-

acter and other blank spaces should not have any cost. A blank space can be due to a misspelling or just because one of the words corresponds to a more generic form of describing the same concept. Non existing data is better than contradictory data.
Common-sense tells us that 'ARTIST A' and 'ARTIST B' are more distant than 'ARTIST' and 'ARTIST B', because in the second situation there's a small probability the data that is lacking corresponds to the data on the other word.If $s1$ and $s2$ share some suffix, then the alignment score will be high.
To normalize the score given by this algorithm the returned value is divided by the size of the string with the smaller length. Even though this normalization makes the algorithm to be insensitive to different string sizes, this situation should not affect the general result of the similarity function. If two words have a great length difference it is possible to change the weight of the field in the similarity function. Although this seems a probable situation, testing with real values showed that this situation it very rare.
Considering that the sources of information are trustworthy and each event occurs on a specific date, and has a specific hour, even if the title or the address of two occurrences only matches partially, the global score can be high. In this situation, if other fields have high scores, it means that the title or the address correspond to poor descriptions or either more generic forms of the same name. Therefore these occurrences should be integrated.
When two occurrences are integrated a new object is created containing the data from both objects. This resulting object represents an aggregation of the other occurrences and for that reason will be denominated as an integrated occurrence. Instead of singular data fields, each one of these will be a list of data items.
After data aggregation is done, the normal occurrences are deleted from the process queue and the new integrated occurrence replaces them. The next item of the queue is compared against all other occurrences, including those that are already integrated occurrences. In the case of a normal occurrence is compared against an integrated occurrence, all values of each field that is part of the integrated occurrence are matched against the value of the corresponding field in the normal occurrence. If there is a similarity a new integrated occurrence is created. This process iterates until there's no more integrations left.

5 Tests and Results

The proposed architecture and techniques were implemented and a system was created to integrate all data concerning concerts that occur in the city of Oporto in Portugal. The set of sources [5] selected is Casa da Musica Agenda, Epilespsia Emocional ,Lastfm events in Porto and Lasfm Events in Portugal. The weights considered for the fields were set concerning each situation.

[5] http://www.casadamusica.com/CulturalAgenda/ , http://www.epilepsiaemocional.org/agenda/ , http://www.lastfm.pt/events?findloc=Porto, http://www.lastfm.pt/events respectively

Figure 4 synthetyzes all settings that were used for the similarity function:

Wheights	'Title'	'Address' + 'AddressDetails'	'Hour'
Hour available	0.5	0.3	0.2
Hour non available	0.3	0.7	....

Fig. 4. Weight table for the similarity function

The graph depicted in Figure 5 shows the number of extracted events for each source along five days.

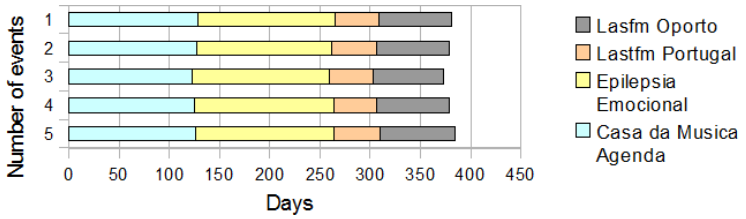


Fig. 5. Number of events extracted for each source during a 5 days period

After running the integration module the number of events for each day decreased considerably, as shown in Figure 6.

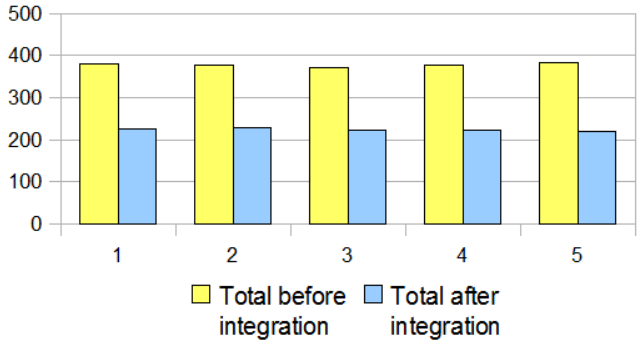


Fig. 6. Graph illustrating the difference between the number of events before and after integration

It is interesting to notice that after integration the final list is much bigger than any list provided by a single source. Comparing the number of events before and after integration it is noticeable a 40% decrease (Figure 6).
During the whole experience only a single false positive was registered. In this case

both ocurrences describe the same address using very distinct names. This situation could be easily corrected by changing dynamically the weights of some fields, and at the end, considering only the best score.
False negatives were not found. A false integration results in data loss, therefore the number of wrong integrations should always be avoided.

6 Conclusions

The presented approach defines a simple architecture for extracting and integrating data concerning events. Even though is focused on a specific domain of information the proposed techniques and methods can be applied to many other contexts.
The extraction module can be far more sophisticated by implementing automatic extraction methods or adding crawling techniques, instead of particularly designed applications. On the other hand, those types of approximations can have lower data granularity.
If the user pretends to integrate events that occur within a specific city the set of event addresses can be considered a finite corpus, enabling the use of vector based techniques, which can offer some advantages.
Even with these possible improvements and with few problems detected, the proposed techniques were successfully tested on a real system, proving that they are valid and robust, despite of being simple.

References

1. Grishman, R.: Information extraction: techniques and challenges. In: In Information Extraction (International Summer School SCIE-97, Springer-Verlag (1997) 10–27
2. Decker, S., Harmelen, F.V., Broekstra, J., Erdmann, M., Fensel, D., Horrocks, I., Klein, M., Melnik, S.: The semantic web - on the respective roles of xml and rdf. IEEE Internet Computing **4** (2000) http://www.ontoknowl
3. Brin, S.: Extracting patterns and relations from the world wide web. In: In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT98. (1998) 172–183
4. Cohen, W., Ravikumar, P., Fienberg, S.: Adaptive name matching. In: in Information Integration. IEEE Intelligent Systems, Sept/Oct. (2003) 2–9
5. Kurtz, S.: Approximate string searching under weighted edit distance. Third South American Workshop on String Processing Recife, Brazil (1996)
6. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. (2003) 73–78
7. Wright, A.H.: Approximate string matching using within-word parallelism. Software Practice and Experience **24** (1994) 337–362
8. Kondrak, G.: Phonetic alignment and similarity. Computers and the Humanities **37** (2003) 273–291
9. Fox, C.: A stop list for general text. SIGIR Forum **24** (1990) 19–21
10. Monge, A.E., Elkan, C.P.: The field matching problem: Algorithms and applications. In: In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. (1996) 267–270

11. Mott, R.: Maximum-likelihood estimation of the statistical distribution of smith-waterman local sequence similarity scores. In: Journal Bulletin of Mathematical Biology. Volume 54 of Lecture Notes in Computer Science., Springer New York (1992) 59–75
12. Bar-yossef, Z., Jayram, T.S., Krauthgamer, R., Kumar, R.: Approximating edit distance efficiently. In: In Proc. FOCS 2004. (2004) 550–559

Improving the Performance of IEEE802.11s Networks using Directional Antennas over Multi-Radio/Multi-Channel Implementation – The Research Challenges

Saravanan Kandasamy[1], Ricardo Morla[1], and Manuel Ricardo[1]
[1] INESC Porto, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 378,
4200-465 Porto,Portugal
{kandasamy, ricardo.morla, mricardo}@inescporto.pt

Abstract. The IEEE802.11s standard is a variety of Wireless Mesh Networks (WMNs), which features infrastructure-less flexible network configurations, is attracting attention as an elemental technology for future ubiquitous networks consisting of various types of nodes built on ad-hoc basis. To solve problems like throughput degradation, delay and fairness, an enhanced Medium Access Control (MAC) protocol may be required taking the advantage of directional antenna (DA) and cross-layer mechanisms. This paper explores in particular the research challenges to improve the performance of WMNs. Analyzing the trend, we are confident towards an enhanced wireless mesh networks performance by means of utilizing directional antennas and cross layer mechanism for the Access Points (APs).

Keywords: IEEE802.11s, Directional Antenna, Multi Radio, Multi Channel

1 Introduction

The IEEE802.11s task group was created by the Institute of Electrical and Electronics Engineers (IEEE) for installation, configuration, and operation of IEEE802.11-based wireless mesh networks (WMNs).WMNs allow for high flexibility in setup and relocation of communication nodes, ubiquitous access, and ease of use at the cost of lower throughput due to interference, high-loss medium, and limited available spectrum. In WMNs, neither predefined infrastructure nor centralized administration is required, as networks can dynamically build by nodes that may be mobile, static or quasi-static. The simplicity of deployment of these networks makes them an attractive choice in scenarios such as disaster recovery, broadband home networking, community and neighborhood networks, military operations, and transportation systems.

The nodes in the WMN are assumed to be equipped with omnidirectional antennas (OA) and, IEEE802.11 Medium Access Control (MAC) standard [1] has been designed only considering this kind of antennas. As the network become larger and denser, the network gets saturated due to the broadcast nature of the technology when the number of users increases with the traditional OA (refer Figure 1). However, with

the rapid advancement of antenna technology, it becomes possible to use directional antennas [2] to improve the capacity of WMNs (refer Figure 2). This paper is an improvement of [3].

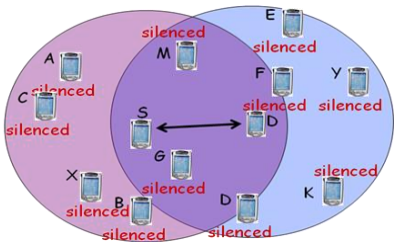


Figure 1 : Node S↔Node D
- Omnidirectional Antenna

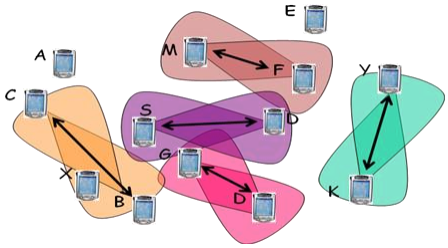


Figure 2 : Node S↔Node D, Node Y↔Node K
- Directional Antenna

Directional antenna (DA) is divided into two categories, one is the traditional directional antenna (refer Figure 3) which is pre-fixed in particular direction and the other is a smart antenna which consist of 3 components; a radiating element, a combining or dividing network and a control unit. The control unit is the intelligence of a smart antenna which is usually implemented using a digital signal processor (DSP). The smart antenna can next be divided into two types, a switched beam antenna and steerable beam antenna [4]. A switched beam antenna (refer Figure 4) combines several directional antenna elements to form up to *N* predetermined directional beams, turned on and off in a determined manner while a steerable beam antenna (refer Figure 5) steers its radiating pattern either electronically or mechanically to focus to an intended direction.

Figure 3 :
Traditional
Directional Antenna

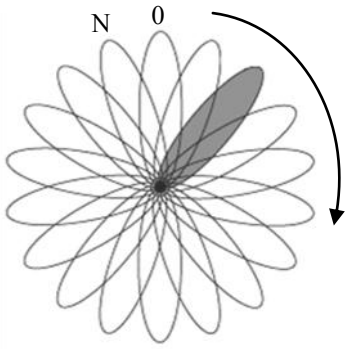


Figure 4 :
Switched Beam Antenna

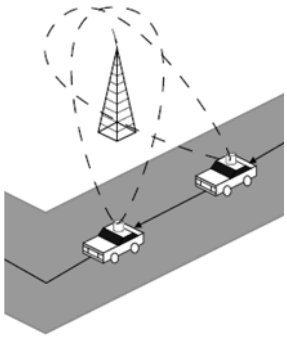


Figure 5 :
Steerable Beam Antenna

Executing real-time applications over IEEE802.11s network are tough due to its mobility uncertainty and fragile radio properties. This is due to real time applications

needed to be delivered within strict quality of service (QoS) requirements. The protocols in the lower layers need to work interactively with the application layer to create application protocols for managing distributed information sharing in WMNs. This requires cross-layer mechanisms through information sharing among application, transport, routing, medium access control (MAC) and physical layers. In this way, the deployed WMN can be self-adaptive to the network.

The rest of this paper is organized as follows: in Section 2, we provide the rationale for this research. Section 3 presents the work carried out in the last 10 years as state of the art in order to emphasize the rationale. In Section 4, we elaborate the research challenges identified through the state of the art, an opportunity for future research work. Section 5 will deliberate on the directional antenna model in IEEE802.11s Network. Finally in Section 6, we conclude the paper.

2 Rationale

Directional antennas offer several interesting advantages for IEEE802.11s networks. For instance,

- by exploiting the gain of the directional antenna, multi small hop transmissions could be reduced to minimal hops and sometimes potentially to a single long hop transmission. This would lower the transmission delay, reduce the number of control signals in the network and reduce congestion due to packet redundancy.
- a node may be able to selectively receive signals from a desired direction. This enables the receiver node to avoid interference that comes from unwanted directions, thus increasing the signal to noise ratio (SNR).
- more users could utilize the network. In an omni-directional antenna scenario (as shown in Figure 1), when Node S communicates with Node D, Node Y would not be able to communicate with Node K as it is in the ***silenced*** region of Node S and Node D, even though the transmission is not directed towards both of them. This does not happen when DA is used (as shown in Figure 2), thus increasing capacity.
- routing performance can be improved using DAs [5] due to its interference reduction capability which minimizes the contention among routes and reduction in the number of routing messages within a WMN.

3 State of the Art

Below we point out some of the recent work on topic related to directional antennas and wireless mesh networks.

3.1 Wireless Mesh Networks

In WMNs, the routing layer needs to work interactively with the MAC layer in order to maximize its performance. The simplest routing metric for WMNs is the hop-count

metric. However, using the hop-count metric leads to suboptimal path selection. Small hop count translates into longer and more error prone individual hops [6]. The use of minimum hop count does not assist to manage the load-balance traffic across the wireless mesh network [7]. This reduces the effective capacity of the WMNs.

The bandwidth availability is harsh for WMNs where the nodes operate over the same radio channel in order to keep the network connected. This results in substantial interference between transmissions from adjacent nodes on the same path as well as neighboring paths, thus, reducing the end-to-end capacity of the network [8].

3.2 Directional Antenna and Directional MAC Protocol

Rappaport [9] described the use of sector antennas on modern cellular base-stations which allow the decreasing of cluster size in order to improve frequency reuse without being afraid of interference. Sectoring at 120 degrees reduces interference significantly and increases capacity by a factor of 1.714.

Nasipuri et al. [10] modified the Request-to-Send (RTS) and Clear-to-Send (CTS) exchange of the MAC protocol in IEEE802.11 networks to support directionality and showed through simulations that as a result, a throughput improvement of 2-3 times over OAs. The primary aim of the work was to minimize routing overhead by using DA elements for propagating routing information as routing overheads from omnidirectional transmissions can be costly. Ko et al. [11] proposed directional MAC (D-MAC), a revamp of IEEE802.11 MAC scheme to support both directional and omnidirectional operation. The D-MAC showed a throughput boost of about 2 times normal IEEE802.11 operation.

Choudhury et al. [12] designed a protocol which uses Multi-hop RTS's MAC (MMAC) to establish links between distant nodes, and then transmits CTS, DATA, and Acknowledge (ACK) packets over a single hop. The results showed that MMAC outperforms IEEE802.11 but the performance depends on the topology and flow patterns. Yi et al. [13] presented an analytical model for evaluating network capacity using DAs. The work showed that with proper tuning, capacity improvements using directional antenna over omnidirectional antennas are improved.

DAs not only improves the network capacity, but they show to be more stable in terms of link quality and not affected by routing metrics. Chebrolu et al. [14] showed that IEEE802.11 long distance links using DAs result in almost “wire-like" characteristics with error rates as a function of the received signal strength behaving close to theory. The time correlation of any packet errors is negligible across the range of time-scales, and links are robust to rain and fog. Under such conditions, routing metrics for wireless links become less and less important.

MAC/DA1 [11] is one of the first efforts to adapt the IEEE802.11 MAC Distributed Coordination Function (DCF) scheme for DAs. Its key feature is the usage of directional RTS frame. On one hand, it narrows the area in which an unintended receiver can overhear the RTS frame and thus significantly relieves the exposed terminal problem. On the other hand, by recording the directions from which the CTS frames are recently overheard and then blocking the antenna elements in the corresponding sectors, a node is further allowed to transmit in the directions that will not collide with other data transmissions, in addition to relieving the hidden terminal

problem. The prerequisite of transmitting a directional RTS or DATA frame is the knowledge of the direction of the intended receiver, which is referred to as the location tracking problem. The solution that MAC/DA1 suggests is to equip every node with GPS support and rely on a beacon protocol for nodes to exchange location information periodically.

MAC/DA2 [10] mechanism exploits the ability of a receiver to determine the direction of an arriving frame in order for the transmitting and the receiving nodes to learn from each other's direction. In contrasts to MAC/DA1, it accomplishes location tracking in on-demand manner, rather than pro-active manner. However, since it uses OAs to transmit RTS and CTS frame, it does not have the benefit of directional RTS frame as in MAC/DA1. MAC/DA2ACK [10] is a modification of the IEEE802.11 MAC DCF specifications with DA support using ACK frame as earlier MAC/DA1 and MAC/DA2 does not include the ACK frame but only RTS, CTS and Data frames are transmitted. Node mobility is expected to degrade the performance of MAC/DA2ACK protocol.

DBTMA/DA [15] splits a single channel into two sub-channels and uses directional busy-tones. It shares the similar feature of the directional RTS frame scheme of MAC/DA1, in that it reserves the network capacity in a finer grain and relieves the exposed terminal problem. By using directional receiving busy tones, it realizes a similar functionality of blocking the corresponding antenna element in the direction from which omnidirectional CTS frame is received in MAC/DA1. Since DBTMA/DA does not rely on the directional RTS frame to solve the exposed terminal problem or to expand effective network capacity, location tracking is only used for transmitting data frame directionally. Therefore, could use an on-demand location tracking mechanism.

3.3 Multi-Radio/Multi-Channel Wireless Mesh Network

The proliferation of IEEE802.11 network in recent years has influenced the drop prices of its RF components tremendously. This lower cost, allows network planners to consider using two or more radios in the same device.

Paramvir Bahl et al. [16], argue that wireless systems that use multiple radios in a collaborative manner dramatically improve system performance and functionality over the traditional single radio wireless systems that are popular today. He shows a median throughput increase by over 70% when two radios are used compared to one radio. Thus confirming multi-radio platform offers significant benefits for wireless mesh networks.

In [17], Raniwala et al. propose an iterative algorithm which aims at assigning channels to radio and routing a predefined traffic profile. He shows that it is possible to achieve a factor of up to 8 improvements with two network interface cards (NIC) in the overall network throughput when compared with the conventional single-NIC-per-node WMN, which is inherently limited to one single radio channel. The performance evaluation demonstrated that the multi-channel wireless mesh network architecture is promising.

3.4 Multi-Radio/Multi-Channel MAC Protocols

MCSMA MAC [18], similar to an FDMA system, the available bandwidth is divided into non-overlapping channels, i.e., n data channels and one control channel. This division is independent of the number of nodes in the system. A node that has packets to be transmitted selects an appropriate data channel for its transmission. When a node is idle, it monitors all the n data channels and all the channels for which the Total Received Signal Strength (TRSS), estimated by the sum of various individual multipath components of the signal, below a sensing threshold (ST) are marked idle channels. When a channel is idle for sufficient amount of time, it is added to the free channel list.

ICSMA [19] is designed to overcome exposed terminal problem present in the single-channel MAC protocol. It is a two-channel system which the handshake process is interleaved between the two channels i.e if a sender sends RTS on channel 1 and if the receiver accepts the request, it sends the corresponding CTS in channel 2 (as shown in Figure 6). If the sender receives the CTS packet, it begins the transmission of DATA packets over channel 1. Again the receiver, if the data is successfully received, responds with ACK packet over channel 2. Figure 7 shows the simultaneous transmission capability between node A and node B. This simple mechanism of interleaving carrier sense enhances the throughput achieved by the two-channel WMNs.

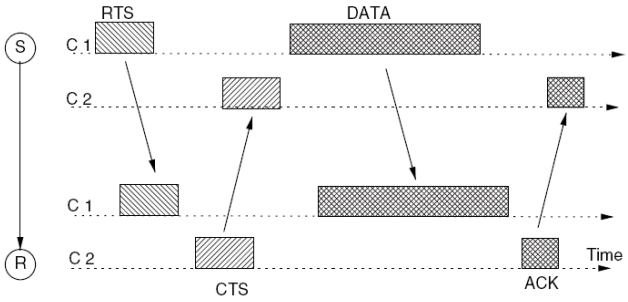


Figure 6 - Interleaved packet transmission in ICSMA

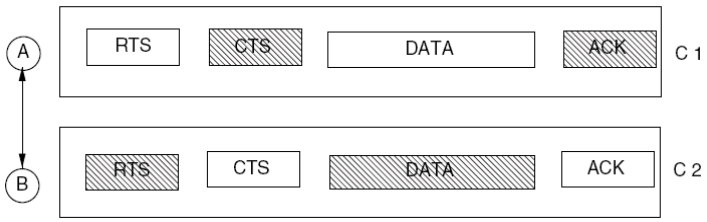


Figure 7 - Simultaneous data transmission between two nodes

2P-TDMA [20] provides an efficient MAC in a single channel, point-to-point, wide area WMN (WAWMN) with multiple radios and directional antennas. The CSMA/CA performs extremely poor in multihop wireless networks such as WMNs [20,21]. Even with the use of directional antennas with high directionality, CSMA/CA

fails to provide simultaneous operation across multiple interfaces. Figure 8 shows an example scenario with two receivers and a central transmitter in a WAWMN. The central node i.e Node 1 has highly directional antennas as that of both Node 2 and Node 3.

Contrary to the notion that the two links, 1→2 and 1→3, can transmit or receive simultaneously, in practical situations it is not possible to provide error-free simultaneous communication when CSMA/CA, in its original form, is employed. The primary advantages of 2P-TDMA protocol include high throughput achieved in a WAWMN and the efficiency in using multiple radios over a single channel. Disadvantages of 2P-TDMA include the inability of the protocol to operate in a general WMN network.

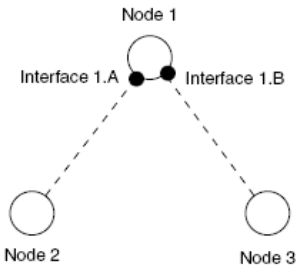


Figure 8 - Topology for 2P-TDMA

3.5 Cross Layer

A cross-layer design approach is one that utilizes information across different layers of the protocol stack for specific improved function. A number of studies over recent years highlighted that cross-layer designs that support information exchange between layers can yield significant performance gains [22-23]. The penalty that has to be paid for deploying cross-layer designs is the complexity and communication overhead.

4 Research Challenges

We identify some of the latest research challenges as below which could be a good opportunity for future works.

4.1 Directional Antenna

The existing IEEE802.11 standard [1] does not gain with the implementation DA and poses additional technical challenges such as hidden terminal problem, deafness, and capture problem [25-27]. A suitable MAC protocol must be designed to best utilize the DA which increases gain in certain direction. While this is seen as an advantage, it may also worsen communications in the direction of transmission. Therefore,

transmitted power should be contained until the destination node, not only to reduce the interfering level at undesired direction but also to save power. A power saving MAC is critical in the cases of battery operated AP or user equipment (UE) and it helps to increase the number of users in the network due to the radio compaction in the system.

4.2 Cross Layer

Cross layer optimization in WMNs are desired as it is impossible to design a universal routing, MAC, multicast or transport protocol that is expected to function correctly and efficiently in all situations. The protocols in the lower layers need to work interactively with the application layer which is a challenge on its own. This requires a cross-layer approach through information sharing among application, transport, routing, medium access control (MAC) and physical layers [22-23]. In this way, the deployed WMN can be self-adaptive to network dynamics and meet end-to-end real-time deadlines of the applications.

4.3 Mobility

Under presence of mobility, dynamic neighbor discovery is a research challenge [10],[26]. AP/UE will come into a mesh and disappears in random manner. Admission and removal of an AP/UE should be fast once discovered.

4.4 Routing

It is a challenge not only to determine the best routing [28] but also a fair routing which communicates with lesser number of hops. This would assist in reducing control signal overheads and congestion in the network hence increasing throughput for good user experience. Ad hoc On Demand Distance Vector (AODV) and Optimized Link State Routing (OLSR) [28] are routing adopted in IEEE802.11s. WMN often a large overlap in radio coverage among nodes, if every node that receives a broadcast packet retransmits it to all its neighbors the neighbors will receive many copies of the same packet. OLSR minimizes this effect by selecting Multi Point Relay (MPR) nodes, which are the only ones that actively retransmit broadcasts. A minimum set of MPRs should be chosen to ensure broadcast coverage for entire mesh network.

4.5 Multi-Radio/Multi-Channel Wireless Mesh Network

Channel allocation is a network-wide process where the allocation of non-interfering channels would lead to significant throughput and media access performance. The channel allocation should consider the number of channels available, the number of interfaces and the technology available. Therefore, techniques such as graph coloring are used for generating channel allocation strategies [29-30]. Wireless channels are

prone to more errors compared to wired-network, thus graceful degradation of communication quality during high channel errors are necessary. In order to achieve graceful quality degradation WMNs need to employ frequency and channel diversity at the expense of additional radio interface at UE instead of loosing full fledge connectivity. By using multiple radio interfaces, the multi-radio or multi-channel system can use appropriate radio switching modules to achieve fault tolerance in communication either by switching the radios, channels by using multiple radios simultaneously.

5 Directional Antenna Model in IEEE802.11s Network

We present the below the progress of this research work which the author is doing presently. There are two ways a directional antenna could be simulated in a IEEE802.11s network, either by feeding in an actual antenna's radiation pattern with respective of its angles depending if it is a switched or steerable beam antenna or another way is by simulating a generic mathematical model. We would use the latter, a cone with spherical cap model, which is a mathematical approximation for simplicity as shown in Figure 9 [4, 24].The beam width of the antenna is 2θ. We will use a switched beam system (as shown in Figure 4) as it is attractive for IEEE802.11s due to its cheaper deployment cost network than a steerable beam system [24].

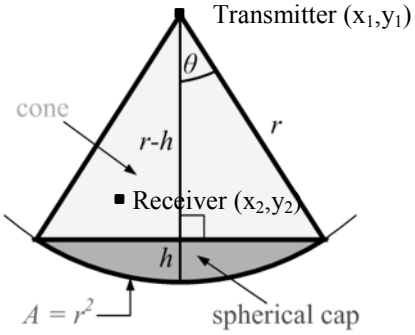


Figure 9 – Directional Antenna Model

When a transmitter transmits using a directional antenna, the receiver should lay inside its three-dimensional radiation beam i.e its solid angle before further communication could be established, returning its received gain in ratio of its solid angle.

angle = $\tan^{-1}$ ($\Delta y/\Delta x$)
angle = angle*180/ π
θ= angle/2

Solid Angle (Sphere) = $dA/r^2 = 4\pi r^2/r^2 = 4\pi$
Solid Angle (Cone with Spherical Cap) = $2\pi\ (1 - \cos\theta)$

Thus, Solid Angle Ratio (SAR)
= Solid Angle(Sphere)/Solid Angle(Cone with Spherical Cap)
= $4\pi/2\pi(1- \cos\theta)= 2/(1-\cos\theta)$

Algorithm,
(1) $Gain_{rx}$ = 0.0;
(2) if(UAngle > LAngle){
(3) if(angle >= LAngle && angle <= UAngle) $Gain_{rx}$ = 1.0;
/*Normal case*/
(4) }else if(UAngle < LAngle){
(5) if(angle >= LAngle || angle <= UAngle) $Gain_{rx}$ = 1.0;
/*e.g LAngle=350 and UAngle=10*/
(6) }else{
(7) $Gain_{rx}$ = 1.0;
/*both == 360 Deg*/
(8) }
(9) return $Gain_{rx}$ *SAR;

6 Conclusion

As the evolution of WMNs technologies continues, IEEE802.11s, directional antenna, directional MAC protocols and cross-layer approaches are being increasingly studied over multi-radio/multi-channel implementations for confronting unintended performance degradations. Not surprisingly, over the last decade those highlighted avenues in this paper have evolved into a hot research topic as presented in the state of the art section. The research activities are still growing due to the immense benefit it could contribute to the WMN field and the research challenges presented in this paper provides future focus of work. With this, we envisage an evolution towards improved wireless mesh networks performance utilizing directional antennas for both the UE and also the AP.

References

1. IEEE P802.11.WLAN medium access control (MAC) and physical layer (PHY) specifications, (1999).
2. Lehne PH, Pettersen M., ”An overview of smart antenna technology for mobile communications systems”, IEEE Communication Surveys 1999; 2(4): 2–13 (1999).

3. Saravanan Kandasamy, “Research Challenges Utilizing Directional Antenna for Improving the Performance of Wireless Mesh Networks", MapTele Workshop 2008,pp15-19, Aveiro,Portugal, (2008).
4. Hongning Dai, Kam-Wing Ng and Min-You Wu, “An Overview of MAC Protocols with Directional Antennas in Wireless Ad-Hoc Networks", International Multi-Conference on Computing in the Global Information Technology, pp 84-91, (2006).
5. Choudhury,Vaidya NH,“Impact of directional antennas on adhoc networks”, IFIP PWC’03, (2003).
6. S.Douglas, J. De Couto, Daniel Aguayo, Benjamin A. Chambers, Robert Morris, ‘Performance of Multihop Wireless Networks: Shortest Path Is Not Enough’, Proceedings of HotNets, (2002).
7. Ashish Raniwala, Tzicker Chiueh. ‘Architecture and Algorithms for an IEEE 802.11 based Wireless Mesh Network’, Proceedings of IEEE Infocom, (2005).
8. J. Jun, M.L. Sichitiu. ‘The Nominal Capacity of Wireless Mesh Networks,’ Wireless Communications, IEEE, Vol. 10, No. 5, pp. 8–14, (2003).
9. T. S. Rappaport, “Wireless Communications: Principles and Practice", Prentice-Hall, 2006.
10. Asis Nasipuri, Shengchun Ye, and Roben E. Hiromoto, “A MAC Protocol for Mobile Ad-hoc Networks Using Directional Antennas,” in IEEE WCNC, Chicago, IL, pp 1214-1219 ,(2000).
11. Young-Bae KO, Jong-Mu Choi, and Nitin H. Vaidya, “MAC Protocols Using Directional Antennas in IEEE802.11 based Ad hoc Networks,” Journal Wireless Communications and Mobile Computing, (2007).
12. R. R. Choudhury, X. Yang, R. Ramanathan, and Nitin Vaidya,"Using directional antennas for medium access control in ad hoc networks", In Proceedings of ACM MobiCom'02, pp 59-70, (2002).
13. Su Yi, Yong Pei, and Shivkumar Kalyanaraman, “On the Capacity Improvement of Ad Hoc Wireless Networks Using Directional Antennas," Proceedings of ACM MOBIHOC, pp. 108-116,Annapolis, (2003).
14. K. Chebrolu, B. Raman, and S. Sen, “Long-Distance 802.11b Links: Performance Measurements and Experience", 12th Annual Int. Conf. on Mobile Computing and Networking (MOBICOM), USA,(2006).
15. Zhuochuan Humg, Chien-Chung Shen, Chavalit Sfisathapomphat, and Chnipom Jalkaeo, “A Busy-Tone Based Directional MAC Protocol for Ad Hoc Networks,“ in IEEE MILCOM, Anaheim. CA, (2002).
16. P. Bahl, A. Adya, J. Padhye, and A. Wolman, “Reconsidering Wireless Systems with Multiple Radios”, ACM Computer Communications Review, vol. 34, no. 5, pp. 39–46, (2004).
17. A. Raniwala, K. Gopalan, and T. Chiueh, “Centralized channel assignment and routing algorithms for multi-channel wireless mesh networks,” ACM Mobile Computing and Communications Review, vol. 8, no. 2, pp. 50-65, (2004).
18. A. Nasipuri, J. Zhuang, and S.R. Das, “A Multi-Channel CSMA MAC Protocol for Multi-Hop Wireless Networks”, Proceedings of IEEE WCNC 1999, pp. 1402–1406, (1999).
19. S. Jagadeesan, B.S. Manoj, and C. Siva Ram Murthy, “Interleaved Carrier Sense Multiple Access: An Efficient MAC Protocol for Ad hoc Wireless Networks”, Proceedings of IEEE ICC 2003, pp. 1124–1128, (2003).
20. B. Raman and K. Chebrolu, “Design and Evaluation of a New MAC Protocol for Long-distance 802.11 Mesh Network”, Proceedings of ACM MobiCom 2005, pp. 156–169, September (2005).
21. S. Xue and T. Saadawi, “Revealing the Problems with 802.11 Medium Access Control Protocol in Multi-Hop Wireless Ad Hoc Networks”, Computer Networks, vol. 38, (2002).

22. Fotis Foukalas, Vangelis Gazis, and Nancy Alonistioti, Cross-Layer Design Proposals for Wireless Mobile Networks: A Survey and Taxonomy, IEEE Communications Surveys & Tutorials, (2008).
23. Ian F. Akyildiz, X.Wang and W.Wang, "Wireless Mesh Networks: A Survey", Computer Networks, (2005).
24. Vishwanath Ramamurthi, Abu S. Reaz, Sudhir Dixit, Biswanath Mukherjee, "Link Scheduling and Power Control in Wireless Mesh Networks with Directional Antennas" , IEEE ICC, Beijing China, May 19-23, (2008)
25. Thanasis Korakis, Jakllari and L.Tassiualas, "CDR-MAC:A Protocol for Full Exploitation of Directional Antennas in AdHoc Wireless Networks", IEEE Transaction on Mobile Computing, Vol 7,No.2,(2008).
26. Romit R.Choudhury,X.Yang, Ram Ramanathan and Nitin H.Vaidya, "On Designing MAC Protocols for Wireless Networks using Directional Antenna",IEEE Transaction on Mobile Computing, Vol 5, (2006).
27. P. Sai Kiran, "A Survey on Mobility Support by Mac Protocols Using Directional Antennas for Wireless Ad Hoc Networks", International Symposium on Ad Hoc and Ubiquitous Computing, pp. 148-153, (2006).
28. T. Clausen and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)." RFC 3626, (2003).
29. Partha Duttay, Sharad Jaiswal, Debmalya Panigrahiz and Rajeev Rastogi, "A New Channel Assignment Mechanism for Rural Wireless Mesh Networks", Infocom (2008).
30. Andrew Chickadel, "Interference Reduction in Wireless Networks Using Graph Coloring Methods", Computer Science Research Symposium 2007, pp 22-29 (2007).

Towards the Optimization of Video P2P Streaming over Wireless Mesh Networks

Nuno Salta[1], Ricardo Morla[1] and Manuel Ricardo[1]

[1] INESC Porto, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 378,
4200-465 Porto,
Portugal
{nsalta, ricardo.morla, mricardo}@inescporto.pt

Abstract. P2P overlay networks are large-scale distributed systems. They are used mainly for data sharing and content distribution and can be more efficient than the traditional client-server data models. On the other hand, emerging Mesh networks are effective solutions for ubiquitous broadband access. High throughput, cost effectiveness, and ease of deployment are key features of these networks. The deployment of P2P systems over mesh networks is the main topic of this paper. Both systems need to form multiple shortest path trees in order to transfer data and the joint optimization of these systems, using a cross layer approach, promises good gains in terms of data throughput and fairness. In this paper, we describe the current solutions on Video P2P Streaming and we discuss some research opportunities to be addressed on our work.

Keywords: P2P, Video, Streaming, WMN.

1 Introduction

In the past years, we have assisted to a change of the Internet paradigm. The global network is evolving from a homogeneous network topology to a emergence of heterogeneous and mobile topology. In addition, WLANs are changing the way people access Internet, supporting the network concept of always connected, anytime, anywhere.

From all the new kinds of computer networks, Peer-to-peer (P2P) is playing a decisive role on today's Internet. The traffic produced by this type of overlay network is now dominant in the total Internet traffic [1].The emergence of P2P networks can be seen as a response to the inefficiency of the traditional server-client model, incapable of meeting the demands of a global network that grows at an increasing pace, where clients can have processing power similar to those of servers. These overlay networks build up in a dynamic, distributed way, with capacity of auto-configuration, among other features. The serverless approach allowed the introduction of new applications and services, with file sharing being the most prominent. Other

services, however, are now emerging and subject of more research such as video broadcast over P2P networks.

Network protocols are commonly divided into several independent communication layers, which interact through interfaces. The most successful network stack is TCP/IP that comprises physical, medium access, network, and application layers. This model was designed based on the wired networks of the original Internet whose physical topology is static. Wireless Networks have characteristics such as high latency, lossy links, and constant-changing capacity; applying the old TCP/IP model to them can lead to suboptimal use of the network resources due to the lack of cooperation between the lower layers (Physical, MAC, and Network). On the other hand, overlay applications such as P2P do not usually consider the conditions of the lower layers such as their current routing tables. This has a special importance in Ad-Hoc networks and mesh networks since the constant changing topology impacts the system performance.

This paper is structured as follows: section 2 presents our motivation and the research scenario, defining research questions to be addressed. Section 3 gives an insight on the technical background. Section 4 shows the current trends on the video P2P streaming area and Section 5 presents the research challenges to be addressed on our work. At last, in section 6 we draw our conclusions.

2 Motivation

Video-over-IP traffic is rising in recent years as the number of Internet users interested in this type of media is increasing. The year of 2008 saw an increase of video streams delivered through professional content sites in 43.4% to 33.5 billion [2] online video views without considering user-generated videos. In middle 2006, Youtube [3] hosted around 45 terabytes of information, having more than 1.7 billion views [4].

The client-server service model is the typical solution for streaming over the Internet. The client establishes a connection to a video server and asks the transmission of a given video content, which is then streamed by the server if the request was successful. One of the most challenging concerns of video streaming server solutions is scalability. To provide a good quality stream, high bandwidths are required. Video source servers should have bandwidth provision that grows with the number of clients. These provisions can have significant costs for deploying server based video stream solutions.

P2P networks break the traditional server-client model to build a distributed system. Since all nodes act as both client and server, they not only download data but also upload to other nodes on the network. This topology enables a more efficient use of the upload bandwidth of the users, reducing in this way the overall bandwidth load of the network. P2P file-sharing applications such as Napster [5] or more recently, Bittorrent [6] Limewire [7], and eMule [8] allow a fast diffusion of data files on the Internet. More recently, P2P applications have emerged to provide video streaming services, both live video or on-demand. Some of the current deployed applications are PPLive [9], PPstream [10], Sopcast, [11], and TVUplayer [12].

Since P2P is an overlay network, it relies on a physical topology. In general, P2P applications do not take into account the physical topology when they choose the neighbors. This can lead to suboptimal routing schemes at the network layer, with lower throughput and higher delays. Figure 1 shows a possible map between the overlay links and the physical networks links. A single overlay hop can translate into multiple physical hops, which leads to higher waste of bandwidth.

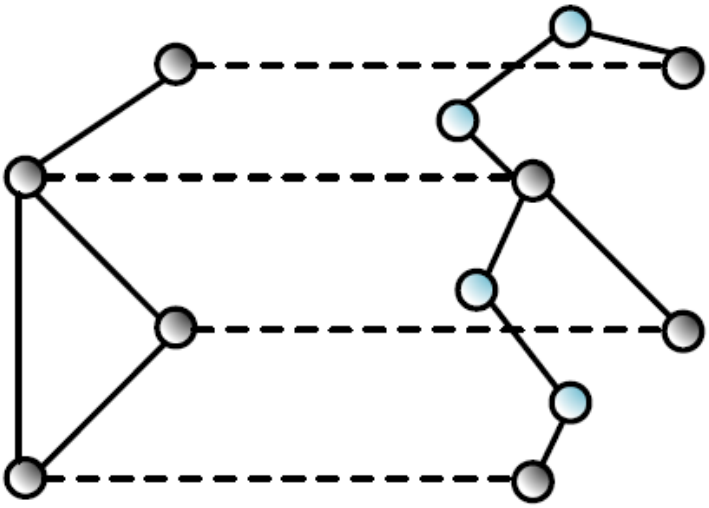


Fig. 1. Logical and physical connections mapping on a P2P system.

In the typical case of the global Internet, is difficult to change the physical topology and better match P2P requirements due the static link connections between the nodes. However, in scenarios where wireless and mobile nodes form mesh networks, the same limitations do not apply, especially if the usage of the P2P resources is confined to the mesh network.

Wireless Mesh Networks, such as 802.11s [13], have a set if features that makes them potentially capable of improving the performance of P2P applications: mesh networks are scalable; the neighbor links can be easily changed on demand; its ad-hoc characteristic enables an intrinsic P2P behavior, which can benefit P2P services, particularly the low-delay required services like video streaming. The ability to change both topologies enables that the overlay and physical layers to be mapped in a more effective way to provide better performance.

Real life scenarios employing P2P video broadcast could benefit from a better exploration of mesh networks capabilities. One of such scenarios is the public transportations as depicted by Figure 2. In this scenario buses would act as mesh

nodes, with capacity to connect with other buses or bus stops, establishing a decentralized network.

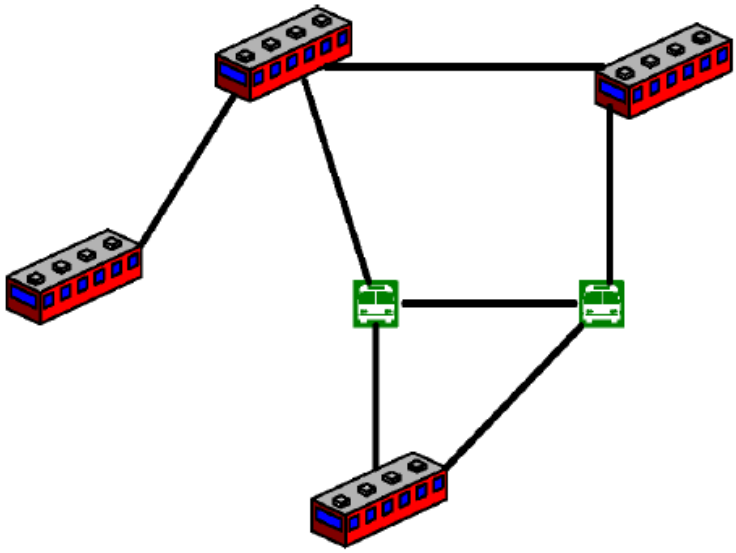


Fig. 2. Possible scenario for WMN implementation based on a public transportation system.

3 Technical Background

In this section, we present a brief technical background on P2P and Wireless Mesh Networks.

3.1 Peer-to-Peer Networks

Peer-to-peer networks are essentially an overlay topology based on a distributed system paradigm, where the end nodes, peers, are not usually constrain to a central organization [14]. Each participating node can operate both as a client and as a server and define direct links to neighbors at the logical application level. This kind of organization promotes equality among the nodes, allowing them to build a robust self-organized overlay network upon the original IP network.

P2P overlay networks can provide a vast list of features such as fault tolerant network architecture; resource sharing; distributed and redundant storage; searching mechanisms without relying on centralized entities; anonymity; high scalability.

In addition, P2P allows the deployment of services present on centralized architectures as hierarchical naming, trust, authentication, and search capabilities. Each peer contributes its capabilities to improve and scale the overlay network as storage space, computational power, and bandwidth. The distributed characteristic of P2P is highly valuable for various scenarios such is file sharing, multimedia distribution, and real time data that can exploit the P2P features to overcome the issues commonly attached to central management systems.

3.2 Video P2P Streaming

P2P Video broadcast is a specific type of P2P applications focused on redistributing video streams over the network. P2P video distribution systems can be classified as tree-based and mesh-based.

In tree-based systems, the overlay structure is typically well defined. Each child peer receives data through their parent peer.

However, this kind of layout is very vulnerable to peers departure. When a peer leaves the system, it will stop the streaming of video to its children peers, having the latter to cope with the loss by finding another parent. Commonly, tree-based systems follow a single-tree model or a multi-tree model.

In single-tree systems, users form a tree, at the overlay level, that has its root on the streaming server. Each participating peer, joins the network at a certain level, receives the video streams from its parent, and is responsible of forwarding the stream to its children. The joining process should be able to balance the tree as much as possible to minimize the delays of the peers at the bottom level. Tree maintenance is also important due its fragility to node departures, which disrupts the stream for children node. To minimize the impact of a leaving node, some mechanisms can be employed like the using of timeouts to detect a departure. This can be managed by a central entity or in a distributed system.

In multi-tree systems, the server divides the original stream into multiple sub-streams, creating this way multiple sub-trees, one for each sub-stream. When a node joins the networks, it tries to connect to all sub-trees needed to retrieve the video. In addition, a peer can be at different levels on each of the sub-trees.

On the other hand, swarm-based systems, try to avoid the former hierarchical architecture, in order to avoid the single-point of failure characteristic of the tree-based systems. Mesh-based systems can maintain multiple parent connections at same time and can change the topology dynamically. Peers can connect randomly to neighbors and can be a parent or a child of a given neighbor on different instances. Several applications follow this model such as [15] [16]. A recent study [17] showed that mesh systems have better performance than tree systems, concerning bandwidth usage, resilience and self-healing.

3.3 Mesh Networks

Wireless Mesh Networks (WMN) is an emerging wireless technology vowed to change the paradigm of network topologies. Internet Service Providers as well as end-

user are showing an increasing interest on this technology because of its broad list of features, namely its robustness and reliability at a low cost deployment.

WMN are dynamically self-organized and self-configurable, where the nodes are capable to establish an ad-hoc topology and maintain the mesh connectivity. This model allows easier network maintenance, reliable service coverage, and sustainable scalability. WMN also employs advanced radio technologies, like multiple radio interfaces and smart antennas, increasing the network capacity.

Some features of WMN are as follows: **increased reliability** - the possibility of having redundant paths eliminates the single point failures and reduces bottleneck links; **low installment costs** - using the traditional infrastructure topology to deploy wireless networks on metro scale scenarios, would require a large number of higher cost access points while using less expensive WMN enabled nodes would reduce the total cost of the network whilst maintaining a high level of connectivity for all nodes; **large coverage area** - with the multi-hop capability of mesh networks, the covered area can be significantly increased; **automatic network connectivity** - WMN enabled nodes are capable of finding and establishing connections to its neighbor, providing network connectivity.

WMN usually comprehends two types of nodes: mesh routers and mesh clients. Besides the common routing capability for gateway functions, mesh routers contain additional functions to support specific mesh networking requirements. Multi-hop communications not only allow larger coverage areas but also the same coverage area of a regular wireless network to be achieved by a mesh router with much less transmission power. The flexibility of WMN can be improved by using mesh routers with multiple interfaces, even for different technologies. Mesh routers are supposed to have low mobility in order to provide a solid network backbone for the mesh clients.

Mesh clients, which also have some routing functionalities, can have a simpler hardware platforms when comparing to mesh routers. Moreover, mesh routers have the capability of integrating the mesh network within various other networks, bestowing WMN of diversity facilities, which makes this technology more than a simple ad-hoc network.

4 Related Work

In this section, we focus on current work done related to P2P and WMNs as well as propose some research directions in this area.

The need to improve P2P networks performance taking into account not only the logical but also the physical topology of the network is a subject already addressed by the community. We can divide the work on this area in two branches: generic P2P and video streaming specific P2P with each one sub-divided in solutions for WMN or for general networks. [18] [19] [20] [21] are works on generic P2P that can provide an insight on how the application network layers can interact, based on cross-layer design, in order to optimize the performance of P2P using cross-layer mechanisms. Moreover, [20] [21] provide information in how to exploit the characteristics of WMN to achieve better performances.

On the other hand, in Video P2P Streaming [22] [23] [24] [25] provide useful information on the special characteristics of Video P2P Streaming, in both WMN and non-WMN scenarios.

5 Research Challenges

In order to improve the overall network performance on P2P video streaming our research work should take into account:

- The characteristics of P2P applications for real-time video streaming over WMN;
- The characteristics of Mesh networks;
- Novel MAC layer approaches used to improve mesh networks performance;
- Node mobility and network mobility.

Analyzing the current solutions, we can find room to introduce improvements to video P2P streaming on WMN. Consider figure 3 that depicts an overlay network based on a swarm Video P2P solutions deployed over a WMN:

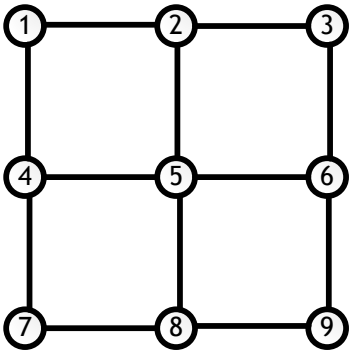


Fig. 3. Mesh network.

Assuming that more than one node will be watching the same stream, is expected that these nodes will need the same video chunks. If node 6 requests a given chuck to the overlay network and node 1 accepts the request, the latter will send the chuck to node 6 after finding a route using a given routing protocol. In the case of the route be 1-4-5-6 and node 4 also needs the same chuck, even if it did not requested the chunk yet, a cross-layer approach could be used to allow the network layer inform the application that a potential needed chunk is being forwarded by the node. This would avoid node 4 to request the same chuck later. We could also extend the previous approach, considering that every node would be monitoring every transmission at link layer and inform the upper layers when a potential needed packet is captured. This approach can have some drawbacks such as higher energy consumption.

Other approach could be clustering the network in order to avoid request flooding on the WMN, where the nodes nearer to the source would have the most recent chunks. This could be particularly interesting on mobility scenarios like in cases when a node is moving away from the source to an area where it could be used as a new source since it has newer chunks.

Based on the presented research challenges, our work has the potential to produce the following original contributions:

- A new routing protocol that meets the objective of improving the performance of the network based on the application and link layers changing characteristics.
- Enhanced P2P video streaming applications over Wireless Mesh Networks.
- Improved P2P mobility in scenarios where peers and mesh nodes move.

6 Conclusion and Future Work

In this paper, we presented the emerging Video P2P streaming solutions over WMN and the motivation of our research work. After showing some technical background, we indicated the current related work and the research challenges on this area.

Our approach will focus on real-time video streaming and should evaluate solutions based on tree or swarm model, support location awareness, minimize duplicated content and improve mobility for an uninterrupted service.

This research should lead to the creation and validation of new routing mechanisms that allows optimizing the available resources and ultimately improving the overall system performance. To meet the aforementioned objective, the problem shall be addressed in three phases: requirements evaluation through an extensive literature review; simulation using suitable software in order to test and validate the preliminary solutions; implementation of a valid solution in a real system in order to assess and confirm the results obtained through the simulations.

References

1. Torrentfreak, http://torrentfreak.com/p2p-traffic-still-booming-071128/ accessed in 2009-02-02
2. Accuestream, http://www.accustreamresearch.com accessed in 2009-02-02
3. Youtube, http://www.youtube.com accessed in 2009-02-02
4. Youtube by the numbers, http://www.micropersuasion.com/2006/08/youtubebythe.html accessed in 2009-02-02
5. Napster, http://www.napster.com accessed in 2009-02-02
6. BitTorrent, http://www.bittorrent.com. accessed in 2009-02-02
7. LimeWire, http://www.limewire.com accessed in 2009-02-02
8. eMule, http://www.emule-project.net accessed in 2009-02-02

9. PPLive, http://www.pplive.com accessed in 2009-02-02
10. PPStream, http://www.tvants-ppstream.com accessed in 2009-02-02
11. SopCast, http://www.sopcast.com accessed in 2009-02-02
12. TVU Networks, http://www.tvunetworks.com accessed in 2009-02-02
13. IEEE draft, 802.11s , 2006.
14. D. Schoder and K. Fischbach,: Peer-to-peer prospects, Communications of the ACM, vol. 46(2), pp. 27–29, 2003.
15. Pai V, Kumar K, K. Tamilmani, V. Sambamurthy, A. Mohr: Chainsaw: eliminating trees from overlay multicast, in The fourth international workshop on peer-to-peer systems, 2005.
16. N. Magharei, R. Rejaie, "Prime: peer-to-peer receiver driven mesh-based streaming," in In Proceedings of IEEE INFOCOM, 2007.
17. N. Magharei N, R. Rejaie, Y. Guo: Mesh or multiple-tree: a comparative study of live p2p streaming approaches, in Proceedings of IEEE INFOCOM, 2007.
18. W. Wu, Y. Chen, X. Zhang, X. Shi, L. Cong, B. Deng, X. Li: LDHT: Locality-aware Distributed Hash Tables, in National Basic Research Program of China, 2007.
19. Z. Xu, R. Min, Y. Hu,: HIERAS: A DHT Based Hierarchical P2P Routing Algorithm, in Proceedings of the 2003 International Conference on Parallel Processing, 2003.
20. E. Conti and G. Turi: A Cross-layer Optimization of Gnutella for Mobile Ad hoc Networks., in In Proc. of the 6th ACM International Symposium on Mobile ad hoc networking and computing, 2005.
21. H. Park, W. Kim, and M. Woo: A Gnutella-based P2P System Using Cross-Layer Design for MANET, in Proceedings of World Academy of Science, Engineering and Technology, vol. Vol. 22, 2007.
22. F. Soldo, C. Casetti, C. Chiasserini, and P. Chaparro: Streaming Media Distribution in VANETs, in IEEE "GLOBECOM", 2008.
23. E. Gurses and A. Kim: Maximum Utility Peer Selection for P2P Streaming in Wireless Ad Hoc Networks, in IEEE "GLOBECOM", 2008.
24. B. Li, G. Keung, S. Xie, F. Liu, Y. Sun, and H. Yin: An Empirical Study of Flash Crowd Dynamics in a P2P-based Live Video Streaming System, in IEEE "GLOBECOM", 2008.
25. Q. Wang, K. Lin, K. Lin, D. Mao, and M. Yang: A Measurement Study of P2P Live Video Streaming on WLANs, in Proc. of IEEE GLOBECOM, 2008.

VoIP AS A TOOL FOR AN EFFECTIVE VOICE COMMUNICATION COST REDUCTION

Tito Carlos S. Vieira[1],
tito@fe.up.pt

[1] Faculty of Engennering of University of Porto

Abstract. The evolution of computer networks had made the integration of phone services in the computer network infrastructure possible and it has became theoretically possible to integrate data and voice communications into the same network, therefore there are many points which will be offered by computer network and which will make VoIP communications possible. Considering that FEUP computer network had the conditionings required to support VoIP we propose to implement a system in order to reduce the total FEUP voice communications costs. Our approach has developed an in-house system based on opensource software and programmed several new modules to permit the extension of features such as billing, statistics and web management. After a pilot-project period which occurred between January 2006 and August 2007, the system was put in production in September of 2007and the results have exceeded all expectations, and proved that FEUP's total cost of voice communications has been reduced by around 70% and the VoIP system has been fully integrated with the traditional telephone system. On the other hand many other functions have become available which has permitted an increase in the conditions of the phone system for FEUP users.

Keywords: VoIP, PBX, SIP, IAX

1 Introduction

In the last years research in voice over Internet (VoIP) was very strong and now there are several implementations available. The Computer Centre (CICA) of the Faculty of Engineering of the University of Porto (FEUP) is a central service with the mission of promoting, supporting and managing the infrastructure of communications technologies (ICT) of FEUP. Like many other technologies VoIP has became an interesting project because it seems there it presented a good opportunity to increase the traditional telephone system and, simultaneously, to reduce the total cost of voice communications. We initiated a project named VoIP@FEUP in the last quarter of 2005 and it evolved to this current state. The main goal was to implement a VoIP solution at FEUP which permits the reduction of the total cost of voice communications. Our first step was to search for commercial solutions that could work along with our current PBX. We approached several international companies that provide VoIP solutions, but either the solution was beyond our budget or it lacked the features we required. Our next step was to approach the opensource community, where we found some interesting software. Of course some of the features we required were missing, but seeing that their code was available, these features could

be developed by us in-house. The decision was thus to developed an in-house solution, based on opensource software, Asterisk[1], which addressed all of our requirements. In this case the risk of that decision was greater because we didn´t know anything about Asterisk and the decision that we had taken was based on the assumption that we would develop the system in that platform. Despite being aware of the capabilities and potential of our technical team, the project did not prove to be easy and we encountered many problems that will be explained in next sections of this paper.

Our expected results were to provide a good solution to give users the capacity to make internal and external calls based on VoIP technologies, in addition to integrating the new system with the traditional telephone infrastructure and to ensure the quality of all calls, as well as to interconnect many national and international VoIP brokers with our systems and, based a knowledge algorithms, to opt for the best way of routing calls for each broker, and finally to create a web page with all the information required for users and managers of the system. Though these interest technological functionalities were considered to be very important, we knew that the main goal was reduce the total cost of voice communications.

This paper is organized as follows: Section 2 presents the technical matter of VoIP systems. Section 3 describes the process of our VoIP implementation and describes its application. The next section presents some results. Finally, we provide some conclusions as well as an outlook of future research which we intend to do.

2 VoIP technologies

Increasing opportunities for converged telephony-Web services motivated the convergence of telephone and data networks. Because IP networks can be relatively inexpensive, network providers are encouraged to build common IP core networks surrounded by various accesses to networks. These accesses to networks (wireless, wireline, cable, etc.) can share the IP core resources, and thus reduce the costs of providing common services to customers with different access devices[2]. VoIP (voice over IP, voice delivered using the Internet Protocol) is a term used in IP telephony to denote a set of facilities for managing the delivery of service information using the internet protocol. VoIP totally depends on TCP/IP/UDP to control call setup and voice transmission. Signaling protocol is used to set up and end calls, carry information required to locate users and negotiate capabilities. The signaling protocol used for this VoIP implementation is Session Initiation Protocol (SIP) [3-5]. The Inter-Asterisk EXchange version 2 (IAX2) protocol provides control and transmission of streaming media over IP networks. IAX2 which facilitates VoIP connections between servers, and between servers and clients that also use the IAX2 protocol [6]. For real time data transmission such as voice and video, Real Time Protocol (RTP) is used to transfer data between end users. RTP provides end-to-end network transport functions suitable for applications transmitting real-time data, such as audio, video or simulation data, over multicast or unicast network services [3]. The basic steps involved in creating an Internet telephone call is the conversion of the signal into IP packets for transmission over the Internet; the process is reversed at the receiving end [7]. VoIP is becoming an important means of supporting telephony to the desktop as it

provides more reduced costs and permits a lot of business functionality. VoIP calls can be free or cost only a small fraction of typical telecommunications service provider charges. VoIP easily provides a rich set of features such a lower routing cost, call forwarding, voicemail, caller ID, call parking, music on hold, PBX to PBX dialing, voice and video conferencing and much more. As most of the functions are software controlled, it is possible to manage calls depending on a user's location, activity, or time of day[3].

2.1 Advantage and disadvantages of VoIP systems

A major advantage of VoIP and Internet telephony is that it avoids the toll charges of ordinary telephone services and VoIP systems can be interconnected via the Internet which is cheaper than the traditional system. On the other hand, two separated infrastructures for data and voice communications are not required, one network will hold both. But here we encounter one disadvantage. If there is no electricity or the network is down, the phones will also be down. One can resort to UPS (Uninterrupted Power Supply) and PoE (Power over Ethernet), to minimize these kinds of problems. But even the former PBX systems required power so this is not a unique VoIP problem.

Another difficulty which was encountered was that emergency number always had to work, but an International broker would not route them. Therefore a connection to the local phone provider always has to exist. Most of the ISP (Internet service providers) provide a phone lines along with the Internet for no extra charge, so by just buying an analogue or digital interface this problem can be easily solved.

The last disadvantage is related to sound quality over the Internet because this cannot be controlled. This situation can be minimized by using QoS (Quality of Service) and negotiated Internet connections. In more desperate situations, we can easily work with VoIP over normal ISDN and analogue lines. But in doing so, we will loose some of the advantages that VoIP brings us. Advantages such as ENUM (Electronic Numbering)[8] or the "anonymous SIP" connection depend entirely on the Internet. The possibility of roaming our softphone remotely also depends on it. Nowdays, due to the appearance of able DSL providers the Internet is available at a very affordable price. The quality of the call also depends on the codec used during that conversation. While some high quality codec's can use up to 100kbps, there are others that use as little as 6kbps and have a reasonable audio quality. So technically speaking that would mean that we could make about 20 phone calls over one old 128kbps ISDN line instead of only two phone calls, which was be allowed by traditional BRI lines.

2.2 Asterisk software

Asterisk software transforms an inexpensive PC architecture server running Linux or UNIX into a reliable, sophisticated, full-featured enterprise telephone system. Because Asterisk is free and runs on an industry standard PC platform, an Asterisk system will cost far less than any traditional, proprietary PBX. With Asterisk, one can

quickly and easily create a sophisticated business telephone system for any enterprise, no matter how large or small. Because it is reliable, free and effective, and due to the fact that it is based on modern Internet protocols, Asterisk will replace many legacy telephone systems in the marketplace. Asterisk is far less expensive and much more effective than any competing telephone system. Asterisk provides all the functionality of a traditional PBX, but it also provides new features and capabilities a legacy PBX can't offer. Because Asterisk is open it is possible to change it and tune whenever needed, unlike legacy systems which only provide closed black boxes with closed interfaces. With Asterisk, this equipment will never be considered to be obsolete from an single-source vendor [9].

3 The VoIP implementation in FEUP

In this section we will show the most important aspects of VoIP implementation in FEUP, beginning with the installation of the actual infrastructure and ending with some of most popular features which were developed by us.

3.1 Installation of the infrastructure

Installing and configuring the Asterisk server was quite easy, but optimizing and discovering small details to make it work properly proved to be somewhat of a challenge. After a few weeks we were finally ready for a trial period. We started with about 40 extensions, some of which used softphones and others which used hardphones of different brands. As telecom brokers we used several of the available VoIP SIP providers such as Betamax. The experience was very positive, not only in terms of the price but also as far as base features were concerned. This was an experimentation process seeing so to verify what did or not work. Testing hardphones, providers, protocols and legacy integration was our main focus. But it wasn't until when we interconnected our PBX with our VoIP Server, that other departments approached us to join the project. Most of our initial users started off using softphones and later changed to hardphones. The ability to register three different phones to the same account and ring all of them at the same time allowed them to play around with different hard and softphones, and at the same time, it provided us with a lot of feedback about features and problems. We soon found out, that some configuration worked well with some phones and worked badly or not at all with others. We slowly discovered special configurations that allowed us to increase the success rate of tested phones. Details like which codes are better suited for each phone and which configuration cannot be used with certain phones, were included in the core of the server.

Our network diagram is showed in Fig. 1. Here, you can see two servers, “cnet-pbx” and “pri-pbx”. In our case this was necessary because of the distance between our datacenter and our PBX, therefore we added pri-pbx to function as a gateway between both systems. Our former means of communicating between both systems was an “ATA SPA 3000”, which is still working and configured as an alternative route for certain users. We use a Vierling 16 slot GSM gateway connected to our PBX

for mobile phone calls. International calls are all placed using our international brokers. Incoming calls were still routed through the old PBX and through 2 BRI ports that were directly connected to the server.

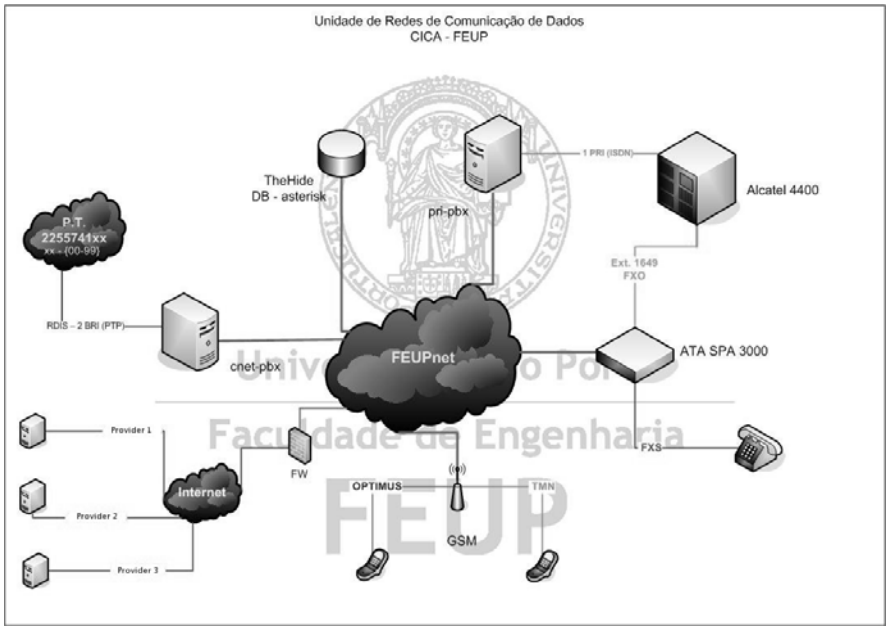


Fig. 1. VoIP infrastructure

3.2 The PolySpeak system

We had created a software package with our VoIP System and we named it PolySpeak. It required a dedicated server to run with minimum requirements which was an Intel Pentium 4 processor and 1 GB of memory. All configurations could be done via a Web Interface, such as simply creating a user as well as the more complex operation of creating trunks. But the web interface is not only used as a system administrator, it has three different access levels, which we called user, superuser and admin. In Fig. 2 in the upper right corner, we can see three icons, which allow us to shift from user option to admin options. These three icons are only visible to the admin user, two of them to the superuser and only one to end-user. The superuser level gives access to all server statistics in the accounting area and some partial information of the server. It allows the user to detect problems and report them to the administrators, but he cannot act on them. Normally superusers are department directors who need access to detailed call logs for accounting purposes. The VoIP server did not normally let the end-user access the server. Nevertheless, we found that providing users with access to some features like click and dial or call logs, would increase the usage of the system. Looking up lost phone calls or a number called three months ago can be performed quite easily.

3.2.1 The User Web Interface

This part of the interface is dedicated to the common user. In Fig. 2, we can see a log of completed, received and lost calls. The icon in front of the numbers allows to immediately launch a call, that can be picked up by any of the users phone. The call box allows one to call any number, login or remote SIP address. The user can view the minimum and maximum rates for the current available destinations. The phonebook is divided into two parts, a public list and a private list. The public list is the entire local user list dynamically created with information of presence. The private list permits the user to add his own entries to the phonebook. Both lists have a click and call feature. The “change PIN” tab permits the user to change his or her PIN. In the case of a supported hard phone, the server will try to synchronize the PIN automatically. The next tab permits the user to define personal options such as language, forward, blacklist, voicemail and phonelock. The “IP phone” tab permits one to set certain options on users hard phone such as fast keys. The last tab permits the user to view the monthly charged phone calls in real-time.

Changing the language in the web interface implies that from then on all interaction with the server will be performed in the chosen language. At the moment we have only two officially supported languages. One is our native language: Portuguese and English was chosen as the second language. However, the structure of the server permits us to add as many languages as we like without having to recompile any code. The same applies to the look. The server supports different themes that can be added to the server.

Strangely, it is not possible to predict how much the phone call will cost. Because the server uses an LCR algorithm, it will try to pass the call on to the cheapest route, but whenever the connection can not be established, it will try to use the next alternative route. Therefore we can see “Minimum Rates” and “Maximum rates” in the “Home” tab. The advantage of this is that the call will always be established, independently of the use of the Internet or ISDN or of any other configured method. If the user uses a supported hard phone it is possible for him to view these rates in real-time as the server tries to connect the call and also to hang-up before the call is established if he does not agree with the price.

Some of the features were added due to user suggestions. One of these suggestions was the ability to define a whitelist and a blacklist, which allows users to control who is allowed to phone them. A small add-on to this feature is the ability to redirect the call to other users instead of providing the caller simply with a busy signal. This is very useful in the case of queues in a Helpdesk and a caller does not want to respect his order of arrival.

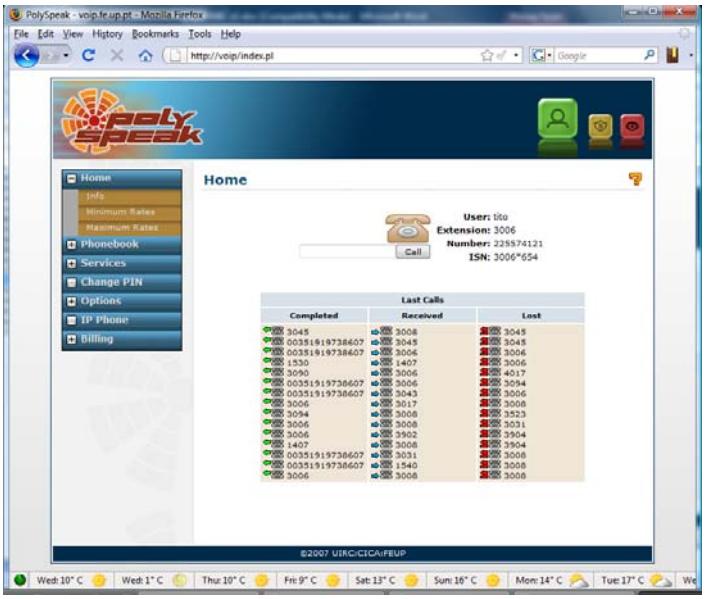


Fig. 2. PolySpeak user web interface

3.2.2 The superuser web interface

Superuser interface adds information about system utilization. It provides a very complete reporting tool and Fig. 3 show one sample of that feature. A superuser is considered to be a user who can view information such as global statistics, ongoing calls and server information. In Fig. 3 we can see the group statistics. In our case groups are organized by departments, by clicking on the department we can see the users who belong to that group as well as details of their calls. Due to the fact that it is not always the department that pays the users bill, we have created another tag which we called “Cost center”, and whose function is similar to the group tag. Another useful statistic is the “Destination” tab. It permits to see where all the calls are destined and in this way, allows to optimize the service providers. The “reporting” tab will allows us to see graphics in different time ranges which contain information about CPU, memory, disk and connections. In the “Calls” tab one can discover ongoing calls and even terminate calls with a simple click. The rest of the existing tabs permit the superuser to view server configuration and broker information such as remote credit. The superuser can view almost everything, but cannot change any kind of configuration. Configurations like enabling accounting have to be performed on the “Admin web interface”.

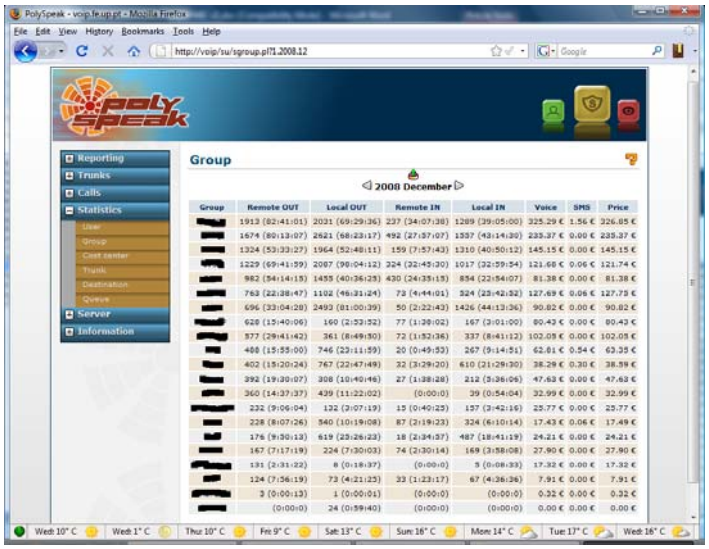


Fig. 3. Calls and costs by department

3.2.3 The Admin Web Interface

Management PolySpeak becomes very easy due to this feature. It permits one to make some complex configurations in an easy way, so the system can be implemented by people who know only a minimum information about VoIP systems and so do not need to be an expert in network infrastructure. The admin user can perform all configuration and maintenance here. Admin Web Interface has a lot of options like: creating backups, updating the server with the latest revision is simple and straightforward. The “System” tab permits one to discover and configure hardware boards such as ISDN BRI or PRI. The “Broker” tab allows one to define dynamic brokers that are analyzed each 30 minutes for the best routes and availability. Creating users is very simple, but when creating mass users one should opt for the SOAP service, which permits the use of scripts. Trunks permit us to interconnect local PBX or VoIP servers. There is also the notion of powertrunks, which permits two server to synchronizes theirs users as if it they where one. Call routing can be done in two ways, routes can be inserted system wide or on a per user basis. The IP Phone database allow us to keep an up to date record of all hard phones and also to enable each phone’s special features used in auto configuration, automatic PIN changes and address book synchronization. Queues, IVR (Interactive Voice Response) can be performed in the “Service” tab. For fine-tuning the server, the “Registry” tab possesses some special options.

With the interface it is possible to automatically configure a lot of features and thus simplifying the configuration process. The most interesting one of all is the broker account. In these cases the server can obtain the current credit and current rates

dynamically. All the information is used by the LCR algorithm in calculating the best route the next phone call will make.

3.3 Migrating to PolySpeak

Migrating to VoIP seemed to be the obvious way to go, but how would it be possible for one to achieve this without causing any major problems and having a reduced budget. We had over 1100 phones on campus and more were needed. The first step had already been taken, which was the interconnection between both systems. The next logical step was that new phones would only be VoIP phones. At the beginning not much interest was demonstrated by the departments due to the cost of the phones. Nevertheless, one of the departments decided to invest in the new system phones and others departments invested when the first results came out. Being a university, teachers and student were eager to try out the system and people approached use everyday asking for information.

3.4 Problems and solutions

There were several aspects that had to be considered before setting up a VoIP system, some of which we learned about the hard way. One of the most important aspects was the Network Infrastructure and QoS. As VoIP used the network instead of the old telephone, care had to be taken so as not to affect telephones due to an incorrect bad usage of the network. Heavy broadcast or uncontrolled multicast can have disastrous consequences for hardphones. An easy solution would be to create a parallel isolated network, but that would increase the cost of the project dramatically. VLAN tagging was a simple solution that made it possible to use one single point of access for the phone and computer. It was important to understand that the phones interface would be the maximum speed available for the computer. So using a phone with a 10Mbps network interface was completely out of the question. We also discovered that phones using this kind of interface were much more sensitive to network problems. There were no phones with 1gbps interfaces yet that we knew of, so this solution implied that speed was limited to 100Mbps. Because phones were normally connected to desktop computers faster speeds would not be required.

Interconnecting both PBX systems was more of a political problem than a technical one. We used one of the PBX's ISDN PRI ports to connect our VoIP server by using a Digium E1 card. Because we used 1XXX and 2XXX extensions in the PBX, we decided to use 3XXX for VoIP. After adding new dialplans to both systems communication seemed to work like a charm, but later on we discovered that the ISDN protocol was not enough for the PBX to consider the VoIP local system. It treated the VoIP system as a remote system, making it impossible to use internal services. After some investigation we discovered that using the ABC or the QSIG protocol would solve that problem.

The SIP protocol worked very well in the local network, but it had some flaws when it came to crossing firewalls. Special attention had to be taken, not only in the case of server configuration, but also in terms of the Firewall. Both configurations had

to be exact in order for audio and video to be available in both directions in a phone conversation. It took some time to fine-tune both configurations to incoming and outgoing SIP calls. These problems did not exist in the IAX2 protocol that worked immediately. The problem with IAX2 was that basically, this protocol is only supported by Asterisk and for the time being, there is no video support for it.

It was very important to make things easy for the end user. Asking them to configure their own hardphone was out of the question, even if it was an easy task to use the phone web interface. Therefore we created the necessary technology to do things automatically. However, there was the problem of PIN changes. When the user changed his PIN in the PolySpeak web interface, the system searched its internal database to discover which type of phone the user was using. If it was one of the supported, the system automatically changed the phones account settings and rebooted the phone.

Another interesting point was quality, by that we mean, which calls we considered needed absolute quality and which calls weren't considered to be a quality issue. International calls even on POTS (Plain Old Telephone system) did not have a very good audio quality. In our experience using Internet brokers improved audio quality. Therefore all our international calls were routed directly through SIP or IAX2 providers around the world. Mobile phone calls had their ups and downs. In our particular case, we used GSM gateways as we had done on our old system. In the future it may be possible for us to negotiate better contracts, when VoIP trunks become more frequent. On the other hand, local country calls were free on some international brokers, but if we wished to have better audio quality we had to pay for them. The decision was not easy and we are still debating this issue. Some people believed that the price was more important, others preferred the quality of the service, never the less both systems were configured so there was a backup in case one of one of them being down.

4 Results and discussion

Our VoIP system, PolySpeak, was developed and implemented in FEUP. PolySpeak was based on opensource software, Asterisk, and we added a lot of innovative functions. Some of them have been presented in Fig. 2 and Fig. 3. The system was placed online in the following site http://voip.fe.up.pt. The users of the system started increasing significantly in 2006 and have continued to increase as Fig. 4 shows.

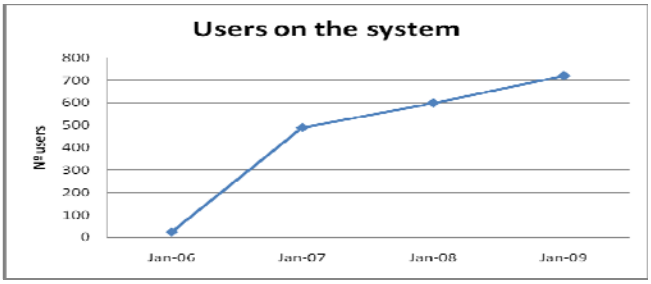


Fig. 4. Users on system

Another important point to analyse is the utilization of the system in terms of hours of conversation. Fig. 5 shows the evolution of utilization since January 2007 and it is clear that there was a significant increase throughout 2007 and 2008 and, on the other hand, in terms of peaks, 1311 hours of conversation occurred in October 2007 and 2095 in October 2008. This was very important because it demonstrates that the infra-structure was appropriately dimensioned and it supported a growth of around 50%.

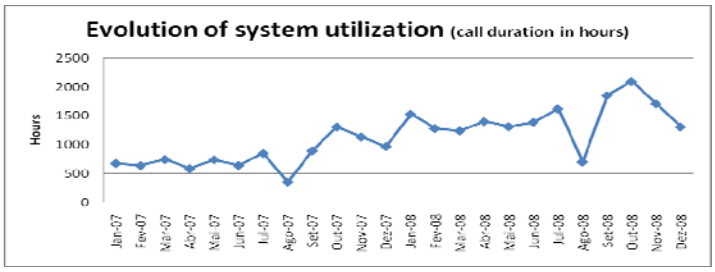


Fig. 5. Evolution of system utilization

Table 1 summarizes the results obtained in VoIP utilization at FEUP since January 2007 and it provides information, organised according to years, the number of calls made and brokers utilization.

Table 1. Utilization of VoIP in FEUP since 2007

Year	Month	Call details			Broker details			
		Cost (*in €*)	Call duration (*in hours)*	Number of calls	ISDN	Betamax	Voip anywhere	Inov8
2007	Jan	258,93 €	678	14919	34	4975		
	Feb	228,67 €	640	14086	45	4383		
	Mar	285,48 €	745	17134	22	5501		
	Apr	236,62 €	588	13615	19	4872		
	May	263,74 €	742	17508	20	5700		
	June	222,97 €	642	16236	6	4924		
	July	259,97 €	849	18101	8	5053	4	
	Aug	109,31 €	351	7284	11	2120	9	
	Sep	317,67 €	896	20043	284	4038	368	
	Oct	1.050,74 €	1311	26095	1569	4118	1971	
	Nov	1.298,28 €	1137	28938	1983	3763	2731	
	Dec	1.244,53 €	971	25096	1907	3027	2585	
2008	Jan	1.572,72 €	1533	35106	2815	3709	3613	
	Feb	1.557,44 €	1283	31429	2323	3945	3581	
	Mar	1.473,67 €	1243	34491	3184	3626	3901	
	Apr	1.494,02 €	1404	37247	4566	3329	758	2346
	May	1.522,09 €	1316	35121	3979	3251		2927
	June	1.544,57 €	1390	35270	5562	1356		3170
	July	1.957,49 €	1623	42416	6749	961		4534
	Aug	1.028,79 €	702	15697	3189	5		2392
	Sep	2.084,94 €	1854	47605	5779		7417	223
	Oct	2.190,81 €	2095	52412	8659		6151	
	Nov	1.898,26 €	1713	43547	5478		7451	
	Dec	1.627,68 €	1311	35834	2544		9244	

Another very interesting point is the analysis of the broker utilisation to make voice calls. Table 1 demonstrated this and in fact here we have the reason for the significant reduction of costs because calls were made with the best route for each call, and the best route was not the same in each case. We used various national and international brokers to make VoIP voice calls and Table 1 also summarizes the number of calls which each of them did in 2007 and 2008.

A most important result is presented in Fig. 6. It shows the evolution of the FEUP phone call cost since we placed the VoIP system in production, which occurred in September 2007. As we can see from the beginning of 2007 to September of that year VoIP was not significant, but in September VoIP calls started growing, traditional voice calls started dropping significantly and the total cost of voice communications was greatly reduced. The total FEUP cost of voice communication was reduced by 70% in comparison to the cost in January 2007 and the homologue in January 2008.

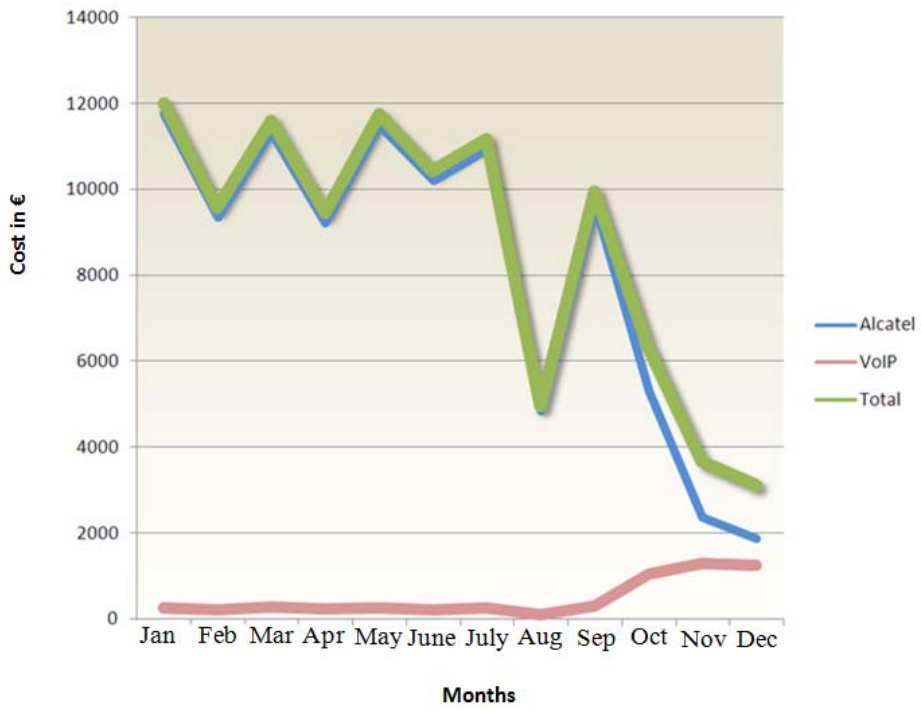


Fig. 6. Evolution of voice calls

5 Conclusions and future work

In this paper we have presented an approach for changing the traditional telephone system to a new one based on IP. We implemented a VoIP infrastructure at the Faculty based on opensource software and we have developed the features which we considered as being important so as to address all of our requirements.

One of our main contributions in this paper is the VoIP strategy to reduce the cost of voice calls. As has been shown in section 3, we reduced FEUP's total cost of communications by about 70% in one year. On the other hand, the strategy which was implemented to encourage users to adopt the VoIP infrastructure demonstrated that it was possible to motivate people to abandon a system which worked well, and which had been used for a number of years and to start using an alternative new system in which factors such as stability and quality were unknown at the outset, without them having to make any difficult decisions.

Another important contribution of this paper is related to sharing our experience of the creation of the VoIP solution and its implementation it in a large organization. In fact when implementing a telephone solution from scratch, VoIP is the solution that provides the best relation between price and features. We chose to encourage a small change to VoIP and the results were very positive and encouraging. Typical VoIP users were more interested in features like mobility, the integration of voice calls with e-mails and its easy access and use. Management was very interested in price reduction and accountability. These were the main reasons why VoIP became so popular. In my opinion VoIP changes our notion of the phone. Previously the phone was seen as being an object on our desks, which could be shared with several other users. Now we see it as a personal means of contact. It began to accompany us wherever we go. It has pleased us to see that our users have started using the system more and more as each day that goes by. We can see this especially in the case of international phone calls. The price rate of an International call is almost the same as a local call, sometimes even cheaper. Previously, in specific types of calls, we paid almost 40 times more than the value per minute which we now pay.

The VoIP solution we made became popular and we received a number of contacts from external Institutions asking us to implement this system there. When these situations occurred and while we were analysing the interest in collaboration with other organization in that area, we decided to register the brand PolySpeak. Some time later, we decided to share our technology and thus some projects were developed: PolySpeak was introduced in a number of Institutions like: *Casa da Música, SA*; CCDRN - *Comissão de Coordenação e Desenvolvimento Regional do Norte*; AMAT - *Associação de Municipios do Alto Tâmega*, and in each of the six city councils; FCNAUP - *Faculdade de Ciências da Nutrição e Alimentação da Universidade do Porto*. We also have some other pilot-projects underway in important Institutions like banks, a national drinks company and a national technology company.

The results achieved prove the utility of the VoIP systems and we consider that this is more than just a passing trend. As far as we are concerned, this represents the future in the voice communication infrastructure.

Future work will be concerned with the integration and development of new PolySpeak features. Video is one of the most interesting features, which we are now

working on and thus we hope to be able to make video-calls in near future. Another interesting feature that we are planning to do is integrating Fax and SMS messages in the PolySpeak systems. Our final planned feature is the integration of VoIP with instant messaging as well as our videoconference system.

References

[1] Digium. "Asterisk open source PBX," 8-1-2009, 2009; http://www.asterisk.org/.

[2] S. R. Ahuja, and R. Ensor, *VoIP: What is it Good for?*: ACM New York, NY, USA 2004.

[3] M. Ahmed, and A. M. Mansor, "CPU dimensioning on performance of Asterisk VoIP PBX," *Proceedings of the 11th communications and networking simulation symposium*, pp. 139-146, 2008.

[4] J. Rosenberg, H. Schulzrinne, G. Camarillo *et al.*, "SIP: Session Initiation Protocol," *RFC Editor United States* 2002.

[5] H. Schulzrinne, and J. Rosenberg, "The Session Initiation Protocol: Internet-Centric Signaling.," *Communications Magazine, IEEE,* vol. 38, no. 10, pp. 134-141, 2000.

[6] M. Spencer, B. Capouch, E. E. Guy *et al.*, "IAX2: Inter-Asterisk eXchange Version 2," 2007.

[7] M. P. Jaiswal, and B. Raghav, "Cost-quality based consumer perception analysis of voice over Internet protocol (VoIP) in India " *Internet Research: Electronic Networking Applications and Policy,* vol. 14, no. Number 1, pp. 95-102(8), 2004.

[8] P. Faltstrom, and M. Mealling, "The E.164 to Uniform Resource Identifiers (URI) Dynamic Delegation Discovery System (DDDS) Application (ENUM)," *RFC Editor United States* 2004.

[9] P. Mahler, *VoIP Telephony with Asterisk*: Signate, 2005.

A Pedagogical Scenario based on the ILEM Model: A Case Study

Dulce Mota[1,2], Carlos Vaz de Carvalho[1]
[1]ISEP – Institute of Engineering, Polytechnic of Porto, {mdm,cvc}@isep.ipp.pt
[2]FEUP – Faculty of Enginnering, University of Porto, pro08016@fe.up.pt
Porto, Portugal

Abstract. In this paper, we present and evaluate a new learning-teaching model, ILEM (*Intelligent Learning Environment Model*) supported by an adaptive hypermedia application. This model operates in a blended scenario, combining *online* and face-to-face approaches. In the former, students interact with an educational adaptive hypermedia application, whereas in the latter students essentially carry out group work. The overall model has the main purpose of maximizing the balance of the analytical, practical, and creative intelligence components in this blended context. The model's theoretical basis is supported on Robert Sternberg's theory: the Triarchic Theory of (Successful) Human Intelligence. The adaptive hypermedia application presents the subject matters and also recommends personalised learning activities classified as analytical, practical and creative. The results obtained in this experiment reveal a maximized balance among the individual capabilities, which is in accordance with the theoretical claims.

Key words: Adaptive Hypermedia Systems, AHA!, Human Intelligence theories, blended learning settings.

1 Introduction

In the present society, the educational system has to deal with a wide range of challenges. These challenges include time and space constraints along with different types of skills demands. The creative skills, for instance, are more and more needed in the professional careers.

In order to cope with these new challenges, the educational system has gradually been developing measures to include novel educational processes in classrooms. However, concerning learning-teaching methodologies, teachers in general still give more attention to some classical skills, namely analytical ones. These skills, according to some contemporary experts in the Intelligence field, are rather insufficient when compared with the new professional and life standards needs. The psychologist Robert Sternberg, author of the Triarchic Theory of (Successful) Intelligence (TTI) [1][2], claims that there are three branches of Intelligence, analytical (conventional one), practical and creative intelligence, and that all of them are required in different

moments or tasks of ones' lives. Furthermore, the author argues that those intelligences should be trained in the classrooms from the very first school years [3].

The fundamentals of our learning-teaching model, ILEM [4] are sustained on the main principles of the above theory. Briefly, ILEM helps teachers to prepare courses centered in triarchic learning activities in a blended context. In turn, students should be engaged doing the proposed activities during a sequence of lessons. Each lesson is divided in both online and face-to-face periods. The former implies that students interact with an adaptive hypermedia application which gives them personalized suggestions about the kind of activity they should do in a particular moment, whereas the latter leads students to do work team. The overall purpose is to train all three individual skills in order to improve students' competence levels.

For developing the adaptive hypermedia application the AHA! (Adaptive Hypermedia Architecture) platform [5] was used. It is a general purpose authoring tool to develop Web-based adaptive applications in very different domains. This framework provides both adaptive content and adaptive navigation, which are very interesting characteristics to educational area. The adaptive application algorithm we developed has the following general behaviour: In each online subject matter, it recommends the triarchic learning activity for which the student's skill has the lowest achievements. In short, the algorithm aims to reach a maximized balance among the three student's skills.

A case study was carried out to evaluate the adaptive application integrated in an undergraduate course in Institute of Engineering, Polytechnic of Porto (ISEP), an engineering school, with 18 students. The results gave us confidence to continue working on this research. Moreover, the results let us confidant to deepen new learning-teaching methods and the creativity field as well.

This article is organized as follows. Section 2 makes a brief introduction to the theoretical basis for the learning-teaching model. Section 3 describes briefly the ILEM model. The main features of the AHA! framework and a description of the possible student-application interactions are presented in Section 4. Section 5 describes the case study carried out in this investigation and presents the experimental results. Finally, Section 6 draws the conclusions and points out some future work.

2 Theoretical basis for Developing the ILEM Model

This brief theoretical introduction aims at pointing out some of the influential perspectives on intelligence issues. We finalise this introduction with the theory that lay the foundations of our research.

The early idea of intelligence as a single ability, which could be translated into a simple number, has been refuted by researchers on modern theories of Intelligence. The complexity and variety of these theories may explain the importance of this area both in educational settings and in people's professional lives.

The theory on Multiple Intelligences [6] by Howard Gardner challenges the traditional assumption that there is only one type of intelligence. According to H. Gardner, nine or ten kinds of intelligence need to be considered. The development of

these intelligences is strictly connected with the social-cultural context where the interactions take place.

Reuven Feuerstein presented the theory of Structural Cognitive Modifiability (SCM) [7]. He argues that intelligence is plastic and changeable, which can be taught, therefore, it is not fixed at birth.

Daniel Goleman [8] refuses the idea that Intelligence Quotient (IQ) is a synonym of being more intelligent or less intelligent. Goleman argues that our emotions play a much greater role in thought, decision making and individual success than it is commonly acknowledged.

More recently, Ceci [9] put forward a bio-ecological model of intellectual development. The roles of various social contexts (schooling, cultural values and social organization) are underlined in that development.

R. Sternberg is the author of the TTI, among others. This is a theory of individuals and of their relations to their internal and external worlds, and to their experiences as mediators of the individuals' internal and external worlds. According to the author, intelligence can be placed into three (triarchic) subcategories that feed into successful intelligence. They are: analytical, practical and creative intelligences. Analytical intelligence is applied to the abilities of analyzing, evaluating, judging, comparing and contrasting. Creative intelligence is related to the ability to design one's own ideas to those of others, or to entirely devise a new concept. Practical intelligence relies on the intellectual capacity of the individual to make informed decisions. R. Sternberg also describes Successful Intelligence as the ability to balance analytical, practical, and creative intelligence and to use these intelligences effectively; in other words, the keys to Successful Intelligence are: "1) the ability to achieve one's goals in life, given one's sociocultural context; 2) by capitalizing on strengths and correcting or compensating for weaknesses; 3) in order to adapt to, shape, and select environments; and 4) through a combination of analytical, creative, and practical abilities." We emphasise the need to capitalise one's strengths and compensate one's weaknesses, circumstances inherent of human beings.

R. Sternberg claims that is necessary an educational system based on his triarchic model in order to achieve success in learning settings and in life as well [10]. In his studies [11], the author concluded that both the triarchic teaching and evaluation contribute for improving students' learning when compared with the traditional teaching model.

Briefly, the triarchic conception of (successful) intelligence emphasises the use of three intelligence components – analytical, practical and creative. Although these are interdependent, they could be applied independently. The question is to known how and when to use them.

Lastly, we believe that the TTI may be a very realistic opportunity to rethink the overall practices of the learning-teaching methods. Our research has been conducted with the TTI theoretical basis together with user-adaptative models supported by computer. Our main objective was to improve learning outcomes for *all* students, getting all of them highly motivated in a learning environment.

The ILEM model has been conceived following that motivation. In the next section, we introduce its main features in order to get a more understandable picture of our learning-teaching scenario.

3 ILEM Model Description

The learning-teaching model we have conceived, ILEM (*Intelligent Learning Environment Model*), follows this main principle: "Balancing analytical, practical, and creative learning activities" both during students' interaction with the adaptive hypermedia application and during the group work. By "learning activity" we mean "any online or presential analytical, practical or creative task", for example, readings, evaluation tests, exercises. In our case, both the reading tasks and exercises fall into the online activities whereas the group work task are connected to the presential activities. In the case of the evaluation tests they are linked to both.

The adaptive application comprises the online component of the model, whereas the group work belongs to the presential one. The two complementary learning scenarios in the classroom aim to improve the student's learning achievement, point out a more individual student's pace with the benefits of a group's rhythm. Besides, the work in groups boosts relationships which foster benefits beyond the particular group assignment work.

Moreover, an assessment component was also added to the model. The online component is sustained in the previous component in order to perform the adaptation of the learning activities to each student.

Figure 1 illustrates the schema of the learning-teaching model. On the base (the conceptual level), we have three important components: the triarchic subject material developed by the teacher/author; the set of attributes needed to the algorithm for student's adaptation; and the adaptive mechanisms. On the top (the implementation level), there are also three components: the online, the presential and the assessment components. As mentioned before, the online application can automatically assess each student. Besides, the teacher needs also to plan the student' triarchic assessment based on the presential learning activities.

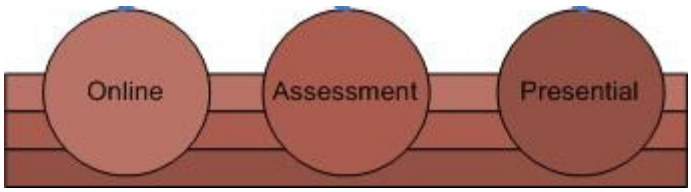


Fig. 1. ILEM model architecture.

4 The Adaptive Hypermedia Application Developed

In the next section, we describe the main characteristics of the AHA! framework and the student-application interactions related to the adaptive hypermedia application developed.

4.1 A brief Overview of AHA!

To develop our adaptive hypermedia application, the AHA! authoring framework was chosen as its functionalities fits to our goals.

AHA! system is an Open Source Web-based platform. It integrates several technologies, namely Java servlets, (x)html, xml and XLST. Its architecture embraces the following main components: the AHA! engine, the WWW server and several Authoring tools. The domain/adaption model and user model are stored as xml files on the server, or in a mySQL database.

AHA! provides both adaptive presentation and adaptive navigation [12]. The main technique to accomplish adaptive presentation is conditional inclusion of fragments/objects. In respect to adaptive navigation, adaptive annotation and hiding of links are the techniques available. The adaptation is accomplished on the basis of information from the user model. AHA! can adapt local as well as remote pages to the user.

As mentioned above, this framework provides several authoring tools. We emphasize the Graph Author tool and the Test Editor tool [13]. The former is a high level tool that allows creating the domain knowledge and the concept relationships, that is, the domain model and the adaptation model (see figure 2). The latter is used to produce multiple-choice tests.

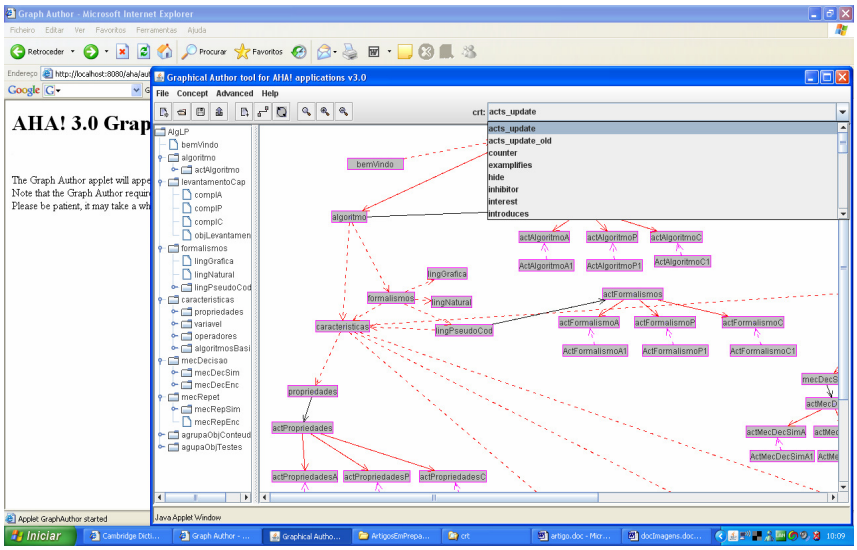


Fig. 2. A snapshot from the Author tool screen.

One import issue about AHA! system concerns the creation of adaptation rules. The author can profit from the built-in rules (the concept relationships types -CRT). However, the author can make his/her own rules depending on the type of adaptation

needed. CRT are defined using templates, just like concepts, and they are produced through xml files.

Finally, creating an AHA! adaptive application consists of defining the domain model including the adaptation model, that is, the domain knowledge and the concept relationships or adaptation rules, and the user model. Moreover, the content of the (x)html pages needs also some attention, specially whether fragments and objects are to be included in the content pages.

4.2 The Student-Application Interactions

The adaptive hypermedia application was applied to for the Algorithms and VisualBasic Programming discipline which is integrated in the Chemistry Course of ISEP school.

At first, each student needs to register in the application by filling out the presentation page (login, password, email are some of the required data). After that, the welcome page is visualized (see figure 3). This page presents the subject matter and the main pedagogical goals. Moreover, there is a link to an online questionnaire. Their result aims to initialize the three student's intelligence components (analytical, practical and creative one). For this purpose, the questionnaire was conceived based on information assembled in [3] author's book.

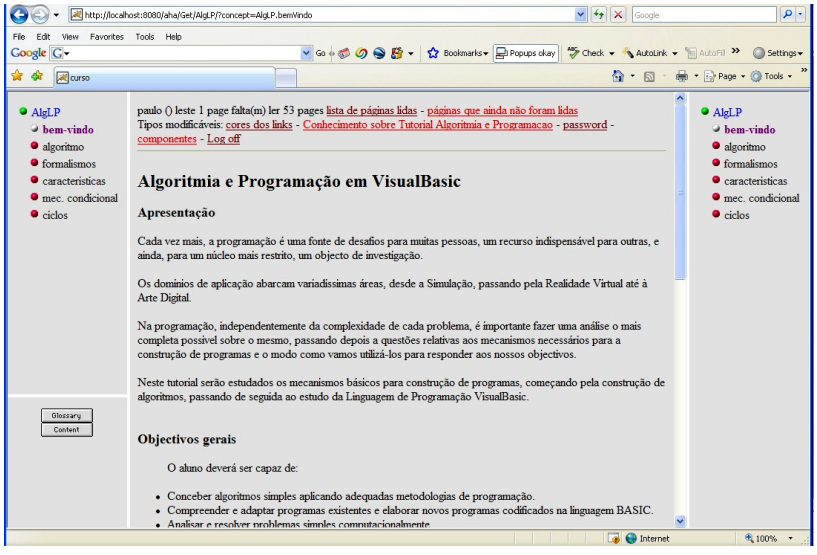


Fig. 3. The hypermedia adaptive application interface.

After these first steps, the student can do several actions, namely:

- selecting content pages;
- solving learning activities;
- changing student's knowledge level related to each concept s/he has visited (the application also updates automatically student's knowledge level based on the interactions between s/he and the application);
- changing student's domain independent attributes (password, link colors);

The icon colors in the content index indicate if a page/concept is suitable to the student, that is, weather the student knowledge is above, or not, a predefined threshold. White color means that a page was already visited; Green color means that a page is suitable to the student; red color means not suitable. A similar process happens with the link colors.

Finally, we show a learning activity layout (see figure 4) recommended to the student by the application. The design of these pages allows both text and images as visualized in figure 4. As mentioned in Section 1, the adaptive application suggests the most suitable learning activity to the student when the activity page, integrated in each subject matter topic, is selected. In this case, the student needs to solve a multiple-choice test (AHA! present only Multiple-Choice Tests). The results of this test are used to update the student model by the algorithm of adaptation.

In the next section, we describe the case study carried out in our research investigation.

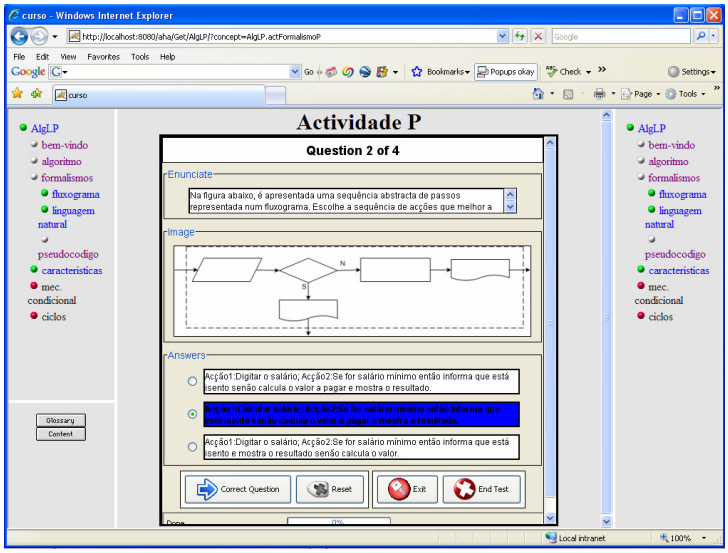


Fig. 4. An online learning activity interface.

5 The Case Study

In this case study, we intend to assess our adaptive hypermedia application which is integrated in the learning teaching model, ILEM. To accomplish that task, we have planned the Algorithms and VisualBasic Programming course in the scope of the Algorithms and Programming (APROG - it aims to teach the elementary concepts in construction of algorithms and their implementation using VisualBasic) discipline. This discipline belongs to the 1st year of the Chemistry Engineering course in ISEP higher school. The research methodology chosen was the qualitative approach. The case study features are presented in the following section.

5.1 The Research Method

The participants consisted of 18 students (10 females and 8 males): 1st year grade in Chemistry Engineering Course and it took place in ISEP higher school in January 2008. All these students have already attended the APROG lectures but they fail in some tests. The majority was not very interested in programming subject and so, their motivation was very low.

It was planned to make two sessions of 2 hours and 30 minutes each. Each session was divided in two parts: in the first half, the teacher invited the students to interact individually with the online application to study a specific topic. As mentioned before, students filled out the intelligence components questionnaire in first place. After that, they began to navigate through the content pages and solve the learning activities recommended by the adaptive application. In the second half, the students worked in group doing the learning activity proposed by the teacher. They finished the session discussing the results obtained by all groups.

The hypothesis evaluated is described in the following way:

"*Adaptive hypermedia applications based in the Triarchic Theory of (Successful) Intelligence can contribute to maximize the balancing of the student's intelligence components in a blended context*."

5.2 Results of the Case Study and Discussion

The first results are presented in table 1. They were obtained in two moments of the study: questionnaire phase and after the two sessions. These results are divided into three branches: analytical (A), practical (P) and creative (C) one, per student. On the left side of table 1, it is presented the students' grades (0-100) obtained from the questionnaire answers, whereas on the right side, the results concern the students' grades after the two sessions. The standard deviation (SD) is also presented in both cases.

The results of the questionnaires show clearly that the creative component has lower grades when compared with the other components.

The bold numbers identify the students that showed great performance since the beginning of the study. Ten of eight students fit in this situation. Five of them, are

female students. There are 8 male students only. Consequently, this student group achieved better grades in relation to the female group.

In general, students' final results in each intelligence component reveal a better performance than the initial ones. In addition, the results point out a balancing among the individual intelligence components which can be proved with the lower values of the standard deviation when compared to the initial ones. We could also realize that some of the students improved on the component they had initially lower grade and had worse performance in the two other components. For instance, the X15 student fit this condition. Nevertheless, the majority of students obtain better results at least in two intelligence components.

Table 1. The grades of the students' intelligence components in two moments of the study.

Stud.	**Questionnaires results**			**S.D.**
	A	P	C	
X01	**30**	**50**	**30**	**9,43**
X02	**30**	**40**	**10**	**12,47**
X03	**60**	**60**	**100**	**18,86**
X04	**50**	**40**	**20**	**12,47**
X05	**40**	**50**	**10**	**17,00**
X06	40	50	20	12,47
X07	40	50	40	4,71
X08	**70**	**50**	**20**	**20,55**
X09	**40**	**10**	**0**	**17,00**
X10	60	50	30	12,47
X11	**60**	**60**	**10**	**23,57**
X12	70	90	60	12,47
X13	**70**	**40**	**30**	**17,00**
X14	30	50	10	16,33
X15	60	70	40	12,47
X16	40	30	50	8,16
X17	**50**	**40**	**20**	**12,47**
X18	10	30	0	12,47

Stud.	**Final results**			**S. D.**
	A	P	C	
X01	**82**	**81**	**65**	**7,79**
X02	**70**	**70**	**64**	**2,83**
X03	**87**	**57**	**100**	**18,01**
X04	**59**	**49**	**68**	**7,76**
X05	**65**	**43**	**67**	**10,87**
X06	72	17	91	31,38
X07	58	87	56	14,17
X08	**77**	**44**	**48**	**14,70**
X09	**78**	**46**	**43**	**15,84**
X10	42	72	52	12,48
X11	**64**	**66**	**90**	**11,81**
X12	43	90	81	20,37
X13	**82**	**80**	**49**	**15,11**
X14	27	74	17	24,85
X15	46	18	56	16,08
X16	65	55	27	16,08
X17	**50**	**61**	**48**	**5,715**
X18	55	50	23	14,06

Table 2 presents the average and the standard deviations of students' intelligence components calculated based on the results of the questionnaires. The female students' average grades have lower grades in all the three components when compared with the male students' ones. We do not have an explanation for this

situation. Perhaps, female students' mood in relation to computers in Chemistry Courses is lower when compared with the male students, which may contribute to lower performances. Further research effort is needed to get more answers.

Table 2. The average and the standard deviation of students' intelligence components obtained from the questionnaires.

Average	A	P	C
Fem.=	46,00	45,00	24,00
Masc.=	48,75	51,25	32,50
S. D.			
Fem.=	18,38	11,79	15,78
Masc.=	15,53	22,32	32,84

Table 3 shows the average and the standard deviations of students' intelligence components after the two sessions. The final average values to each intelligence component – analytical, practical and creative one – have improved significantly when compared with the ones at the beginning of the experiment. Adding this fact to the results achieved at the end of the two sessions, it seems that our hypothesis has succeeded.

Furthermore, we realize that the creative component improved significantly, which can demonstrate that the creative thinking can also be trained.

Table 3. The average and the standard deviation of students' intelligence components after the two sessions.

Average	**A**	**P**	**C**
Fem.=	58,40	56,10	45,90
Male=	67,25	62,38	73,25
S. D.			
Fem.=	15,68	20,35	17,87
Male=	16,86	22,77	20,32

We can also observe that the male students achieved better grades than the female students. One explication can be related to the fact that female students from non-technological courses may fill less comfortable with computers. We could observe that the male students worked in a more independent way with the adaptive application whereas the female students interrogated the teacher more often.

Some students have improved the intelligence component they had poor grade in the initial of the study but got worst grades in the other components. We do not have

an explication to this situation, except the fact that the adaptive application recommended more often learning activities connected with the student's intelligence component with poor performance.

In conclusion, the global results of the case study point out a balancing of the three intelligence components in the majority of the students. The average final results have improved when compared to the ones at the beginning and the standard deviations have improved too. We could also realize that students were very engaged, during the experimental sessions, doing their learning activities. As each session has a two-part format, students interact individually with the adaptive application and then they do tasks in group, may please to a larger number of students in accordance to their learning preferences.

For the reasons explained above, we claim that the hypermedia adaptive application, included in the ILEM, can contribute to improve the students' achievement in a blended context.

6 Conclusion and Future Work

In this paper, we have described a case study to evaluate our adaptive hypermedia application which is integrated in the learning teaching model, ILEM. The aim was to analyse a new learning-teaching model, ILEM, supported by an adaptive hypermedia application and based in the Triarchic Theory of (Successful) Intelligence, can contribute to maximize the balancing of the student's intelligence components in a blended context.

We have begun to introduce the theoretical foundations, namely the main principals of the Triarchic Theory of (Successful) Intelligence, and then, we have described the underlined methodology to plan the course: "Algorithms and VisualBasic Programming". We emphasize that the adaptive application recommends learning activities most suitable to each student based on the student's intelligence grades.

The overall results sustained the hypotheses of the case study and give us motivation to continue on this research direction.

Finally, we believe that our proposed model, ILEM, including the adaptive hypermedia application, provide a flexible learning-teaching environment and contributes to improve students' achievements.

Our further research is twofold. Firstly, we intend to improve the functionalities of the adaptive hypermedia application, namely, augmenting the diversity of means to assess students' intelligence components. Secondly, we want to investigate pedagogical methods and techniques deeply, especially those that concern the creative field. We argue that the creative field has more and more importance in our day-to-day living.

References

1. Sternberg, R.: The Theory of Successful Intelligence. In: Revista Interamericana de Psicologia/International Journal of Psychology, vol. 39, n. 2, pp. 189-202. (2005)
2. Sternberg, R.: Beyond IQ: A triachic theory of human intelligence. New York, USA: Cambridge University Press (1985)
3. Sternberg, R., Grigorenko, E.: Inteligência Plena: ensinando e incentivando a aprendizagem e a realização dos alunos. Editora ArTMed (2003). Translation of: Teaching for successful intelligence: to increase student learning and achievement. SkyLight Training and Publishing Inc. (2000)
4. Mota, D.: Um método de adaptabilidade de conteúdos multimédia num contexto semi-presencial. Master Dissertation, Faculdade de Engenharia da Universidade do Porto, Portugal (2008)
5. De Bra, P., Smiths, D., Stash, N.: Creating and Delivering Adaptive Courses with AHA!. In: Proceedings of the first European Conference on Technology Enhanced Learning, EC-TEL 2006 Springer LNCS 4227, pp. 21-33 (2006)
6. Gardner, H.: Frames of mind: The theory of multiple intelligences. New York: Basic Books (1983)
7. Feuerstein, R.: Instrumental Enrichement: An Intervention Program for Cognitive Modificability. Baltimore: University Park Press (1980).
8. Goleman, D.: Emotional Intelligence. New York: Bantam Books (1995)
9. Ceci, S. J.: On Intelligence: A bioecological treatise on intellectual development. Harvard University Press (1996)
10. Sternberg, R.: Raising the achievement of all students: Teaching for Successful Achievement. In: Educational Psychology Review, vol. 14, n. 4 (2002)
11. Sternberg, R., Torff, B. e Grigorenko, E.: Teaching Triarchically Improves School Achievment. In: Journal of Educational Psychology, vol. 90, n. 3, pp. 374-384 (1998)
12. Brusilovsky, P.: Adaptive hypermedia. In: User Modeling and User-Adapted Interaction Journal, Springer Netherlands, vol. 11, n. 1/2, pp. 87-110, (2001)
13. Romero, C., Martin-Palomo, S., De Bra, P., Ventura, S.:An Authoring Tool for Web-Based and Classical Tests. In: G. Richards (Ed.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education., pp. 174-177, (2004)

Competence gap analysis in the Skills Recognition process using Treemaps

Teresa Mota[1]

[1]University of Porto, Faculty of Engineering, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal,
teresa.alvesmota@gmail.com

Abstract. Recognition, Validation and Certification of Competencies (RVCC), is playing an important role in the qualification of young people and adults, through the formal recognition of skills, knowledge and competences gained through work experience, informal learning and life experience. During the process candidates are assisted by technicians and trainers to identify and recognise acquired skills/competences, to gather supporting evidence in order to demonstrate them. A commission then validates the candidate's skills and identifies any gaps, in which case it recommends additional training to be followed before final certification can be awarded. The whole process is oriented by a set of key competencies that works as a reference framework. Competence matching allows the identification of the gap between candidate's skills and the set of required key competences. Visualization information techniques, such as treemaps, can be used as a tool for gap analysis revealing the areas where the current level of the candidate can be improved.

Keywords: Skills recognition, Information Visualization, Treemaps.

1 Introduction

Portuguese government has launched a national system for Recognition, Validation and Certification of Competences [1], implemented through a network of centers

known as RVCC[1] Centers [2]. The RVCC system aims to formally validate the skills acquired by adults through work experience, informal learning and life experience. This process develops over a series of sessions during which candidates are assisted by technicians and trainers concerned to identify and recognise acquired skills/competences, to gather supporting evidence, and to demonstrate them. The candidates, individually and in small groups, identify, evaluate and reflect on their life experiences. They collect evidence showing their knowledge and begin to organize a portfolio which is the instrument for assessment of each individual [3]. The recognition of skills is therefore an organized and coherent process of identification and evaluation of knowledge and skills acquired by adults in contexts of non-formal and informal learning, revealed by a ”navigation" through their life stories.

A commission then validates the candidate’s skills and identifies any gaps, in which case it recommends additional training to be followed before final certification can be awarded.

To obtain a certification, the adult with the support of trainers, assess and build upon the work done in the first phase of the process, matching the skills shown by each adult with a set of key competencies. Thus the wide range of experience, knowledge and skills identified in the first phase is "mapped" to a list of key skills, skills necessary in contemporary society, to be formally recognized and validated.

This paper proposes a tree representation of the key competences, in order to allow a better understanding of their distribution within the areas and competence units in which they fall. The visualization of both acquired and required key competences within a tree representation is described as method for identification of the gap between candidate’s skills and the set of required key competences. Visualization information techniques, such as treemaps, can be used as a tool for gap analysis revealing the areas where the current level of the candidate can be improved.

This paper presents treemaps as a tool to provide gap analysis in the skills recognition process, since they provide an "elegant" view of hierarchical/tree structures, through an efficient use of available space on the screen, in contrast to long lists of items.

This paper is organized as follows: section 2 presents trees of Knowledge as an existing solution for competence visualization; section 3 provides an overview of tree visualization techniques and introduces treemaps; section 4 provides an overview of the key competences model; section 4 gives an example on how treemaps can be used for competence gap analysis.

2 Trees of knowledge

A review on previous efforts within competence visualization and gap analysis showed that trees of knowledge are the only solution that addresses the problem described in this paper, although focusing on the knowledge of a group instead of individuals. Trees of knowledge are a representation of knowledge or skills within a group of people, providing continuous review and update of the knowledge capital of the group. They are intended to highlight the skills and knowledge, broader than

[1] RVCC - Recognition, Validation and Certification of Competences

covered by conventional education systems. Its space is not determined by a default referential of competences, but by the organization of knowledge and skills of the individuals who are part of the group [4].

Tree of knowledge were proposed by Michael Authier and Pierre Levy in 1992 [5]. These researchers were also the founders of Trivium Soft, which developed the software tool known as Gingo, for the creation of Trees of Knowledge, and then a new product, called See-K, integrating Gingo in a web platform, along with a tool for creating Umaps [6], which are maps built on "Text mining" technology, which lets you analyze the content of documents, evaluation of evidence, notes, among other documents. Figure 1 shows some examples of trees constructed with Gingo, representing a summary of the skills of several groups.

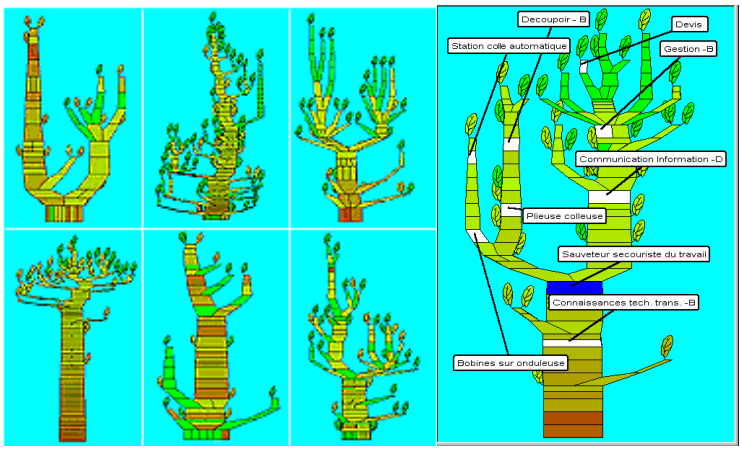


Fig. 1. Trees of knowledge build up with Gingo Software [6].

The tree is composed of several blocks, each block represents a competence or knowledge that at least one element of the group holds. The colour of the block provides information on the number of individuals that hold that competence [6].

The shape of the tree reflects the structure of the global knowledge of the group and identifies future directions for development and improvement, that is, the shape of the tree allows us to identify the areas where there are fewer skills, and where there are opportunities for growth and training of other individuals and so on. Most people share the skills that are located in the trunk. Those o hold skills within a branch of the tree are we can say that are a specialized community. Each individual can also view their own skills integrated in the tree, as show on the right tree of Figure 1, and see with who they share expertise within the group.

3 Treemaps

A large amount of information available, in various fields, is hierarchically structured and can be represented in the form of trees. The simple structures are easy to analyze

and understand, however this is not the case with more complex hierarchical structures and large, that is, trees with a high number of nodes and levels.
There are several techniques for the visualization of hierarchical data structures. Treemaps are proposed in this paper as the selected technique for the visualization of competences and gap analysis since it provides efficient utilization of space on the screen and the ability to represent structural information and content simultaneously in an elegant way.

There are two major categories of techniques for visualization of hierarchical structures: Node-link techniques and space-filling techniques. Figure 2 shows a hierarchical structure visualized according to these two types of techniques.

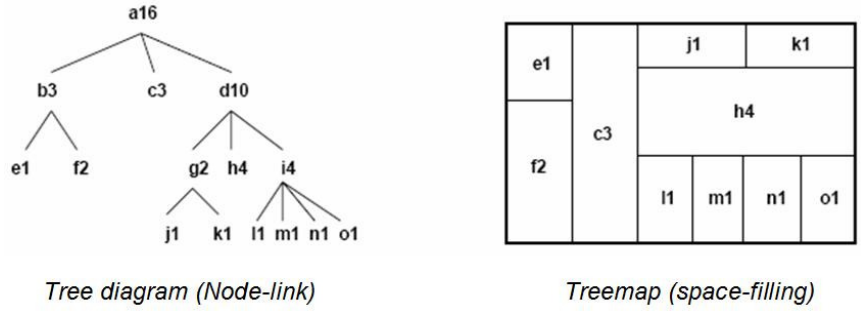


Fig. 2. Node-link techniques versus space-filling techniques.

Node-link techniques represent hierarchical structures by its nodes and branches or links between the nodes. For large data structures, these techniques do not allow an efficient use of available space on screen; it takes several screens to have a full view of large trees. There are various solutions developed in the field of view of the information they seek to minimize the waste of space on the screen and make it easier to navigate on representations of the node-link type, with examples of these solutions the Space Tree [7] *Star Tree* as hyperbolic browser [8].

Space-filling techniques represent hierarchical structures in a compact form, without waste of space on the screen, by a recursively splitting the initial available space. Treemaps are a visualization technique, developed by Ben Shneiderman [9], that maps hierarchically structured information into a rectangular 2D display in a space-filling manner. Unlike other methods, treemaps use the complete available space. The entire structural and content information is drawn in a single panel. The drawing area is partitioned into rectangles representing tree nodes. Rectangles which are inner nodes are also partitioned in order to contain their child nodes. This is recursively done until the whole tree is drawn.

Treemaps are an effective technique in the visualization of hierarchies in which each node has a numerical weight or attribute associated with it, such as the size or the space occupied by directories on a hard disk, which was the base problem in treemaps development [10]. The weight of a node can be used to influence the size of the rectangle on the screen, therefore a node with a low weight may be represented by a very small rectangle. Furthermore the fill colour of the rectangle can provide information about the node or the type of content associated with it.

The basic concept of a treemap includes the following aspects: subdivision of recursive initial rectangle; the size of each sub-rectangle is the size of the node; the colour can match the information on the node. The rectangles can have an aspect ratio (the ratio between length and width of the rectangle) variable. If this is close to one, are almost square, and therefore easily comparable. The layout of the rectangles depends on the used algorithm for division.

The original algorithm of division and also the simplest one is the *slice and dice*. However this algorithm generates rectangles with different aspect ratio. *Squarifed treemaps*, using a different algorithm for subdivision, try to eliminate these disadvantage dividing the space in regions with aspect ratio very close to one, that is the most similar possible to a square (see Figure 3). A disadvantage of this algorithm is that natural data order is not kept, generating unstable maps when changes in data occur [11]. Shneiderman has subsequently developed ordered treemaps, suitable for complex structures, and leading to layouts with low aspect ratio while maintaining the natural order of the data. The *Strip* algorithm provides treemaps with less square rectangles near but maintaining the natural order of the data, and as a result the treemap is much more stable as they data updates occur.

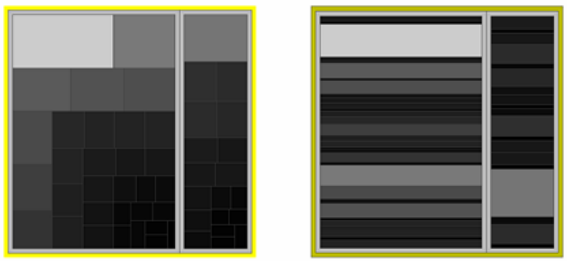


Fig. 3. Comparison between *squarifie*d layout (left) and *Slice and Dice* layout [12].

The different treemaps division algorithms can be compared using the following metrics:

- Aspect ratio of rectangles;
- Stability of the map in terms of size and position of rectangles;
- Order of the elements, respecting the natural order of the data;
- Readability of the hierarchical structure.

The different algorithms optimize one or more of the above mentioned metrics, like shown in the comparison between three division algorithms, presented in Table 1.

Table 1. Comparison between three different division algorithms.

	Order	Stability	*Aspect ratio*	Readbility
Slice and Dice	+	+	-	-
Squarified	-	-	+	+
Strip	+/-	+/-	+/-	+/-

The original treemap concept has been changed not only according to the division algorithm, as we have seen so far, but also according to usability aspects like user satisfaction regarding the appearance of the map and ease of understanding of the hierarchical structure of data. Examples include the Cushion treemaps [13], and the radial treemaps [14]. New treemap development also concern the extension of the treemap concept to a three-dimensional space (see Figure 4). Examples of 3D treemaps are *Step Tree* [15] and *Beam Trees* [16] where nodes are represented by the overlapping of cylindrical elements.

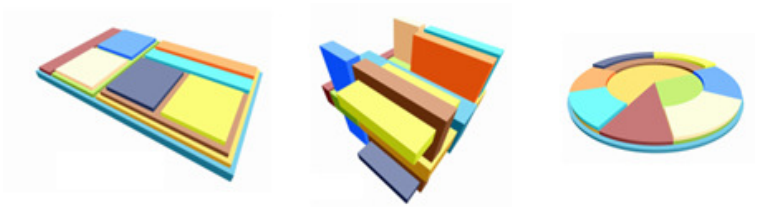


Fig. 4. 3D treemaps [17].

4 Competences Model

The RVCC system comprises two educational levels: basic and secondary level. The official specification of Key Competences for Secondary Educational level [3] is organized in three Key Competence areas: Society, Science and Technology (STC), Culture, Language and Communication (CLC), Citizenship and Professionality (CP). Each area is divided into several Competence Units (CU), each one generated from a Generator Core (GC), related with Reference Domains (RD). Each competence Unit is composed of several skills, evidenced through a set of Evidence Criteria (EC). Figure 5 shows the structure of the official specification of the Key Competences for Secondary Educational level [3].

CLC Culture, Language and Communication							STC Society, Science and Technology							CP Citizenship and Professionality							
GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC8
CU1	CU2	CU3	CU4	CU5	CU6	CU7	CU1	CU2	CU3	CU4	CU5	CU6	CU7	CU1	CU2	CU3	CU4	CU5	CU6	CU7	CU8
RD1 - Private Context																					
EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1
EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2
EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3
RD2 - Professional Context																					
EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1
EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2
EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3
RD3 - Institutional Context																					
EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1
EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2
EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3
RD4 - Macro-structural Context																					
EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1	EC1
EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2	EC2
EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3	EC3

Fig. 5. Official specification of Competences for secondary educational level.

The model proposed to represent this structure is presented in the diagram of Figure 6.

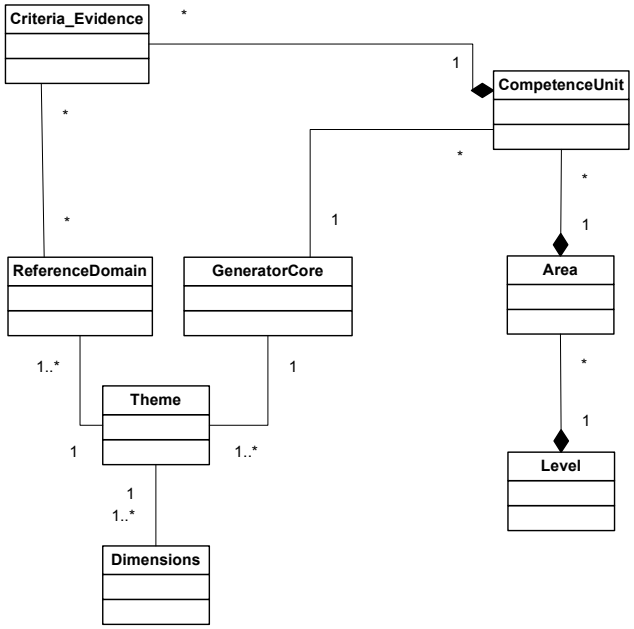


Fig. 6. Competences model.

This model is composed of the following structural elements:

- Dimensions - sets of Competence Units.
- Generator Core (GC) - theme from which evidences can be generated/produced.

- Reference Domains (RD) - contexts of action: private, professional, institutional and macro-structural.
- Theme - life situation in which skills are generated, implemented and observed.
- Competence Units (CU) – combination of consistent competences.
- Evidence Criteria (EC) – elements for which the adult must produce evidences.
- Evidence – Element that demonstrates the acquisition of a competence or part of it.

5 Competence gap analysis

The visualization model proposed in this section intends to show the acquired individual competence profile, providing simultaneously a perception of the gap between the acquired and required competence profile. Treemap 4.1.1 software, developed by the HCI Laboratory at the University of Maryland, is proposed as a tool to display, compare and explored interactively large competence profiles in the same framework.

A comparative study conducted by Alfred Kosba [18] provides a comparison of the several techniques for visualization of hierarchical data structures: *Treemap* (TM*)* – *treemaps*; *Sequoia View* (SV) – cushion treemaps; *Beam Trees* (BT) – beam tree; *Star Tree* (ST) – hyperbolic trees; *Tree Viewer* (TV) – botanical trees; *Windows Explorer* (EX)- conventional browser view. In this study 15 tasks, in the test hierarchy, were given to 48 user's representative of the target group. According to the quantitative results (average task completion times, correctness of answers, and perceived easy of use and effectiveness - user satisfaction), Treemap turned out to be the best visualization system overall in this study.

In the RVCC process the hierarchical data structure to visualize, i.e. the acquired competence profile, is composed of the evidence criteria (CEs) for which the individual has gathered supporting evidence in order to demonstrate it. This structure is provided to the Treemap 4.1.1 visualization tool in a table (see Table 2) resulting from the junction of several other tables, implemented according to the conceptual model described in Figure 6 of the previous section.

Table 2. Hierarchical data structure - acquired competence profile.

Level	Area	Competence Unit	Domain of reference	Criteria for Evidence
4	CP	UC1	DR1	CE1
4	CP	UC3	DR2	CE1
4	CP	UC3	DR2	CE2
4	CP	UC3	DR2	CE3
4	CP	UC5	DR1	CE3
4	CP	UC5	DR2	CE3
4	CP	UC5	DR4	CE1
4	CLC	UC1	DR1	CE1
4	CLC	UC3	DR2	CE1
4	CLC	UC3	DR2	CE2
4	CLC	UC3	DR2	CE3
4	CLC	UC5	DR1	CE3
4	CLC	UC5	DR2	CE3
4	CLC	UC5	DR1	CE1
4	CLC	UC5	DR2	CE3
4	STC	UC1	DR1	CE1
4	STC	UC3	DR2	CE1
4	STC	UC3	DR2	CE2
4	STC	UC3	DR2	CE3
4	STC	UC5	DR1	CE3

Each line of this table represents evidence and each column represents an attribute (area, unit of competence, criteria of evidence) for that evidence. Each attribute value on the table corresponds to a node of the tree, as shown in the diagram of Figure 7.

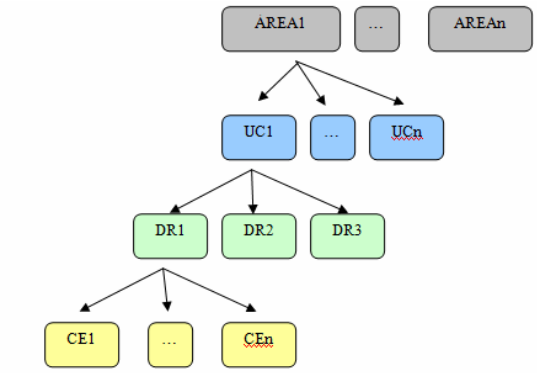


Fig. 7. Node-link representation of the hierarchical data structure in Table 2.

Figure 8 shows a treemap corresponding to the structure presented in Table 2. The colour of the rectangles is allocated to the attribute *Area* while the size is associated with the percentage of Criteria of evidence (CEs) for which the candidate supplied an evidence.

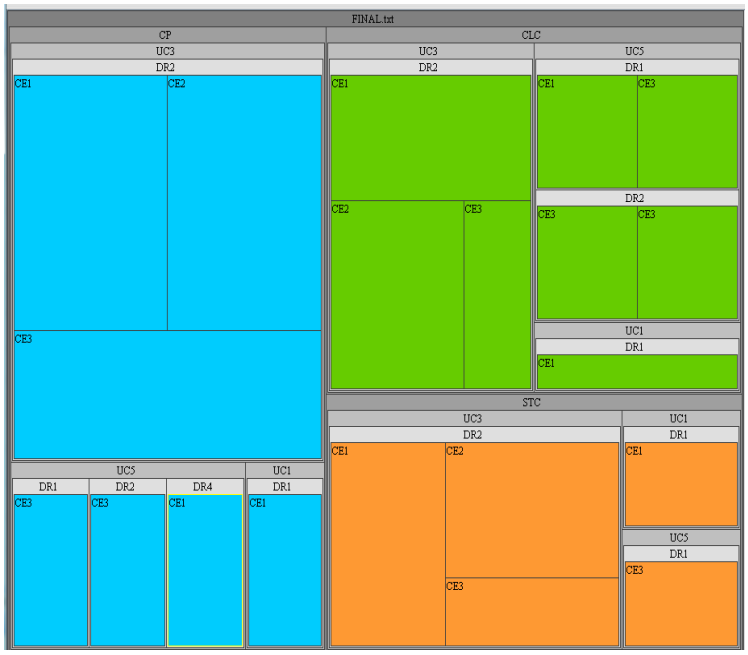


Fig. 8. Treemap representing the hierarchical data structure of Table 2.

Figure 9 shows the treemaps corresponding to the structure presented in Table 2, for each one of the presented division algorithms: Slice and dice, Squarified and Strip. As it can be seen both Squarified and Strip algorithm produce good results in terms of readability. It is easy to understand which areas have more criteria of evidence met. The aspect ratio of the rectangles in the squarified treemap is better, allowing a better comparison, but the order and position of rectangles is not maintained when the data is updated. Since data order is not relevant for this study squarified treemaps are a good choice for implementation.

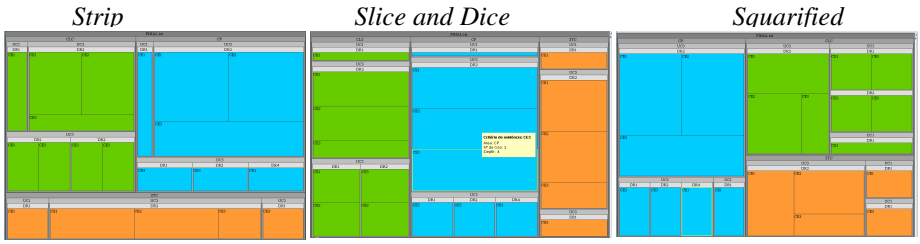


Fig. 9. Treemaps for data structure of Table 2, with different division algorithms.

Competence gap analysis can be achieved if the treemap shows not only the CEs with evidence but also those CEs who lack evidence, highlighting the difference between the two types of CEs. This can be done in two ways:

- Determining the initial available space for subdivision as a function of the total percentage of CEs with evidence, therefore satisfied. In this case a bigger treemap corresponds to a lower competence gap.
- Adding to the hierarchical structure a node corresponding to the CEs that lack evidence, therefore are not satisfied. The addition of new a attribute, *status,* splits the tree in two main groups: satisfied CEs and not satisfied CEs like shown in Figure 10, where the white area represents the competence gap.

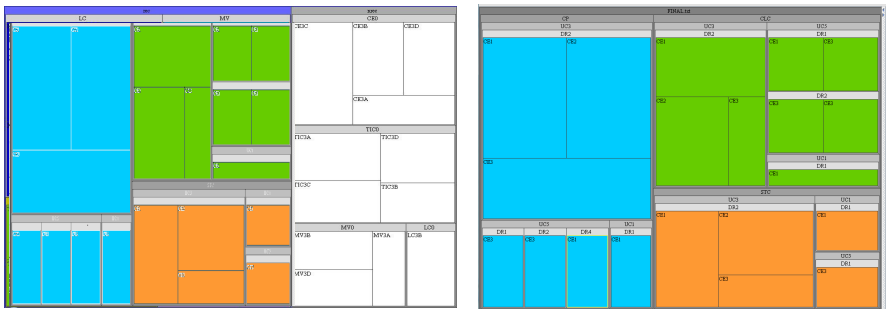


Fig. 10. Treemaps representing the hierarchical data structure of Table 2.

6 Conclusions

The recognition and validation of skills is a process which aims to certify skills acquired through life and work experience. The visual representation of skills is an added value to the skills recognition process, since it allows the candidates to feel the constructive process, and have an idea of their strengths and weaknesses.

Visualization information techniques, such as treemaps, can be used as a tool for gap analysis revealing the areas where the current level of the candidate can be improved.

Of the several visualization techniques for hierarchical data structures treemaps have many advantages such as the efficient utilization of space on the screen and the ability to represent structural information and content simultaneously in one single panel. Treemaps could also be used in a future work in the data mining field, as a tool for visual knowledge extraction (clustering, etc), within de RVCC system.

References

1. ANQ. *Sistema Nacional de Reconhecimento , Validação e Certificação de Competência*. 2007 [cited 2007]; Available from: www.anq.gov.pt.

2. M.Educação and M.TSS. *Iniciativa novas oportunidades*. 2006 2006 [cited; Available from: www.novasoportunidades.gov.pt.
3. DGFV, *Referencial de competências-chave para a educação e formação de adultos - Nível Secundário*, D.G.d.F. Vocacional, Editor. 2006.
4. Sens, A. *Arbor & Sens*. [cited; Available from: http://www.arbor-et-sens.org.
5. Chancerel, J.-L. and B. Collot. *Presentation theorique et methodologique de L'approche des "Abres de connaisances"*. 2007 [cited 2007]; Available from: http://www.arbor-et-sens.org/.
6. Lebrun, C., *De Gingo à See-K- mais toujours les arbres de connaissances*. 2002, Trivium Soft.
7. Plaisant, C., J. Grosjean, and B.B. Bederson, *SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation*. Human-Computer Interaction Laboratory, 2002.
8. Rao, J.L.a.R. *Visualizing Large Trees Using the Hyperbolic Browser*. in *CHI96*. 1996.
9. B.Shneiderman and B.Johnson, *Treemaps: A space Filling Aproach to the visualization of hierarchical information structures*. Readings in Information Visualization - Using vision to think. 1999: Morgan Kaufmann Publishers.
10. Shneiderman, B. *A history of treemap research at the University of Maryland*. 1998b [cited; Available from: http://www.cs.umd.edu/hcil/treemap-history.
11. Shneiderman, B. and M. Wattenberg, *Ordered Treemap Layouts*. 2001.
12. Engdahl, B., *Ordered and unordered treemap algorithms and their applications on handheld devices*. 2005, Royal Institute of Technology: Stockolm.
13. Wijk, J.J.v. and H.v.d. Wetering, *Cushion Treemaps: Visualization of Hierarchical Information*, in *IEEE Symposium on information visualization 1999*. 1999.
14. Stasko, J. and E. Zhang, *Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations*. 2000.
15. Bladh, T., D.A. Carr, and J. Scholl, *Extending Tree-Maps to Three Dimensions: A Comparative Study*. Lecture Notes in Computer Science. Vol. 3101/2004. 2004: Springer Berlin / Heidelberg.
16. Ham, F.v. and J.J.v. Wijk, *Beamtrees : Compact Visualization of Large Hierarchies*. Information Visualization, 2003. **2**(1): p. 31-39(9).
17. Schulz, H.-J., M. Luboschik, and H. Schumann (2006) *Interactive Poster: Exploration of the 3D Treemap Design Space*. **Volume**,
18. Kosba, A., *User Experiments with Tree Visualization Systems*, in *IEEE Symposiumon Information Visualization* 2004.