CoMIC'06

Conferência de Metodologias de Investigação Científica

ACTAS

9 de Janeiro de 2006

Faculdade de Engenharia
da Universidade do Porto

Universidade do Porto
FEUP Faculdade de Engenharia

Programa de Doutoramento em Engenharia Informática
Faculdade de Engenharia da Universidade do Porto, 2006

Actas da CoMIC'06

1ª Conferência em Metodologias de Investigação Científica

9 de Janeiro de 2006

Programa de Doutoramento em Engenharia Informática
Faculdade de Engenharia da Universidade do Porto

Patrocínios

Prefácio

O Programa de Doutoramento em Engenharia Informática (ProDEI) da FEUP teve início neste ano lectivo de 2005/06. Do plano de estudos que compõe a parte curricular consta a disciplina MIC - Metodologias de Investigação Científica, cujos objectivos principais são fazer com que os alunos assimilem, os processos, metodologias e práticas associados à Investigação Científica em vários domínios, incluindo a Informática, assim como aumentar a sua capacidade de produção de texto científico em formato adequado.

Com um formato misto, baseado em tutoriais e seminários multidisciplinares, a disciplina culminou com a realização da CoMIC - Conferência de Metodologias de Investigação Científica. Esta destinou-se a funcionar, figurativamente, como um laboratório de prova dos conceitos apreendidos pelos alunos: estes são os autores dos artigos, a comissão de organização e a comissão científica, devidamente acompanhados pelos docentes da disciplina e de outros docentes que se comprometeram a ajudar na revisão dos artigos. Com uma divulgação alargada da CoMIC, surgiu mesmo a hipótese de se juntarem artigos com proveniência de outras escolas, desde que publicados num contexto semelhante: por alunos de doutoramento em áreas de Informática.

Este volume reúne os 11 artigos publicados nesse contexto, reunidos, de acordo com os assuntos versados, em quatro sessões técnicas da conferência. A heterogeneidade dos temas é grande, reflectindo a abrangência das áreas científicas do ProDEI, assim como é diverso o nível de profundidade das apresentações, este último facto decorrente de diferentes estados dos trabalhos conducentes a doutoramento por parte dos vários intervenientes. Apesar dos artigos publicados já demonstrarem os interesses actuais dos doutorandos, a principal preocupação evidenciada pelos autores não foi a natureza e o detalhe científico das questões abordadas mas, principalmente, a redacção e a forma de exposição.

Uma análise de conteúdos dos artigos publicados permite uma classificação num sistema de coordenadas a três dimensões que ajuda a discriminar as várias abordagens. No eixo Generalidade versus Específico, verificamos que todos, excepto dois (artigos sobre programação orientada a aspectos e sobre modelação de processos de negócios) tendem a debruçar-se sobre problemas mais específicos.

No eixo Estado-da-arte versus Hipóteses próprias, verifica-se um equilíbrio, com seis artigos ousando já aventar novas hipótese a provar, e cinco mais baseados em relatórios de trabalho relacionado com o tema escolhido. Esta situação reflecte sem dúvida a assimetria da situação dos estudantes, tendo alguns iniciado os trabalhos antes do início do ProDEI.

Finalmente, no eixo Teoria versus Aplicações verifica-se uma acentuada preferência (oito artigos) por temas mais aplicados, dois artigos de índole mais teórica, e um outro (sobre Contratos Electrónicos) casando uma parte mais teórica com uma mais aplicativa.

Os docentes de MIC, em face dos bons resultados obtidos, agradecem o empenho de todos quantos participaram nesta realização que, esperam, tenha contribuído para uma melhor apreensão dos temas tratados ao longo da disciplina, nos domínios da pesquisa científica e da escrita de documentos relacionados.

Eugénio Oliveira e A. Augusto de Sousa,
(Docentes de MIC - Programa de Doutoramento em Engenharia Informática)

Programa

Sessão Técnica 1

Sessão Técnica 2

Sessão Técnica 3

Sessão Técnica 4

Sessão Técnica 1

Data Mining e Extracção de Informação

Applying Data Mining Techniques to Football Data from European Championships

Sérgio Nunes[1] and Marco Sousa[2]

[1] Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
sergio.nunes@fe.up.pt
[2] zerozero.pt
http://www.zerozero.pt
msousa@zerozero.pt

Abstract. Data Mining is the process of finding new, potentially useful and non trivial knowledge from data. Football is a popular game worldwide and a rich source of data. Gathering only part of this data we are able to collect hundreds of cases. In this paper we describe an exploratory work where we use Data Association Rules, Classification and Visualization techniques to find patterns in datasets from several European championships. For each one of these techniques, different hypotheses were stated. For Association Rules and Visualization, our hypothesis was that we would be able to find non trivial knowledge and confirm several known patterns. For Classification, our hypothesis was that we would be able to classify matches according to their results based on the available history. Our findings didn't confirm our hypotheses to their full extent as expected. Our exploratory work confirmed several well known patterns in football and highlighted borderline cases. Among the several techniques used, visualization produced the best results.

1 Introduction

1.1 Context and Motivation

Data Mining (DM) is commonly viewed as a specific phase in the Knowledge Discovery in Databases (KDD) process. Currently, *Data Mining* is an overloaded term used to mean several concepts. We consider DM to be the application of machine learning techniques to extract implicit, previously unknown, and potentially useful information from data [12]. Nevertheless, during this paper, we will sometimes use this term to refer to the whole process of KDD. The exponential increase in the amount of data that exists stored in electronic databases has fostered the growth of this field. A simple search for "data mining" in any popular web search engine will return several millions of hits [1].

[1] In December 2005, a search in Google Search returns more than 17.900.000 hits.

Football is a very popular game worldwide, it was invented in England in the XIX century and is now played regularly by more than 240 million people according to Fédération Internationale de Football Association (FIFA) [8]. Football is also known as soccer, or association football, in some countries, namely in the USA.

The motivation for this project arouse from an opportunity to work with a large database of football data. This data was provided by zerozero.pt [4], an independently maintained website that gathers and presents data from several football championships worldwide. The data granularity varies significantly among championships. Two main datasets were used. The first dataset includes the 2004/05 edition of the Portuguese championship and was chosen because it is the one with the highest level of detail and the lowest levels of missing values and erroneous data. The second dataset includes all matches played in six European countries, including Portugal, for the last 50 years [2]. Although rich in the total number of cases, this dataset has very few attributes available.

In this paper we present an exploratory work where we apply several DM techniques to the chosen datasets in search of existing patterns. We expect to find patterns that relate the events in a non trivial fashion. If these patterns are found, they can provide valuable insight to the people involved directly or indirectly in the match. An example of application would be the development of a decision support system to be used during the match. Another example application would be the use of this information to aid in the selection of referee or locations for each match.

We are not aware of any published work where these specific DM techniques are applied to football data to discover or confirm existing patterns. Research found in this area is mainly related to robot soccer and autonomous agents [10]. In this case, data mining modules were developed to provide adaptive agent behavior in dynamically changing environments using automata data. Considering the use of DM in other sports besided football, the work published by Bhandari et al. in 1997 [6] describes Advanced Scout, a PC-based data mining application used by the NBA coaching staffs to discover interesting patterns in basketball game data.

1.2 Paper Structure

The Cross-Industry Standard Process for Data Mining (CRIPS-DM) [7], an European Community developed standard framework for data mining tasks, identifies six generic phases in the life cycle of a data mining project. In this work, these phases are used to structure the paper. The first phase, called **Business Understanding**, focuses on understanding the project objectives and requirements from a business perspective and setting a preliminary plan to achieve the objectives. This has been covered in Section 1.

In Section 2, the next two phases of the CRISP-DM process are covered, **Data Understanding** and **Data Preparation**. The Data Understanding phase starts

[2] For some countries we have all the matches played since the XIX century.

with the initial data collection and proceeds with activities in order to get familiar with the data. Also included in this phase are activities related to the analysis of quality problems in the data. The Data Preparation phase covers all activities to construct the final dataset from the initial raw data. Also in this section, initial obvious results are presented.

In Section 3, the **Modeling** phase of the CRISP-DM process is covered, several modeling techniques are applied and tuned for best performance. We use Data Association Rules, Classification and Visualization to mine the datasets.

Finally, the **Evaluation Phase** is covered in Section 4. The results from the previous section are organized, presented and discussed. In this section our work is viewed in the light of our initial hypotheses.

In the CRISP-DM framework, one last phase of deployment is identified. This phase wasn't included in our work since it wasn't one of our goals.

2 Datasets

2.1 Data Preparation and Exploration

The raw data was collected from zerozero.pt's [4] main database. The data is stored in a relational database management system and was exported to flat CSV files using PHP scripts and SQL. The initial exploration and preparation of the CSV files was done using R [2], an open-source language and environment for statistical computing and graphics.

The two datasets used are described in the following sections. Initial data explorations are also described.

2.2 Portuguese Championship 2004/05 Events

All existing events from the 2004/05 edition of the Portuguese football championship were exported. This edition of the Portuguese championship included 18 teams that performed a total of 306 matches, there were 711 goals scored and a total of 1.771 cards shown by the referees.

The exported data includes information about the players in each match, substitutions made during the match, the time and location of the match and information about the teams and players when the match happened. For instance, for each team, there is information available about the number of points, goals scored and goals conceded since the beginning of the championship. On the other hand, for each player and besides demographic data, there is information about the number of goals scored and cards received since the beginning of the championship.

The final dataset has more than 17.000 cases, each one with more than 50 features. Each case represents an event (see Table 2). For each event, the available features are summarized and explained in Table 1.

Football occurrences stored in the original database were analyzed and normalized to fit a standard representation. In this standard representation, each

Table 1. Features available in the Portuguese Championship dataset.

Group	Related Features
Event	Related to each event: type, minute and half within the match.
Match	Related to the match being played: date, start time, score, TV channel transmitting, referee, number of spectators and total overtime granted.
Teams	Related to each team involved in the match: name, coach and current position, number of points, victories, defeats and draws in the championship.
Location	Related to the place where the match takes place: stadium and city.
Player	Related to the player involved the event: name, age, playing position, nationality, birth country, weight and height.

event has only one player associated. Hence, all occurrences were split in multiple simpler events. For example, a substitution corresponds to 2 events - one associated with the player leaving, another associated with the player entering. Another example is the initial line up of the teams, that correspond to 36 events - 11 starter events and 7 substitute events from each team. In the end, 9 types of normalized events were identified and characterized. These types are depicted in Table 2.

The final dataset has very few errors or missing values. This was one of the factors considered to choose this dataset. In fact, the 2004/05 edition of the Portuguese championship is the most complete one in zerozero's database. Existing errors, missing values and outliers were easily detected using simple statistical tools, namely boxplots. These records were deleted, no attempt was made to fill in or correct the data.

An initial exploration of the data was performed using statistical tools. A density chart for event types was plotted (see Figure 1). It is interesting to note that:

- Substitutions only start to occur at the end of the first half, being rare at the beginning of the match.
- There is a strong peak of substitutions near the minute 45. This corresponds to the substitutions performed at half time.
- The number of cards shown increases during the match with peaks at the end. Red cards and second yellow cards have a very high peak near the end of the match.
- During the first half of the match, the number of double yellow cards is very low and is surpassed by the number of red cards.
- The number of goals doesn't exhibits peaks but increases in the end of the match.

Table 2. Event types in the Portuguese Championship dataset.

Event Type	Description
Starter	Represents a starter player included in the initial lineup. For each match there are 22 events of this type, 11 for each team, occurring in the minute 0 of the match.
Substitute	Represents a substitute player for the match. For each match there are 14 events of this type, occurring in the minute 0 of the match.
In	Represents the exiting of a player during a substitution.
Out	Represents the entering of a player during a substitution.
Yellow	Represents the showing of an yellow card to a player.
Second Yellow	Represents the showing of the second yellow card to a player.
Red	Represents the showing of a direct red card to a player.
Goal	Represents the scoring of a standard goal.
Penalty	Represents the scoring of a penalty.
AutoGoal	Represents the scoring of an auto goal.

Although a football match starts at minute 0 and ends near minute 90, in Figure 1 the various lines begin before and end after these values. This is a result of the smoothing performed by the density function available in R.

2.3 European Matches

The second dataset contained information about the championships and matches from several European countries. The countries included were: Portugal (15.382 matches since 1934), England (43.730 since 1888), Spain (19.846 since 1930), Italy (17.680 since 1946), France (22.702 since 1933) and Germany (13.406 since 1963). Although a large number of cases (matches) were collected (132.749 in total), few features were available for all matches for all countries. The features included in this dataset are shown in Table 3

In Figure 2, the three major teams in Portugal were plotted by year and by final position. Each team was drawn with a different shade of gray. It is evident the predominance of these three teams in the history of the Portuguese championship. A more detailed analysis of this figure reveals that:

- FC Porto has the most irregular path. Prior to the 80s several fluctuations in the final position achieved are evident. While Benfica has the most overall consistency.
- The 50s were dominated by Sporting, the 60s, 70s and part of the 80s were dominated by Benfica and, since the middle of the 80s, FC Porto has won

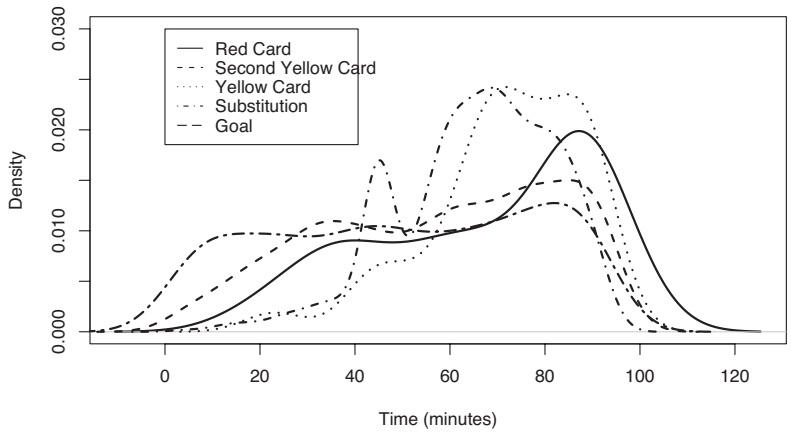


Fig. 1. Density plots for event types.

Table 3. Features available in the European Championships dataset.

Feature
Visited and Visiting team's name.
For each match, the number of goals scored, the number of goals suffered and the winner.
Country's name, year and decade of the match.
For each team, the number of goals scored and suffered for each specific championship (total, in and out).
For each team, the number of points, victories, draws and defeats for each specific championship (total, in and out).

most of the championships. A density plot for each first place for each team clearly reveals this pattern.
– For each team, exceptional bad seasons are evident - FC Porto (40s, 1969) and Benfica (2000).
– The two championships won by none of these three teams are easily spotted, 2000 (Boavista) and 1945 (Belenenses).

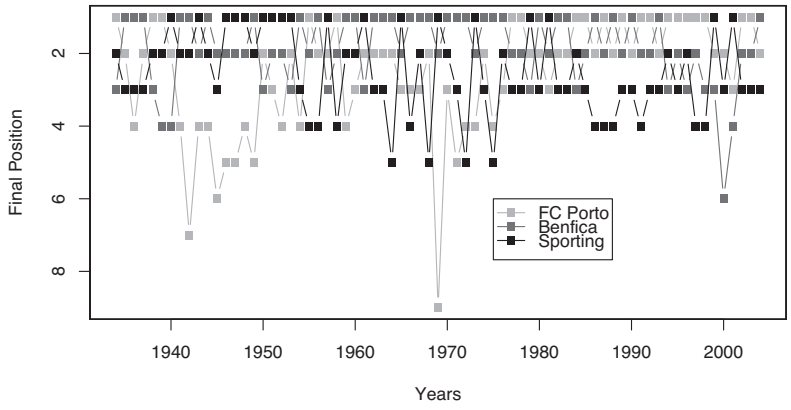


Fig. 2. Final positions for the three major teams in Portugal (1934-2004).

Among countries, density plots for the matches along the years reveal interesting patterns. In Figure 3, density plots for England, France and Portugal are shown. Before 1920 and after 1940 the two World Wars are evident in the plots for England and France. For Portugal, the increase in the number of matches is visible.

3 Modeling

3.1 Association Rules

Mining for association rules is a DM technique that enables the finding of frequent patterns, associations, correlations or casual structures among sets of items. This task was performed using two different open-source software tools, Weka [3] and AlphaMiner [1]. Due to the low number of attributes in the European championships dataset, only the Portuguese dataset was used. In this case, after the discretization of numerical variables, a total of 40 nominal attributes in 16.900 cases were available.

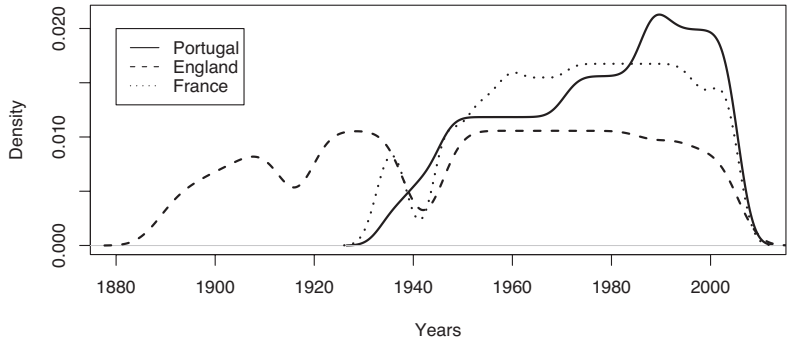


Fig. 3. Total number of matches density.

We used the Apriori algorithm [5] to search for association rules. Three types of metrics were used with different minimum values: Confidence (75%), Lift (1.5) and Leverage (0.1). For each one of these metrics, a minimum support of 25% was set and a maximum of 100 rules were produced.

Having a reasonable number of attributes and a high number of cases yielded high expectations towards the finding of patterns. Nevertheless, after exhaustive exploration, no interesting or unexpected rules were found. Only trivial rules were identified, for example: *"Matches that start between 15:30 and 16:30 are on Sundays"* (84% conf.) or *"Matches that are not transmitted on TV are on Sundays"* (80% conf.).

3.2 Classification

Classification is a DM technique for mapping objects into predefined classes. Classification was performed using Weka's implementation of the C4.5 algorithm [9], named J48. This technique was used only with the second dataset. In this case it is possible to set interesting, realistic and useful goals. Despite having many more attributes, the first dataset is less interesting as a classification problem. In this case, simple tests have showed that, for example, predicting the end result of a match is quite trivial since we have all the events for that match.

With the second dataset, including match results from several European countries, we set the goal of classifying each match according to the final result. Three match results are possible for the visited team: victory, defeat or draw. A very small set of attributes was used, namely the name of both teams and the year of the championship. The dataset was also split by country and several runs of the classification algorithm were performed with different values for the confidence factor (C) and the minimum instances per leaf (M). The values used

were: C (0.05, 0.1, 0.5, 1, 10) and M (1, 5, 10, 20, 50). Each model was tested using a training set (70%) and a test set (30%).

For Portugal, the best model (C=0,05 and M=50) was able to correctly classify 59,81% of the test set instances. This score was obtained with two simple rules:

- When the visiting team is "FC Porto", "Benfica" or "Sporting" the result is **defeat**.
- In every other case the result is **victory**.

In this model, no matches were classified as "draw". With a trivial classifier, based on the frequency of each result in the Portuguese Championship (victory 54%, draw 23%, defeat 22%), we have a success rate of 54% classifying every match as "victory". Thus, we can state that our classifier only slightly improves this result, being able to correctly classify 5% more cases.

Nevertheless, only for Portugal we were able to surpass the results achieved by the trivial classifier. In each of the remaining five countries, the best rules simply classified every match as "victory". Thus achieving a success rate equal to the one accomplished with the simple statistical classifier. This can be explained by the predominance of only three teams in the Portuguese Championship. In all other countries there is a greater balance among the various teams, making classification based on a small set of attributes a harder task.

3.3 Visualization

Visualization techniques make use of graphics to produce multiple observations of the data. Of the methods used in this work, this is the most exploratory since no rules are defined on how to conduct research. Visualization is mainly developed for human observation and allows multiple insights into the same data. Visualization can be used to simply view outputs from the application of other techniques or to explore the initial input. In this work, several plots were drawn using R with an exploratory mindset. In this section we show and comment those that are most revealing or unexpected.

Although several experiments were made with the first dataset, visual results were below our expectations. Hence, only explorations with the second dataset are presented. In Figure 4, first places among countries are plotted. Each line represents one country and each year is depicted in the X axis. For each team that won the championship, a different color was used. Different shades of gray were used since they provide an easily comprehended scale to human observation [11]. It is important to note that we only have all the matches, from the start of each championship, for Portugal, England and France. Nevertheless, the following observations are possible:

- Portugal has a very low diversity in the number of teams that won the championships.

– England has the highest diversity on the teams that won the championship. The 50s mark a clear separation on the teams that commonly won the championship.
– Interruptions, mainly due to the World Wars, are easily spotted among championships.

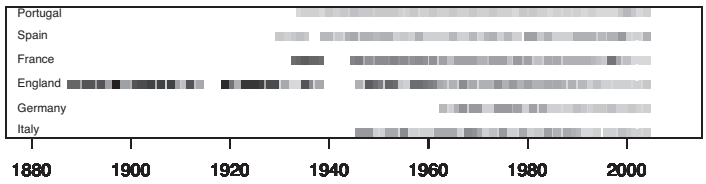


Fig. 4. First placed teams in European championships.

An alternative visual display of the three major teams in the Portuguese championship was produced. In Figure 5 teams are plotted by year and by total points achieved, instead of their final position (as in Figure 2). Although the final positions aren't so clear, more information is available in this second graphic. We are able to see the evolution in the total number of points along the years, reflecting the evolution in the number of teams. Also visible is the increase in the mid 90s, as a consequence from the changing of the rules (victories worth 3 points instead of 2). Excellent seasons are easily spotted, namely Benfica's (1971, 1972) and FC Porto (1995, 1996). Also interesting to note are the bad overall seasons as compared to neighbor championships. For example, in 2004 the winning team achieved fewer points than the third team in several of the previous years.

We've also performed a visual analysis of the evolution of match results for each year in each country. Each match result was plotted in a 2D graphic with the X axis being the goals received and the Y axis the goals scored. These results were then grouped by year and the year's centroid was calculated. In Figure 6 and 7 these centroids are plotted for Portugal and England. Different shades of gray were used for each year, so that the time dimension was visible in the figures. Although similar in recent years, these figures show that match results in England have fewer variations. In Portugal, significant differences between the older matches and the more recent matches are impressive. These analyses were also performed for the other countries and we concluded that France, Germany and Italy exhibit a pattern similar to England's, while in Spain the pattern is more similar to Portugal's.

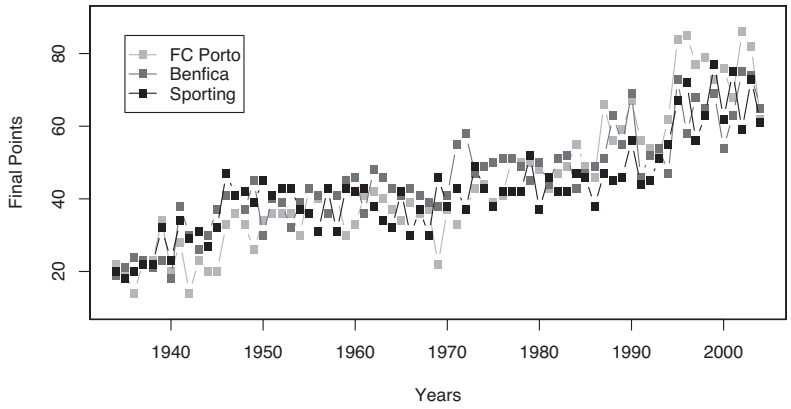


Fig. 5. Total points by championship for the three major teams in Portugal (1934-2004).

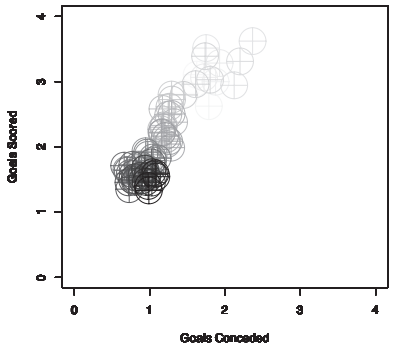


Fig. 6. Centroids for match results in Portugal (1934-2004).

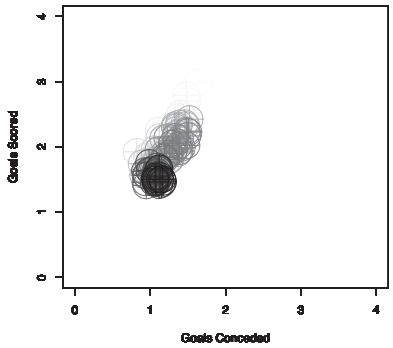


Fig. 7. Centroids for match results in England (1988-2004).

This type of centroid plots were also used to analyze data within each country. Two different analyses are shown for the Portuguese championship. In the first example (Figures 8 and 9), each plot represents the team's match result according to four dimensions: time, goals scored, goals conceded and place. Time is represented using different shades of gray, lighter colors portrait older matches. Goals scored and goals conceded are depicted in the plot's axis. Finally, the place of the match is distinguished using different symbols for each centroid, matches at home are plotted using a circle while matches away are plotted with a square. These plots were produced for every team in the Portuguese championship. Benfica and Boavista were chosen because their plots reveal contrasting evolutions in each team's match results. While Benfica had a greater change in home matches, Boavista had an even greater change in away matches.

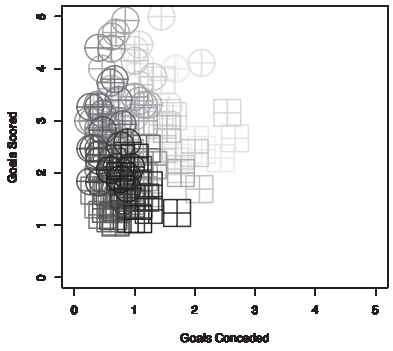


Fig. 8. Centroids for Benfica matches in the Portuguese Championship (1934-2004).

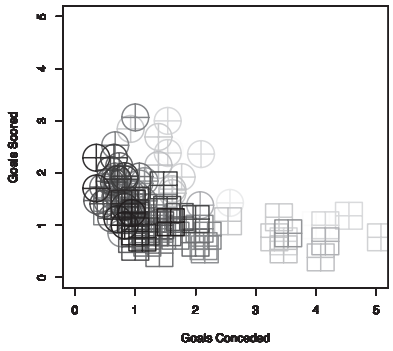


Fig. 9. Centroids for Boavista matches in the Portuguese Championship (1934-2004).

Finally, a similar type of graph was used to compare two teams. In Figure 10 and 11 two examples are shown. For each two teams, the most common match results are shown using different sizes for each point. It is important to note that these plots represent only the matches of Team A versus Team B, not Team B versus Team A. In the examples shown, two different patterns are visible. As expected, in matches against Belenenses, FC Porto concedes fewer goals and the results are concentrated in the "victory side" of the plot. With Benfica, while victories still dominate, draws are more frequent and the amplitude of goals scored is much lower.

4 Conclusions

Our initial expectations were that we would be able to find non trivial knowledge from the available datasets. After several explorations only existing strong sus-

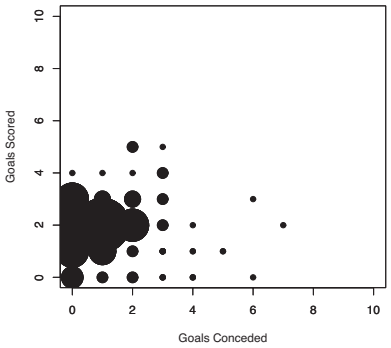


Fig. 10. FC Porto versus Benfica in the Portuguese Championship (1934-2004).

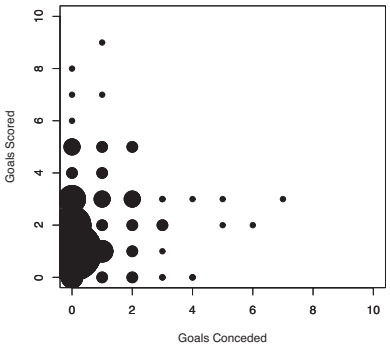


Fig. 11. FC Porto versus Belenenses in the Portuguese Championship (1934-2004).

picions were confirmed. Although we were able to extract knowledge from these datasets, no important and unexpected result was revealed, thus our initial hypothesis was partially refuted. It is important to refer that our hypothesis was partially refused for these two datasets where, although a significant amount of cases is available, the number of attributes is limited. We believe this is the main reason for the bad results obtained with Association Rules and Classification.

With Classification we were able to produce a model for the Portuguese championship that returns better results than a pure probabilistic classifier. This can be explained by the high predominance of three teams that exists in this championship.

The good results obtained with visualization were unexpected. We believe that the high number of numerical attributes and the existing knowledge of the domain greatly justifies this success. Most of the graphics produced emerged as a way to see patterns that were already known in advance. While the other two techniques search for patterns with few inputs from domain experts, with visualization human intervention is necessary during the decision process.

Data preparation is a very time consuming step in the KDD process. Gathering and preparing data to be used with the different algorithms occupied a significant part of the whole process. More than two thirds of our work was invested in data preparation. The word *mining* clearly reflects the nature of the whole KDD process. A lot of time is spent searching for patterns, adjusting parameters in the algorithms and drawing graphics, to find out that only a minimum part of this work is useful in the end. The results obtained are directly related to the time invested in the work.

Several tools were used to perform the data mining tasks. AlphaMiner was found to be very well designed for a knowledge discovery work. Tasks are graphically shown and the steps are evident, useful for the kind of work developed while following an exploratory path. Although being graphically intuitive, this

tool offers less KD methods than Weka and can't cope with large volumes of data as well as Weka. R is an excellent statistical software tool, it is able to perform calculations on large datasets and provides a large repository of packages with extra features.

As future work, we suggest additional exploration of visualization techniques and, if possible, the gathering of more attributes to allow the use of other data mining techniques with improved success. Due to the characteristics of our datasets, we also suggest the use sequential pattern analysis algorithms for finding association rules.

References

1. AlphaMiner. Available from: `http://www.eti.hku.hk/alphaminer/` [cited 2005-11-28].
2. The R Project for Statistical Computing. Available from: `http://www.r-project.org` [cited 2005-11-28].
3. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available from: `http://www.cs.waikato.ac.nz/ml/weka/` [cited 2005-11-28].
4. zerozero.pt :: Porque todos os jogos começam assim... Available from: `http://www.zerozero.pt` [cited 2005-11-28].
5. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules, 12–15 1994.
6. Inderpal S. Bhandari, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, and Krishnakumar Ramanujam. Advanced scout: Data mining and knowledge discovery in NBA data, 1997.
7. The CRISP-DM consortium. CRISP-DM 1.0 - Step-by-step data mining guide, 2000. Available from: `http://www.crisp-dm.org/CRISPWP-0800.pdf` [cited 2005-11-28].
8. FIFA. FIFA Survey: approximately 250 million footballers worldwide, 2000. Available from: `http://www.fifa.com/fifa/survey_E.html` [cited 2005-11-28].
9. J. Ross Quinlan. C4.5: programs for machine learning, 1993.
10. Lev Stankevich, Sergey Serebryakov, and Anton Ivanov. Data Mining Techniques for RoboCup Soccer Agents. In *AIS-ADM*, pages 289–301, 2005.
11. Edward R. Tufte. The Display of Quantitative Information, 1983.
12. Ian H. Witten and Eibe Frank. Data Mining: pratical machine leaning tools and techniques with Java implementations, 2000.

A expansão de conjuntos de co-hipónimos a partir de colecções de grandes dimensões de texto em Português

Luís Sarmento

Faculdade de Engenharia da Universidade do Porto
las@fe.up.pt

Resumo. Neste artigo iremos apresentar dois métodos para a expansão de conjuntos de co-hipónimos usando exclusivamente informação extraída a partir de uma colecção de texto em português de grandes dimensões. Os métodos baseiam-se na hipótese de que é possível explorar com sucesso a enorme redundância de informação existente em tais colecções recorrendo a algoritmos relativamente simples. Estes métodos operam de uma forma análoga ao conhecido sistema Google Sets, e num dos casos são alcançados tempos de execução muito reduzidos. Iremos enquandrar os dois métodos desenvolvidos numa estratégia mais ampla de construção de recursos léxico-semânticos para a língua portuguêsa e iremos posicioná-los relativamente a trabalhos realizados para outras línguas. Serão apresentados detalhadamente os algoritmos desenvolvidos, e para cada um deles serão apresentados e discutidos os resultados experimentais, comparando as suas limitações e vantagens. Abordaremos em seguida algumas questões relativas à avaliação deste género de métodos e destacaremos a necessidade de desenvolver recursos para esse efeito. Serão em seguida discutidas algumas limitações que derivam da indeterminação associada co-hiponímia e alguns dos problemas intrínsecos às abordagens que apresentamos. Terminaremos apresentando possibilidades de trabalho futuro.

1 Introdução

Em linguística é comum distinguir dois tipos de relações entre as palavras: relações paradigmáticas e relações sintagmáticas. Segundo [6] as relações paradigmáticas estabelecem-se entre palavras que partilham um determinado paradigma, seja este semântico, gramatical, morfológico ou outro, possuindo assim um série de características em comum que as permitem agrupar num determinado conjunto. Por exemplo, poderemos dizer que os pares (vermelho, azul), (cor, vermelho), (bom, mau), (ouvir, falar) agrupam palavras relacionadas por um derminado tipo de relação paradigmática. Por outro lado, as relações sintagmáticas são aquelas que se estabelecem entre palavras que co-ocorrem ao nível de um sintagma, podendo não existir qualquer tipo de relação paradigmática entre elas. Por exemplo "sentar" – "cadeira" ou "viagem – avião" estão relacionados sintagmáticamente, visto que co-ocorrem normalmente no interior de um determinado sintagma (ex: "sentar numa cadeira" ou "viagem de avião"). Note-se que em ambos os tipos de relações existe

17

obviamente uma forte ligação semântica entre as palavras, sendo que por isso nem sempre é simples destinguir entre os dois tipos de relações.

Considerando apenas as relações paradigmáticas do tipo semânticas, podemos enumerar as relações de sinonímia, de antonímia (ou mais genericamente de contraste), de hiponímia e de meronímia. Estas são relações habitualmente empregues na construção de rede léxicais do tipo WordNet, pelo que, dada a inexistência de um recurso deste género para o Português, e sendos tais recursos extremamente difíceis e custosos de produzir manualmente, torna-se particularmente interessante estudar métodos (semi-)automáticos para a identificação e compilação deste tipo de relações.

De todas as relações paradigmáticas semânticas anteriormente enumeradas, a relação de hiponímia reveste-se de particular interesse pois é a base da organização geral de um recurso como o WordNet. A hiponímia é a relação estabelecida entre um elemento e a classe mais geral onde esse elemento se inclui. Por exemplo "vermelho" é hipónimo de "cor", ou "cão" é hipónimo de "mamífero". A hiponímia possui uma uma relação inversa denominada hiperonímia, pelo que tudo o que for enunciado para a hiponímia aplica-se em sentido inverso à hiperonímia. Podemos então afirmar que "cor" é hiperónimo de "vermelho" e "mamífero" é hiperónimo de "cão". A relação de hiponímia (e logo a de hiperonímia) apresenta a propriedade transitiva, isto é, se "cão" é hipónimo de "mamifero" e "mamífero" é hipónimo de "animal", então "cão" é hipónimo de "animal".

Da relação de hiperonímia deriva-se a relação de co-hiponímia, ou seja a relação que se estabelece entre as palavras que possuem um hiperónimo comum. Assim, o conjunto (vermelho, azul, amarelo) está ligado por co-hiponímia, já que todos os seus elementos possuem (pelo menos) um hiperónimo comum ("cores primárias"). De uma forma aproximada, pode-se afirmar que a co-hiponímia une elementos "semelhantes" segundo um determinado critério, por vezes implícito ou até desconhecido à partida.

No resto do artigo, começaremos por justificar o interesse no desenvolvimento de mecanismos automáticos para a pesquisa de co-hipónimos, nomeadamente no seu interesse para a posterior pesquisa de relações de hiperonímia/hiponímia e também para acelerar processos de enriquecimento semi-automático de recursos léxico-semânticos. De seguida, apresentaremos alguns trabalhos relacionados, com os quais tentaremos comparar a nossa aproximação. Serão em seguida apresentadas 2 técnicas que permitem o alargamento de conjuntos de co-hipónimos, partindo de um conjunto reduzido de elementos semente, e analisados os seus resultados.

2 Motivação

A pesquisa e o alargamento automático de conjuntos de co-hipónimos possui um grande interesse prático, nomeadamente como forma de auxiliar a construção de recursos léxico-semânticos tais como por exemplo o WordNet. A obtenção automática de co-hipónimos poderá acelerar bastante o processo de adição de novas palavras, ou novas ligações ao recurso, já que permitirá, a partir de um conjunto de palavras conhecidas e já incluídas no recurso (ex: frutos como "morango", "banana", "maçã"), obter novos candidatos que partilhem directa ou transitivamente as mesmas relações de hiponímia (ex: "laranja", "cereja", "pêssego", "ananás", "melão", etc.).

Apesar de esta aproximação exigir sempre a validação manual, todo o processo de compilação de novos elementos e de descoberta de relações beneficia de uma aceleração considerável.

Em segundo lugar, e tal como demonstrado em [7] a obtenção de grupos de co-hipónimos potencia a obtenção dos respectivos hiperónimos (podem ser obtidos vários hiperónimos relacionados). É sabido que a determinação de relações de hiperonímia em corpora pode ser feita com algum sucesso para elementos frequentes usando certos padrões léxico-sintácticos tais como por exemplo "(umluma) [HIPERÓNIMO] como (alo) [HIPÓNIMO]" [4]. Contudo, para elementos (neste caso os hipónimos) mais raros a probabilidade de ocorrência dos referidos padrões reduz-se substancialmente, sendo por isso muito dificil encontrar evidência significativa da relação hiperonímia-hiponímia em causa. A obtenção alternativa de conjuntos de co-hipónimos permite aliviar este problema, quer por propagação automática das relações de hiperonímia já conhecidas ao restantes elementos do conjuntos de co-hipónimos, quer por fusão das evidências de hiperonímia recolhidas individualmente para cada um dos elementos do conjunto de co-hipónimos.

3 Trabalho Relacionado

Este trabalho foi inicialmente inspirado num sistema experimental fornecido pelo motor de pesquisa Google denominado Google Sets e que se encontra disponível em http://labs.google.com/sets. Este sistema recebe como parâmetros de entrada um conjunto de elementos fornecidos pelo utilizador e tenta expandir esse conjunto até a um máximo de 15 elementos "semelhantes". Apesar de não existir muita informação disponível acerca do modo de funcionamento concreto do Google Sets - nem uma avaliação do seu desempenho - este parece apoiar-se essencialmente na grande quantidade de dados que o Google tem nas suas bases de dados, sobre os quais aparenta utilizar alguma técnica de agrupamento ou de processamento "data-driven" para a obtenção dos elementos semelhantes. O sistema parece ter sido preparado apenas para lidar com inglês, não respondendo actualmente a pedidos de expansão de conjuntos em português. Por inspecção visual, o funcionamento deste sistema parece ser muito positivo, embora com a limitação do tamanho máximo do conjunto resultado a 15 elementos, o que talvez possa ser considerado uma indicação indirecta dos limites de precisão da aproximação adoptada.

Em [1] é descrita uma técnica que através de agrupamentos sucessivos das co-occorências entre palavras se consegue obter evidência de relações sintagmáticas e paradigmáticas em alemão. Os autores mostram como, através de agrupamentos sucessivos das co-ocorrências calculadas sobre um corpus de mais de 100 milhões de palavras, se consegue separar as relações sintagmáticas das paradigmáticas, e como numa segunda se torna fase possível discriminar as relações de hiperonímia e de co-hiponímia. A avaliação dos resultados foi realizada comparando os resultados obtidos com informação acerca das relações presente no GermaNet e com pares de palavras relacionadas compilados manualmente.

Em [8] foi calculada a matriz de co-ocorrências com uma janela de comprimento 2 para 1 milhão de palavras do British National Corpus, removendo palavras-função e

ignorando a anotação morfo-sintáctica. Em seguida usou-se esta informação para, dada uma palavra "semente" e aplicando a medida de distância vectorial "city-block", obter o conjunto de palavras semelhantes. Desta forma, o autor foi capaz de obter palavras relacionada paradigmaticamente, já que a pesquisa efectuada permite encontrar palavras que possuam perfís de co-ocorrência semelhantes, o que segue indirectamente a definição de relação paradigmática. Os resultados obtidos foram comparando com o desempenho humano na tarefa de selecção de sinónimos do TOEFL, tendo sido reportados resultados equiparáveis. Na nossa opinião a avaliação realizada pelo autor é pouco rigorosa não permitindo tirar conclusões acerca do desempenho efectivo deste método. Para além disso, esta aproximação não permite distinguir qual o tipo de relação semântica efectivamente obtida entre as palavras.

Um aproximação mais sofisiticada encontra-se descrita em [5] onde através da utilização de um *parser* se construem tripletos de relações gramaticais da forma (palavra1, relação, palavra2). Foram processados vários corpora de texto jornalístico em língua inglesa totalizando cerca de 100 milhões de palavras dos quais foi possível gerar 56.5 milhões de tripletos. Usando uma medida derivada da Informação Mútua sobre a lista de tripletos, foi computada a semelhança entre todos os pares de nomes, verbos e adjectivos/adverbos, para obter um *tesaurus* composto por vários conjuntos de semelhantes (ou em alguns casos sinómimos). Os resultados obtidos foram comparados com a informação contida no WordNet e o *tesaurus* Roget. Os autores concluiram que esta aproximação permite construir *tesaurus* em larga escala, embora não permita diferenciar entre os diferentes sentidos das palavras.

Outro trabalho relevante, embora com uma orientação ligeiramente diferente do nosso objectivo, vem descrito em [10] onde se descreve um sistema que recorre à pesquisa na Web através do motor de pesquisa Altavista para obter sinónimos. O sistema utiliza uma bateria de pesquisas na Web e pondera os resultados obtidos usando a medida PMI-IR (adaptada da medida Pontwise Mutual Information) para obter um medida de "sinonímia" entre palavras. Esta aproximação é extremamente interessante já que compensa os problemas da escassez de dados que normalmente afectam a performance de certos métodos quando lidam com palavras raras, utilizando a imensidão da Web. O autor reporta que os resultados deste sistema na tarefa de identificação de sinónimos do TOEFL foram superiores à média obtidos por humanos.

4 Descoberta de Co-hipónimos Usando o BACO

O objectivo do nosso trabalho consiste na exploração de métodos para expansão de um conjunto de co-hipónimos, que se baseiem apenas em informação inferida a partir de uma colecção de texto de grandes dimensões. Por outras palavras, pretendemos que, a partir de um conjunto de elementos fornecidos como entrada, o sistema seja capaz de encontrar numa colecção de documentos de texto outros elementos que lhes sejam semelhantes. Por exemplo, para um conjunto de entrada como (vermelho, verde, azul) espera-se ser capaz de encontrar várias outras palavras que se refiram a nomes de cores. Com este objectivo experimentamos dois métodos diferentes para expandirmos listas de co-hipónimos, cada um dos quais explorando uma propriedade

heurística distinta associada à co-hiponímia. Iremos em seguida descrever cada um dos métodos.

4.1 Pesquisa Usando Contextos Léxicais de 3 Palavras

O primeiro método consiste na observação de que os co-hipónimos deverão ocorrer em contextos léxicais muito semelhantes, já que pela própria definição de co-hiponímia possuem um hipónimo comum do qual herdam a maioria das propriedades e por isso também as ligações léxicais que podem estabelecer com outros elementos. Por exemplo, se pesquisarmos num corpus de grandes dimensões contextos onde ocorre a palavra "morango" poderemos quase certamente encontrar algo como "[com compota de] morango" ou "[um batido de] morango". Inversamente, se no mesmo corpus pesquisarmos quais as palavras que ocorrem nos contextos léxicais "[com compota de] X", ou "[um batido de] X" parece natural que, se estes contextos léxicais ocorrerem, a palavra X terá grande probabilidade de se referir a um fruto, ou eventualmente a outro alimento.

Seguindo esta ideia decidimos implementar uma pesquisa sobre a base de texto BACO [9] que permite efectuar pesquisas rápidas sobre contextos léxicais. O BACO (BAse de Co-Ocorrência) é uma base de dados gerada a partir de colecção web WPT03 (ver http://poloxldb.linguateca.pt/) e possui informação relativa a mil milhões de palavras provenientes de textos da web portuguesa. Esta informação encontra-se armazenada em várias tabelas diferentes, cada uma optimizada para um determinado tipo de pesquisa. Para além de permitir a pesquisa ao nível da frase, o BACO possui também tabelas que armazenam sequências de 2, 3 e 4 palavras, vulgarmente conhecidas com n-gramas. As tabelas de n-gramas são particularmente apropriadas para a pesquisa dos contextos léxicais onde ocorre uma dada palavra (neste caso contextos até 3 palavras), ou, inversamente, para a pesquisa de palavras que ocorrem num determinado contexto.

No sentido de máximizar a informação de contexto, decidimos realizar as pesquisa sobre a tabela de 4-gramas, permitindo assim a pesquisa de contextos com comprimento de 3 palavras. A tabela de 4-gramas do BACO possui o seguinte esquema: wpt_4_gramas(p_1, p_2, p_3, p_4, f, d). A informação armazenada nesta tabela é uma sequência de quatro palavras (p_1, p_2, p_3 e p_4), o número de vezes que essa sequência foi encontrada (f), e o número de documentos (d) da colecção WPT03 nos quais a referida sequência ocorre. Durante toda a experimentação, e para acelerar o processo de testes, foi utilizado um sub-conjunto da tabela de 4-gramas que se refere apenas de 370 mil documentos, cerca de 25% do total dos documentos disponíveis no BACO e contendo 92.835.207 tuplos. A tabela 4-gramas completa possui cerca de 273 milhões de tuplos, mas não foi usada nestas experiências por razões de desempenho e de espaço em disco.

Muito suncintamente, o algoritmo para a expansão de hipónimos consiste numa primeira fase, em pesquisar a tabela de 4-gramas para encontrar os contextos léxicais nos quais ocorrem os exemplos de entradas, e, numa segunda fase, é realizada uma pesquisa complementar, sendo desta vez pesquisadas as palavras que ocorrem nos referidos contextos. O contexto léxical considerado nestas experiências é constituido pelas 3 palavras anteriores à palavra fornecida como semente. Este não o único

contexto possível havendo pelo menos mais três. Contúdo se entrarmos em consideração que a pesquisa de co-hipónimos se refere à pesquisa de nomes e que, tal como demonstrado nos exemplos anteriores, procuramos especialmente ocorrências em que o nome ocorre como modificador no interior de um sitagma nominal, então a escolha das três palavras anteriores parece adequar-se melhor à pesquisa dessas estruturas. Apesar disso, não temos dados concrectos que nos permitam afirmar isto peremptoriamente e a investigação desta questão será alvo de futuro trabalho.

Foram considerados relevantes apenas os contextos com os quais ocorrem um determinado número mínimo de elementos do conjunto inicial, para que os referidos contextos sejam suficientemente correlacionados com o conjunto exemplo. Por outro lado, um contexto léxical que co-ocorra com demasiadas palavras (para além do conjunto inicial) é considerado demasiado genérico pelo que também é invalidado. De todos os candidatos a co-hipónimos assim obtidos removem-se aqueles que são artigos, preposições, e outras palavras ou siglas muito frequentes e que não podem ser à partida considerados co-hipónimos válidos. De notar que são apenas pesquisados candidatos com um única palavra, o que é certamente uma limitação desta implementação. No entanto, se considerarmos que um recurso como o WordNet é constituido maioritariamente por palavras simples, podemos ainda assim considerar que se trata de uma ajuda para a construção de recursos semelhantes.

O algoritmos em pseudo-código será o seguinte:

0: Incialização
  seja S ={s_1,s_2,... s_n} o conjunto de co-hipónimos semente (pelo menos 1)
  seja L o número minimo de co-hipónimos que um dado contexto têm de cobrir
  seja M o número máximo de palavras com que um contexto válido pode co-ocorrer
  seja P o conjunto de palavras proíbidas (artigos, preposições, etc.)
1: para cada s_i elemento de S
    pesquisar na tabela 4-gramas o contexto c_k = (p_1,p_2,p_3) para os quais p_4=s_i
    recolher c_k no conjunto de contextos C, incrementando a sua contagem
2: para cada ck do conjunto C cuja contagem seja igual superior a L
    pesquisar na tabelas de 4-gramas lista de [p_4] para os quais (p_1,p_2,p_3) = c_k
    se tamanho da lista [p_4] < M então contexto é válido
      adicionar cada elementos da lista p_4 ao conjunto resultado R, incrementando a sua contagem
3: Retornar R, o conjunto de candidatos a co-hipónimo. Ordenar por contagem excluindo os elementos pertencentes aos conjuntos P e S.

Na seguinte tabela, apresentam-se os resultados da execução deste algoritmo para várias configurações de conjuntos "semente" iniciais. O valor L indica o número de elementos do conjunto inicial com os quais um determinado contexto léxical tem de co-ocorrer para ser considerado válido e o valor #C indica o número de contextos léxicais distintos encontrados que verificam tal condição. Os valores entre parêntesis indicam o número de contextos léxicais válidos que existem em comum entre o elemento proposto e os elementos do conjunto inicial (as reticências entre os candidatos apresentados indicam séries de candidatos considerados correctos mas omitidos por questões de brevidade). Os elementos em itálico indicam ocorrências erradas.

#	Conjunto inicial	L	#C	Resultado
1	amarelo, vermelho, azul	3	44	verde (26), branco (22), preto (19), cinza (14), castanho (14), ... violeta (6), prata (5), *escuro* (4), dourado (4), *fe* (4), ... *pele* (4), *cores* (3), ... *liso* (2), *carvalho* (2), marrom (2), ... *terra* (2), *iluminado* (2), *54* (2), *brasil* (2), *pobre* (2)
2	granito, mármore, basalto	3	3	*betão* (2), *vidro* (2), *papel* (2), *pedra* (2), *madeira* (2), *material* (2)
3	whiskey, rum, gin	2	6	vodka (4), vinho (3), tequila (3), porto (3), cerveja (2), licor (2), sumo (2), coca-cola (2), verdelho (2), whiskie (2), tinto (2), aguardente (2), conhaque (2), *jack* (2), *neoplast* (2), uisque (2), água (2), *coca* (2), *plástico* (2), champanhe (2), cachaça (2), champagne (2)
4	porto, braga, aveiro	3	210	coimbra (141), lisboa (137), *vila* (126), *castelo* (115), leiria (110), viseu (110),... almada (51), guimarães (49),... *cidade* (5) ... régua (2), *avaliação* (2), *recrutamento* (2), *municípios* (2), *editorial* (2), gorazde (2), *gás* (2), *coliseu* (2), alvor (2), inhambane (2)

Tabela 1. Resultados usando o método da pesquisa por contextos léxicais de 3 palavras

4.2 Breve Análise de Resultados

Numa breve análise a estes resultados, ressaltam as seguintes observações:

1. Há uma grande variância entre o número de contextos válidos encontrados para cada conjunto inicial e consequentemente entre o número de candidatos a co-hipónimos retornados. Nitidamente, para certos conjuntos os dados são mais esparsos. No caso particular do conjunto (graníto, mármore, basalto) foram encontrados apenas 3 contextos válidos em contraste com o conjunto (porto, braga, aveiro) onde foram encontrados 210 contextos válidos.
2. Na maior parte dos casos analisados, os elementos do topo da lista retornada são de facto alguns do co-hipónimos esperados. Contúdo, em certos casos verifica-se a presença de co-hipónimos mais distantes e que por isso não correspondem ao esperado, como foi o caso do conjunto 2 para o qual, como já mencionado, não foi aparentemente possível encontrar contextos léxicais válidos em número suficiente. Reduzindo o parâmetro L para 2, isto é reduzindo a exigência relativamente à especificidade do contextos léxicais válidos, o resultado torna-se extremamente ruidoso.
3. Em alguns casos é possível encontrar na lista de candidatos não só co-hipónimos, mas também os hiperónimos (“cores”, ”material”, “cidade”).
4. Dada a limitação deste método em apenas permitir a pesquisa de uma palavra, alguns candidatos apresentados são de facto apenas parte de um eventual candidato

multi-palavra. Por exemplo, nos conjunto 3 o candidato "jack" provavelmente se refere a "jack daniels", assim como "coca" se deverá referir a "coca cola".

5. Este método apresenta algum sensibilidade relativamente a certos problemas de ambiguidade muito típicos da língua portuguesa. Por exemplo, nos resultado da lista (whiskey, rum, gin) encontramos "plástico" o que parece um resultados perfeitamente estemporâneo. Após inspecção mais detalhadas verificamos que se a causa deste resultado advém de ter sido considerado válido o padrão "uma garrafa de X" que é instânciavel não só para bebidas – ex: "uma garrafa de tequila" - como também para materiais – ex: "uma garrafa de plástico".

Como observações mais gerais podemos ainda afirmar que este método é relativamente elegante do ponto de vista algorítmico pois a sua implementação decorre directamente da definição de co-hipónimos adoptada. Apesar disso, apresenta graves problemas de eficiência e não é facilmente escalável. Os exemplos que apresentamos demoraram entre 10 e 30 segundos para o caso de conjuntos iniciais contendo elementos pouco frequentes, como é o caso do conjunto 2 e 3 ou mais de 5 minutos para conjuntos com elementos muito frequentes, como o conjunto 4. Finalmente, o algoritmo não apresenta nenhum mecanismo de controlo automático relativamente ao parâmetro L, que tem de ser por enquanto controlado manualmente.

4.3 Pesquisa por Co-ocorrência de Co-hipónimos em Coordenação

O segundo método desenvolvido baseia-se na observação simples de que os co-hipónimos são muitas vezes referidos no contexto de uma coordenação. Por exemplo, "as minhas cores preferidas são o vermelho e o azul" ou "os materiais usados foram mármore, ferro e granito". Estas coordenações, que aparentam ser frequentes, sugerem ser possível extrair informação útil acerca da co-hiponímia. Foram por isso selecionados manualmente para uma janela de 4 palavras 12 padrões que estivessem tipicamente associados a coordenações. Na seguinte tabela apresentam-se os padrões selecionados, sendo que X e Y indica a posição dos supostos co-hipónimos.

Formula	P1 X P3 Y	X P2 Y P4	X P2 P3 Y
Padrões	, X e Y , X ou Y , X , Y	X , Y , X , Y e X, Y ou	X e o Y / X e a Y X ou o Y / X ou a Y X , o Y / X , a Y

Tabela 2. Padrões de 4 átomos associados à co-hiponímia.

Refira-se que esta lista não é de forma alguma exaustiva e a validação destes padrões não passou de algumas experiências realizadas sobre a base de dados. A pesquisa directa de tais padrões sobre a tabelas de 4-gramas revelou-se muito pouco eficiente já que muitos dos átomos associados aos padrões de coordenação (i.e a virgula, o "e", o "ou", etc) são muito frequentes, não permitindo por isso tirar grande vantagem dos índices de pesquisa associados às tabelas. Por esse motivo resolveu-se criar uma tabela auxiliar contendo apenas uma sub-selecção dos tuplos da tabela de 4-gramas que correspondem aos padrões apresentados em cima. A seguinte tabela

indica o número de tuplos que foi possível recolher para cada um dos referidos padrões:

Padrão	Tuplos recolhidos
, X e Y	179415
, X ou Y	25.203
, X , Y	399.013
X , Y ,	428.746
X , Y e	202.619
X, Y ou	28.941
X e o Y	112.746
X e a Y	153.477
X ou o Y	6.824
X ou a Y	13.083
X , o Y	207.068
X , a Y	271.152
Total	**2.028.287**

Tabela 3. Número de tuplos recolhidos para cada um dos padrões utilizados

Através desta tabela, a pesquisa de co-hipónimos de um derminado elemento torna-se na simples tarefa de pesquisar elementos nas posições X e Y com os quais o elemento base co-ocorra na posição complementar. Ao contrário do método anterior, este método não possui nenhum parâmetro de qualidade que permita filtrar elementos ruidosos durante o processo de pesquisa. Na verdade, o único critério que foi considerado para tentar julgar a qualidade dos candidatos é o número de padrões com os quais o candidato e os elementos do conjunto inicial co-ocorrem. Desta forma, associado a cada candidato é possível associar um valor que pode variar entre 1 e 24 por cada elemento exemplo fornecido. O valor 1 ocorre quando o candidato a co-hipónimo co-ocorre apenas uma vez com um padrão numa determinada posição (X ou Y), e o valor 24 ocorre quando o candidato co-ocorre nos 12 padrões em ambas as posições (X e Y).

Para permitir uma comparação entre os dois métodos, foram repetidas as experiências com os mesmo conjuntos iniciais. A próxima tabela apresenta resumidamente os resultados obtidos (as reticências entre os candidatos apresentados indicam séries de candidatos considerados correctos mas omitidos por questões de brevidade).

#	Conjunto inicial	Resultado
1	amarelo, vermelho, azul	verde (48), preto (39), branco (38), laranja (28), rosa (23), cinza (18), castanho (18), violeta (13), cinzento (11), negro (11), lilás (11), *cor* (11), … cores (6), … transparente (4), azulão (4), *champanhe* (4), *sol* (4), *céu* (3), castanha (3), mediterrâneo (3), alaranjado (3), *camisa* (3), *claro* (3), púrpura (3), âmbar (3)…
2	Granito, mármore, basalto	*madeira* (9), pedra (8), calcário (7), *bronze* (7), *cimento* (6), *vidro* (5), xisto (5), *cantaria* (4), *tijoleira* (4), ardósia (3), arenito (3), barro (3), gesso (3), calcários (3), travertino (3), *tabaco* (2), *ouro* (2), *cellano* (2), (2), quartzo (2),…

3	whiskey, rum, gin	Vodka (8), conhaque (3), tequila (3), rumpi (2), *tabaco* (2), *limão* (2), calvados (2), *creme* (2), bourbon (2), *curaçau* (2), *anseios* (2), *açúcar* (2), brandy (2),…
4	Porto, braga, aveiro	lisboa (31), coimbra (28), leiria (24), gaia (23), viseu (22), setúbal (22), évora (21), guimarães (21), guarda (19), *minho* (19),… *algarve* (16),… *madeira* (14), …*portugal* (13), … cidade (13), … *sporting* (12), *benfica* (12) … (várias centenas de candidatos)

Table 4. Resultados do método de pesquisa por contextos de coordenações

4.4 Breve Análise de Resultados

A primeira observação que resulta da aplicação deste método é a de que este gera muito mais candidatos que o anterior. De facto, para todos os conjuntos exemplo foi possível obter várias dezenas de candidatos.

Mais uma vez também, os candidatos do topo da lista podem ser considerados correctos, havendo no entanto uma situação na qual isto não se verfica claramente, nomeadamente para o conjunto (granito, mármore, basalto). Neste caso, e tal como já verificado na experiência com o método anterior, os resultados apresentam alguns candidatos diferentes daqueles que se esperaria (tipos de rocha), o que, como iremos discutir em seguintes secções, advém de algumas ambiguidades intrínsecas à relação de co-hiponímia, e das limitações inerentes aos métodos apresentados no tratamento de dados esparsos.

Verifica-se também que grande parte dos candidatos que se podem considerar errados, são de facto palavras relacionadas sintagmaticamente (isto é pertencentes contexto comum). Este problema não era tão notório no método anterior. Por outro lado, pelo que nos foi dado a observar nestas experiências, este método é muito mais resistente a certos problemas da ambiguidade como os os verificados na situação "garrafa de gin" vs. "garrafa de plástico". Esta robustez é perfeitamente compreensivel já que ocorrência de coordenações entre "gin" e "plástico" parece absolutamente improvável pois são conceitos que ocorrem em dimensões quase ortogonais. Apesar da sua simplicidade, este método apresenta a enorme vantagem de ser capaz de eliminar estas ambiguidades.

Um ponto positivo deste método é a sua velocidade de execução. Pelo facto de se ter reduzido o espaço de pesquisa de cerca de 92 milhões de tuplos para a cerca 2 milhões de tuplos, a pesquisa de candidatos é executada em menos de 5 a 10 segundos, para as mesmas condições de hardware. Este desempenho permite utilizar este método como auxílio em tempo-real na construção de recursos lexico-semânticos. As grandes desvantagens deste método estão relacionadas com a inexistência de um parâmetro eficiente para controlo da qualidade dos candidatos obtidos, e com a dificuldade de filtrar as relações sintagmáticas.

5 O Problema da Avaliação dos Resultados

Antes de avançar com a discussão dos resultados, há um ponto que é importante abordar e que se prende com a avaliação dos resultados obtidos por estes dois métodos. Até agora temos vindo a apresentar resultados mas em nenhum dos casos fornecemos medidas concretas desempenho (como a Precisão ou a Abragência) calculadas relativamente a um padrão. De facto, a avaliação rigorosa dos resultados obtidos é problemática para o nosso caso, facto pelo qual os comentários acerca das qualidades ou defeitos dos métodos apresentados se baseiam na simples inspecção visual dos resultados.
O problema da avaliação de métodos de obtenção de palavras semelhantes a partir de texto foi já estudado anteriormente [3] tendo sido sugeridas duas técnicas:

1. A utilização de um padrão semântico, tal como um *tesaurus* ou outro recurso léxico-semântico já existente, para poder verificar se os métodos são capazes de "descobrir" correctamente relações comparando-as com as que já são conhecidas e que estão já codificadas explicitamente.
2. A utilização de um dicionário de definições lingua corrente a partir do qual é possível inferir se as relações descobertas são válidas. A ideia é para que dois conceitos / palavras relacionadas as suas definições no dicionário deverão apresentar um elevado grau de sobreposição.

Infelizmente, para o caso da lingua portuguesa não conhecemos recursos - como *tesauros* ou dicionários de definições - publicamente disponíveis para o efeito. Assim sendo restam-nos duas alternatives: (i) executar a avaliação manual dos resultados obtidos ou (ii) construir padrões especializados através da compilação e organização de informação de várias fontes. A primeira opção é sempre viável, mas não pode ser considerada uma metodologia satisfatória a longo prazo pois a subjectividade e flutuação temporal de critérios impede a correcta avaliação da evolução dos métodos e a sua comparação. A construção de um recurso destinado a permitir a avaliação destes métodos, apesar de poder parecer um trabalho excessivo parece também ser a única forma de conseguir realizar avaliações consistentes a médio e longo prazo. Como tal, e depois do desenvolvimento destas experiências, foi tomada a decisão de planear um tal recurso que sirva de padrão para a avaliação de métodos de pesquisa de co-hipónimos e que deverá ser utilizado em futuros desenvolvimento desta linha trabalho.

6 Discussão

Apesar de não termos efectuado uma avaliação efectiva dos métodos apresentados é útil fazer algumas observações. Em primeiro lugar, ambos os métodos foram capazes de expandir os conjuntos iniciais com co-hipónimos razoavelmente relevantes, em particular para conjuntos contendo elementos bastante frequentes. Como se pode verificar, os topos das listas expandidas são na maior parte dos casos co-hipónimos que se podem considerar válidos. Contudo, o mesmo já não ser verifica para

conjuntos iniciais cujos elementos não ocorram tão frequentemente, como foi o caso do conjunto (granito, mármore, basalto). A título ilustrativo, e considerando toda a colecção WPT03, a palavra "basalto" ocorre apenas 460 vezes em 339 documentos, enquanto que a palavra "amarelo" ocorre 18.186 em 12.336 documentos e a palavra "aveiro" ocorre 147.851, em 65.601 documentos. Estes valores reflectem a natural dificuldade em se lidar com dados esparsos, o que sugere que melhores desempenho obrigarão à eventual utilização de técnicas de suavização de dados.

Em segundo lugar, convém referir que apesar destas diferenças nos valores das frequências das palavras semente, ambos métodos fazem uso de medidas de qualidade (ou de limiares de filtragem) baseados no número de co-ocorrências distintas entre os candidatos e os contextos léxicais / padrões pesquisados, dependendo por isso da frequência apenas indirectamente. A grande vantagem de basearmos as nossas medidas de qualidade no número de co-ocorrência distintas, e não na frequência de cada uma dessas co-ocorrência, é podermos evitar parte dos problemas associados à duplicação de documentos que existe naturalmente em colecções web. A existência de documentos duplicados enviesa todos os parâmetros derivados do valor da frequência, pelo que a utilização deste parâmetro pode ser problemática.

Um ponto que ficou por explorar prende-se com a utilização de medidas de qualidade alternativas, também baseadas no número de co-ocorrências, mas que fossem capazes de ponderar a importância relativa de uma determinada co-ocorrência. De entre estas medidas, destaca-se a Informação Mútua [2] cuja aplicação poderia permitir a identificação dos contextos léxicais mais discriminantes, reduzindo o impacto dos contextos muito genericos que por co-ocorrerem com um elevado número de palavras também geram muitos candidatos inválidos.

Outra questão que não foi abordada, relaciona-se com o tamanho do contexto léxical utilizado neste métodos. Por razões de viabilidade, todas as pesquisas centram-se num contexto léxical de 3 palavras. Os resultados obtidos sugerem que este contexto parece ser suficientemente informativo para a tarefa a que nos proposemos, mas não foi possível recolher qualquer evidência de qual seria a evolução do desempenho dos métodos com o aumento (para 4 ou 5 palavras) ou com a redução (para 2 ou mesmo 1 palavra) do contexto léxical. E mesmo considerando manter o tamanho do contexto léxical em 3 palavras no caso do primeiro método, ficou também por explorar o impacto da escolha da posição da janela em torno dos candidatos: foram consideradas as 3 palavras anteriores ao candidato, mas o que aconteceria se fossem as 3 palavras posteriores? Ou se se considerassem as duas palavras anteriores e uma posterior?

Contúdo, a exploração sistemática destas questões teóricas exige, como já referido anteriormente, o desenvolvimento de técnicas e recursos objectivos para a avaliação de desempenho, pelo que essa deverá ser uma prioridade futura.

7 Indeterminação Instrínseca à Co-hiponímia

Em alguns dos resultados obtidos, ambos os métodos apresentaram resultados que fogem um pouco ao que seria de esperar: parte dos candidatos retornados não tinham aparentemente o nível de especialização compatível com o conjunto semente,

parecendo assim um pouco deslocados dos restantes co-hipónimos. Uma desta situações foi particularmente visível para o conjunto (granito, mármore, basalto) em que se verificou a presença de candidatos que nada tinham a ver com rochas, tal como intuitivamente se esperaria, nomeadamente "madeira", "vidro" ou "betão". Estes resultados estão relacionados com uma característica intrinseca da noção de co-hiponímia que é a multiplicidade de hiperónimos comuns que podem unir os referidos co-hipónimos. Neste caso, apesar de implicitamente termos assumido que o hiperónimo natural fosse "rocha", não é menos verdade que a partir dos mesmo exemplos também seria possível inferir como um hiperónimo válido "materiais de construção". Poderiamos eventualmente arranjar ainda outros hiperónimos comuns mais ou menos especializados, ou com um maior ou menor grau de sobreposição com os hipónimos apresentados, e que serviriam igualmente para agrupar estes 3 co-hipónimos. Note-se que os hiperónimos alternativos não estão relacionados entre sí apenas por especialização, já que a sua relação pode ser também de sobreposição parcial.

Estas características de indeterminação associadas à co-hiponímia dificultam a formulação do problema da pesquisa de co-hipónimos e, de certa forma, alastram o problema também à pesquisa dos vários hiperónimos possíveis. Evidentemente que o problema assim colocado torna-se muito difícil de resolver pelo que se torna necessário encontrar meios alternativos para conseguir refinar as noções de proximidade entre os co-hipónimos, ou para tentar identificar os contextos semânticos mais restritos em que eles se podem conjugar (ex: "natureza", "construção", "escultura", "cantaria", etc.). Como trabalho futuro, pretende-se vir a estudar formas de separar os vários contextos semânticos associados a um grupo de palavras, mesmo que não seja possível determinar com exactidão o que cada um desses contexto semânticos pode ser. Admite-se que sendo possível discriminar / separar contextos semânticos alternativos (quaisquer que estes sejam), será também possível melhorar as técnicas de pesquisa de co-hipónimos pela selecção apenas dos contextos semânticos interessantes, algo que os métodos apresentados são incapazes de fazer correctamente.

Conclusões

Neste artigo apresentamos dois métodos alternativos para a expansão de conjuntos de co-hipónimos. Verificamos qu e, com técnicas relativamente simples e fazendo uso das quantidades massivas de texto que agora estão disponíveis, é de facto possível obter resultados satisfatórios na pesquisa de elementos semelhantes aos fornecidos como exemplo ao sistema. Os dois métodos apresentam diferentes níveis de robustez relativamente a ambiguidades típicas do português e à possibilidade de ruído proveniente de palavras relacionadas sintagmaticamente com os elementos do conjunto inicial. Foram apontadas as actuais limitações associadas às possibilidades reais de avaliação deste género de sistemas e concluiu-se que é necessário desenvolver urgentemente recursos para esse fim. Foram também levantadas algumas questões relacionadas com a necessidade de utilização de técnicas de suavização de dados que poderão melhorar o desempenho dos métodos de pesquisa semelhantes aos

apresentados em situações onde os dados são mais esparsos. Foi também apontado interesse em explorar medidas de qualidade alternativas, como por exemplo a Informação Mútua, no sentido de ajudar a descriminar ligações mais relevantes entre os exemplos e os seus contextos léxicais. Reflectiu-se também sobre a necessidade de desenvolver métodos capazes de discriminar contextos semânticos, para lidar mais eficientemente com os problemas de indeterminação associados à própria noção de co-hiponímia.

Como conclusão final podemos dizer que, mesmo considerando todas limitações destes métodos, e em particular o facto de nesta implementação apenas se lidar com palavras simples, este métodos demonstraram ter potencial para serem utilizados na construção ou na expansão de recursos léxico-semânticos para a língua portuguesa.

Referências

1. Chris Biemann, Stefan Bordag, Uwe Quasthoff: Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurences. In: Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisboa, Portugal (25 May 2004).
2. Kenneth Church, William Gale, Patrick Hanks, Donald Hindle: Using statistics in lexical analysis. In: Uri Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. New Jersey: Lawrence Erlbaum (1991) pp. 115-164
3. Gregory Grafenstette: Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. Corpus processing for lexical acquisition. MIT Language, Speech and Communication Series. MIT Press Cambridge, MA, USA (1996) pp. 205 – 216.
4. Marti A. Hearst: Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics. Nantes, France (1992) pp. 539 - 545.
5. Dekang Lin: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL 1998, Montreal, Vol. 2 (1998) pp. 768–773.
6. M. Lynne Murphy: Semantic Relations and the Lexicon: antonymy, synonymy and other paradigms. University Press, Cambridge (2003)
7. Patrick Pantel and Deepak Ravichandran: Automatically Labeling Semantic Classes. The Proceedings of HLT-NAACL. Boston, MA (2004) pp. 321-328
8. Reinhard Rapp: The computation of word associations: comparing syntagmatic and paradigmatic approaches. In: Proceedings of the 19th international conference on Computational linguistics. Taipei, Taiwan (2002)
9. Luís Sarmento and Luís Cabral: BACO – A large database of text and co-occurrences. In preparation (2005).
10. Peter D. Turney: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning. Lecture Notes in Computer Science; Vol. 2167 (2001) pp. 491-502

Sessão Técnica 2

Técnicas de Negócio Electrónico e Sistemas Multi-Agente

Electronic Institution: an E-contracting Platform for Virtual Organizations

Henrique Lopes Cardoso

LIACC – NIAD&R, Faculty of Engineering, University of Porto
R. Dr. Roberto Frias, 4200-465 Porto, Portugal
hlc@fe.up.pt

Abstract. Automated tools that assist contract drafting are mostly focused on the representation of contract documents. Multi-agent systems have been applied in the e-business domain, namely for information discovery and contract negotiation. Work on contract monitoring and enforcement is less explored. In this paper we start from these two observations to expose our efforts towards the development of tools that enable the computational representation of contracts and furthermore their monitoring and enforcement. We are mostly interested in Virtual Organization settings, where groups of agents representing different business entities form consortiums that must be regulated by appropriate norms. We are pursuing the concept of an Electronic Institution as a platform providing a normative environment and a set of e-contracting related services. Within this environment, contracts are represented through norms.

1 Introduction

The representation of legal contracts in computer systems has been sought by the research community [5]. However, most attempts (e.g. [10]) have focused on the contract document rather than the contract agreement. That is, in many cases the approach is to represent, in computer systems, the structure and information of the signed document for human consumption.

Successful attempts towards a computable representation of contracts, allowing for automated tasks such as contract monitoring, are still missing. This is especially the case when considering complex contractual relationships such as those related with the formation of consortiums among different organizations.

Considering the whole e-contracting lifecycle (comprising the stages of information discovery, contract negotiation, and contract execution), automated tools have been developed for the first two stages. In fact, most currently available support to e-business is devoted to the first phase: we can find typical e-market functions such as yellow-page support, customer aggregation mechanisms and recommender systems.

Contract negotiation has been addressed in a number of research projects. In [20] an approach is presented concerning the formation of Virtual Organizations (VO): temporary consortiums of different organizations that "pool their resources to meet short-term objectives and exploit fast-changing market trends" [6]. However, the pro-

posed negotiation protocol is more concerned with the selection of the partners that will compose the VO than with the contractual agreement that is to be implemented.

The subject of virtual organizations/enterprises is gaining increasing importance in the B2B world, where players are becoming more focused on their core businesses and rely on outsourcing and dynamic consortiums. This can lead to complex relationships, in which partners' compliance must be assessed.

The Multi-Agent Systems (MAS) paradigm has been applied in the domain of e-business automation, namely in the three stages of e-contracting identified above. Agents are typically used as a means of encapsulating the individual interests of different business entities.

While agent theory describes agents as autonomous self-interested entities, preferably interacting in open environments, the application of MAS in real-world scenarios has risen a concern in the MAS research community: the need to regulate agent interactions. Two complementary lines of research have been developed since. The field of normative multi-agent systems explores the design of environments where interacting agents can be usefully regarded as governed by norms [14]. Agents are subject to these norms, which influence their decision making. Therefore, besides their goals, agents must take into account the norms that apply to them.

Another important research concept is that of an *Electronic Institution* (EI): a framework providing a regulated and trustable environment by enforcing norms and providing specific services [7, 17].

In our work we are applying these two research directions to the concern identified above: the representation of contracts formalizing VO relationships in a computable fashion, and the development of automated tools for contract monitoring and enforcement. We perceive contracts as being composed of norms that contractual agents are subject to. The EI framework provides a set of services that address the whole e-contracting life-cycle, including contract execution. The EI also provides a normative background that facilitates the establishment of contractual agreements.

The rest of the paper is organized as follows. Section 2 discusses some references in the literature regarding the computational handling of contracts. It also explores the representation of norms in contracts. Section 3 introduces the EI concept and provides information about our efforts in pursuing its development as a platform providing a normative environment and a set of e-contracting related services. Section 4 concludes.

2 E-contracting

Contracts are specifications for the behavior of a group of agents that jointly agree on a specific business activity. Contracts are used as a means of securing transactions between the involved parties, forming a normative structure that explicitly expresses their behaviors' interdependencies. *Electronic contracts* are virtual representations of such contracts. The aim of e-contracting is to improve the efficiency of contracting processes, supporting an increasing automation of both e-contract drafting and execution.

The components of a contract include the identification of participants, the specification of products and/or services included and a discrimination of actions to be performed by each participant. These actions are usually accompanied with time and precedence constraints. Typified business relations can recurrently use pre-formatted contract templates. In this case, contracts usually have a set of identified *roles* to be fulfilled by the parties involved in the relationship.

2.1 Representing E-contracts

Approaches towards a computational representation of contracts have been made. A normative conception of contracts is often used for contract representation. Formal models of norms rely on deontic logic, embracing the notions of obligation, permission and prohibition. Extensions to the original work on deontic logic have been made so as to allow its practical use, namely approaches to handle norm violations (such as the application of sanctions, also known as contrary-to-duties [13]), and considering the use of conditional and temporal aspects [8].

Languages for representing norms in contracts have been proposed. In [8] a logical formalism for describing interaction in an agent society, including social norms and contracts, is presented. Focusing on contract representation, it emphasizes on conditional obligations with deadlines. Other approaches to contract representation through norms include [21], considering normative statements that can comprise obligations, permissions and prohibitions. Sanctions are seen as obligations or prohibitions activated by the violation of another obligation. Also, [16] proposes the inclusion of obligations, permissions and sanctions in a contract specification language.

A common line in these approaches is the specification of deontic operators that are dependent on certain conditions and that have associated deadlines. For instance, [21] uses the notion of normative statement formally represented as:

$$\varphi \rightarrow \theta_{s,b}\,(\alpha < \psi)$$

where

φ is an activation condition
θ is a deontic operator (obligation, permission or prohibition)
s and b are, respectively, the subject and beneficiary of θ
α is the action to perform or the state of affairs to bring about
ψ is a deadline

An approach to contract representation based on event calculus can be found in [15]. This representation incorporates how a contract is to be fulfilled (that is, which events initiate or terminate obligations), making it a heavier structure.

In [4] the authors propose modeling contracts as processes, that is, state diagrams where transitions correspond to the execution of actions by the parties, considering also temporal elements.

Contracts can have different forms, ranging from simple contracts used to buy a product to complex contracts defining complex interactions between parties [2]. However, most of the research literature devoted to e-contract automation simplifies con-

tracts to the former type, defining one time relationships between a client and a supplier. After the delivery and payment phase, the parties are assumed to be no longer related. Little attention has been given to contracts that result from a Virtual Enterprise formation process. Nevertheless, as argued in [5], the construction of automated tools that deal with legal contracts is mostly helpful in complex contracting settings, such as long-term trading agreements and multi-party relationships (as is the case of a VO).

2.2 Monitoring and Enforcing Contracts

The execution of an e-contract consists on the parties following the norms they committed to when signing the contract. If any deviations from the prescribed behavior should occur, sanctions can be applied as specified in the contract or in its normative system of reference. However, the parties involved will typically not voluntarily submit themselves to such penalties. Therefore, appropriate mechanisms are needed to monitor and enforce norm execution. Only a trusted third party can enable the necessary level of confidence between the parties involved in a business relation.

The automation of contract monitoring and enforcement is challenged by the presence of complex legal issues and subjective judgments on agent compliance. Nevertheless, approaches have been made in some research projects.

The involvement of a third party in e-contract execution is generally claimed. In [16] a supervised interaction framework is proposed, where a trusted third party is included as part of any automated business transaction. Agents are organized in three-party relationships between two contracting individuals (a client and a supplier) and an authority that monitors the execution of contracts, verifying that errant behavior is either prevented or sanctioned. This authority enables the marketplace to evaluate participants, keeping reputation records on the basis of past business transactions.

In [21] a *contract fulfillment protocol (CFP)* is proposed, a collaborative protocol based on the normative statements' lifecycle. The idea is that, since contractual relationships are distributed, there is a need to synchronize the different views each agent has about the fulfillment of each contractual commitment. Agents communicate about their intentions on fulfilling contractual norms, allowing their contractual partners to know what to expect from them.

An alternative norm representation can be found in [22], where norms are defined as having explicitly associated violation conditions, means of detecting those violations, and sanction and repair measures. Therefore, in this case norms have a heavy structure, making the monitoring and enforcement process dependent on each individual norm.

We are unaware of an approach that considers the monitoring of contracts that represent a VO activity, taking into account the cooperation efforts that each partner is supposed to practice during the VO's lifetime. In the next section we introduce the Electronic Institution concept as a computational framework where such contracts can be created and monitored.

3 Electronic Institutions

Human societies are governed by institutions providing services or regulating the way citizens interact. The same approach has been proposed as a means to regulate the interaction among software agents. The *Electronic Institution* (EI) concept [7] represents the virtual counterpart of real-world institutions.

The benefit of an EI resides in its potential to assure legitimacy and security to its members, through the establishment of norms [7]. An EI provides an environment that regulates the relationships between software agents. Some approaches have considered such an environment as a constraining infrastructure [9], where the institution imposes the actions that agents may perform, thereby defining an interaction protocol that agents must follow. We do not follow such a restrictive scenario.

We consider that besides enforcing norms, institutional services should be provided to assist the coordination efforts between agents which, representing different real-world entities, interact with the aim of establishing business relationships. In our perspective, an EI is thus a comprehensive framework that provides a set of institutional services covering the formation and operation of VOs, while assuring norm enforcement through the imposition of sanctions and reputation mechanisms.

One of the main roles of the EI is to provide trust by working as a third-party that enables partners to engage in (automated) business interactions. The provided services compose a coordination framework that assists the interaction of software agents representing different organizations or business units.

We may summarize the main goals of an EI as follows: (1) to support agent interaction as a coordination framework, making the establishment of business agreements more efficient; and (2) to provide a level of trust by offering an enforceable normative environment. Therefore, our perspective regards an EI not as an end *per se*, but as a means to facilitate both the creation and the enforcement of contracts between agents.

3.1 Institutional Services

A number of agent-based institutional services are provided (see figure 1, where we omitted typical e-market facilities, such as registration and white/yellow page support).

Negotiation mediation services are provided to assist the formation of a VO. This includes the utilization of appropriate negotiation protocols (such as [20]) and contract templates, which are instantiated with the outcome of the negotiation process.

When addressing open environments with no centralized design, it may well be the case that agents representing different organizations use different domain ontologies. In order to enable a meaningful negotiation, ontology matching services must be put into place [19].

The mentioned services are used by different organizations, which can be seen as potential partners in a future VO. A subset of these, according to the outcome of a negotiation process, will become partners in a new VO.

Contracts resulting from successful negotiations are registered in the EI through a notary service, responsible for validating them according to institutional norms. The

execution stage is assisted by providing services that monitor the carrying out of contractual commitments by each VO partner. The VO contract defines cooperation efforts between the involved agents, and includes specific interactions during a certain time frame.

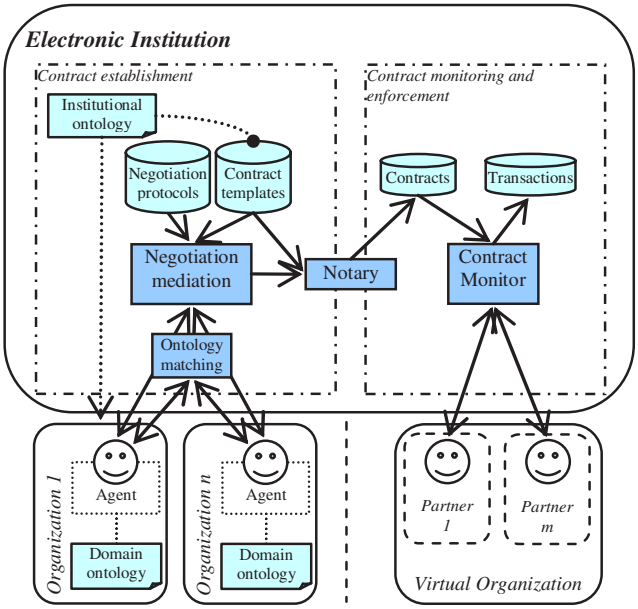


Fig. 1. Services in an Electronic Institution (adapted from [17])

Every agent intending to use an institutional service must be registered as a member. Agents have, inside the EI's boundaries, a record of reputation concerning their observance to past contractual relationships. This public information may be used by other agents, in the future, when choosing appropriate business partners. Agents' reputations may also be used, if not as a ruling out factor, at least when deciding the level of detail a contract should have.

3.2 Normative Environment

As mentioned before, one of the main aims of the EI is to provide a level of trust through an enforceable normative environment. As we are concerned with the possibility of commitment creation at run-time through the establishment of contracts, our environment has a flexible normative structure (unlike other EI formalizations such as [9]). Contractual norms are used to represent agents' commitments.

A norm-aware environment can operate either preventively (making unwanted behavior impossible) or reactively (detecting violations and reacting accordingly) [22]. In order to cope with the autonomous nature of agents, our approach considers norms as regulations that agents may or may not abide to.

Norms prescribe the expected behavior of agents, specifying states of affairs that *must* be brought about by an agent before a certain deadline. Therefore, we consider *obligations* as the means to express the prescription of behavior norms. Our basic norm definition is therefore based on the following EBNF description:

<Norm> ::= <Situation> "→" <Prescription>
<Situation> ::= {<Cond> "∧"} <Cond> {"∧" ¬<Cond>}
<Prescription> ::= {<Obligation> "∧"} <Obligation>
<Obligation> ::= obligation(<Agent>, <Fact>, <Deadline>)

Norms prescribe behavior by specifying what obligations come about when a specific situation is accomplished. The situation is characterized by conditions related with the state of a particular contractual relationship. The prescribed obligations indicate what facts an agent is supposed to bring about by a certain deadline. In the case of sanctions, the situation is characterized by the violation of another norm.

While being based on the notion of conditional obligations with deadlines, this representation shows how norms may be represented using a rule-based approach (see subsection 3.3).

Agents will not voluntarily submit themselves to associated penalties in case of deviation. Therefore, appropriate mechanisms are needed to enforce norm compliance. It is the EI's responsibility to maintain the normative state of the environment, taking into account the compliance or non-compliance of agents regarding their applicable norms. This is done through a contract monitoring and enforcement service. Contracts are monitored by employing rules that detect the fulfillment and violation of obligations, based on the occurrence of facts and on the passage of time. When agents fail to comply with their obligations they expose themselves to punishments, either direct (e.g. sanctions) or social (e.g. reputation records). Contracts are enforced by applying predicted sanctions in case of non-compliance, by affecting the agents' reputation and, ultimately, by preventing their access to institutional services.

A normative environment should be embodied with a set of norms applicable in the absence of further information. An important concept in contract law theory is the use of "default rules" [3], which exist with the intent of facilitating the formation of contracts, allowing them to be underspecified by defining default clauses or default values. The most useful case for this is in defining contrary-to-duty situations [13], which typically should be not likely to occur. For this reason, such situations are normally not dealt with in each contractual agreement, and agents usually recur to legislative systems that define default procedures [4]. Default regulations provide a normative background in which agents can rely to build their contractual commitments.

Taking into account our stated goal of providing assistance to VO formation, we developed a normative framework [18] that considers three hierarchical layers of norms: *institutional*, *constitutional* and *operational* (figure 2). While institutional norms may be applicable to all agents inside the EI, constitutional norms apply to agents taking part in a VO, and operational norms specify the operationalization of such organizations. Default norms may be defined for each of these layers.

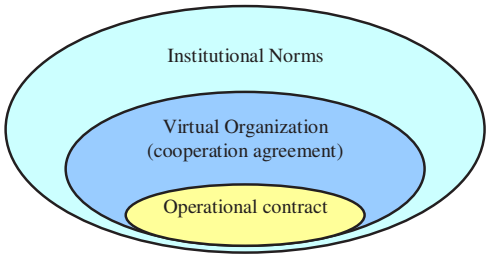


Fig. 2. Norms in an Electronic Institution (from 18)

While we have defined our basic norm representation, a definition of contractual norms that fits each of the layers of norms identified above is still at a preliminary stage.

3.3 Implementation

We are in the process of implementing a first EI prototype integrating the different services illustrated in figure 1. Our implementation is based on the Jade platform [12].

Regarding the normative environment, our norms obviously lend themselves to a rule-based representation. The monitoring of norms is implemented by appropriate rules that detect the fulfillment and violation of obligations, also allowing for the chaining of norms within a contractual relationship.

Since the normative environment is based on the occurrence of facts, the obvious solution towards its implementation is by using a forward-chaining production system. Therefore, we are pursuing the development of the normative environment (including the norm monitoring and enforcement services) using the Jess shell [11]. Jess is a rule engine that very efficiently applies rules to data. Our knowledge base consists of rules and norms. The working memory includes the facts that describe the normative state.

Jess has a number of features that allow us to implement our normative environment in an efficient and EI-integrated fashion. It includes the use of frame-based approaches and the possibility to organize norms in different modules, which is appropriate to manage the complexity of our normative framework. Jess also connects easily with Java, allowing us to define institutional procedures not amenable to a declarative representation. For instance, we may define a rule that triggers a notification procedure whenever a new obligation arises. The set of institutional rules and procedures implement the contract monitoring and enforcement service.

4 Conclusions

The agent technology roadmap [1], by AgentLink III, identifies as key problem areas the development of infrastructures for open agent communities, as well as the need for trust and reputation mechanisms. Electronic institutions address the needed infrastruc-

tures. Norms, electronic contracts and their enforcement are pointed out as means to achieve trust in open environments. Our work is motivated by the need to develop services that assist the coordination efforts between agents which, representing different real-world entities, interact with the aim of establishing virtual organizations. In order to be trustful, a VO needs to be regulated by appropriate norms.

The work already developed concerns the design of the EI platform in order to integrate different services, including ontology-based services [19], negotiation mediation [20], and contract monitoring. We also conceptualized a framework of norms that takes into account the need to regulate VO agreements.

A basic norm representation was defined; a definition of contractual norms that allow us to define VO agreements is still at a preliminary stage. As identified in [18], some characteristics of such complex settings that should be addressed are: the ongoing nature of VO relationships (as opposed to one-shot purchase operations); the existence of interactions that are continuously repeated in time; the support for and regulation of the exit and entrance of partners during the VO lifetime; and the handling of monetary transfers, such as profit distribution.

We intend to test the applicability of our approach through illustration with case-studies. Furthermore, the contract representation to develop shall be compared with other approaches.

References

1. AgentLink III (2004). *Agent Technology Roadmap: Overview and Consultation Report*. http://www.agentlink.org/roadmap/index.html
2. Angelov, S. & Grefen P. (2001). *B2B eContract Handling – A Survey of Projects, Papers and Standards*. University of Twente: CTIT Technical Reports.
3. Craswell, R. (2000). Contract Law: General Theories. In Bouckaert, B. & De Geest, G. (eds.), *Encyclopedia of Law and Economics, Volume III: The Regulation of Contracts*, Edward Elgar, Cheltenham, pp. 1-24.
4. Daskalopulu A. & Maibaum T. (2001). Towards Electronic Contract Performance. *Legal Information Systems Applications, 12th International Conference and Workshop on Database and Expert Systems Applications*, IEEE C. S. Press, pp. 771-777.
5. Daskalopulu A. & Sergot M. J. (1997). The Representation of Legal Contracts, *AI and Society*, 11 (1 & 2), pp. 6–17.
6. Davulcu, H., Kifer, M, Pokorny, L.R., Ramakrishnan, C.R., Ramakrishnan, I.V., & Dawson, S. (1999), Modelling and Analysis of Interactions in Virtual Enterprises, *Proceedings of the 9th International Workshop on Research Issues on Data Engineering: Information Technology for Virtual Enterprises (RIDE 1999)*, IEEE Computer Society, pp. 12-18.
7. Dignum, V., & Dignum, F. (2001). Modelling agent societies: co-ordination frameworks and institutions. In P. Brazdil & A. Jorge (eds.), *Progress in Artificial Intelligence: Knowledge Extraction, Multi-agent Systems, Logic Programming, and Constraint Solving*, LNAI 2258, Springer, pp. 191-204.
8. Dignum, V., Meyer, J.-J., Dignum, F. & Weigand, H. (2003). Formal Specification of Interaction in Agent Societies. In Hinchey, M., Rash, J., Truszkowski, W., Rouff, C. & Gordon-Spears, D. (eds.), *Formal Approaches to Agent-Based Systems*, Springer, pp. 37-52.

9. Esteva, M., Padget, J. & Sierra, C. (2002). Formalizing a language for institutions and norms. In Meyer, J.-J. & Tambe, M. (eds.), *Intelligent Agents VIII*, Springer, pp. 348-366.
10. Field, S., & Hoffner, Y. (2005). Dynamic Contract Generation for Dynamic Business Relationships. In G.D. Putnik & M.M. Cunha (eds.), *Virtual Enterprise Integration: Technological and Organizational Perspectives*, Idea Group Inc., pp. 207-228.
11. Friedman-Hill, E. (2003). *Jess in Action*. Manning Publications Co.
12. Java Agent DEvelopment Framework. http://jade.tilab.com
13. Jones, A. & Carmo, J. (2001). Deontic logic and contrary-to-duties. In Gabbay, D. (ed.), *Handbook of Philosophical Logic*, Kluwer, pp. 203-279.
14. Jones, A., & Sergot, M.J. (1993). On the Characterisation of Law and Computer Systems: The Normative Systems Perspective. In J.-J. Meyer & R.J. Wieringa (eds.), *Deontic Logic in Computer Science: Normative System Specification*, Chichester, England: John Wiley & Sons, pp. 275-307.
15. Knottenbelt, J., & Clark, K. (2005). Contract-related Agents. *Sixth International Workshop on Computational Logic in Multi-Agent Systems (CLIMA VI)*. London, England.
16. Kollingbaum, M. J., & Norman, T. J. (2002). Supervised Interaction – Creating a Web of Trust for Contracting Agents in Electronic Environments. In C. Castelfranchi & W. Johnson (eds.), *Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, New York: ACM Press, pp. 272-279.
17. Lopes Cardoso, H., Malucelli, A., Rocha, A.P. & Oliveira, E. (2005). Institutional Services for Dynamic Virtual Organizations. In Camarinha-Matos, L. M., Afsarmanesh, H. & Ortiz, A. (eds.), *Collaborative Networks and Their Breeding Environments – 6th IFIP Working Conference on Virtual Enterprises (PRO-VE'05)*, Springer, pp. 521-528.
18. Lopes Cardoso, H. & Oliveira, E. (2004). Virtual Enterprise Normative Framework within Electronic Institutions. In Gleizes, M.-P., Omicini, A. & Zambonelli, F. (eds.), *Engineering Societies in the Agents World V*, Springer, pp. 14-32.
19. Malucelli, A., Palzer, D. & Oliveira, E. (2005). Combining Ontologies and Agents to help in Solving the Heterogeneity Problem. In *Proceedings of the International Workshop on Data Engineering Issues in E-Commerce (DEEC2005)*, IEEE Computer Society, pp. 26-35.
20. Oliveira, E. & Rocha, A. P. (2000). Agents Advanced Features for Negotiation in Electronic Commerce and Virtual Organisations Formation Process. In Dignum, F. & Sierra, C. (eds.), *Agent Mediated Electronic Commerce: The European AgentLink Perspective*, Springer, pp. 78-97.
21. Sallé, M. (2002). Electronic Contract Framework for Contractual Agents. In R. Cohen & B. Spencer (eds.), *Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence*, Springer, pp. 349-353.
22. Vázquez-Salceda, J., Aldewereld, H. & Dignum, F. (2004). Implementing norms in multiagent systems. In Lindemann, G., Denzinger, J., Timm, I. J. & Unland, R. (eds.), *Multiagent System Technologies*, Springer, pp. 313-327.

Dissecting the Business Process Modelling fields: a concept maps approach

Célia Talma Martins[1,2]

[1] LIACC-NIAD&R, Faculty of Engineering, University of Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[2] ISCAP, Rua Jaime Lopes Amorim, s/n, 4465-S. Mamede de Infesta, Porto, Portugal
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
talma@fe.up.pt

Abstract. The multi-disciplinary nature of business process modelling with its different perspectives/views (business, organization, software and systems development,...) raised several objects/fields of study such as web services, languages, standards, business rules and architectures among others. In our approach we consider an environment (breeding environment) where a set of enterprises exist and maintain a set of social relationships, mainly trust relationships, that can be mobilised to join resources and collaborate to compete for a business opportunity. The definition of business processes, from a business point of view, implies a close articulation with the software systems components available to support parts of those processes. In this field of research there are many similar terms used in different contexts with different meanings, many different approaches whose intervention objects must be clearly identified as well as many descriptive and execution languages whose scope and goals must be well understood. The clarification of the business process modelling field is beneficial both for the business architects and for information systems architects. An approach based on the conceptualization domain, more specifically on the concept maps approach, is an effective way to achieve such a clarification. This work resulted in a conceptual map of the business process modelling field that enabled a critical analysis and the clarification of the relationships between business originitated concepts and software systems ones.

1 Introduction

The heterogeneous, complex and continuously changing field of business networking (e.g. B2B, collaborative networks, etc.) raises several issues related with the management of Inter-Organizational Business Processes that satisfy a particular consumer's need: the establishment of shared visions and goals, process and activities coordination, resource allocation and distribution, information systems and information technology inter-operability, are examples of important issues addressed both by

research and practice. The articulation of business activities distributed over a set of organizations is an important research topic that has been addressed by various disciplines, in particular management and computer science. In the several phases in which IOBP management can be decomposed (definition, configuration, execution, maintenance) the modelling activities are of utmost importance.

Business Process Modelling is a well established research and practice field (thought immersed in different research topics such as Enterprise Modelling or Information Systems Architectures, to name just two in opposite sides of the BPM spectrum), embraced in a first moment by the management and industrial communities and in a second moment by the computer science and information systems communities. Somewhere in between, we can identify the workflow management community. Business Process Modelling (BPM) is still an ongoing research topic. In fact, BPM is a research challenging issue specially focusing on the expression of interdependencies among business processes, information systems components and the emerging web technologies. The main objective of BPM is to provide a better understanding of how to express the business processes, their strategies and their behavior. Business models provide ways of expressing business processes or strategies in terms of business activities and collaborative behavior so we can better understand the business process and the participants in the process. Models are helpful for documenting, for comprehending complexity and for communicating complexity. Recently, BPM has gained a new breath pushed by the technological development in the area of internet/web technologies: web processes, service oriented architectures, semantic web. Although dealing with the same object of study - the organizational/business process - the terminology used by both communities can sometimes be confusing. This happens because of the use of the same terms referring to different concepts (different here is a continuum from "slightly" to "completely" different), or the use of different terms referring to the same concepts.

In this paper we undertake a conceptual analysis of the main fields dealing with BPM with the goal of clarifying conceptually the uses of BPM in the management and computer science fields. We intend to provide researchers and practitioners in these fields with a tool that helps them in understanding the BPM concepts and their relationships. Also an important goal (and the first aim of our work), is to set up a solid conceptual basis for interdisciplinary research in this area.

Section 2 will address briefly concept maps. Section 3 will provide a deep analysis of Business Process Modelling fields, describing the approach that we have followed to build the conceptual map, and presenting the obtained concept map as well as some concluding remarks. Section 4 enumerates the related work and Section 5 points out our future research and open issues.

2 Conceptual Mapping

Conceptual maps are an effective way of representing complex concepts and messages in a clear and understandable way. Conceptual maps are simple and practical

knowledge representation tools that allow the representation of knowledge in the form of a graph. The concepts are represented trough boxes (nodes) and the relations between them are represented by lines (arcs) connecting the related boxes [5]. Conceptual maps are structured in a hierarchical way, where the most general concepts lie in the root of the tree and, and as we descend the structure, we find the more specific ones. Concept maps have been demonstrated to be an effective means of representing and communicating knowledge.

Through a concept map we can identify the scope of the subject, the relative importance of information and ideas, and the way this information is related through the concepts in the conceptual map.

In many disciplines various forms of concept map are already used as formal knowledge representation systems, for example: semantic networks in artificial intelligence, bond graphs in mechanical and electrical engineering, CPM and PERT charts in operations research, Petri nets in communications, and category graphs in mathematics [6].

Concept Maps can also be used well to summarise information, to consolidate information from different research sources, to think through complex problems and as a way of presenting information that shows the overall structure of your subject. Concept Maps are also very quick to review - it is easy to refresh information in your mind just before it is needed by glancing at one [6].

3 Mapping the BPM Fields

3.1 Conceptual Analysis

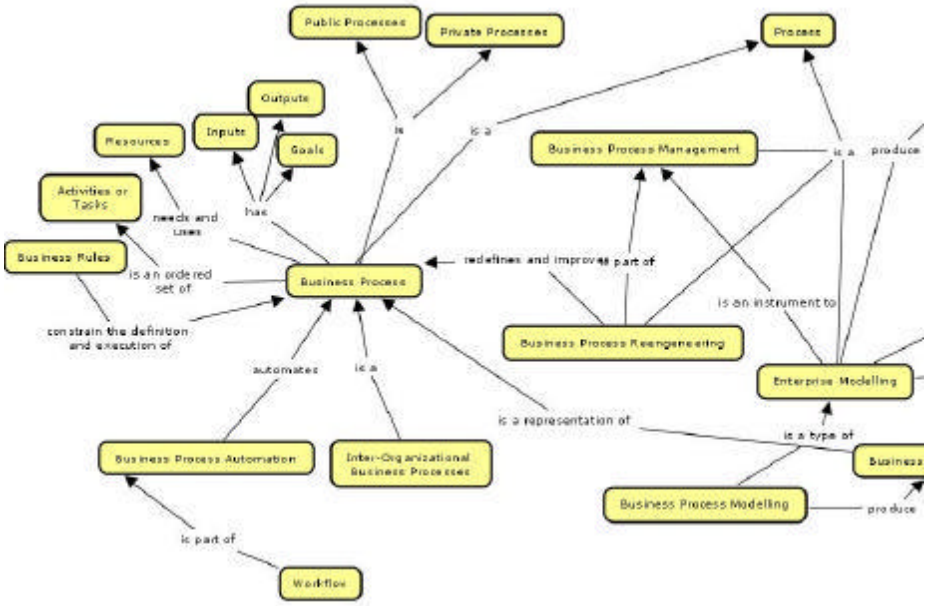
The definition of IOBP, from a business point of view, implies a close articulation with the software systems components available to support business process automation. In this field of research there are many similar terms used in different contexts with different meanings, many different approaches whose intervention objects must be clearly identified as well as many descriptive and execution languages whose scope and goals must be well understood. The clarification of the business process modelling field is beneficial both for the business architects and for information systems architects. An approach based on the conceptualization domain, more specifically on the conceptual maps approach, is an effective way to achieve such a clarification. This work has resulted in a conceptual map of the business process modelling field that enabled a critical analysis and the clarification of the relationships between business original concepts and software systems ones.

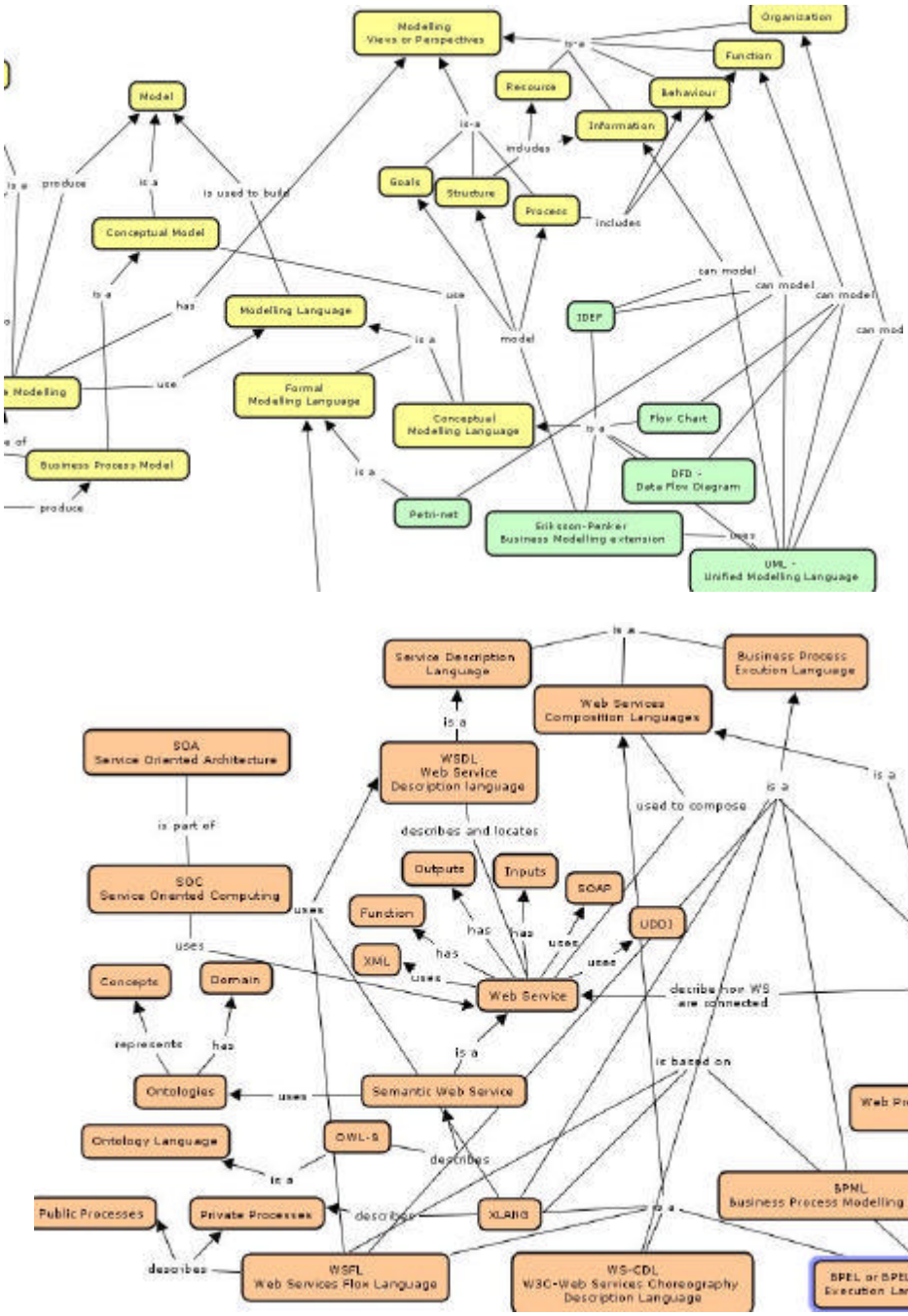
One of the works that we based on to build this concept map was on the Athena's Project [1]. The results of these works "focus on providing the means (languages, methodologies and tools) to engineer the enterprise and to show this specification as an enterprise model. These means are specifically designed for collaborative enterprises, providing different views over the model, allowing the exchange of models

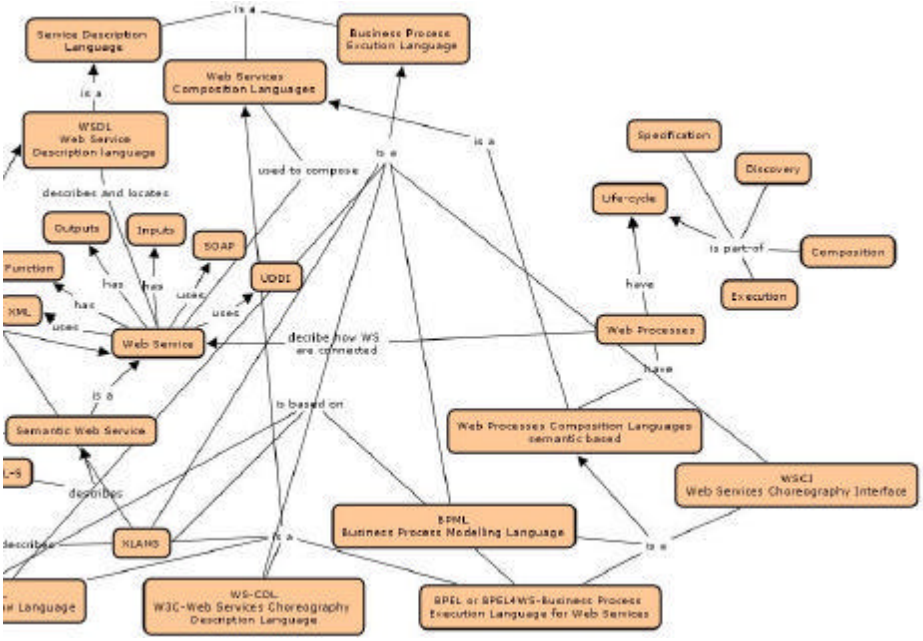
between collaborating enterprises, and providing the necessary flexibility to adapt to the changing environment forced by external entities. “

The presented conceptual map tries to highlight the interdependencies among business processes and, information’s systems components and the adopted technologies.

This conceptual map is very condensed because of the innumerous concepts that are related the BPM covers. We will present it in several pictures for a better vision and understanding.




Modelling Views or Perspectives
Organization
Function
Behaviour
Resource
Information
Model
Goals
Structure
Process
includes
produce
is a
is used to build
Conceptual Model
has
Modelling Language
use
model
IDEF
can model
Formal Modelling Language
Conceptual Modelling Language
Flow Chart
Business Process Model
DFD - Data Flow Diagram
Petri-net
Silkxson-Parker Business Modelling extension
UML - Unified Modelling Language
Service Description Language
Business Process Execution Language
Web Services Composition Languages
SOA Service Oriented Architecture
WSDL Web Service Description language
used to compose
is part of
describes and locates
SOC Service Oriented Computing
Outputs
Inputs
SOAP
UDDI
Function
XML
uses
Concepts
Domain
Web Service
describe how WS are connected
represents
Ontologies
Semantic Web Service
is based on
Ontology Language
OWL-S
describes
Public Processes
Private Processes
XLANG
BPML Business Process Modelling Lan
WSFL Web Services Flow Language
WS-CDL W3C-Web Services Choreography Description Language



As we can see from the cmap approach Business Process models can have different goals. The classic vision of Business Process Modelling divides BPM in four different views, as represented in the cmap: Functional, Behavioural, Organizational and Information [2]. This traditional approaches lack the adaptability and agility of current web based business environments. There is a lack from high level modelling methods to lower levels implementation methods.

[8] Business Process Management refers to the monitoring, measurement, controlling and optimizing business activities using automation technologies. The Business Process Automation is a subset of Business Process Management concerned with the modelling and automating of individual processes. Business Processes Modelling includes the description of the structure and behaviour of an organizational activity such as process activities flow, the role of its actors, the rules actors use and the information's needs that these actor have.

Business Process Reengineering concerns with the redefinition of a process for better compliance, faster speed of execution using automation technologies.

In fact a business process involves multiple actors (people, business units,…), concurrent activities, explicit synchronization points (e.g. some task cannot start until several other concurrent tasks are complete) and end-to-end flow of activities. Business processes consist of partially ordered activities that correspond to the operations of their defined business in order to achieve their common goal. The information structure for a business process can be defined as a network of activities performed by resources so as to transform inputs into outputs [7]. An activity is an element that performs a specific function within a process. Activities can be as simple as sending or

receiving a message or as complex as coordinating the execution of other processes and activities [9].

[11] Technologies supporting collaborative communities must operate efficiently in an open environment with practically no geographical, cultural, and technical limits. This type of environment is characterised by the fact that participants are autonomous, i.e. they can come and go, and act independently and self-contained. For a specific purpose they may be willing to participate in loosely coupled communities, taking some role and responsibility and/ or providing some services. In such communities, they may negotiate and fix some agreements, perform some tasks, provide and/or access some information, and access or offer some resources, while others are restricted to their own use. Consequently technologies used in such environments must support loose coupling, autonomy, and flexibility on the one hand, and agreement making, trust and security on the other hand.

Currently research in BPM did acquire a new vision with web services. Web Services facilitate machine-to-machine interactions, they are self-describing, modular applications that can be published, located, and invoked across the Web.

A landscape of languages and techniques for web service composition has emerged and his continuously being enriched with new proposals. We will try to highlight the differences, capabilities and limitations between the different Business Process Modelling Languages [1]:

Language	Description
BPEL, BPELWS or BPEL4WS	**Business Process Execution Language (BPEL)** defines a notation for specifying business process behavior based on web services (is entirely defined in XML). Combines the both features of WSFL and XLANG.A business process is composed of several steps called activities.
BPML	Is a meta-language for the modeling of business processes BPML aims to provide a comprehensive means of specifying the processes of an enterprise, allowing that complete business processes to be embedded as activities within other flow models.
BPMN	BPMN stands for Business Process Management Notation and was developed by the Business Process Management Initiative (BPMI) to provide a notation that is understandable by all users, from the business analysts that create the initial drafts of the process to the technical developers responsible for the technological implementation of these processes. BPMN fulfils the gap between business process design and business process implementation.
WSCI	A mature XML language for web services choreography, or the statefull process-oriented interactions of web services among multiple participants.

WSCL	WSCL allows the business level conversations or public processes supported by a Web service to be defined. WSCL specifies the XML documents being exchanged, and the allowed sequencing of these document exchanges. WSCL conversation definitions are themselves XML documents and can therefore be interpreted by Web Services infrastructures and development tools.
XLANG	Provides language constructs to describe behavioral aspects of web services and combining those services to build multi-party business processes. At the intra-service level, XLANG extends WSDL language by adding a behavior element that defines the list of actions that belong to the service and in what order they should be performed. XLANG is a notation for the automation of business processes based on Web Services for the specification of message exchange behavior among participating Web Services.
WSFL	Allow complete business processes to be embedded as activities within other flow models. The flow model specifies the execution sequence between component services.
OWL-S	Is an ontology language for describing semantic web services

[2] Processes are relationships between inputs and outputs, where inputs are transformed into outputs using a serious of activities, which add value to the inputs.

Business process models are mainly used to learn about the process, to make decisions on the process or to develop business process software. Some business process models are better suited depending on the specific purpose.

The main process modelling techniques used before were Flowcharts, Data Flow Diagrams (DFDs), Gantt charts, IDEF techniques, Coloured Petri-Nets (CPN), GRAI-GIM techniques, workflow techniques and UML among others.

We are particularly interested in the definition of Inter-Organizational Business Processes in the context of a Virtual Enterprise based to pursue of a business opportunity. However collaborative business processes need new methods that enable flexible business processes (section 2 gives a brief description of the existing methods and languages that enable collaborative business process modelling).

Modelling Collaborative Enterprises requires specific constructs and methodologies, and requires a certain level of wide-consensus for exchanging and merging behaviors of several entities into an orchestration operation. Flexibility and ability for a quick adaptation constitute the key for establishing collaboration among enterprises.

To fulfill this gap, some new approaches for Business Process Modelling were developed; the most promising are BPEL4WS, WS-CDL, BPML, WSCI, WSFL, XLANG and WSDL. All of these languages use Web Services based. Web Services is an emerging technology for building complex distributed systems focusing on interoperability that allows enterprises to describe the internal structure of their processes and how they can be invoked and composed; and also allows supported interactions between business partnerships based on the exchange of messages

However, each enterprise in the Virtual Enterprise probably has their processes described in different Enterprise Modelling Languages, which increases the degree of complexity for the exchanging of knowledge between these enterprises.

If we look carefully to the concept map we can see that Business Processes and Web Services both have common concepts (Inputs and Outputs).

A number of standards have been proposed over the past years for the process composition (WSFL, XLANG, BPML, WSCL and WSCI), however, these languages lack semantic expressivity, which has guided to the actual on going initiatives of standardization: the Business Process Execution Language for Web Services (BPEL4WS) and the WebServices Choreography Description Language (WS-CDL).

Web Process Composition is the task of combining and linking existing Web Services and other components to create new processes.

BPEL4WS defines a notation for specifying business process behavior based on Web services. Business processes can be described in two ways. Executable business processes model actual behavior of a participant in a business interaction. Business protocols, in contrast, use process descriptions that specify the mutually visible message exchange behavior of each of the parties involved in the protocol, without revealing their internal behavior.

[9] Ontologies are expected to play a central role to empower Web Services with semantics. The combination of these powerful concepts (i.e. ontologies and web services) has resulted in the emergence of a new generation of web services called semantic web services. One important challenge is service composition that refers to process of combining different web services augmenting their value.

3.2 Discussion

There are a lot of concepts with different meaning around Businesses Process Modelling and the conceptual map is the best way for us to clearly figure out how these concepts fit together and take some conclusions on this matter.

The existing methodologies to design business processes are naive in modelling Web application aspects (like information, transactions, and navigation patterns).

Technologies supporting collaborative communities must operate efficiently in an open environment with practically no geographical, cultural, and technical limits. This type of environment is characterised by the fact that participants are autonomous, i.e. they can come and go, and act independently and self-contained. For a specific purpose they may be willing to participate in loosely coupled communities, taking some role and responsibility and/ or providing some services. In such communities, they may negotiate and fix some agreements, perform some tasks, provide and/ or access some information, and access or offer some resources, while others are restricted to their own use. Consequently technologies used in such environments must support loose coupling, autonomy, and flexibility on the one hand, and agreement making,

trust and security on the other hand. The following outlines some technologies that are considered to specifically support such environments.

[13] Each network enterprise has its own private methods of process modelling methods (Petri-Net, UML, DFDs, and so on) and tools. Due to the lack of common interfaces and mapping-methods, neither can tools interact with each other nor can the methods be transformed into one another. To extract information relevant to the network from these "private processes", a collaboration specific view is generated, providing all or at least some information (white-box) or in a black-box manner with no indications about their realization (only the interfaces of the private process are described). Private processes must be protected from external insights but at the same time integrated into the whole collaborative process for the extended approach of Collaborative Business Process Management.

4 Conclusions and further work

This paper provides a survey of the current research in BPM trough a conceptual map. We have tried to highlight the deep relationships between the most important concepts in the field of Business Process Modelling that, as we may conclude, requires new forms of flexibility in today's business changing environment. In our approach we consider a Breeding Environment where a set of enterprises exist and have the intention to cooperate with each other in order to maintain a set of social relationships, mainly trust relationships, that can be mobilized to join resources and collaborate to compete for a business opportunity. A Breeding Environment represents an association of enterprises that have the intention to cooperate with each other in order to establish a long term cooperation agreements and an interoperable infrastructure. When a Business Opportunity is detected a subset of this Breeding Environment can be selected to accomplish these Business Opportunity. Our goal is to apply a Multi-Agent System (MAS) in the discovery and inter-organizational articulation of individual public Business Processes. Web services coordination between organizations must be preceded by the definition of the IOBP. When we deal with the setup of temporary collaborations of enterprises to take advantage of a given business opportunity, one fundamental step is to articulate individual BP in order to achieve a set of inter-organizational processes that satisfy the business goals. Research work on MAS applications did not favor this problem, and we believe that this is an important research direction. Semantic interoperability is crucial to assure a meaningful interaction, communication and cooperation among the heterogeneous agents and services.

Factors like: distributed system architecture, reactivity to changes, interoperation among heterogeneous systems, resource management and intelligent decision making are some of the advantages of using a Multi-Agent System.

The enactment of IOBP is fundamental for enterprises to create new partnerships efficiently and in a quick way. To accomplish this a few requirements are needed:

Define a Business Process in a detailed and comprehensive way is a complex task because of the dynamic environment they are involved in, such as complex business rules and policies, abnormally action from the involved partners, among others.

The definition of IOBP in a Virtual Enterprise is still a challenging issue for the research community. In spite of the several approaches being made in this area, there seems to be an agreement that the agent-based systems are the most promising technology that address the IOBP life cycle, since they can effectively support agility by adapting themselves to the continuous environment changes. Furthermore, agent technology provides enterprises the ability to learn both from their individual behavior and from the cooperative relations with others. One of our future directions is to improve the definition of the IOBP using a learning approach within social networks modelling to optimize the IOBP selection by adding social relations parameters of negotiation.

The development of such a decision support system requires a set of concepts that provides a systematic way to define IOBP. IOBP possess characteristics that require different design approaches then the one founds in traditional systems. One of these characteristics is that IOBP are emergent and cannot be defined a priori. IOBP include intensive interactions where experts merge tacit and explicit knowledge to create and exchange ideas in order to identify the next process step.

This paper presented a conceptual analysis of the main fields dealing with BPM with the goal of clarifying conceptually the uses of BPM in the management and computer science fields, clarifying the BPM concepts and their relationships. Also an important goal (and the first aim of our work), is to set up a solid conceptual basis for interdisciplinary research in this area.

4 References

1. Athena Project, Enterprise Modelling in the Context of Collaborative Enterprise, Work Document Wd.A1.1.2, February 2005.
2. Aguilar-Savén, 2003, Business Process Modelling: Review and framework, International Journal of Production Economics, April, 2003.
3. Axenath, B., Kindler, E., Rubin, V.,The Aspects of Business Processes: An Open and Formalism independent Ontology, Technical Report, University of Paderborn, 2005.
4. Barros, A., Dumas, M., Oaks, P., A Critical Overview of the Web Services Choreography Description Language (WS-CDL), http://www.bptrends.com/, March 2005.
5. Brian R. Gaines and Mildred L. G., Concept Maps as Hypermedia Components, Shaw Knowledge Science Institute University of Calgary, Alberta, Canada.
6. J. Cañas, R. Carff, G. Hill, M. Carvalho, M. Arguedas, T. C. Eskridge, J. Lott, R. Carvajal, Concept Maps: Integrating Knowledge and Information Visualization, in Knowledge and Information Visualization: Searching for Synergies, S.-O. Tergan, and T. Keller, Editors. 2005. Heidelberg / New York: Springer Lecture Notes in Computer

Science.Aguilar-Savén, 2003, Business Process Modelling: Review and framework, Inter-national Journal of Production Economics, April, 2003.
7. Kazanis, P., Ginige, A., "Asynchronous Collaborative Business Process Modelling Through a Web Forum", Seventh Annual CollECTeR Conference on Electronic Commerce. Melbourne, VIC, Australia in association with ACIS 2002, October 2002.
8. Leeming, N., 2005, "Business Process Management Implementation", The Journal of Enterprise Architecture, 1, 69-91. Global Enterprise Architecture Organisation. Otago University Press.
9. Medjahed, B., Bouguettays, A., A Multi-level Composability for Semantic Web Services, IEEE Transactions on Knowledge and Data Engineering, July 2005.
10. Papazoglou, M. Web Services and Business Transactions , World Wide Web: Internet and Web Information Systems, 6, 49-91, Netherlands, 2003.
11. Tshammer, V., Collaborative Commerce – Trends and Technology Potentials, First DEEDs Policy Group Meeting Brousells, 2001
12. Bernners-Lee, T., Handler, J., Lassila, O., The Semantic Web, Scientific Am., vol. 284, nº5, pp.34-43, May 2001.
13. Vanderhaeghen, D., Zang, S., Hofer, A., Adam, O. "XML-based Transformation of Business Process Models - Enabler for Collaborative Business Process Management", 2004. (Svirkas and Roberts, 2003)
14. Verma, K., Sheth, A., Miller, J. Creating Web Processes using BPEL4WS, LSDIS Lab, Com-puter Science, 2004.

Resolução de Conflitos na Marcação Automática de Reuniões

António Nabais

anabais@ipca.pt

LIACC/FEUP – Laboratório de Inteligência Artificial e Ciências da Computação – Universidade do Porto
Http://www.ncc.up.pt/liacc/, Tel.: 351-22-5081315, Fax: 351-22-5081315

EST/IPCA – Escola Superior de Tecnologia - Instituto Politécnico do Cavado e do Ave, Barcelos, Portugal
Http://www.ipca.pt, Tel.: 351-253-802260, Fax: 351-253-802261

Resumo. A marcação automática de reuniões tem sido largamente estudada nos últimos 20 anos. Os trabalhos recentes nesta área trataram o problema como um sistema distribuído, sem controlo fixo centralizado, adequado à utilização de um Sistema Multi-Agente (SMA). Os agentes actuam de forma autónoma, comunicando e negociando com outros agentes, tendo em atenção as preferências e disponibilidades dos utilizadores.

Neste artigo é apresentada a formulação do problema de Marcação Automática Distribuída de Reuniões, e uma perspectiva sobre alguns dos trabalhos anteriores neste tema, para justificar a abordagem SMA. São identificadas várias técnicas de negociação e resolução de conflitos, com vista à resolução deste problema.

O artigo analisa ainda algumas das implementações práticas existentes, de forma a determinar os pontos fortes e fracos de cada abordagem e produzir uma arquitectura para uma implementação prática futura. Pretende-se que as restrições em relação ao número máximo de agentes envolvidos sejam relaxadas, e que a eficácia computacional e qualidade da solução sejam as melhores possíveis.

Da análise efectuada resulta um conjunto de conclusões relativas ao modelo de comunicação, negociação, heurísticas e função de avaliação (individual e colectiva), que melhor se adaptam a este problema.

Palavras-Chave: Marcação Automática de Reuniões, Sistemas Multi-Agente, Negociação, Resolução de Conflitos

1 Introdução

Neste artigo é descrita a marcação de reuniões, usando um sistema distribuído envolvendo vários agentes.

A marcação de uma reunião é uma tarefa frequente na actividade humana, envolvendo a marcação do local, data e hora comuns a um grupo de pessoas, em que se podem encontrar para realizar determinada actividade. Através da troca de informação, procura-se uma possível optimização das datas relativamente ao maior número possível de participantes.

No quotidiano, a marcação de uma reunião é uma tarefa naturalmente distribuída, de natureza dinâmica e combinatória [2]. Além disso, é comum que um grupo de pessoas possua restrições e indisponibilidades que produzam preferências em conflito com outros membros do grupo.

Globalmente, a marcação de reuniões é um processo fastidioso, iterativo e consumidor de bastante tempo. Por isso, a automatização da marcação é importante, não só porque poupa tempo e esforço humanos, mas também porque pode levar a uma marcação mais eficiente [10].

Este artigo apresenta uma formulação do problema de Marcação Automática Distribuída de Reuniões (DMS), baseada no trabalho de Sen e Durfee [5, 7]. A partir daqui é efectuada uma análise sobre alguns dos trabalhos relacionados com este tema, de forma a resumir a forma como outros autores abordaram a utilização de sistemas Multiagente, negociação entre os agentes e resolução distribuída dos conflitos.

Analisam-se algumas implementações práticas existentes, incluindo soluções comerciais. As aplicações que implementam meras agendas electrónicas com processamento centralizado não foram consideradas. Interessa considerar soluções baseadas em verdadeiros agentes autónomos.

A partir daqui são retiradas conclusões relativas ao modelo de comunicação, estratégia de negociação, heurísticas utilizadas e função de avaliação (utilidade), de forma a projectar uma implementação de um SMA para este problema que permita remover algumas restrições em relação ao número máximo de agentes envolvidos e melhorar a eficácia computacional e qualidade da solução.

2 Descrição Formal do Problema

O processo de marcação de reuniões é composto por um conjunto de reuniões e um grupo de participantes. Dado o conjunto de reuniões n, e k participantes, Sen and Durfee [5], representaram o problema de marcação por:

$$S = (A,M)$$

Onde: $A = \{1,2,3,...k\}$, é um grupo de participantes;
$M = \{m_1,m_2,m_3,...m_n\}$, é um conjunto de reuniões;

A janela de tempo pode ser representada como um dia e uma hora <D,H>, o conjunto da janela chama-se **intervalo tempo**.

Cada reunião representa o conjunto de atributos:

$$m_i = (A_i, h_i, l_i, w_i, S_i, a_i, d_i, f_i, T_i),$$

Onde: $A_i \subseteq A$, é o conjunto de agentes que participam na reunião;
$h_i, \in A_i$, é o agente que funcionará como "host" desta reunião;
l_i, é a duração da reunião;
$0 < w_i, \leq 1$, é o peso ou a prioridade da reunião;
S_i, é um conjunto de tempo início no seu calendário para esta reunião.
Se $|S_i| = 1$, significa esta reunião está constrangida;
Se $|S_i| = \{\}$, significa esta marcação de reunião pode começar;
a_i, é a proposta $<D\ a_i, H\ a_i>$ do "host" para esta reunião m_i;
d_i, é duração $<D\ d_i, H\ d_i>$ do "host" para esta reunião m_i;
f_i, é a proposta $<D\ f_i, H\ f_i>$ que converge.
T_i é o tempo intervalo da proposta final, representado por:
$\{<dia_i, hora_i>, <dia_i, hora_i+1>, \ldots..<dia_i, hora_i+ l_i -1>\}$

Cada agente tem o seu próprio calendário pessoal, no qual figuram as suas reuniões já marcadas. O calendário tem de estar associado ao processo de marcação de cada nova reunião e pode ser representado como:

$$C_j = \{<D_s,0,X_{s,0}>,<D_s,1,X_{s,1}>,\ldots,<D_s,L-1,X_{s,L-1}>,$$
$$<D_{s+1},0,X_{s+1,0}>,\ldots,<D_{s+1},L-1,X_{s+1,L-1}>,\ldots,$$
$$<D_e,0,X_{e,0}>,\ldots,<D_e,L-1,X_{e,L-1}>\}$$

Onde:
D_s é data início de calendário;
D_e é data fim de calendário;
L é número de horas por dia, e

$$X_{x,y} = \begin{cases} m_i \text{ Se } j \in A_i \text{ e } <D_{x,y}> \in T_i \\ nil, \text{ nos restantes casos.} \end{cases}$$

A reunião m_i pode marcar-se quando:

$$\forall_j, j \in A_i, \text{ e } \forall y,z, <D_y, H_z> \in T_i, <D_y, H_z, m_i> \in C_j.$$

3 Trabalhos Relacionados

Tem havido muitos trabalhos de pesquisa científica na área da Marcação de Reuniões. É de destacar o trabalho de Sandip Sen e Edmund Durfee da Universidade de Tulsa e da Universidade de Michigan [5, 6, 7], que trabalharam com sistemas Multiagente, efectuaram a formulação vista acima e procuraram soluções distribuídas baseadas em protocolos de negociação.

Estes autores não consideraram um atributo adicional, que é o local da reunião. Restrições de espaço, de salas ou até de cidade, são importantes na marcação de muitas reuniões e seriam uma adição interessante a ser considerada.

Para Sycara e Garrido [8], a marcação de reuniões é uma tarefa distribuída onde cada agente sabe as preferências e disponibilidades do utilizador agindo em seu proveito. Os autores apresentaram um sistema distribuído de marcação de reuniões sem um controlo fixo centralizado, ou seja, não existe um agente especial de controlo. Assim, todo e qualquer agente é capaz de marcar uma reunião e de negociar com outro, considerando as preferências individuais e disponibilidade do calendário, através de um protocolo dinâmico e de um mecanismo de coordenação [8, 9].

Nos suas experiências, os autores referidos consideraram o processo de negociação/marcação entre três agentes, consideraram um calendário com 3 dias, 3 horas por dia, e uma duração com tempo inicio múltipla de 30 minutos.

Este sistema mediu os parâmetros de eficiência e qualidade de reunião.

- Se a eficiência é constante, a qualidade de reunião decresce quando a densidade do calendário está a aumentar.
- Se a eficiência é melhor quando a densidade de calendário é menor ou igual a quatro tarefas preenchidos.

Foi apresentada uma alternativa [9] baseada no modelo de comunicação de restrições e de preferências entre os agentes.

Os agentes são capazes de negociar e relaxar restrições para procurar atingir o acordo sobre as marcações com utilidade alta. Usando este modelo, os agentes também podem reagir, rever a marcação e responder com alterações dinâmicas à evolução das situações.

O sistema COSMA (COopertive Schedule Management Agents) desenvolvido no Centro Alemão de Inteligência Artificial DFKI [4] pretende desenhar agentes com planos locais, metas, conhecimentos, compreensão e modelo do mundo conceptual e deixá-los interagir de forma cooperativa.

Este sistema foi aplicado para marcações de reunião. A negociação das datas das reuniões envolveu vários aspectos, começando pela selecção de um conjunto apropriado de mensagens e a especificação sequencial das mensagens via protocolo, depois o desenvolvimento de um modelo temporal que sustenta a classificação e comparação das propostas para guiar as decisões que serão feitas quando usarem uma estratégia

específica. Os efeitos do comportamento local dos agentes com uma negociação foram uma alocação cooperativa das reuniões dentro do calendário dos participantes.

A principal contribuição deste trabalho consiste no desenvolvimento de um modelo de negociação que suporta a comunicação em diferentes níveis desde o simples envio de mensagens até diferentes estratégias de negociação que conduzam a diferentes mecanismos de cooperação a nível social [4].

4 Implementações práticas

Descreve-se de seguida alguns dos SMA propostos para a resolução do problema DMS, dando destaque às implementações práticas dos autores referidos no capítulo anterior.

Sistema de Marcação de Reuniões da Carnegie Mellon University (CMU)

O factor primordial no desenvolvimento deste sistema é manter a privacidade da informação, dando prioridade a uma verdadeira autonomia e independência dos agentes. Assim, neste sistema, os agentes somente sabem a sua própria informação, isto é, preferências sobre a reunião e informação de calendário, podendo no entanto trocar alguma informação no processo de negociação [9].

Desta forma, uma reunião consiste em três atributos fundamentais:

- Data
- Início
- Duração

O processo de negociação desenrolar-se-á até que os agentes envolvidos cheguem a um consenso sobre os três parâmetros acima mencionados.

Basicamente, neste sistema, cada agente tem capacidade de relaxar três restrições diferentes relacionadas com o tempo: data, tempo de início e duração. Além disso, cada agente tem pesos, valores compreendidos entre 0 e 1, que indicam como relaxar cada uma dessas restrições de tempo.

A estrutura de agente é modular, isto é, cada módulo funciona como um processo independente, que pode enviar e receber mensagens de forma assíncrona [8].

Os agentes comunicam e negoceiam, sendo cada agente capaz de relaxar as suas preferências em função da medida do conflito na negociação.

Cada agente que recebe uma proposta pode aceitar ou rejeitar, respondendo com uma mensagem. Quando o agente a aceita, ele pode partilhar o valor da prioridade que este agente tem atribuído ao intervalo de tempo aceite.

Existe um agente coordenador, escolhido aleatoriamente, cujo papel é gerar uma proposta inicial e recolher mensagens de resposta para cada proposta. Quando a proposta foi aceite por todos os agentes, o coordenador calcula a utilidade do grupo e comunica-a aos restantes agentes, indicando que a proposta foi aprovada.

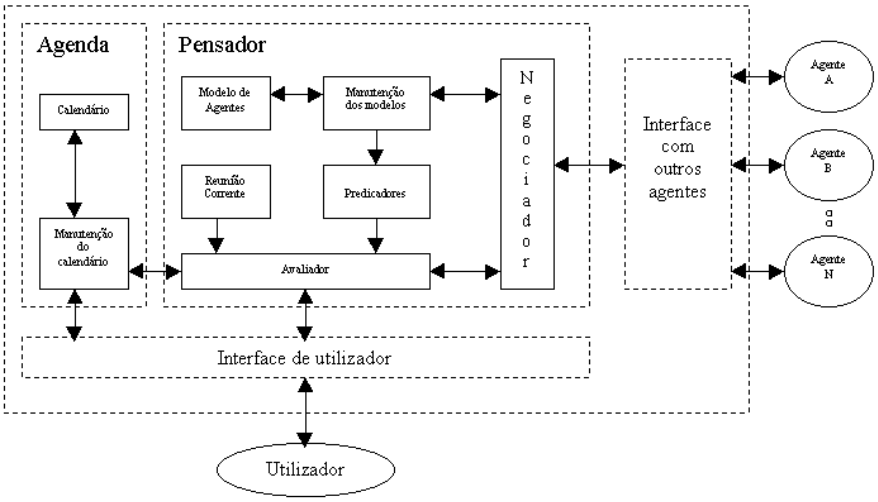


Fig. 1. Estrutura dos Agentes na Agenda do CMU

Se pelo menos um agente rejeitar a proposta, o coordenador tem de seleccionar e enviar uma nova proposta, eventualmente relaxando as suas restrições. Este processo repete-se até haver acordo entre todos.

A "agenda automática distribuída" da Universidade de Tulsa

Os trabalhos de Sen & Durfee [5, 6, 7] focaram sobretudo a resolução do problema da marcação de reunião usando o agente organizador ("host") fixo e centralizado que é capaz de comunicar com todos os outros agentes para marcar as reuniões. Os restantes intervenientes em cada reunião são chamados *convidados.*

Quando uma reunião é solicitada pelo utilizador do agente *organizador*, é pesquisado o calendário local para encontrar um intervalo que satisfaça as restrições. Esse intervalo é anunciado para os *convidados* como proposta para a reunião. Os restantes agentes pesquisam os seus calendários de reuniões e respondem com um subconjunto

O sistema distribuído de marcação de reunião utiliza agentes cuja estrutura é seguinte:

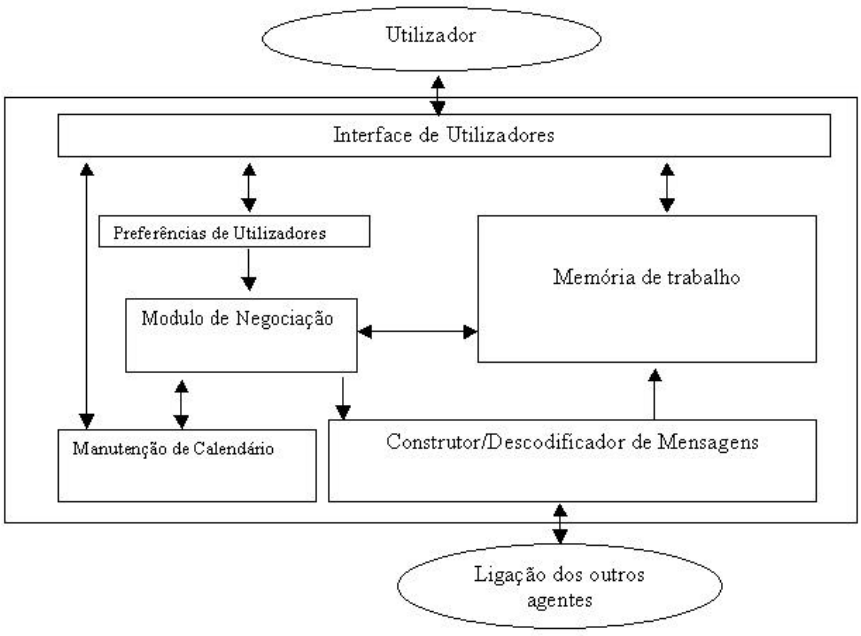


Fig.2 Estrutura de agente de Sandip Sen

O utilizador interage com o sistema através do *Interface de Utilizadores*, que permite introduzir pedidos de reuniões e preferências. O utilizador pode também verificar quais as reuniões marcadas pelo agente e negociações em curso.

As *Preferências de Utilizadores* armazenas as preferências, prioridades para os diferentes tipos de reuniões, etc.

A *Memória de Trabalho* contém as estruturas de dados e valores temporários de reuniões em processo de negociação.

O *Módulo de Negociação* usa as preferências do utilizador para trocar propostas, que serão depois armazenadas na *Memória de Trabalho*.

A *Manutenção do Calendário* permite ao módulo de interface e ao módulo de negociação aceder e alterar a agenda do utilizador, que depois será comunicada através do *Interface de Utilizadores*.

O *Construtor/Descodificador de Mensagens* é utilizado para comunicar com o sistema de e-mail usado por esta implementação para comunicar com os outros agentes. Este módulo converte propostas em mensagens e vice-versa.

COSMA do DFKI

O processo de negociação de COSMA consiste em cinco ingredientes: Formato de mensagens; Protocolos; Avaliação; Negociação; Cooperação.

Fig.3 Estrutura dos ingredientes de negociação do COSMA

O protocolo de COSMA, é implementado como um autómato de estados, onde cada estado representa uma acção de um agente (acção de comunicação) e as transições representam as mensagens recebidas.

A ideia base é simples: a negociação começa com a mensagem de inicialização (tipo ARRANGE) do agente iniciador (que propõe a reunião e cria a primeira mensagem), especificando a prioridade da reunião, a lista de participantes e a primeira proposta do intervalo de tempo de marcação da reunião.

Se pelo menos um dos participantes rejeitar a reunião, esse agente tem de participar a rejeição a todos os participantes, terminando a negociação.

Se alguns participantes responderem com uma sugestão de modificação da proposta inicial, têm de notificar todos os restantes, propondo uma alteração de tempo, que será calculada baseando-se nas suas preferencias e nas sugestões de modificações recebidas.

Se todos os agentes aceitam a proposta, ou refinam o tempo proposto de forma a que os diferentes refinamentos se enquadrem num intervalo de tempo aceitável, o agente tem de fixar o horário para o início da reunião dentro desse intervalo e tem de comunicá-lo a todos os outros através duma mensagem de confirmação.

Pode acontecer que os participantes concordem com o intervalo de tempo, mas que não consigam encontrar suficiente sobreposição para os diferentes refinamentos a efectuar. Neste caso a proposta é considerada incompatível.

O nível de negociação está directamente construído sobre os níveis de protocolo e de avaliação. O nível de protocolo determina o contexto comunicacional para a decisão. O de avaliação providencia um critério de escalonamento para a realização das decisões. No contexto de uma proposta num protocolo de alto nível, um agente decide puramente sobre a existência de um sub-intervalo com uma utilidade positiva.

MARSiMA

Desenvolvido por Li Xin, no âmbito da sua tese de mestrado, utiliza um sistema multi-agente para a implementação de uma agenda electrónica para marcação automática de reuniões.

Para a negociação entre os agentes foi utilizado um algoritmo genético, integrando aspectos de aprendizagem automática [10]. O algoritmo de negociação considera a agenda corrente e integra as crenças que o agente possui sobre os demais agentes.

Durante a negociação, é utilizada uma função de "crossover", permitindo relaxar as restrições iniciais do agente, conduzindo a uma possível convergência. Neste processo é permitida a diminuição da utilidade do próprio agente, mas obtém-se uma maior satisfação global.

A comunicação entre os agentes é feita utilizando a linguagem KQML e a plataforma de agentes JATLite [10].

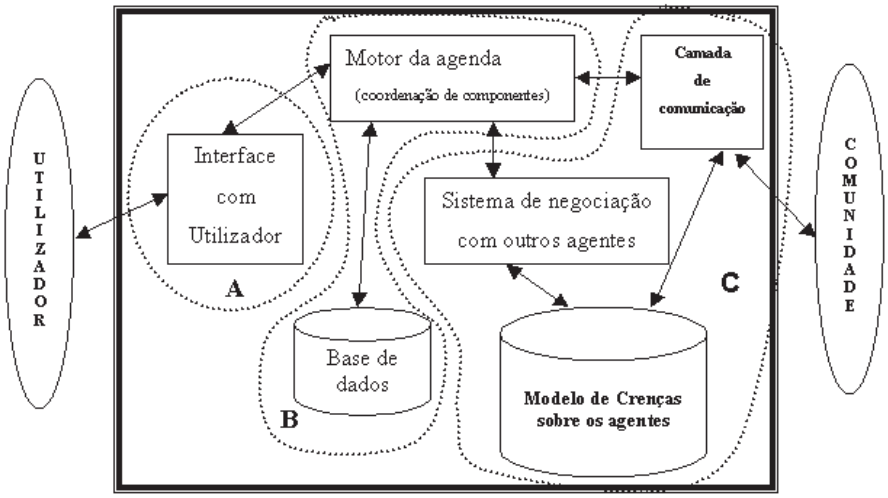


Fig.4 - Arquitectura do agente MARSiMA

5 Arquitectura proposta para os agentes

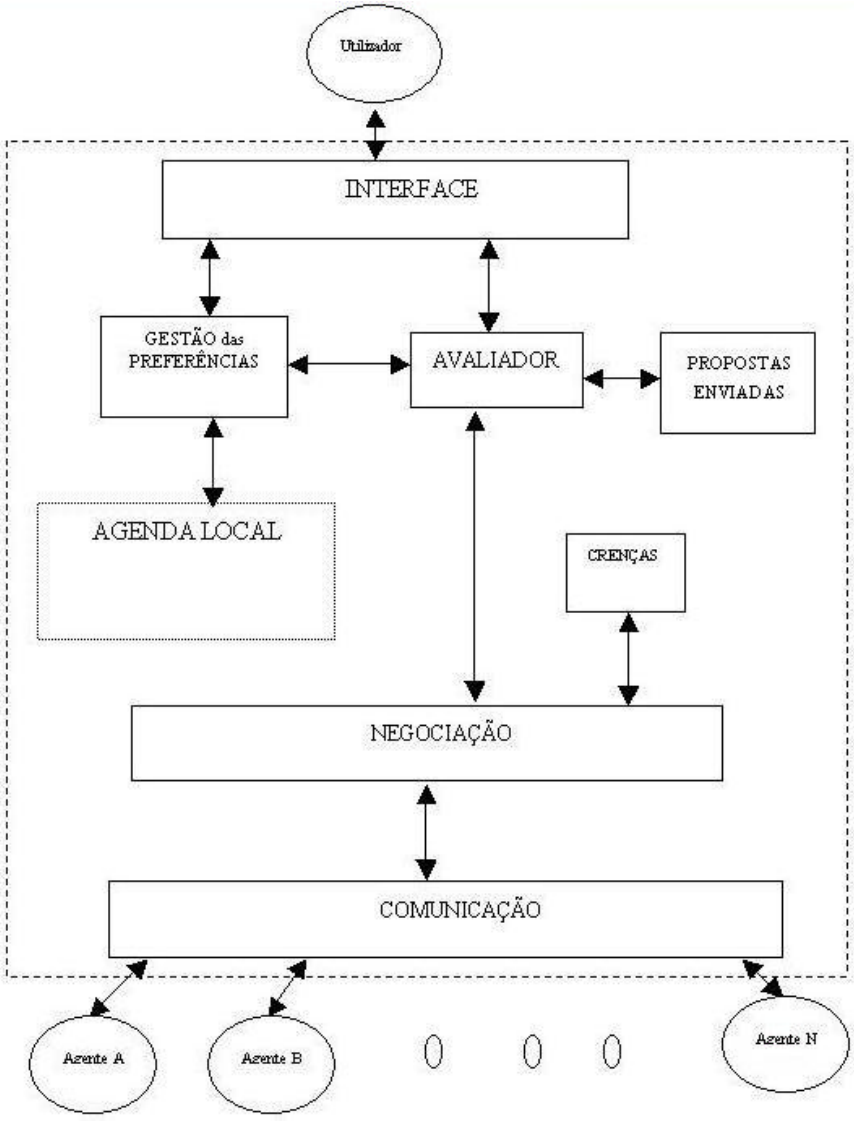


Fig.5 - Estrutura proposta para os agentes

Interface

Permite a interacção com o utilizador. É possível consultar as reuniões já marcadas, pedir uma nova reunião e acompanhar o processo de negociação. O utilizador associado a cada agente pode indicar as suas preferências (restrições suaves) e indisponibilidades (restrições rigorosas). As reuniões já agendadas farão parte deste grupo, como é natural.

Gestão das Preferências

Em conjunto com a Agenda Local, irá fazer permitir a consulta e manipulação dos diferentes tipos de restrições deste agente, que servirão de base à formulação de novas propostas.

Agenda Local

Serve para armazenar e manipular o calendário deste agente, contendo reuniões já marcadas. Este módulo comunicará à interface eventuais conflitos das preferências com o calendário e a marcação de uma nova reunião a partir da negociação.

Propostas Enviadas

Para cada negociação em curso é armazenada a lista das propostas já enviadas. Irá conter também outros dados temporários relativos a reuniões cuja marcação ainda não foi concluída. Esta informação permite o uso de heurísticas para pesquisar novas propostas a enviar.

Crenças

Contém informação recolhida das propostas recebidas, permitindo ao agente orientar as propostas que irá enviar e eventualmente implementar negociação estratégica, conforme descrito em [1].

Avaliador

É responsável pela geração de propostas, que o módulo de Negociação depois seleccionará. Durante a negociação os agentes procuram os horários livres que podem propor. Esta procura é direccionada por uma função de utilidade que indica o valor de prioridade para cada intervalo de tempo (de acordo com as preferências individuais da reunião e do peso do relaxamento). Assim, a distância pesada entre o intervalo j e as preferências do agente K, tendo em conta os pesos de relaxamento, pode ser definida pela seguinte função [8]:

$$WDist(\vec{I}^{\,j}, \vec{P}^{k}, \vec{W}^{k}) = \sum_{i=1}^{3} [\vec{W}^{k} \times Dist(\vec{I}^{\,j}, \vec{P}^{k})]$$

$WDist$ é a distância pesada entre o intervalo j e as preferências do agente K.

$\vec{I}_j$, é o vector do intervalo j do calendário com três atributos: uma data, uma hora inicial e uma duração.

$\vec{p}^{\,k}$, é o vector das preferências do agente k com três atributos: uma data, uma hora inicial e uma duração.

$\vec{W}^{k}$, é o vector dos pesos atribuídos pelo utilizador para relaxar as três variáveis ao longo do processo de negociação.

O indíce i indica o atributo dessa reunião: 1 é a data; 2 a hora inicial e 3 a duração.

$Dist\,(\vec{I}_i^{\,j}, p_i^{\,k})$, é a distância entre $\vec{I}_i^{\,j}$, e $p_i^{\,k}$, ou seja, o número de possíveis diferentes instâncias do atributo *i* entre o valor do atributo do intervalo *j* e o valor do atributo que é o mais preferido pelo agente *K*.

Negociação

A partir das propostas recebidas, este módulo decide se a proposta é aceite, recusada ou ajustada, passando esta informação aos outros agentes. Cada proposta em conflito terá um determinado nível de relaxamento, representando quanto este agente está disposto a perder da sua própria utilidade, para aumentar a utilidade global do sistema.

Este módulo recolhe informação das propostas recebidas, que armazena sob a forma de crenças. Esta informação será utilizada para direccionar a busca de novas propostas.

Comunicação

Utilizando os recursos da plataforma JADE [11], este módulo irá comunicar com os restantes agentes do sistema para a marcação das reuniões.

6 Conclusões e Trabalho Futuro

Este artigo apresentou o problema de Marcação Automática Distribuída de Reuniões e efectuou uma análise sobre alguns dos trabalhos relacionados com este tema.

A formulação apresentada para o problema, tem servido de base a muitos trabalhos da área e é adequada ao trabalho que se pretende desenvolver. Poderia ser expandida para incluir como propriedade adicional o local da reunião.

Foram analisados alguns dos trabalhos mais importantes desta área, incluindo contribuições que se tornaram pratica comum em todas as aplicações desenvolvidas. Destaca-se o trabalho de Sandip Sen [6].

Das implementações práticas destes trabalhos podemos concluir que os autores deram muita importância a questões como a interface e o mecanismo de comunicação com os outros agentes. No projecto proposto estes pontos tornam-se secundários devido à utilização de linguagens como o Java e a plataforma de agentes JADE [11]. Isto permite concentrar esforços no desenvolvimento dos sistemas de negociação e resolução de conflitos mais adequados, com vista a melhorar a qualidade da solução obtida.

A arquitectura apresentada para os agentes simplifica alguns pontos e permite a implementação de técnicas muito promissoras como a alteração de reuniões já agendadas [3] e o uso de estratégias de negociação [1]. Neste último caso é importante o equilíbrio entre a informação privada de cada agente e a informação enviada aos restantes agentes. Se demasiada informação for divulgada, não há espaço para estratégias de negociação. Se pouca informação for divulgada, o sistema pode nunca convergir para uma solução.

Pretende-se brevemente implementar e sistema descrito e retirar dados experimentais que permitam verificar estas conclusões.

Referencias

1. E. Crawford and M. Veloso: Learning to Select Negociation Strategies in Multi-agent Meeting Scheduling. In: *12th Portuguese Conference on Artificial Inteligence (EPIA 2005)*, p.584-595, 2005
2. A. B. Hassine, X. Défago and T.B. Ho: Agent-based Approach to Dynamic Meeting Scheduling Problems. In: *Proceedings of the Third Conference on Autonomous Agents and Multi-agent Systems (AAMAS'04)*, p.1130-1137, New York, USA, 2004.
3. P. J. Modi and M. Veloso: Bumping Strategies for the Multiagent Agreement Problem. In: *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS'05)*, p.390-396, Utrecht, Netherlands, 2005.
4. P. Sablayrolles and A. Schupeta: Conflict Resolving Negotiation for COoperative Schedule Management Agents (COSMA), In: *DFKI Research report TM-93-02*, 1993
5. S. Sen and E. H. Durfee: A Formal Study of Distributed Meeting Scheduling: Preliminary Results. In: *ACM Conference on Organizational Computing Systems*, p. 55-68, 1991
6. S. Sen and E. H. Durfee: Developing an Automated Distributed Meeting Scheduler. In: *IEEE Expert, vol.12, no. 4*, p.41-45, 1997
7. S. Sen and E. H. Durfee: A Contracting Model for Flexible Distributed Scheduling. In: *ACM Conference on Organizational Computing Systems*, p. 55-68, 1991
8. K. Sycara and L. Garrido: Multi-agent Meeting Scheduling: An Experimental System. In: *V Congresso Iberoamericano de Inteligencia Artificial*, p.104-113, 1996
9. K. Sycara and J. Liu: Distributed Meeting Scheduling. In: *Sixteenth Annual Conference of the Cognitive Society*, 1994
10. L. Xin: MARSiMA - Marcação Automática de Reuniões usando um Sistema Multi-Agente. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, 2000
11. JADE - Java Agent DEvelopment Framework. In: *http://jade.tilab.com/papers/2003/WhitePaperJADEEXP.pdf*, (consultado em Dezembro de 2005)

Sessão Técnica 3

Engenharia de Software e Aplicações

Proposta para um Web Feature Service Temporal

Artur Rocha[1], Alexandre Carvalho[2,1]

[1]INESC Porto, Rua Dr. Roberto Frias 378, 4200-465 Porto
[2]FEUP, Rua Dr. Roberto Frias 378, 4200-465 Porto
{artur.rocha,alexandre.carvalho}@inescporto.pt

Resumo: Este artigo descreve uma aproximação à inclusão de suporte para lógica espaço-temporal em *Web Feature Services* (WFS). Este tipo de *Web Services* permite a interrogação de fontes heterogéneas de informação geográfica, num ambiente distribuído, retornando o resultado em formato *Geography Markup Language* (GML). No entanto, não é possível obter, a partir destes WFS, múltiplos estados para cada entidade (*feature*) geográfica, que resultariam da combinação de filtros espaciais e literais com predicados temporais. O uso destes predicados permite restringir o conjunto de estados contemplados no cálculo da resposta, ou mesmo apresentar um resultado espacial coalescido sobre uma dimensão temporal. Uma vez que a GML 3.1.1 já caracteriza os tempos válidos da informação geográfica, é viável a estruturação temporal das respostas devolvidas pelos WFS, sendo no entanto necessário enriquecer a sintaxe dos seus pedidos com a capacidade de utilizar predicados e operadores temporais. Este artigo propõe alterações à norma WFS, mais concretamente ao nível da linguagem de filtragem que utiliza, e apresenta uma solução efectiva para o suporte espaço-temporal nos SGBD subjacentes.

1 Introdução

1.1 Caracterização do suporte temporal nos SGBD

As bases de dados tradicionais são utilizadas para armazenar informação inter-relacionada num domínio do mundo real ou sintetizado. No que respeita aos aspectos temporais, os SGBD capturam a essência temporal correspondente à representação mais recente dos "factos" armazenados.

No domínio dos Sistemas de Informação Geográfica (GIS), os SGBD que lhe estão subjacentes manipulam também bases de dados do tipo referido, pelo que este tipo de tecnologia (GIS) é utilizada para armazenar e manipular o estado mais recente dos factos espaciais. A actualização do estado mais recente de um facto acarreta assim a perda da informação que descreve o seu estado anterior.

No sentido de permitir a gestão, numa base de dados, dos aspectos temporais dos factos, Date [1] considera duas aproximações: uma semi-temporal e outra verdadeiramente temporal. Na primeira, a captura da essência temporal de dados históricos é realizada com recurso a atributos temporais do tipo *timestamp*. Na

segunda, cada facto deve ter associado um intervalo que indica o período de tempo em que esse facto foi considerado válido no domínio representado, podendo ainda ser necessário manter o registo temporal da sua actualização na base de dados. Resulta daqui que uma aproximação temporal *full-fledged*, passa pela representação dos factos sob dois domínios temporais designados, respectivamente, por domínio dos tempos válidos [3] e pelo domínio dos tempos de transacção [4].

Importa referir que, segundo Date [1], a aproximação semi-temporal pode conduzir a graves problemas e dificuldades, particularmente na interrogação à base de dados, na medida em que as *queries* realizadas tomam graus de complexidade muito elevada. Estes factos resultam da inadequação dos SGBDs não temporais em fornecer o suporte necessário para que a realização destas operações [5], por exemplo, tipos de dados temporais, operadores temporais (operadores de Allen [7], entre outros) e funções temporais [6].

Este trabalho recorre pois à utilização de um SGBD com suporte espaço-bitemporal, em desenvolvimento. A opção por esta solução resulta do facto das implementações temporais mais sólidas da actualidade, TimeDB [17] e TEMPOS [18], não possuírem suporte para dados espaciais.

1.2 Acesso interoperável a Informação Geográfica

A informação geográfica (IG) possui características únicas de referenciação, pelo que pode funcionar como um excelente elemento aglutinador de informação, ainda que esta resida em sistemas geograficamente distribuídos, que não tenham tido na sua concepção a preocupação de interoperarem. Essas capacidades de referenciação, aliadas à crescente utilização da *World Wide Web* como meio preferencial de ligação entre sistemas distribuídos, impulsionaram o desenvolvimento acelerado dos GIS e despertado a sua comunidade de utilizadores para a resolução das questões de interoperabilidade inerentes à heterogeneidade que os caracterizam.

O Consórcio OpenGeospatial (OGC) assumiu um papel preponderante na resolução dos referidos problemas de interoperabilidade, tendo produzido especificações abstractas como o modelo de referência OpenGIS e de implementação como o *Web Map Service* [14] (WMS), o *Web Feature Service* [13] (WFS) e a *Geography Markup Language* [10] (GML) entre muitas outras.

Tanto o WMS como o WFS foram concebidos para suportar o acesso remoto e interoperável a IG, mas enquanto o primeiro (WMS) permite aos clientes a sobreposição de mapas (representações da IG em formato de imagem) provenientes de várias fontes de informação geográfica, o WFS permite a interrogação destas mesmas fontes apresentando o resultado codificado na forma de GML. Se o primeiro é útil para representar grandes volumes de informação de uma forma leve e eficiente (uma imagem que os representa), é no segundo que encontramos o suporte necessário para a elaboração de *queries* com filtros que poderão conter predicados e operações espaço-temporais.

1.3 Suporte temporal nas normas OpenGIS

Actualmente, a dimensão temporal é tratada ainda de forma muito incipiente, pelos WFS e pelos serviços de catálogo (*Catalogue Service* [12]), que encaram esta componente como mais um atributo literal, utilizado nas filtragens que implementam.

Uma vez que a GML contém, na versão 3.1.1, um extensivo suporte temporal (na dimensão *validtime*, uma vez que não inclui tempos de transacção), é possível utilizar as primitivas temporais *gml:TimeInstant* e *gml:TimePeriod*, por forma a obter na resposta a um pedido WFS, *features* dinâmicas (*DynamicFeature* e *DynamicFeatureCollection*) que utilizam a propriedade *history* para expressar o seu desenvolvimento temporal, à custa de *time slices* (*gml:TimeSlice*) que capturam a evolução da *feature* ao longo do tempo.

Poder-se-ia pensar que este suporte é suficiente para lidar com a questão temporal, passando a responsabilidade de efectuar cálculos com lógica temporal (*validtime cross-product*, *validtime selection*, *validtime projection*, entre outros) [3] para a aplicação cliente. No entanto, é opinião dos autores deste artigo, que a transposição da complexidade necessária para o suporte destas operações, do SGBD para a aplicação cliente, é desencorajadora da sua utilização, para além de se tornar extremamente ineficiente. Para além de os SGBD serem muito mais eficientes a realizar estas operações (possuem *cartridges* espaciais; lógica, operações e índices temporais), na maioria dos casos seriam pedidos ao WFS (e consequentemente transportados pela rede) muito mais dados (*features* geográficas) do que o necessário, uma vez que estes poderiam já vir coalescidos sobre a sua dimensão temporal.

Assim, torna-se necessário dotar os WFS da capacidade de filtrar os pedidos recorrendo a predicados e operadores temporais. Ao fazê-lo, estaremos automaticamente a enriquecer as capacidades dos serviços de catalogo [12], uma vez que ambos remetem para a *Common Query Language* (CQL) a definição de uma gramática para filtros. Neste artigo são propostas alterações à *Filter Encoding Implemtation Specification* [11] que é uma codificação em XML derivada da CQL.

2 Definição do suporte temporal

Considere-se uma relação *P*, não temporal, com o esquema (*nome*, *capital*, *fronteira*), que permite representar o estado de países. Neste esquema, o atributo *fronteira* possibilita uma representação poligonal bidimensional, enquanto que os atributos *nome* e *capital* correspondem a uma designação textual. Considere-se ainda que a instância de *P* contém três tuplos que representam o estado actual de três países (Portugal, Polónia e República Checa).

Esta relação *P*, temporalmente designada por *snapshot*, permite capturar apenas o estado actual dos factos. Por exemplo, a mudança da cidade num destes três países, acarreta uma operação de actualização do tuplo correspondente, do atributo *capital*, que passa a conter a designação da nova capital. Neste processo ocorre que a designação anterior (que se mantinha válida até ao instante de actualização) é perdida. O mesmo acontece com a alteração dos limites da fronteira de um desses países, facto

que corresponde, na instância da relação P, à substituição da descrição geométrica da fronteira para esse país, pela descrição geométrica da nova fronteira.

Se o objectivo consiste em manter a informação de histórico acerca de países, então, uma das soluções consiste em temporalizar os factos. Para este efeito, neste artigo adopta-se o modelo *Bitemporal Conceptual Data Model* (BCDM), proposto por Jensen *et al*. [2], embora desse modelo apenas seja contemplado o domínio dos tempos válidos.

Assim, utilizando o BCDM no domínio dos tempos válidos, a uma relação *R,* caracterizada pelos seu conjunto de (A_1, A_2, ... A_n) é acrescentado um atributo temporal T^v ficando $R^v = (A_1, A_2, ... A_n \mid T^v)$. Cada tuplo desta relação $\{a_1, a_2, ... a_n \mid t^v\}$ contém, associado ao conjunto de valores não temporais (a_i), um valor temporal do domínio dos tempos válidos (t^v). No BCDM cada valor de t^v pode conter um conjunto de instantes e/ou de intervalos passados, em que o facto foi considerado válido, ou futuros, quando que se pensa que um facto será válido futuramente.

A transposição deste modelo conceptual para um modelo lógico de base de dados envolve que sejam criados, para cada facto representado, tantos tuplos quantos os intervalos e instantes distintos, que caracterizam o valor de t^v de um tuplo pertencente a uma relação R^v expressa através do modelo BCDM.

Isto significa que a versão unitemporal da relação P corresponde a P_v = (*nome*, *capital*, *fronteira,* t^v), onde t^v representa o intervalo em que os atributos não temporais são considerados válidos. Numa base de dados unitemporal, o atributo de tempos válidos, t^v de um tuplo é representado através de dois valores numéricos que significam os *chronons* [16] limitadores da validade do facto representado nesse tuplo. A título de exemplo considerem-se alguns dos tuplos pertencentes à relação unitemporal P_v:

{'Portugal', 'Coimbra', <geometria$_1$>, [1143-1200)}

{'Portugal', 'Lisboa', <geometria$_1$>, [1200-1250)}

{'Portugal', 'Lisboa', <geometria$_1$>, [1250-1300)}

{'Portugal', 'Lisboa', <geometria$_2$>, [1300-forever)}

Estes quatro tuplos capturam a realidade de que Portugal teve como capital a cidade de Coimbra entre [1143-1200), e que durante esse tempo as fronteiras do país apresentavam a configuração geométrica expressa por `geometria`$_1$. Já a partir de 1200 e até 1300 a capital do país foi Lisboa e durante esse período manteve a configuração da sua fronteira. No entanto, esta informação encontra-se representada através de dois tuplos, temporalmente contíguos. Finalmente, a partir de 1300 a fronteira sofreu alterações passando a ter a configuração `geometria`$_2$ que, pelo facto do limite superior ser um instante futuro, indeterminado, faz deste tuplo o facto actual, isto é, o facto que à data actual se mantém verdadeiro.

Esta abordagem temporal dos factos proporciona que os SGBDs temporais possam manipular múltiplos estados de factos e possibilitam respostas para questões que envolvem cálculos sobre cada estado, mas também cálculos que se realizam sobre vários estados. Se considerarmos a relação P_v contendo factos representados apenas sobre Portugal, é possível realizar questões espaço-temporais, tais como:

1 - Quais as datas em que Coimbra deixou de ser a capital?

2 - Em que períodos foi Lisboa capital?

3 - Qual é a configuração actual da fronteira Portuguesa.

Se considerarmos a relação P_v contendo factos sobre os países referidos, é possível realizar questões espaço-temporais, tais como:

4 - Para o séc. XV, que países eram vizinhos da Polónia?

5 - Alguma vez Portugal e a Polónia foram vizinhos (fronteiras adjacentes)?

6 - Quando existiu a Polónia?

7 – Ao longo dos tempos, qual foi a maior área territorial da Polónia?

Importa referir que, segundo Snodgrass [5], estas questões podem ser realizadas com recurso a SQL em SGBDs não temporais. No entanto, o autor refere que, apesar da linguagem SQL ser efectivamente poderosa para responder a questões que envolvem o estado actual (numa perspectiva *snapshot*), não proporciona o suporte adequado para *queries* temporais, alterações temporais e definição de restrições temporais. Este autor refere a extrema dificuldade de determinar, em SQL, os resultados correctos para as colunas temporais. Segundo Snodgrass, a solução passa por transferir estes cálculos para o SGBD, através de extensões temporais ao SQL.

1.4 Descrição da semântica temporal

Considerando um SGBD temporal e as relações P e $P_{v,}$, é importante referir que a resposta ao conjunto de questões levantadas envolve diferentes semânticas temporais, nomeadamente, *upward compatibility*, *temporal upward compatibility*, semântica sequencial e semântica não sequencial.

Por exemplo, a questão 3 pode ser realizada quer sobre P, quer sobre P_v, envolvendo o mesmo SQL:

```
SELECT fronteira FROM P WHERE nome = 'Portugal'
```

Quando a pergunta é realizada sobre a relação P a semântica temporal envolvida é *upward compatibility*. Neste caso um SGBD temporal responde a questões realizadas em SQL *standard* sobre relações não temporais como se fosse um SGBD não temporal. Este facto proporciona a migração de aplicações que trabalham em SQL standard para SGBDs temporais sem a necessidade de alteração do código.

Quando a pergunta é realizada sobre a relação P_v , a semântica temporal envolvida é *temporal upward compatibility.* A resposta a esta questão é não temporal e corresponde ao cálculo efectuado apenas sobre o estado actual da relação P_v. Todos os outros estados correspondem a factos que, na actualidade, já não são válidos, e são, portanto, ignorados.

Na semântica sequencial, a informação de cada estado do resultado é calculada exclusivamente com base na informação de um só estado, da relação P_v. Estas questões envolvem o uso do predicado temporal, `VALIDTIME`, que pode ser seguido por um período de interesse. Por exemplo:

```
VALIDTIME PERIOD [1400-forever) SELECT * FROM Pv WHERE capital =
'Lisboa'
```

Esta questão é similar à questão 2, mas restringe o período de interesse a partir do ano 1400.

Na semântica não-sequencial a informação de cada estado pertencente ao resultado é calculada a partir de múltiplos estados da relação P_v. Adicionalmente, a lógica

utilizada envolve operadores relacionais não temporais. Por exemplo, na questão 7, com o SQL temporal:

```
NONSEQUENCED VALIDTIME SELECT validtime(P_v),
max(sdo_geom.sdo_area(fronteira, 0.005)) FROM P_v WHERE nome =
'Polónia';
```

só é possível determinar qual é a maior área se forem realizadascomparações com as áreas que constam nos vários estados (tuplos), para os factos que dizem respeito à Polónia. As questões que utilizam semântica não-sequencial obrigam à utilização de um predicado temporal, `NONSEQUENCED VALIDTIME`, que pode ser seguido por um período de interesse.

3 Suporte temporal em WFS

De modo a dotar um WFS com o suporte temporal descrito torna-se é necessário que o WFS disponibilize a utilização de:

- predicados temporais que definem a semântica temporal com que o GIS temporal irá calcular a resposta;
- operadores temporais que actuam sobre o *validtime* de cada tuplo/*feature* contida no GIS temporal;
- operações temporais que actuam sobre resultados quando estes contêm dimensão temporal, como por exemplo, a operação de *coalesce*;
- suporte para a caracterização temporal dos resultados devolvidos pelo WFS aos clientes, isto é de *features* com um *validtime* (já presente no GML 3.1.1).

O suporte de predicados temporais, adicionados a um pedido, realizado por um cliente, prende-se com a necessidade de definir a semântica temporal com que a resposta deve ser calculada. Os predicados temporais contemplados são `SEQUENCED` e `NONSEQUENCED`, sendo que ambos podem ser seguidos por um intervalo temporal de interesse, tendo esse intervalo semânticas distintas nas duas situações. Na primeira situação – `SEQUENCED` - a inclusão de um intervalo tem o significado de restringir o universo de *features* candidatas a serem incluídas na resposta àquelas cujo *validtime* está contido nesse intervalo. Na segunda situação – `NONSEQUENCED` - a inclusão de um intervalo após o predicado tem o significado de caracterizar temporalmente as *features* que resultam do pedido.

A omissão de predicados temporais corresponde à resolução da pergunta com recurso a operadores relacionais não-temporais, sem utilizar o suporte temporal descrito nesta secção.

A utilização do predicado `SEQUENCED` activa a utilização dos operadores de álgebra relacional temporal [3]. Finalmente, a utilização do predicado `NONSEQUENCED` resulta numa utilização dos operadores relacionais não-temporais, que podem actuar sobre o *validtime* de cada feature, encarando-a desprovida de semântica temporal. O resultado de um pedido calculado com esta semântica é não-temporal excepto se, conforme foi referido, se definir um intervalo temporal após o termo `NONSEQUENCED`.

A utilização dos operadores temporais de Allen [7], tais como *meets*, *contains*, *precedes* e *overlaps*, permite restringir as *features* que resultam de um pedido. Estes

operadores actuam sobre o *validtime* de cada *feature* confrontanto-o com instantes e intervalos. Por exemplo, o pedido para obter o nome da cidade que se manteve capital de Portugal, durante o período [1155-1160), não havendo neste período qualquer alteração de outros atributos, tem como filtro:

```
<ogc:Filter>
  <ogc:Sequenced/>
    <ogc:TContains>
      <ogc:QueryValidtime>ValidTime</ogc:QueryValidtime>
      <gml:TimePeriod>
        <gml:begin>
          <gml:TimeInstant>
            <gml:timePosition>1155</gml:timePosition>
          </gml:TimeInstant>
        </gml:begin>
        <gml:end>
          <gml:TimeInstant>
            <gml:timePosition>1160-31-12T23:59:59.999</gml:timePosition>
          </gml:TimeInstant>
        </gml:end>
      </gml:TimePeriod>
    </ogc: TContains>
</ogc:Filter>
```

Se esta pergunta for realizada sobre a relação unitemporal P_v, a resposta consistirá no tuplo onde está representado o facto de que Coimbra foi capital de Portugal, porque o *validtime* desse tuplo contém (ogc:TContains) o *validtime* definido na query – [1155-1160).

Também se considera útil a inclusão da operação de *coalesce* no suporte temporal a incluir no WFS pela possibilidade de solicitar resultados coalescidos pela dimensão dos tempos válidos, isto é, para a obtenção dos resultados numa forma em que os intervalos *validtime* são maximizados para *features* que se intersectam temporalmente e são *value-equivalent* para atributos não temporais [8]. Por exemplo a operação de *coalesce* da relação unitemporal P_v resulta em:

{'Portugal', 'Coimbra', <geometria$_1$>, [1143-1200) }

{'Portugal', 'Lisboa', <geometria$_1$>, [1200-1300) }

{'Portugal', 'Lisboa', <geometria$_2$>, [1300-forever) }

O Filtro usado nesta situação deve corresponder a:

```
<ogc:Filter><ogc:Sequenced><ogc:Coalesce/></ogc:Sequenced></ogc:Filter>
```

Onde os tuplos de P_v com o *validtime* de [1200-1250) e de [1250-1300), pelo facto dos seus atributos não temporais, *nome*, *capital* e *fronteira*, serem equivalentes (quando comparado os valores atributo a atributo) resultam num tuplo coalescido temporalmente, [1250-1300).

Finalmente, considera-se necessário contemplar que os resultados devolvidos por um WFS, possam conter, para cada *feature* devolvida, um intervalo temporal *validtime*. Para este efeito, conforme foi referido na introdução, a versão 3.1.1 da especificação GML [10], contempla o suporte para a caracterização temporal de uma *feature* geográfica ou de um outro qualquer objecto, através da inclusão do *schema temporalReferenceSystems.xsd,* que por sua vez inclui o *schema temporal.xsd*. Neste último, está já definido o elemento *gml:validTime*, que compreende a caracterização temporal de um instante e de um período, por meio de múltiplas formas de representação destas grandezas temporais.

4 Especificação das alterações à implementação de um WFS

Propõe-se que o suporte para predicados temporais, operadores temporais e funções temporais seja adicionado à especificação *Filter Encoding Implemtation Specification* [11] (*Filter*), sendo esta a definição de uma codificação em XML para a realização de expressões de filtragem, baseada na definição *Backus Naur Form* [15] (BNF) da OCG *Common Catalog Query Language* [12] (CQL), tal como é definida na OGC *Catalogue Service Implementation Specification* [12]. Fundamenta-se esta opção no facto da *Filter*:

- ser utilizada na constituição de filtros que são parte integrante de perguntas realizadas a um WFS [13], através de um pedido *GetFeature* (Figura 1). Destas perguntas resultam conjuntos de *features*, que obedecem aos critérios definidos na filtragem.
- ser utilizada nos pedidos *LockFeature, GefFeatureWithLock,* bem como nas operações *Update* e *Delete*, que estarão disponíveis se o WFS apresentar capacidades transaccionais (WFS-T). De facto, verificou-se que a definição dos pedidos identificados compreende a utilização de Filtros de modo a restringir o conjunto de *features* alvo da operação ou do pedido.
- ser utilizada tanto no *Catalogue Service* como no *Web Feature Service*: actualmente, o suporte temporal previsto consiste somente na utilização de filtros literais sobre os metadados que foram utilizados para catalogar os OpenGIS Web Services (OWS) existentes. Através desta funcionalidade é possível, por exemplo, inquirir um *Catalogue Service* sobre os WFS e os WMS que disponibilizam dados hidrológicos sobre Portugal, entre 1 de Janeiro de 1996 e 31 de Dezembro de 2003.

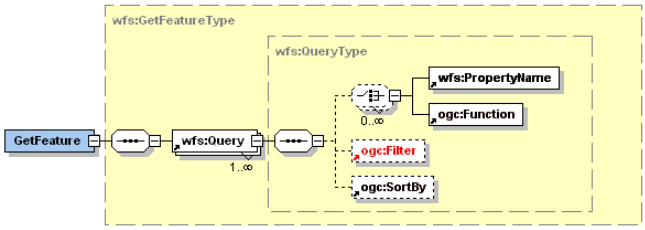


Figura 1: *schema* de um pedido *GetFeature*

Torna-se ainda necessário propor alterações à componente do WFS Capabilites, (Figura 2), que descreve as suas capacidades de filtragem, por forma a informar o cliente que o WFS em questão possui capacidades de filtragem utilizando predicados, operadores e operandos na dimensão unitemporal *validtime*.

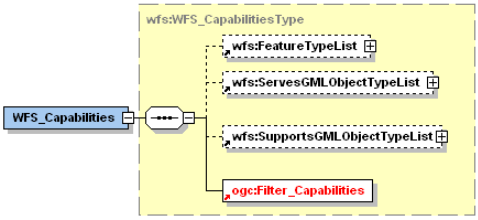


Figura 2: *schema* do pedido *getCapabilities*

Ao nível do *ogc:Filter*, o *schema* permite uma escolha de operações (ver Figura 3) de filtragem espaciais (*ogc:spatialOps*), de operações de filtragem de comparação (*ogc:comparisonOps*), de operações de filtragem pela identificação de features (*ogc:_Id*) e de operações lógicas (*ogc:logicOps*), sendo que estas últimas permitem a composição de várias operações através dos operadores *ogc:and*, *ogc:or* e *ogc:not*.

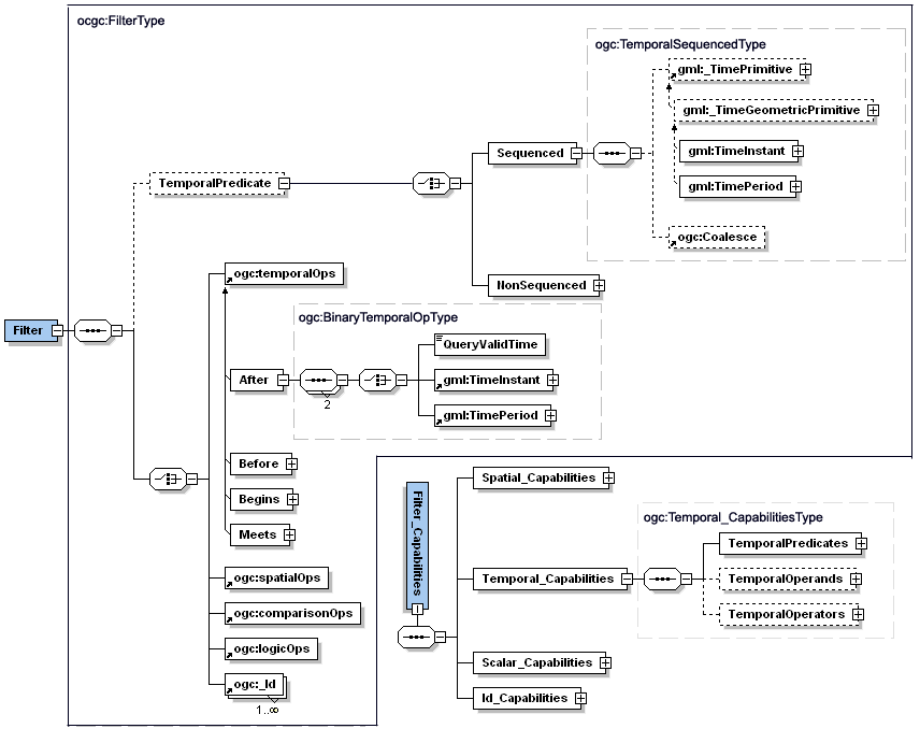


Figura 3: Proposta de alteração dos *schemas Filter* e *Filter_Capabilities*

As operações espaciais permitem filtragem espaciais sobre as propriedades geográficas de uma *feature*, por exemplo, *ogc:BBOX*, *ogc:Disjoint* e *ogc:DWithin.*

As operações de comparação, por exemplo, *PropertyIsEqual*, *PropertyIsLike*, *PropertyIsLessThan*, permitem realizar filtragem de *features* pela comparação entre o valor da propriedade e um ou mais literais (dependendo do número de operandos que a comparação permite).

As operações lógicas compreendidas no *ogc:Filter* permitem realizar operações lógicas binárias ou unárias entre operandos que podem ser operações espaciais, de comparação ou mesmo outras operações lógicas.

Finalmente, o *ogc:Filter* contempla a filtragem de uma *feature* pelo seu atributo identificador, através da definição dos elementos *ogc:FeatureId* (*deprecated*) e *ogc:GmlObjectId.*

Sobre esta especificação propomos:

- preceder o conjunto das operações acima descritas por um novo elemento, opcional, designado *ogc:TemporalPredicate;*

- acrescentar ao conjunto de operações acima um novo elemento que consiste nas operações temporais disponibilizada, por exemplo, *ogc:After*, *ogc:Before*, *ogc:Meets*, *ogc:Begins* e *ogc:TContains*. Nesta última (tal como na operação *ogc:TOverlaps*) a designação é precedida pela letra *'T'* para distinguir das operações espaciais com o mesmo nome;
- definir nas operações lógicas mais um tipo de operandos: as operações temporais;

O novo elemento *ogc:TemporalPredicate*, opcional, é composto por uma escolha entre os elementos *ogc:Sequenced* e *ogc:NonSequenced*. A utilização do primeiro equivale à especificação de semântica temporal `SEQUENCED` no cálculo da resposta, e pode conter dois elementos opcionais: um *gml:_TimePrimitive* e um elemento *ogc:Coalesce*. A utilização do *gml:_TimePrimitive*, que pode ser um *gml:TimeInstant* ou um *gml:TimePeriod* tem o significado de restringir temporalmente o cálculo da resposta às *features* cujo *validtime* está contido nesse elemento temporal. A utilização do elemento *ogc:Coalesce* significa que os resultados devem vir *coalescidos* sobre a dimensão dos tempos válidos.

O elemento *ogc:NonSequenced* equivale à especificação de semântica temporal `NONSEQUENCED`, pelo que o SGBD subjacente deve calcular a resposta usando lógica não temporal, tratando o atributo *validtime* das *features* como um qualquer atributo não temporal. O elemento *ogc:NonSequenced* pode ainda conter um elemento *gml:_TimePrimitive*. Nesta situação, apesar do cálculo ser à custa de lógica não temporal, cada *feature* do resultados deve ser temporalizada com um *validtime* equivalente ao elemento representado pela *gml:_TimePrimitive*. Este é um processo de promover resultados *snapshot* a resultados *validtime*.

Na constituição de um filtro, a ausência do elemento *ogc:TemporalPredicate* tem o mesmo significado dos WFS não temporais, mantendo assim *backward compatibility*.

Por último, relativamente ao *ogc:Filter*, cada operação temporal definida no elemento *ocg:TemporalOps* é um operador de Allen [7]. Nestes operadores binários, propõe-se que os operandos sejam *ogc:BinaryTemporalOpType*, cujo tipo é restringido a *gml:TimeInstant* ou *gml:TimePeriod*.

Ao nível do *schema* do *ogc:Filter_Capabilities* a proposta de alteração é uma consequência das alterações definidas para o *ogc:Filter*, isto é, no *Filter_capabilities* propõe-se adicionar mais um elemento, *ogc:TemporalCapabilities*, que lista as capacidades temporais do filtro, e que consistem nos predicados temporais (*ogc:TemporalPredicates*), nos operandos temporais (*ogc:TemporalOperands*) e nas operações temporais (*ogc:TemporalOperators*). Neste *schema* está ainda definido que a utilização de capacidades temporais obriga à utilização de um predicado temporal.

A Figura 3 resume as propostas apresentadas neste artigo, de alteração à definição dos *schemas ogc:Filter* e *ogc:Filter_Capabilites*, no sentido de incluir o suporte temporal na dimensão de tempos válidos. Todos os *schemas* de base utilizados poderão ser encontrados no endereço http://schemas.opengis.net/. As alterações propostas, baseiam-se na versão 1.1.0 da Filter e poderão ser obtidas na sua totalidade a partir do endereço http://gis.inescporto.pt/schemas/filter/1.1.0/.

5 Resultados esperados

Por forma a ilustrar o que seria esperado de um filtro espaço-temporal, que seria utilizado num pedido WFS, considere-se o seguinte exemplo.

```
<ogc:Filter>
  <ogc:Sequenced>
        <ogc:Coalesce/>
  </ogc:Sequenced>
  <ogc:and>
    <ogc:during>
      <ogc:QueryValidtime>ValidTime</ogc:QueryValidtime>
      <gml:TimePeriod>
        <gml:begin>
          <gml:TimeInstant>
            <gml:timePosition>1901-01-01</gml:timePosition>
          </gml:TimeInstant>
        </gml:begin>
        <gml:end>
          <gml:TimeInstant>
            <gml:timePosition frame="#ISO-8601">
              2000-31-12T23:59:59.999
            </gml:timePosition>
          </gml:TimeInstant>
        </gml:end>
      </gml:TimePeriod>
    </ogc:during>
    <ogc:BBOX>
      <gml:PropertyName>Fronteira</gml:PropertyName>
      <gml:Envelope srsName="http://www.opengis.net/gml/srs/epsg.xml#4326">
         <gml:lowerCorner>13.345 31.213</gml:lowerCorner>
         <gml:upperCorner>35.345 42.573</gml:upperCorner>
      </gml:Envelope>
      </ogc:BBOX>
  </ogc:and>
</ogc:Filter>
```

Este filtro utiliza lógica temporal sobre a dimensão *validtime* e restringe as *features* resultantes àquelas cujo *validtime* contém o século XX (no calendário Gregoriano, ISO-8601) e cuja localização geográfica se encontra compreendida na *bounding box*, [{13.345 31.213}, {35.345 42.573}], expressa em coordenadas geográficas no SRS WGS84 (EPSG:4326). Pretende-se ainda que os resultados devolvidos pelo WFS sejam coalescidos sobre a dimensão dos tempos válidos.

6 Conclusões

Este artigo apresenta uma proposta de inclusão de suporte unitemporal `VALIDTIME` nos filtros de WFS. Este suporte caracteriza-se pela utilização de predicados, operandos e operadores temporais, sobre a dimensão dos tempos válidos e deve estar contemplado na definição dos *schemas ogc:Filter* e *ogc:Filter_Capabilites*. Estes *schemas* são ainda utilizados pelo *Catalogue Service*, por forma a filtrar os OpenGIS Web Services (OWS) registados num serviço de catálogo pelos seus metadados, providenciando assim um suporte temporal transversal a todos os OWS considerados.

A obtenção dos resultados (*Features* e *FeatureCollections*) temporais encontra-se garantida, na medida em que o GML 3.1.1, disponibiliza suporte temporal através de tipos temporais (*TimeInstant* e *TimePeriod*) e de *features* com atributos dinâmicos.

As alterações propostas serão validadas por recurso a uma implementação específica de um "Temporal WFS" por forma a confirmar os testes realizados sobre a implementação espaço-temporal já realizada ao nível do SGBD.

7 Bibliografia

[1] Date, C., *An Introduction to Database Systems*, 7th edition, Addison-Wesley, Capítulo 22, pp. 730-768, 2000

[2] Jensen, C., *Temporal Database Management*, PhD. Thesis, 2000

[3] Snodgrass, R., Böhlen, M., Jensen, C., Steiner, A., *Adding Valid Time to SQL/Temporal*, SQL/Temporal Change Proposal, ANSI X3H2-96-501r2, ISO/IEC JTC1/SC21/WG3 DBL MAD-146r2, 1996.

[4] Snodgrass, R., Böhlen, M., Jensen, C., Steiner, A., *Adding Transaction time to SQL/Temporal*, SQL/Temporal Change Proposal, ANSI X3H2-96-502r2, ISO/IEC JTC1/SC21/WG3 DBL MAD-147r2, 1996

[5] Zaniolo, C., Ceri, S., Faloustos, C., Snodgrass, R. Subrahmanian, V., Zicari, R*., Advanced Database Systems*, Morgan Kaufman, 1997.

[6] Steiner, A., *A Generalization Approach to temporal Data Models and Their Implementations*, Phd, Zurich, 1998

[7] Allen, J. *Maintaining Knowledge about Temporal Intervals*, Comm. ACM 26(11), 1983.

[8] Böhlen, M., Snodgrass R., Soo, M., *Coalescing in Temporal Databases*, Proceedings of International Conference on Very Large Databases, 1996.

[9] ISO 19108:2002, *Geographic Information – Temporal Schema*, Ref: N1224, 2002

[10] ISO/TC 211/WG 4/PT 19136, OGC GML RWG, *Geographic information – Geography Markup Language (GML)*, versão 3.1, 2004

[11] OGC 04-095, *Filter Encoding Implementation Specification*, versão 1.1.0 , Vretanos, P., 2005

[12] OGC 04-021r3, *OGC Catalogue Services Specification*, versão 2.0.0, Nebert, D., Whiteside , A., Vretanos, P., 2005

[13] OGC 04-094, *OGC Web Feature Service Implementation Specification*, versão 1.1.0, A., Vretanos, P., 2005

[14] OGC 04-024, *Web Map Service*, versão 1.3, Beaujardiere, J., 2004

[15] ISO/IEC 14977:1996, *Information technology - Syntactic metalanguage - Extended BNF*, 1996

[16] Jensen, C., *et al.*, *A consensus glossary of temporal data base concepts*, ACM SIGMOD Records, vol. 23, 1994

[17] Steiner, A., *A Generalization Approach to temporal Data Models and Their Implementations*, Phd, Zurich, 1998

[18] Dumas, M., Fauvet, C., Scholl, P., *TEMPOS: A Temporal Database Model Seamless Extending ODMG*, LSR-IMAG, University of Grenoble, Phd, 2000

The Case for Aspect Oriented Programming

André Restivo <arestivo@fe.up.pt>

Faculdade de Engenharia da Universidade do Porto

Abstract. Aspect Oriented Programming (AOP) deals with what are called cross-cutting concerns. AOP practitioners believe that single abstraction frameworks (like OOP) are not sufficiently powerful to separate cross-cutting concerns. They also state that the tangling of concerns is one of the major contributors to the complexity of large software applications. This paper will show some typical examples of concerns that are difficult to separate from the main core of the application. It will also exemplify how AOP can be used to describe each one of those concerns in a separate and natural form. Finally an overview of where AOP research is heading towards will be described.

1 Introduction

One of the major goals of Software Engineering has always been achieving a clear Separation of Concerns (SoC). By separating each one of the software components into well defined units, a number of advantages, like higher re-usability of coding, unit testing and many others, arise. The emergence of Object Oriented Programming (OOP), as opposed to Procedural Oriented Programming (POP), was clearly a step forward in the attempt to achieve an easier and more comprehensible programming environment.

However, practice shows that even following the most rigid OOP prerogatives, total SoC is not always possible to achieve, because some concerns seem to get inevitably tangled throughout the code [1]. This happens because OOP is a single abstraction paradigm, meaning that we are forced into organizing the code following a single perspective. The core concerns of the applications force other, less important, concerns (cross-cutting concerns) to get scattered throughout the application. It feels like trying to solve the *Rubik's Cube* by solving one face at a time[1].

Aspect Oriented Programming (AOP) advocates that it is possible to build upon the OOP paradigm creating a richer framework[2]. AOP is a new programming technique where *aspects* (cross-cutting concerns) are captured in their own units of modularity and then *weaved* together, and in this way creating the application through a process of composition [2]. Weaving can be done at many

[1] When solving the *Rubik's Cube* one normally solves all faces in parallel. If we try to solve one face at a time, solving the second face will destroy the first one.

[2] AOP is not an evolution of the OOP paradigm and can also be easily applied over the POP paradigm.

levels but at the moment it is mainly seen as an integrating part of the compiling process. At compilation time, aspects and units are weaved together creating a tangled version of the source code, then this code is compiled as usual generating a binary for the application (see Figure 1).

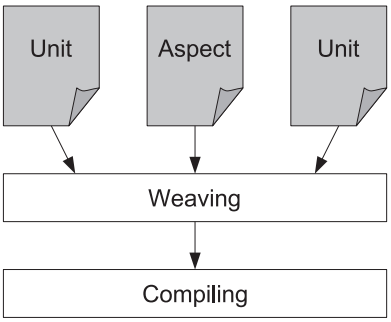


Fig. 1. The weaving process

1.1 Implementation

In order to be possible to specify where units and *aspects* should be weaved together a set of points in the code where weaving can occur must be specified. These are called *joinpoints*. Each AOP language can define its own set of *joinpoints*. For instance, *AspectJ*, an AOP language based in Java and the current de facto standard for AOP, defines the following *joinpoints: Method call* and *execution*, *Constructor call* and *execution*, *Read/write access to a field*, *Exception handler execution,* and *Object* and *class initialization execution*. In the remaining of this section we will see what other characteristics are implemented by AspectJ in order to transform the Java language into a AOP language.

As cross-cutting concerns normally are scattered throughout the application, most of the times we want to weave an *aspect* to several different *joinpoints*. *AspectJ* introduced the notion of *pointcut designators* which allow the filtering of *joinpoints*. A *pointcut designator* refers to several *joinpoints* and can be defined with the help of wildcards.

In AspectJ the element that defines how an application behaviour is altered in order to implement a certain *aspect* is the *advice*. Advices are code blocks that execute implicitly whenever a certain *joinpoint*, belonging to the pointcut associated to it, is reached (see Figure 2). AspectJ defines three types of advices: *after*, *before* and *around*. *After* advices run just after the joinpoint, *before* advices run before the actual joinpoint code is executed and *around* advices have control over the joinpoint execution.

Finally, the last element introduced by *AspectJ* is the *aspect*. An *aspect* combines *pointcuts* and *advices* composing a single cross-cutting unit.

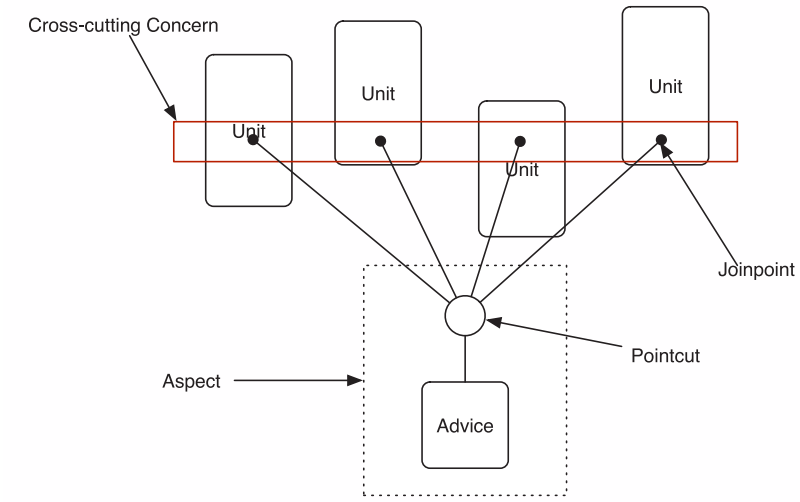


Fig. 2. Joinpoints and Pointcuts

1.2 Advantages

Literature identifies the following as the main advantages of using AOP [3]:

Explicitness *Cross-cutting concerns* are explicitly captured by aspects;
Reusability. It is possible through a single *aspect* to describe crosscutting concerns common to several components;
Modularity. Since *aspect*s are modular crosscutting units, AOP improves the overall modularity of an application;
Evolution. Evolution becomes easier since implementation changes of cross-cutting concerns occur locally within an *aspect* and save the need to adapt existing classes;
Stability. Special AOP language support makes it possible to express generic *aspects*, which will remain applicable throughout future class evolution;
Pluggability. Since *aspects* are modular, they can be easily plugged in and out of an application.

1.3 Structure

In the next Section some typical examples of cross-cutting concerns will be described as well as how they can be implemented using AOP. Some of the current issues being researched at the moment will be depicted in Section 3. And finally conclusions will be drawn in Section 4.

2 Examples of Cross-Cutting Concerns

The most typical example that can be found in AOP introduction papers and tutorials is the *Logging* concern example. This fact is not that peculiar as this example has all the major characteristics for a good application of AOP techniques. To follow the trend this section will start by describing how AOP can be used in that particular scenario and then other possible uses of AOP will be discussed.

2.1 Logging

Most software systems write information about its actions to log files. Several reasons make *logging* the perfect candidate for AOP.

Due to its own nature, logging code gets scattered throughout the application even if a good architecture is used. Besides that, logging is something that tends to change often.

Most logging is done using a specific logging package. The need of changing the logging package should be something that developers should anticipate, if the logging code is spread all over the application these changes could prove lengthy and complicated.

The most important factor off all is that logging, normally, is not a core concern of the application. This makes it even more important to separate this code from the code that really matters. For example, imagine a *class* named *HTTPConnection*, responsible for handling an HTTP Connection from a single client. If it is necessary to log each connection made by a client, and probably it will, possible places to put that code would be the *HTTPConnection* constructor or the class *connect* method. However that would subvert the objective of those functions. Class constructors only should know how to create a new instance of the class and should not be *concerned* with the logging *aspect* of the application, the same can be applied to the *connect* method.

In order to separate the *connection handling* code from the *logging* code, one only needs to *weave* the logging code into the *joinpoint* defined by the *HTTPConnection* connect method. Listing 1.1 shows how this can be done by defining an *aspect* named *HTTPConnectionLogging*. This aspect uses an *advice* that executes *before* any *call* to the connect method of the *HTTPConnection* class. This *advice* then simply uses the local variable *logger,* defined in this same *aspect,* to write a new line to the log file.

This will allow the HTTPConnection class to remain free from any code concerning the logging *aspect* of the application.

2.2 Security and Authentication

Security plays a major role in many software applications. It is also a good candidate for AOP as its scope crosses the entire application.

AOP could be used, for example, to control objects owners and permissions in a non pervasive way[4]. This could be accomplished by associating an *aspect*

Listing 1.1. Logging Aspect implemented with AspectJ

```
1 aspect HTTPConnectionLogging {
2     Logger logger = Logger.getDefaultLogger();
3
4     before() : call(void HTTPConnection.connect(IP clientIP,
          Date date))
5     {
6         logger.notice(date.toString()+": "+clientIP);
7     }
8 }
```

to each constructor of the objects whose access has to be restricted, this *aspect* would automatically assign ownership to the current system user.Another *aspect* would be associated to each restricted access method of those same objects, and would verify the ownership of the object (See Listing 1.2).

Another security concern that could be tackled with AOP would be encryption. One could develop an application without thinking about how data would be transmitted in a safe way and then add an *aspect*, around each message sending method, that would encrypt data before it is sent. The same could be done for each message receiving method. This would allow developers to focus in the core concerns of the application and would even allow changing the encryption algorithm very easily.

2.3 Caching

In many cases database access is the bottleneck of an application. One way of dealing with this issue is to cache data that is retrieved very often. Caching can also be seen as a cross-cutting concern as it spreads throughout the code and can be removed without any loss in functionality.

The implementation of data caching with AOP can be done at several levels. One could, for example, weave an *aspect* around each query to the database, weave an *aspect* around a generic query method, or even weave an *aspect* around each object retrieval method (See Listing1.3). These *aspects* would then be responsible for verifying if the object, or query, is cached and if not proceed with the operation.

2.4 Contract Enforcement

Design by Contract (DbC) was first introduced by B. Meyer in 1992[5]. DbC specifies that modules should have formalized obligations, described as sets of rules, regarding their interaction with other modules.

At first glance DbC seems another perfect candidate for AOP implementation. For example, one could create an *aspect* that would verify pre and post-conditions in a certain method (See Listing 1.4).

Listing 1.2. Authentication Aspect implemented with AspectJ

```
1  aspect OwnerManagement perthis(this(Article)){
2          String owner;
3
4          after(): execution(Article.new(..)){
5                  owner = Authentication.getUser() ;
6          }
7  }
8
9  aspect Authorization(){
10         pointcut restrictedAccess():
11                 execution(* Article.update(..)) ||
12                 execution(* Article.delete(..));
13
14         void around() : restrictedAccess(){
15                 if(! OwnerManagement.aspectOf(thisJoinPoint.
                       getThis()).owner.equals(Authentication.
                       getUser()))
16                 System.out.println("Access␣Denied␣!") ;
17                 else proceed() ;
18         }
19 }
```

Listing 1.3. Caching Aspect implemented with AspectJ

```
1  aspect Caching {
2      CacheManager manager = CacheManager.getDefaultManager();
3
4      around() : call(void Account.retrieveAccount(int accountId
           ))
5      {
6          if (manager.isCached("Account", accountId))
7                  return manager.getCachedObject("Account",
                       accountId);
8          else
9                  proceed();
10     }
11 }
```

Listing 1.4. Contract implemented with AspectJ

```
1 aspect MathContract {
2         before : (double x): sqrt(x) {
3                 if ( x < 0 )
4                         throw new IllegalArgumentException("
                              negative not allowed");
5         }
6 }
```

However, the same B. Meyer argues *aspects*, although having some advantages like allowing easy contract reutilization, are not as powerful and safe to use as a proper contract oriented implementation [3]. This happens because contracts developed using AOP do not support inheritance based refinements. Meyer also claims that contracts are not cross-cutting concerns.

This debate about AOP applicability is an important one. As with all development techniques, applying AOP in the wrong scenarios could lead to an unnecessary increase of the system complexity. When thinking about using AOP one should analyze:

- if the concern being implemented is really a cross-cutting concern;
- if it is possible to implement the concern in a modular form using OOP without adding too much complexity;
- if the concern is a core concern of the application (normally secondary requirements are better candidates for AOP);
- if reusability and pluggability are important issues for this particular concern;
- if changes in the units this concern connects with will also imply changes in the concern itself or if the concern depends semantically on the units;

Only after this kind of reasoning one could expect to extract the benefits described in Section 1.2 from AOP.

2.5 Flavours

Several software developing companies have products that come in different *flavours*. Sometimes these different *flavours* have more or less functionalities (and different prices) and sometimes they are adapted to different types of customers (or even tailored to a specific customer). Either way having different versions of the same application has a tremendous impact in the software developing process as changes to the base version must be tested against each one of the different especial versions.

If the differences between these versions can be encapsulated as *aspects*, then by simply removing and adding *aspects* we could have different versions with different capabilities making the development process much simpler.

3 Future Developments

AOP is still a very immature paradigm and a lot of work still needs to be done to bring it into mainstream programming. It took about about 20 years for OOP to become the dominant programming methodology (from Simula 67 and Smalltalk to C++). AOP stands now like OOP did in the mid sixties. In this section some of the current research topics in AOP will be described.

3.1 Tools

The non-linearity of AOP makes debugging harder than in classical OOP programming. This happens because code being run is not the same that as been written by the programmer as the weaving process already has cross-cutting concerns tangled throughout the code. For example, stack traces will be more difficult to understand. AOP programmers need new tools and methods for debugging their applications. Besides that, unit-testing can not be applied to cross-cutting concerns in the same way it is applied to units developed with the OOP paradigm. So, some research into how unit-testing methods and applications can be adapted to the AOP world has to be done.

The integration of AspectJ into the Eclipse IDE[3] (AJDT) has been a big step forward in the AOP evangelizing process [6]. For example, with AJDT Visualizer, one can see how a cross-cutting concern is scattered throughout the code (See Figure 3). However further research must be done into what new tools AOP developers will need and how they can be integrated into existing IDEs.

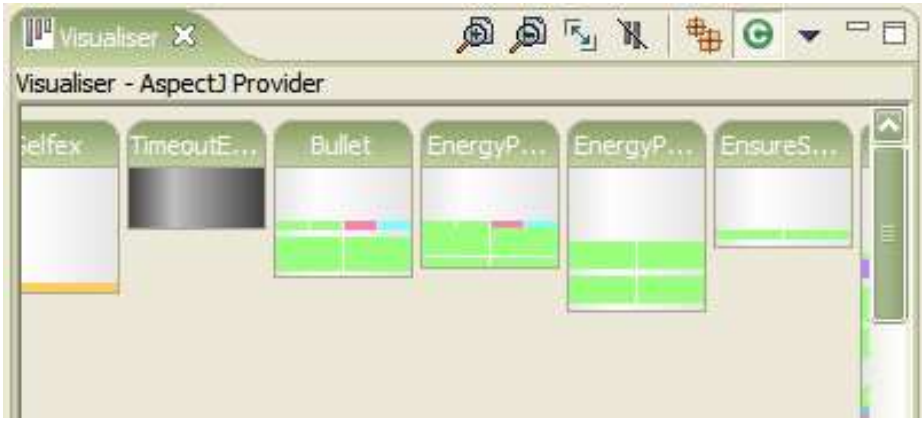


Fig. 3. AJDT Visualizer

[3] Integrated Development Environment

3.2 Software Design

The Design Patterns [7] revolution created an universal language that made thinking and talking about OOP implementation solutions a lot easier. AOP poses two new challenges in this area: are there any AOP specific design patterns, and are there some OOP design patterns that are easier implemented using AOP [8,9].

Refactoring is another major trend in the software development field. First introduced by M. Fowler[10], software refactoring is a set of methods that help developers correct software design mistakes in existing code, normally by transforming the existing code into Design Patterns. Refactoring of existing OOP code into the AOP paradigm is a new field where research is already under way [2].

3.3 Documentation and Specification

A lot of effort has been done in the area of software documentation including, for example, automatic code documentation. With the introduction of AOP and *aspects*, code documentation will have to evolve and accommodate the concepts introduced by the new paradigm.

Finally, *aspects* should not be represented only at code level but also at higher levels of abstraction. The currently most used software modeling language, UML[4], isn't prepared to represent crosscutting concerns in its various diagrams. Some work has already been done in extending UML to incorporate these kind of concerns [11,12] but some more research must be done in this field. Tools supporting these new extensions allowing modeling, code generation and reverse engineering must be created.

4 Conclusions

This paper has shown that AOP can be used to encapsulate cross-cutting concerns into modular units in a way that OOP is not capable of. Examples have been used to explain how cross-cutting concerns can be modeled using AOP.

It has also been shown that the applicability of AOP is not always unanimous. When using AOP the developer must rationalize if the aspect being modeled is really a cross-cutting concern and if using AOP will help in the developing process or increase the complexity of the system without any real gain.

By describing several future developments in the field, AOP has been described as an interesting research field with many possible research topics.

Acknowledgements

The author would like to thank Ademar Aguiar for the inspiring introduction to the world of *Aspect Oriented Programming* and Sérgio Carvalho for the endless discussions on AOP scenario applicability.

[4] Unified Modeling Language

References

1. Kiczales, G., Lamping, J., Menhdhekar, A., Maeda, C., Lopes, C., Loingtier, J.M., Irwin, J.: Aspect-Oriented Programming. In Akşit, M., Matsuoka, S., eds.: Proceedings European Conference on Object-Oriented Programming. Volume 1241. Springer-Verlag, Berlin, Heidelberg, and New York (1997) 220–242
2. Monteiro, M.: Refactorings to Evolve Object-Oriented Systems with Aspect-Oriented Concepts. PhD thesis, Universidade do Minho (2005)
3. Balzer, S., Eugster, P.T., Meyer, B.: Can Aspects Implement Contracts? In: Proceedings of RISE 2006 (Rapid Implementation of Engineering Techniques). (2006)
4. Win, B.D., Joosen, W., Piessens, F.: (Developing Secure Applications through Aspect-Oriented Programming)
5. Meyer, B.: Applying "Design by Contract". Computer **25**(10) (1992) 40–51
6. : (AspectJ Development Tools (AJDT)) http://www.eclipse.org/ajdt/.
7. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design patterns: elements of reusable object-oriented software. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1995)
8. Hachani, O., Bardou, D.: (Using Aspect-Oriented Programming for Design Patterns Implementation)
9. Hannemann, J., Kiczales, G.: Design Pattern Implementation in Java AspectJ. In: Proceedings of the 17th Annual ACM conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA). (2002)
10. Fowler, M.: Refactoring: Improving the Design of Existing Code. Addison-Wesley Professional (1999)
11. Kandé, M.M., Kienzle, J., Strohmeier, A.: (From AOP to UML: Towards an Aspect-Oriented Architectural Modeling Approach)
12. Suzuki, J., Yamamoto, Y.: Extending UML with Aspects: Aspect Support in the Design Phase. In: ECOOP Workshops. (1999) 299–300

Métodos para reconstrução automática de objectos fragmentados

António José Marques
FEUP, Universidade do Porto
Rua Dr. Roberto Frias s/n
4200-465 Porto, Portugal
ajm@ispgaya.pt

Resumo A reconstrução de objectos fragmentados é uma necessidade em âmbitos diversos. Áreas como as da arqueologia, investigação criminal e medicina legal, entre outras, têm abordado a resolução do problema com o auxílio de sistemas computacionais. Contudo, as soluções apresentadas revelam-se tipicamente orientadas à resolução de problemas específicos. Assim, a diversidade de métodos desenvolvidos parece contrastar com a dificuldade de determinação de um produto abrangente e capaz de gerar soluções de modo eficaz. No presente procurar-se-á atingir objectivos a três níveis. Numa primeira fase, tentar-se-á recolher trabalhos representativos de esforços para a resolução do problema. Numa segunda fase, serão seleccionados os que se considerem mais relevantes para análise. Após análise comparativa dos trabalhos seleccionados, tentar-se-á utilizar o conhecimento recolhido para a proposta de um método que permita abordar o problema de forma genérica e eficaz.

1 Introdução

O problema que se pretende analisar com o trabalho apresentado poderá ser genericamente descrito como o da reconstrução auxiliada por computador de objectos desconhecidos que foram quebrados ou rasgados, resultando num grande número de fragmentos irregulares. Para início de resolução do problema enunciado, deverá ser sempre obtida uma representação em formato digital dos fragmentos a considerar. Partindo dessa informação, deverá ser desenvolvido o tratamento adequado à determinação ou verificação das características do objecto original, assim como a posição e orientação de cada fragmento no mesmo. Durante o processo, terão de ser feitas todas as verificações necessárias à minimização da probabilidade de obtenção de falsas soluções. Se aparentemente o problema é de simples resolução, na prática, há um conjunto diverso de aspectos responsáveis pela complexidade que depois se verifica. Em particular:

– O espaço de pesquisa da localização de um dado fragmento no objecto original pode ter até 6 graus de liberdade. Repare-se que, no espaço 3D, a colocação de qualquer fragmento no objecto original poderá estar sujeita a translações em 3 eixos, tal como a rotação em torno de qualquer dos 3 eixos do sistema de referência.

- No caso de um documento rasgado ou de um objecto plano fragmentado, como um mosaico, apenas se verifica o desconhecimento de 3 destes valores. Respectivamente, dois associados ao deslocamento ao longo de um plano e um terceiro associado à rotação em torno de um eixo perpendicular ao mesmo.
- Quando se trate de um objecto de características puramente tridimensionais -como uma peça de escultura, todos os graus de liberdade terão de ser considerados para determinação da posição original relativamente aos restantes.
- As superfícies digitalizadas são frequentemente imprecisas e geradoras de ruído.
- Aspectos como rebarbas geradas no rasgar de papel ou tecidos e estilhaços demasiado pequenos para que possam ser considerados como fragmentos concretos são aspectos introdutores de ruído, por implicarem um ajuste impreciso dos fragmentos.
- Fragmentos pequenos que tenham sido realmente perdidos ou deformações provocadas por erosão ou exposição a factores físicos ou químicos como elevadas temperaturas, corrosão ou sujeição a esforços mecânicos, são outros aspectos responsáveis pela necessidade de tratamento do processo como algo inerentemente impreciso.
- A quantidade de informação a processar pode ser muito elevada, obrigando a enormes recursos computacionais para resolução do problema em tempo útil. Neste caso e em particular, a digitalização de fragmentos de objectos tridimensionais pode levar a situações do tipo descrito.
- Não existe, tipicamente, nenhum aspecto genérico que possa à partida ser utilizado para redução do âmbito de pesquisa de possíveis soluções.
- Com excepção de métodos que utilizem associação a bases de dados de objectos expectáveis com pormenores concretos que possam ser procurados nos fragmentos para determinação inicial de uma localização e posição aproximadas, verifica-se o condicionalismo referido.
- Existência de fragmentos com superfícies sem correspondência directa com outras superfícies.
- Necessidade de detecção de intersecção de superfícies num âmbito global, para eliminação de falsas soluções.
- Não só é frequente a indisponibilidade de todos os fragmentos de um objecto original, responsável por situações como a descrita, como podem ainda existir imiscuídos fragmentos de outros objectos.

O último caso, para além de frequente em alguns cenários como os arqueológicos, pode tornar ainda mais complexa a resolução do problema. Repare-se que, em qualquer dos casos, qualquer solução apresentada terá de contar para além da inexactidão intrínseca, com objectos incompletos ou fragmentos a descartar, indutores de soluções falsas.

2 Trabalho relacionado

Para início do desenvolvimento foram recolhidos vários trabalhos existentes. Foi possível constatar, desde logo, a grande quantidade de artigos publicados abor-

dando o assunto. Porém, esse conjunto de artigos surgiu associado a um relativamente restrito número de investigadores. Sendo cada grupo de investigação afecto tipicamente a um método, a divisão e afectação de trabalhos a grupos de análise foi algo simplificada.

2.1 Métodos baseados na análise e comparação directa de linhas superficiais de fractura

Um dos trabalhos mais representativos deste grupo é, certamente, o descrito em [1] e [2], entre outros. Neste caso, o problema é abordado como um problema clássico de reconstrução de um *puzzle* 2D. O princípio essencial baseia-se na comparação directa entre todas as peças, na tentativa de encontrar peças que, quando associadas, forneçam um encaixe de perfeição superior à definida num dado limiar. A comparação, contudo, é feita apenas através de uma linha limítrofe da superfície de fractura. Os princípios e requisitos enunciados para aplicação do método são:

- Os objectos a reconstruir têm uma superfície bem definida e relativamente plana.
- A superfície a reconstruir considera-se dividida em fragmentos ideais.
- Os fragmentos ideais são separados por linhas de fractura, sendo estas curvas irregulares de largura zero.
- Dois fragmentos consideram-se adjacentes, se partilharem parte de uma mesma linha de fractura.
- As fracturas dividem igualmente o contorno da superfície original do objecto em linhas de borda.

A solução consiste na determinação de um grafo de adjacências representativo da rede de fracturas.

A aplicação do método está sujeita a um conjunto de etapas devidamente definidas:

- Aquisição e separação das imagens dos fragmentos.
- Segmentação e extracção dos contornos.
- Filtragem dos contornos em várias escalas de resolução.
- Codificação dos contornos por sequências de valores de curvatura.
- Análise estatística dos contornos.
- Identificação de segmentos similares nas escalas mais grosseiras.
- Localização e refinamento dos pares similares em escalas mais finas.
- Alinhamento geométrico óptimo dos segmentos semelhantes.
- Apresentação gráfica dos resultados.

Pela análise dos princípios enunciados, é fácil verificar que o âmbito de aplicação do método é relativamente restrito. Por estar limitado à utilização em objectos de características planas, elimina imediatamente um grande conjunto de cenários de possível utilização. No entanto, este mostrou representar uma importante base científica para o desenvolvimento de trabalhos posteriores.

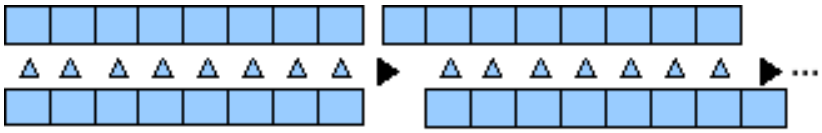


Figura 1. A aplicação do método implica um elevado número de operações de comparação. Para determinação de um possível encaixe adequado entre cada par de fragmentos sendo desconhecida a posição relativa de ambos, torna-se necessário testar todas as hipóteses simulando o deslize entre ambos. Para cada situação de deslocamento relativo é efectuada nova comparação de cada par de valores amostrados.

De facto, são inúmeros os trabalhos posteriores a citar o trabalho em questão. No âmbito da imunidade a ruído, este demonstrou ser igualmente um método especialmente vulnerável. Dependendo essencialmente da informação contida em detalhes submilimétricos de linhas de fractura superficiais, especialmente susceptíveis a imperfeições introduzidas pela criação de estilhaços e a deformações por erosão, o âmbito de utilização do método revela-se especialmente restringido. Repare-se que este aspecto assume relevância superior ainda na área da arqueologia, onde os fragmentos podem ficar longos períodos de tempo expostos a factores ambientais diversos e potenciadores de fenómenos de deformação por desgaste. A necessidade de intervenção humana é reduzida. Após a obtenção da representação em formato digital de cada fragmento, o sistema é completamente autónomo no processo de obtenção de possíveis soluções. A carga computacional envolvida, porém é elevada. Relembre-se que o método consiste na comparação sucessiva de cada par de fragmentos, por combinação dos mesmos entre si. Para comparação de cada par é necessário um conjunto de comparações da sucessão de valores de curvatura das linhas de fractura envolvidas, com desfasamentos progressivos entre as duas peças. A operação de comparação, em si, é igualmente complexa, por não se tratar de uma comparação exacta mas de uma tarefa de determinação de erros sucessivos e acumulados ao longo do processo. Embora a utilização de técnicas de programação dinâmica permita uma eliminação progressiva de não-candidatos em escalas de resolução inferior, menos exigentes em termos computacionais pela menor quantidade de informação processada, as operações de filtragem e determinação de novos contornos em graus de detalhe superior são exigentes. Para mais, o esforço necessário para a procura de soluções aumenta de forma acentuada com o número de fragmentos:

- O número total de comparações base entre pares de possíveis candidatos aumenta exponencialmente com o número de fragmentos.
- Quando o nível de fragmentação leve a fragmentos de dimensões reduzidas, a quantidade de informação contida nas linhas de fractura torna-se menor. Assim, o recurso à análise em níveis de pormenor superior torna-se mais frequente, fazendo aumentar as necessidades computacionais.

A possibilidade de geração de falsas soluções é controlada através do erro máximo admissível na comparação de linhas de fractura. Contudo, este valor

pode ser igualmente crítico. Se um valor excessivamente restritivo pode invalidar soluções verdadeiras pela existência de ruído, um valor demasiadamente permissivo poderá levar a falsas soluções. Pela inexistência de um mecanismo global de detecção de sobreposições, o problema surge ainda mais acentuado.

Outros trabalhos semelhantes minimizam os problemas do anterior pela introdução de mecanismos específicos. Em[3], é introduzida a detecção de sobreposição de partes dos fragmentos durante a tentativa de comparação, para eliminação precoce de falsos candidatos. Em [4], foi tentada a substituição da combinação de programação dinâmica e análise a níveis submilimétricos pela utilização de comparações em escala única. Como informação extra, foi introduzida a componente de cor em cada ponto amostrado para obtenção de valores de curvatura. Assim, dois factores actuam sempre em simultâneo: -a informação geométrica na linha de fractura e a de cor ao longo da mesma. Das últimas abordagens, contudo, não resultam avanços verdadeiramente significativos. No primeiro caso é possível obter uma maior eficiência do algoritmo, reduzindo o trabalho computacional. No segundo, a mesma vantagem é conseguida, mas à custa de um aspecto que torna o âmbito de utilização ainda mais restrito. De facto, quando aplicado a objectos cerâmicos de cor quase uniforme ou a objectos em que a coloração inicial esteja demasiado degradada, poder-se-á ainda prever um incremento de soluções falsas. Mantêm-se contudo a sensibilidade ao ruído e o reduzido âmbito de aplicação apenas a objectos de características essencialmente bidimensionais. É possível ainda encontrar, como extensão aos métodos

Figura 2. Um exemplo de uma falsa solução, obtida pela aplicação do método utilizado em [4]

apresentados, a adaptação dos mesmos princípios à reconstrução de objectos 3D. Em [5], pelo registo de valores de torção da linha de fractura, introduz-se a componente necessária à verificação de continuidade geométrica na junção de dois fragmentos em comparação. Adicionalmente, para verificação de validade das soluções, é introduzido um mecanismo de verificação global com possibilidade de *backtracking*. Esta nova funcionalidade pode evitar falsas soluções, podendo ser especialmente interessante para os casos em que haja imiscuidade de fragmentos pertencentes a objectos estranhos.

2.2 Métodos destinados à recuperação de objectos tridimensionais de características básicas conhecidas

O desconhecimento das características do objecto original é um dos aspectos que pode dificultar a obtenção de soluções simples. Quando se trabalha no espaço tridimensional, qualquer tentativa de posicionamento relativo de um fragmento fica sujeita aos seis graus de liberdade referidos na introdução. A forma mais simples de os restringir é, certamente, a de aplicação de regras emergentes de informação conhecida *a priori*. A esta abordagem prestam-se especialmente objectos de olaria recolhidos e estudados no âmbito da arqueologia. Por serem objectos obtidos por modelação sobre um prato rotativo, as normais à sua superfície intersectam sempre o eixo de rotação que lhes deu origem[6]. Assim, pelo alinhamento de todos os fragmentos recorrendo à sobreposição da projecção do eixo de rotação de cada com o dos restantes, obtém-se um âmbito de trabalho limitado apenas a dois graus de liberdade [7].

Figura 3. Determinação do eixo de rotação original, por intersecção das normais à superfície

É possível encontrar diversas soluções publicadas utilizando o princípio enunciado. Os autores de [8] e [9] consideram que quando um objecto com as características consideradas se fragmenta, se geram tipicamente linhas de fractura que se cruzam em junções que poderão ser do tipo "T"ou "Y". A aceitação de

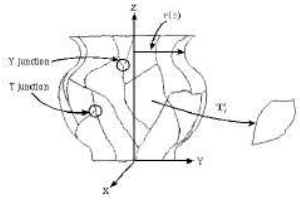


Figura 4. Características geométricas sempre consideradas e representação de junções do tipo "T"e "Y", segundo os autores de [8] e [9].

tal limitação é importante, porque possibilita a definição de pontos de início de

comparação entre linhas de fractura. Ao contrário do que sucedia nos métodos referidos na secção anterior, em que era obrigatório o teste de junção numa grande quantidade de posições relativas para cada par de fragmentos (fig.1), aqui, considera-se um posicionamento inicial previsto, reduzindo acentuadamente o número de comparações a efectuar. Ainda ao contrário do método apresentado em [2] em que era feita a análise sobre um enorme conjunto de pontos recolhidos numa escala submilimétrica, a quantidade de informação recolhida é diferente mas muito mais reduzida. Concretamente, é utilizado um conjunto de parâmetros onde figuram:

- Informação sobre a linha de fractura
- Posição relativa ao eixo de rotação
- Curva de perfil

O método utilizado para reconstrução do objecto prevê um início de processamento no qual se considera intervenção humana para identificação de vértices nas linhas de fractura. A caracterização de cada linha de fractura será efectuada com recurso a cinco pontos pertencentes à mesma linha, equidistantes entre si e partindo de cada vértice. É então o alinhamento dos cinco pontos que é testado para verificação da qualidade da junção entre dois candidatos. A aplicação do método prossegue com as acções de:

- Estimar o eixo original e a curva de perfil para cada fragmento.
- Para cada par de fragmentos, estimar todos os alinhamentos aceitáveis. Cada par formará uma *configuração*. Para cada *configuração* serão verificadas possíveis sobreposições, para eliminação de falsas soluções.
- Para cada configuração válida aperfeiçoar o alinhamento e estimar o eixo/curva de perfil para a nova situação. O objectivo do aperfeiçoamento é a maximização do parâmetro de estimativa máxima de similaridade (*Maximum Likelihood Estimation*). Na prática, tal significa a minimização da soma dos erros no acoplamento pelas linhas de fractura e de posicionamento face aos estimados no objecto original.
- O conjunto de configurações obtidas é registado numa tabela, juntamente com o custo associado (erro determinado face à situação ideal) a cada.
- Tenta-se então a associação de pares de configurações ou de associações a fragmentos.
- Repetem-se os últimos 4 passos até obtenção do objecto final.

Em [10] apresenta-se um outro método com diferenças relativamente ao anterior. O processo prevê a criação de *propostas*, consistindo estas em possíveis associações de fragmentos em que ambos tenham um vértice coincidente e um lado alinhado. É feita então uma classificação ordenada de cada *proposta*, sendo considerados 4 critérios:

- Alinhamento dos eixos de rotação de cada um.
- Sobreposição dos eixos. Numa situação ideal, deverão estar perfeitamente sobrepostos.

– Somatório do quadrado das distâncias entre pontos associados em cada linha de fractura.
– Alinhamento dos vectores tangentes à superfície em cada fragmento.

Em [6] e [7], para além das características puramente geométricas de alinhamento, a verificação e aperfeiçoamento das junções é conseguido pela aplicação do Algoritmo *Iterative Closest Point*[1].

2.3 Métodos baseados em associação de superfícies

Na procura de soluções para o problema da reconstrução de objectos fragmentos foi possível encontrar trabalhos que, não tendo sido desenvolvidos para o efeito, poderão apresentar métodos aplicáveis de forma eficaz. Notavelmente, em [11] e [12] é enunciado um método para associação de superfícies sem restrições. O método é descrito como capaz de estimar os parâmetros de transformação necessários para que seja minimizado o somatório das distâncias entre pontos contíguos de duas ou mais superfícies tridimensionais. Se se verificar a possibilidade de aplicação a um grande número de superfícies sem que a carga computacional se torne demasiadamente elevada e simultaneamente a capacidade para rejeição de tentativas de associação de superfícies não adjacentes no objecto original, então, deverá ser possível a sua aplicação para o ajuste de superfícies de fractura. Conseguir-se-á assim uma outra forma de recuperação de objectos fragmentados a partir dos seus fragmentos.

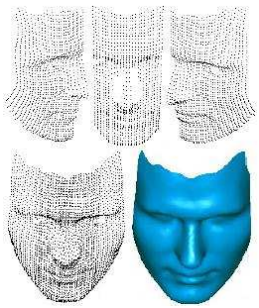


Figura 5. Exemplo de aplicação do método descrito. A partir da definição de 3 superfícies, o algoritmo tem a capacidade de identificar as partes comuns e aplicar a transformação que as maximize a sua aproximação.

Fast Random Sample Matching of 3d Fragments[13] é o nome de um artigo considerado que introduz um conceito diferente para a associação de superfícies.

[1] Dispondo de 2 conjuntos de pontos A e B representando duas superfícies, tais que A seja um subconjunto de B, ou *vice versa*, o algoritmo *ICP* procura pares de pontos mais próximos entre os dois, estimando a transformação que os alinhe. O processo continua iterativamente até à convergência máxima.

Ao invés de utilizar métodos de aproximação sucessiva como o *ICP* ou o *LS3D*, aplica uma variante do denominado *Random Sample Consensus* (*RANSAC*). Os passos para aplicação do método completo surgem descritos como:

- Seleccionar aleatoriamente um ponto *Pi* da superfície de *A* e um ponto *Pj* da superfície de *B*. Um contacto de *A* e *B* fazendo coincidir *Pi* e *Pj* gera uma situação limitada a 3 graus de liberdade (rotação em torno dos 3 eixos). Dois outros graus de liberdade podem ser restringidos pela orientação das normais a cada superfície de contacto, respectivamente em *Pi* e *Pj*. Neste momento, o único grau de liberdade está associado à rotação em torno das normais.
- Seleccionar um segundo par de pontos capaz de definir, em conjunto com o anterior, uma proposta de junção adequada.
- Estimar a qualidade da junção.
- Repetir os passos anteriores memorizando o conjunto de pontos responsável pela geração da solução, até que um limite de qualidade seja atingido - resultando num sucesso, ou um determinado limite de tempo se esgote - resultando numa falha assumida.

3 Considerações adicionais

Pela análise dos métodos apresentados verifica-se novamente a dificuldade em tratar o problema da reconstrução de objectos fragmentados de uma forma genérica e eficaz. Os métodos apresentados em 2.1 estão limitados à utilização na recuperação de objectos com uma superfície plana. Este não é, indubitavelmente, o âmbito mais diverso de aplicação. Com tão grandes restrições à sua aplicação, contudo, seria de esperar uma grande eficácia. Na realidade, não é o que se observa. O método descrito em [1] e [2], pela profundidade da informação disponibilizada, expõe claramente a sua incapacidade para lidar com ruído. Sendo apresentado como candidato à utilização no âmbito da arqueologia - onde os fragmentos estiveram frequentemente expostos durante longos períodos de tempo a factores ambientais, tal aspecto é essencial. Mesmo em condições regulares como as consideradas nos testes apresentados, a sua capacidade para determinar grafos de adjacência completos - o objectivo máximo do método - ficou longe de ser demonstrada.

Os tempos de execução para determinação de cada solução - compreendendo cada cenário aproximadamente uma centena de fragmentos, estiveram sempre contidos na área dos vários milhares de segundos. Mesmo considerando alguma penalização por desactualização ou desadequação da plataforma computacional utilizada, a gama de valores parece estar acima do aceitável. Seria de esperar que a variante descrita em [4], por trabalhar com um volume de informação muito inferior, se mostrasse mais eficiente em termos de carga computacional. Contudo, mesmo considerando a utilização de uma plataforma modesta (de base *Intel PII/300*), são referidos pelo autor tempos entre 1 e 3 segundos para obtenção de soluções, para casos simples de duas peças. Relativamente ao método anterior, este mantém ainda as restrições restantes.

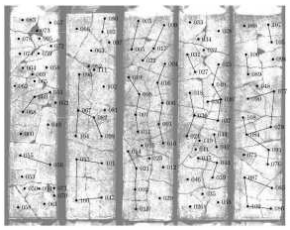


Figura 6. Resultado da aplicação do método à reconstrução de painéis cerâmicos fragmentados intencionalmente.

Os métodos baseados no conhecimento à priori de características geométricas do objecto original - na prática aplicáveis a objectos de olaria obtidos por modelagem sobre um prato rotativo, mostraram-se genericamente eficientes em vários aspectos:

- Imunidade a ruído.
- Capacidade de geração virtual de fragmentos indisponíveis.
- Capacidade de identificação de fragmentos estranhos ao objecto original - em especial, os pertencentes a objectos de características geométricas marcadamente diferentes.
- Baixa exigência em termos de recursos.
- A possibilidade de intervenção humana era considerada em [8], mas apenas de forma limitada e numa etapa inicial.

No entanto, a restrição de aplicação referida e comum a todos estes métodos torna-os aplicáveis, quase em exclusivo, a apenas um subconjunto dos casos do âmbito da arqueologia.

Os métodos direccionados à associação de superfícies não são de forma garantida solução para o problema. Mais difícil que determinar e alinhar as áreas comuns a 2 ou 3 superfícies, sabendo que elas existem, pode ser para estes métodos a identificação de falsos positivos. No cenário típico para o problema enunciado há dezenas ou centenas de fragmentos expondo uma enorme quantidade de superfícies, tais que só cada par específico pode gerar uma solução correcta. O método descrito em [13] parece capaz de lidar com o problema para uma pequena quantidade de fragmentos. Sê-lo-á para o caso oposto?

4 Conclusões

Esta não foi, ainda, uma análise exaustiva de todos os métodos aplicáveis. Dos muitos trabalhos e respectivas variantes encontradas e analisadas, foram seleccionados os que pareceram mais representativos. Contudo, este é um trabalho em curso, onde a orientação ideal para o objectivo final está ainda a ser aperfeiçoada. Durante o processo de recolha de informação, foram considerados numerosos trabalhos que sugeriram possíveis ideias alternativas. Numa primeira fase foi considerada a possibilidade de redução de linhas de fractura a sequências de valores

discretos invariáveis no espaço, para aplicação de algoritmos de *approximate string matching* - muito desenvolvidos ultimamente para a área da genética. Foi abandonada esta hipótese, pela previsível sensibilidade ao ruído que apresentariam as soluções. Considerou-se, então, a possibilidade de aplicação de algoritmos de verificação grosseira do ajuste de superfícies, associados a mecanismos de verificação global com capacidade de *backtracking*. Veio a verificar-se posteriormente que tal não consistia em novidade[5]. Uma terceira hipótese considerada foi a da redução das superfícies de junção, assumidamente irregulares, a *octrees*[2]. A comparação de secções de *octrees* poderia ser eficiente em termos computacionais, mas estava condicionada a questões de posicionamento relativo[3]. Entre outras ideias, a reutilização de métodos utilizados na pesquisa rápida de informação com base em critérios geométricos foi igualmente pensada. Artigos como [14] foram analisados, mas rapidamente se verificou que o propósito era diferente do pretendido. Mais que encontrar e posicionar superfícies que idealmente seriam perfeitamente simétricas, estes métodos pretendem a localização rápida de elementos com características grosseiras comuns.

Sendo este - como antes referido - um trabalho em curso, não existem ainda conclusões finais. No entanto, pelos casos analisados,[13] parece apresentar a melhor abordagem para o atingir de um método que seja simultaneamente versátil no âmbito de aplicação, eficiente na utilização de recursos computacionais e capaz de gerar soluções interessantes.

Referências

1. da Gama Leitão, H.C.: Reconstrução automática de objetos fragmentados. PhD thesis (1999)
2. da Gama Leitão, H.C., Stolfi, J.: A multiscale method for the reassembly of two-dimensional fragmented objects. IEEE Trans. Pattern Anal. Mach. Intell **24**(9) (2002) 1239–1251
3. Smet, P.D.: High-precision recomposition of fragmented 2-d objects. (2000)
4. Amigoni, F., Gazzani, S., Podico, S.: A method for reassembling fragments in image reconstruction. In: ICIP (3). (2003) 581–584
5. Kong, W., Kimia, B.B.: On solving 2D and 3D puzzles using curve matching. In: Proc. of CVPR, Hawaii, USA, IEEE, Computer Society (2001)
6. Kampel, M., Sablatnig, R.: Virtual reconstruction of broken and unbroken pottery. In: 3DIM. (2003) 318–325
7. Kampel, M., Sablatnig, R.: On 3d mosaicing of rotationally symmetric ceramic fragments. In J., K., M., P., M., N., eds.: Proc. of 17th International Conference on Pattern Recognition, Cambridge, UK. Volume 2., IEEE Computer Society (2004) 265–268
8. Cooper, D.B., Willis, A., Andrews, S., Baker, J., Cao, Y., Han, D., Kang, K., Kong, W., Leymarie, F.F., Orriols, X., Velipasalar, S., Vote, E.L., Joukowsky, M.S., Kimia, B.B., Laidlaw, D.H., Mumford, D.: Assembling virtual pots from 3d

[2] As *octrees* constituem um método de representação de volumes, por decomposição espacial recursiva em octantes.

[3] Em [12] é utilizado um processo do mesmo tipo para optimização do método *LS3D*

measurements of their fragments. In: VAST '01: Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage, New York, NY, USA, ACM Press (2001) 241–254
9. Cooper, D.B., Willis, A., Andrews, S., Baker, J., Cao, Y., Han, D., Kang, K., Kong, W., Leymarie, F.F., Orriols, X., Velipasalar, S., Vote, E.L., Joukowsky, M.S., Kimia, B.B., Laidlaw, D.H., Mumford, D.: Bayesian pot-assembly from fragments as problems in perceptual-grouping and geometric-learning. In: ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3, Washington, DC, USA, IEEE Computer Society (2002) 30297
10. Andrews, S., Laidlaw, D.H.: Toward a framework for assembling broken pottery vessels. In: Eighteenth national conference on Artificial intelligence, Menlo Park, CA, USA, American Association for Artificial Intelligence (2002) 945–946
11. Gruen, A., Akca, D.: Least squares 3d surface and curve matching. PandRS **59**(3) (2005) 151–174
12. Gruen, A., Akça, D.: Fast correspondence search for 3d surface matching. In: Workshop Laser Scanning 2005, Enschede, Netherlands. (2005)
13. Winkelbach, S., Rilk, M., Schonfelder, C., Wahl, F.: Fast random sample matching of 3d fragments. (2004) 129–136
14. Sundar, H., Silver, D., Gagvani, N., Dickinson, S.: Skeleton based shape matching and retrieval (2003)

Sessão Técnica 4

Sistemas Distribuídos e Redes

Ensuring Cooperation with Routing Protocols in Mobile Ad-hoc Networks

João P. Vilela and João Barros

Laboratory of Artificial Intelligence and Computer Science,
University of Porto,
Porto, Portugal,
jvilela@dcc.online.pt barros@ncc.up.pt,
WWW home page: http://www.dcc.fc.up.pt/~barros

Abstract. We consider the security of routing protocols for Mobile Ad-hoc Networks (MANETs). We present a classification of routing protocols for MANETs, followed by a brief description of the four base routing protocols as identified by the IETF's Mobile Ad-hoc Networks working group. Afterwards, focusing on the Optimized Link State Routing (OLSR) protocol, we provide a taxonomy of attacks and vulnerabilities and present some of the current schemes to tackle them. Based on that knowledge, we propose a new security scheme that rewards nodes that comply with the routing protocol specifications.

1 Introduction

As a self-organized network without central administration or fixed infrastructure, mobile ad-hoc networks (MANETs) have claimed much attention from the scientific community. The successful operation of an ad-hoc network requires a minimum amount of cooperation between nodes in the network. This requirement is particularly prominent with respect to the discovery and establishment of routes within the network. Therefore, security solutions to secure routing protocols beyond those of the infrastructured/wired paradigm are necessary to ensure communication within these kind of networks.

The goal of this paper is to provide an overview of the state-of-the-art of routing protocols for MANETs and generalized security solutions used to strengthen most of them. We also make an in depth analysis of security issues of a case-study protocol and describe a contribution that we have proposed to make it more secure.

The rest of the paper is organized as follows. As an introduction to the subject, in Section 2, we present a classification of routing protocols for MANETs and a description of protocols that fit in some of the categories identified. Afterwards, in Section 3 we present an overview of the operation of a case-study protocol, identify its main vulnerabilities and present a brief overview of the current security solutions for it. Then we describe our own proposal to secure the aforementioned protocol. The paper concludes with Section 4, which enlightens the main advantages of the proposed solution.

2 Routing Protocols for Mobile Ad-hoc Networks

The goal of this section is to present the main routing protocols for MANETs with sufficient detail to enable a generic understanding of the state-of-the-art in this field and envision the security issues that are traversal to most of them.

Routing protocols for MANETs can be classified as proactive/table-driven, reactive/on-demand or hybrid according to their philosophy.

1. **Proactive routing protocols** have the advantage of making routes immediately available when needed, albeit at the cost of higher amount of routing control traffic exchange. Each node maintains global topology information which has to be updated frequently in order to assure accurate network state information;

2. **Reactive routing protocols** reduce the periodical exchange of routing control traffic at the cost of a route acquaintance delay. These routing protocols acquire the necessary path to a destination only when needed by running an appropriate path-finding algorithm;

3. **Hybrid routing protocols** combine the best features of the two previous categories. Nodes are clustered based on their distance to others or the particular geographical region they are in. For nodes within a certain specified domain, a table-driven approach is used while for nodes beyond this domain an on-demand approach is preferred.

The IETF's Mobile Ad-hoc Networks (manet) working group has identified the following four base routing protocols for use in ad-hoc networks [1].

Proactive/table-driven:

– Optimized Link State Routing (OLSR) Protocol [2]
OLSR is a proactive link-state routing protocol. OLSR uses flooded information about the network to evaluate the best next-hop for every destination and routes are immediately available when needed. OLSR offers, in fact, more than a pure link state routing protocol by (i) reducing the size of control packets through the declaration of only a subset of links and neighbors and (ii) minimizing flooding through the use of a set of selected nodes to diffuse messages to the network. The general idea is that a node communicates with other nodes only through a chosen subset of nodes, thus inducing a reduction on the amount of exchanged control traffic. To guarantee full connectivity in the network, the subset of nodes must be selected in a way that all two-hop neighbors can be reached through them.

– Topology Broadcast Based on Reserve Path Forwarding (TBRPF)
TBRPF is a proactive link-state routing protocol in which each node computes a *source tree* (providing paths to all reachable nodes) based on partial

topology information stored in its topology table. To minimize overhead, each node reports only part of its source tree to neighbors. TBRPF consists of two modules: the neighbor discovery module and the routing module. The neighbor discovery module allows each node to quickly detect neighbors with bidirectional links, link breaks and changes (e.g. becoming unidirectional). It uses so called *differential HELLO* messages which only report changes in the status of links. This results in much smaller messages than those of other link-state protocols. The routing module uses a combination of periodic and differential updates to keep all neighbors informed of the reported part of the source tree (RT). While periodic updates (less often and larger) inform new neighbors of RT, differential updates (more regular, but smaller) ensure the fast propagation of topology changes to all affected nodes.

Reactive/on-demand:

- Dynamic Source Routing (DSR) [4]
  DSR is an on-demand protocol, i.e. it reduces the exchange of control messages by finding routes only when needed. The major difference between this and other on-demand routing protocols is that it does not require nodes to exchange periodic *hello* messages to inform other nodes of their presence. The operation of this protocol is based on establishing routes by flooding *RouteRequest* packets in the network. If the a node receives a *RouteRequest* and is not the intended receiver of the packet, it rebroadcasts it to all its neighbors, otherwise it responds with a *RouteReply* packet which carries the route traversed by the *RouteRequest* packet to the origin.

- Ad-hoc On-Demand Distance Vector (AODV) [3]
  The major difference between AODV and DSR is that DSR uses source routing in which a data packet carries the complete path to be traversed. In AODV, the source and intermediate nodes store the next-hop information for each flow of data packet transmission and are allowed to send *RouteReply* packets to the source. As an on-demand protocol, if there is no route available for the destination, the source node floods a *RouteRequest* packet in the network. AODV singularity in the on-demand context arises from using a destination sequence number to determine an up-to-date path to the destination (a node updates its path information only if the destination sequence number of the current packet received is greater than the one in the last received packet).

3 Case-study: Optimized Link State Routing Protocol

The goal of this section is to present the OLSR protocol, identify its main vulnerabilities and cover some of the security solutions proposed for it. Afterwards, we describe a security scheme we have proposed based on rewarding nodes that comply with the routing protocol specifications.

3.1 Brief Overview of OLSR

OLSR is a proactive link-state routing protocol. Following the proactive protocol philosophy, OLSR has the routes immediately available when needed. As a link state protocol, OLSR uses flooded information about the network to evaluate the best next-hop for every possible destination.

OLSR offers, in fact, more than a pure link state protocol, because it provides the following features:

- *reduction of the size of control packets* by declaring only a subset of links with its neighbors who are its *multipoint relay selectors* (MPR selectors);
- *minimization of flooding* by using only a set of selected nodes, called *multipoint relays* (MPRs), to diffuse its messages to the network (only the multipoint relays of a node retransmit its broadcast messages).

The use of MPRs for message transmission results in a scoped flooding instead of full node-to-node flooding, thus inducing a reduction of the amount of exchanged control traffic. See for example Fig. 1, where the node *A* communicates with the three leftmost nodes only by the MPR *M2*, while he could do it by two distinct nodes – as it would happen in a regular full-flooding routing protocol. OLSR is particularly suitable for large and dense networks, because the optimization procedure based on multipoint relays works best in those cases.

There are two types of control messages in OLSR: HELLO and TC messages.

1. HELLO messages are periodically broadcasted by each node, containing its own address and three lists: (i) a list of neighbors from which control traffic has been heard but no bi-directionality has been confirmed, (ii) a list of neighbors with which bi-directionality has already been confirmed, and (iii) a list of neighbors which have been selected to act as MPRs for the originator node. These messages are only exchanged between neighboring nodes but they allow each node to have information about one and two-hop neighbors which is later used in the selection of the MPR set.
2. TC messages are also emitted periodically by nodes in the network. These messages are used for diffusing topological information to the entire network. A TC message contains the list of neighbors who have selected the sender node as a MPR (MPR selector set) and a sequence number associated to the MPR selector set.

The intent of multipoint relays is to minimize the flooding of the network with broadcasted packets by reducing duplicate retransmissions in the same region. Each node selects a set of its neighbor nodes that will retransmit its packets. This set of nodes is called the *multipoint relay set* of that node and can change over time, as indicated by the selector nodes in their HELLO messages. The node which chooses the multipoint relay set is a *multipoint relay selector* for each node in the set.

Each node selects its MPR set in a way such that it contains a subset of one-hop neighbors covering all the two-hop neighbors. Additionally, all two hop neighbors must have a bi-directional link to the selected MPR set. The smaller the multipoint relay set, the more efficient the routing protocol.

OLSR determines the routes to all destinations through these nodes, i.e. MPR nodes are selected as intermediate nodes in the path. The scheme is implemented by having each node periodically broadcast traffic control information about the one-hop neighbors that selected it as a multipoint relay (or, equivalently, its multipoint relay selectors). Upon receiving information about the MPR selectors, each node calculates and updates its routes to each known destination. Consequently, the route is a sequence of hops through multipoint relays from the source to the destination. The neighbors of any node which are not in its MPR set receive and process the control traffic but do not retransmit it.

3.2 Main Vulnerabilities

In a proactive routing protocol, each node has two tasks to accomplish [7]: (i) correctly generate the routing protocol control traffic (this way giving correct information to the other nodes on the network) and (ii) correctly relay the routing protocol traffic on behalf of other nodes (this way allowing for the control traffic to reach every node in the network). Thus, an attack on the routing protocol must result as the corruption of one of this tasks by some node. This can be accomplished by four main actions:

1. *Fabrication of false routing messages:* A node generates regular routing control traffic messages containing false information or omitting information of the current state of the network.
2. *Refuse of control traffic generation/relay:* A node refuses to generate its own routing control traffic or refuses to forward other node's control traffic (as he is expected).
3. *Modification of routing control traffic:* A node does relay other node's traffic but modifies it to insert wrong information or omit information from the network.
4. *Replay attacks:* A node listens to routing control traffic transmissions on the network and later on injects possibly wrong and outdated information in the network.

Table 1 gives a taxonomy of OLSR security vulnerabilities and provides examples of attack actions based on the network illustrated in Fig. 1.

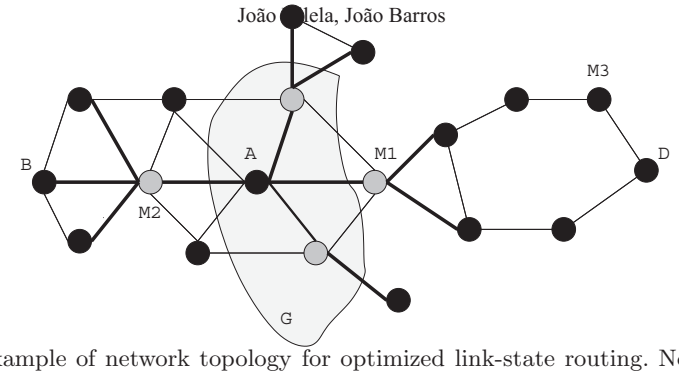


Fig. 1. Example of network topology for optimized link-state routing. Nodes in gray are multipoint relays of node A; light edges represent the connections between nodes; dark edges identify the used links between A and all of its two-hop neighbors through the selected multipoint relay set. M_i denotes a malicious node, D is the destination node and G defines a group of nodes.

Attack	Method	Example	Target	Result
Identity spoofing	Fake HELLO	M_3 generates HELLOs pretending to be A	All nodes	MPR nodes of M_3 will present themselves as last-hop for node A, resulting in conflicting routes to node A.
Link spoofing	Fake HELLO	M_1 generates HELLOs advertising bi-directional links to most of A's two-hop neighbors	Specific node	A chooses M_1 as its main MPR[4] which allows M_1 to intercept and modify most of A's traffic
	Fake TC	M_1 generates TCs advertising D as his MPR selector, directly to G[5]	Group of nodes	Distance between M_1 and D will be deemed to be one hop, thus M_1 will become the main bridge between G and D
	Routing table overflow	M_1 generates many TCs containing non-existing nodes in the MPR set[6]	All nodes	The routing table algorithm will lose a lot of time calculating false routes
Traffic relay/ generation refusal	Drop packets	After becoming a preferential relay choice for A or G[7], M_1 drops packets received from them	Specific node Group of nodes	Loss of connectivity / Degradation of communications
	Refuse to generate control traffic	M_1 is selected as MPR for A and does not advertise that information to the network	Specific node	Node A unreachable, degradation of communications
Replay attacks	Traffic replay	M_1 sends to other nodes "old" previously transmitted[8] TC or HELLO messages	All kinds	Outdated, conflicting and/or wrong information enters the network which may cause defective routing
Wormhole	Protocol disobedience	M_2 tunnels traffic between A and B without the modifications presumed by the routing protocol	Specific nodes	An extraneous inexistent link between A and B is fully controlled by M_2

Table 1. Taxonomy of OLSR security vulnerabilities with examples based on Fig. 1 (M_i - malicious node, A - attacked node, D - destination node, G - group of nodes); [4] Because the smaller the MPR set is, the more efficient the OLSR results are; [5] M_1 is one hop away from G nodes; [6] I.e. declaring non-existing nodes and links; [7] It may use e.g. the described link spoofing techniques; [8] The messages can also be correctly authenticated.

3.3 Current Security Solutions for OLSR

Several security extensions to OLSR have been proposed [7, 5, 8, 9]. They cover a sizeable number of problems identified in Table 1, but consensus only has been found in a few of them. Namely (i) the use of signature and key management systems to ensure the integrity and authenticate the sender of routing control traffic and (ii) timestamps to deal with the replay of old messages. For the remaining issues, different techniques have been proposed. In the case of link spoofing by compromised nodes, the techniques vary from establishing a line of defense (between trusted and untrusted nodes) [7], to the transmission of a cryptographic message in conjunction with routing control traffic [8, 9]. For incorrect traffic relaying, proposals are based on detecting misbehavior based upon the number of packets sent and received by each node or by the usage of geographical positioning [8].

Although these proposals solve some of the key security issues, it is our belief that improvements can be made mainly because of the assumptions and technical drawbacks of the aforementioned proposals. Thus, while adopting some of the generally accepted schemes for tasks such as avoiding replay attacks or guaranteeing integrity and authentication, we propose a scheme based on rewarding nodes that cooperate with the routing protocol to tackle some of the security issues and avoid the problems found in the current schemes.

3.4 Overview of our Security Proposal

The main goal of our proposal is to reward nodes that comply with the routing protocol, either by generating correct routing control messages or by correctly forwarding other node's routing control traffic. For this purpose, we add the following new elements to the regular OLSR operation:

1. *rating table* – a local table were each node holds information about the behavior of its one and two-hop neighbors;
2. *complete path message (CPM)* – a message used by a node to convey the path traversed by a message through the network to another node;
3. *warning message* – a message used to notify neighbor nodes of potential misbehavior of a node.

The operation of the proposed modification to OLSR is based on determining node's misbehavior through two detection mechanisms: (i) detection of misbehavior through direct observation of the transmissions of other nodes and (ii) detection of misbehavior through analysis of CPMs.

Schemes based on detection of misbehavior through direct observation of the transmissions of other nodes have already been proposed [10, 13], but this measure by itself is a unreliable criteria to classify nodes cooperation level. The novelty of our proposal is a scheme to correlate the unreliable information obtained through direct observation of a node transmissions with the reliable information obtained through the CPMs.

The direct observation is done by having each node to listen to its MPR transmissions, thus detecting if it relays messages. If he does, its general classification is increased, otherwise it is decreased.

As we do not have guarantees about the accuracy of the information obtained through the direct observation, the analysis of the CPMs is used to detect those cases in which we could potentially punish a well-behaving cooperative node. The general procedure is as follows.

1. As expected by the operation of the routing protocol, a node floods a Topology Control (TC) message to diffuse topological information to the entire network;
2. From time-to-time, each node sends a CPM back to the origin as a response to this TC message containing the full path traversed by it;
3. When the source node receives the CPM, it compares the information stored about the topology (gotten as result of interaction with neighbor nodes) with the information obtained in the CPM (gotten as result of interaction with a random node).
4. If the comparison favors the information obtained by a neighbor node, its rating is increased; otherwise it is decreased.
5. This rating classification is then used to classify nodes in categories of traffic allowance. Nodes from high categories receive a better treatment in traffic relay than the nodes from low ones.

4 Discussion

Our main concern with this proposal is to provide a new security scheme to solve some of the open security issues of routing protocols for MANETs. Thus, for well studied issues we assume the use of the generally accepted schemes. Namely, for the identity spoofing issue we assume a distributed certification authority [1, 5] is available, and for replay attacks a timestamp scheme can be relied upon.

Our scheme provides a way to successfully solve the following issues:

- *Link spoofing* causes malicious nodes to be penalized in their ability to communicate because they are detected by the correlation of the correct information obtained through the CPMs and the bogus information announced by the malicious node;
- *Traffic relay refusal* can be detected by a correlation of the CPMs received, the probability of a node sending a CPM and the network density (e.g. in a very dense network a node floods a TC message; the probability of a node sending a CPM in response is 50% and none CPM message is received causes immediate suspicion);

Moreover, our scheme presents a simple way to solve typical problems (see e.g. [10–13]) related to the stimulation of cooperation among nodes: (i) a method to classify nodes based on the correlation of the error-prone detection of neighbors retransmissions with the paths traversed by messages sent to the network

is proposed; (ii) we are able to detect elaborated attacks like using power control to fool the source node that a packet has been retransmitted while actually it does not get to destination; (iii) nodes are not able to falsely accuse or praise other nodes without colluding with a considerable amount of nodes.

As part of our ongoing work we are now studying how to tune the scheme proposed to real-case network scenarios to evaluate its behavior when applied to the Mobile Ad-hoc Networks environment.

References

1. D. Dhillon, T. S. Randhawa, M. Wang, and L. Lamont, *Implementing a fully distributed Certificate Autorithy in an OLSR MANET*, Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2004) (Atlanta, Georgia, USA), March 21–25 2004.
2. P. Jacquet, P. Mühlethaler, T. Clausen, A. Laouiti, A. Qayyum, and L. Viennot. Optimized Link State Routing protocol for ad hoc networks. In Proceedings of the IEEE International Multitopic Conference (INMIC 2001), Pakistan, 2001.
3. Charles E. Perkins and Elizabeth M. Royer. Ad-hoc On-Demand Distance Vector Routing. WMCSA '99: Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications, Washington, DC, USA, 1999.
4. D. B. Johnson and D. A. Maltz. Dynamic Source Routing in Ad Hoc Wireless Networks. Mobile Computing, Kluwer Academic Publishers, vol. 353, pp. 153-181, 1996.
5. C. Adjih, D. Raffo, and P. Mühlethaler, *Attacks against OLSR: Distributed key management for security*, 2005 OLSR Interop and Workshop (Ecole Polytechnique, Palaiseau, France), July 28–29 2005.
6. D. Raffo, *Security schemes for the OLSR protocol for ad hoc networks*, Ph.D. thesis, Université Paris, 2005.
7. C. Adjih, T. Clausen, P. Jacquet, A. Laouiti, P. Mühlethaler, and D. Raffo, *Securing the OLSR protocol*, In Proceedings of Med-Hoc-Net, June 25-27 (Mahdia, Tunisia), 2003.
8. C. Adjih, T. Clausen, A. Laouiti, P. Mühlethaler, and D. Raffo, *Securing the OLSR routing protocol with or without compromised nodes in the network*, Tech. Report INRIA RR-5494, HIPERCOM Project, INRIA Rocquencourt, February 2005.
9. D. Raffo, C. Adjih, T. Clausen, and P. Mühlethaler, *An advanced signature system for OLSR*, SASN '04: Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks (New York, NY, USA), ACM Press, 2004, pp. 10–16.
10. S. Buchegger and J.-Y. Le Boudec, *Performance analysis of the confidant protocol*, MobiHoc '02: Proceedings of the 3rd ACM International Symposium on Mobile Ad-hoc Networking & Computing (New York, NY, USA), ACM Press, 2002, pp. 226–236.
11. L. Buttyán and J.-P. Hubaux, *Enforcing service availability in mobile ad-hoc wans*, MobiHoc '00: Proceedings of the 1st ACM International Symposium on Mobile Ad-hoc Networking & computing (Piscataway, NJ, USA), IEEE Press, 2000, pp. 87–96.
12. L. Buttyán and J.-P. Hubaux, *Stimulating cooperation in self-organizing mobile ad hoc networks*, Mob. Netw. Appl. **8** (2003), no. 5, 579–592.

13. S. Marti, T. J. Giuli, K. Lai, and M. Baker, *Mitigating routing misbehavior in mobile ad hoc networks*, MobiCom '00: Proceedings of the 6th Annual International Conference on Mobile computing and networking (New York, NY, USA), ACM Press, 2000, pp. 255–265.

Networking Solutions for Sensor Networks

Pedro Brandão[1], João Barros[1]

CS Department, University of Porto & LIACC
Rua do Campo Alegre, 823
4150-180 Porto - Portugal
{pbrandao,barros}@ncc.up.pt

Abstract. Sensor networks are currently the focus of active research in a considerable number of fields. Because of their many applications, ranging from forest surveillance to anti-terrorist protection, from medical monitoring to crops inspection, from traffic sensing to environment control, sensors are gaining a big momentum in our every day life. The deployment of these technologies encompasses some problems related to communications, application development, lifetime of the network, and security. In this article, we present a set of current proposals that address some of the most relevant technical issues.

1 Introduction

Sensor networks have caught the attention of several scientific communities and are in the course of becoming ubiquitous components in our every day life. Sensors are able to detect presence, monitor the environment (temperature, humidity, wind, etc), track the well being of a person (measuring blood pressure, blood components levels, etc), among many other tasks [1] – the scope of applications is limited only by our imagination.

These wireless sensor networks made of tiny, low-cost devices capable of sensing the physical world and communicating over radio links, are significantly different from classical wireless networks like GSM or wireless LANs: (a) the design of a sensor network is strongly driven by its particular application, (b) sensor nodes are highly constrained in terms of power consumption and computational complexity, and (c) since the network is dense and the nodes share a common objective – to gather and convey information – cooperation can be used to enhance the network´s efficiency.

In general terms, a sensor network can be viewed as a collection of transmitters that observe multiple correlated sources of information, encode the picked up data and cooperate to send this information possibly with multiple hops over the wireless medium to a remote fusion center for further processing.

Several research challenges arise from these emerging sensor networks [2]: how to route data in a-priori unknown network or an ever changing one; how to exploit the correlation structure of the data and the sensors' ability to cooperate in order to increase data throughput, save energy, and improve data analysis; how to combine application-driven requirements with sensor resources so as to optimize

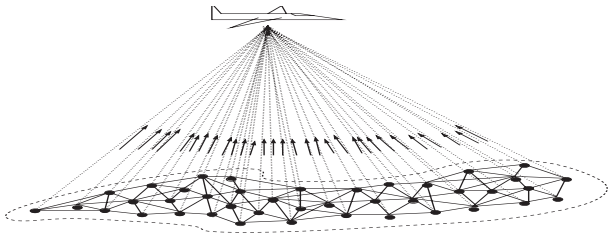


Fig. 1. A sensor network picks up measurements on a physical process evolving in space and transmits this data to a mobile data gathering unit.

the final result; how to secure the identification and the data transmission on a sensor network.

With this article we intend to provide an overview of some of the problems and current proposals for solutions. We will address routing and the definition of middleware for sensor networks. In section 2 we will introduce sensor networks and some of its specificities. Section 3 addresses some of the proposals for routing in sensor networks, whereas section 4 discusses three approaches for designing middleware in sensor networks. We conclude by presenting some thoughts on the current status of sensor networks.

2 What about sensor Networks?

As mentioned in section 1 sensor networks aim to provide different information through small, low price, low power and low computation devices. Ranging from pulse (heartbeat) sensors to video cameras, there are all sorts of *feelings to be sensed*. Many sensors incorporate more than one data acquisition module, which enables them to form networks that can serve more than one purpose ([1]).

As sensors have small resources they usually limit their activity to collecting measurements and transmitting the data to a base station that is a more resourceful node (a PDA, a laptop, etc). In some approaches this base station is mobile (a patrol node) and collects data by entering the transmission region of the sensors (as illustrated in fig. 1). In others, some sensors act as gateways and the information travels several sensor hops until it reaches the base station.

Naturally, the aforementioned mode of operation requires defining routes, and using routing protocols. Some proposals define hierarchies where sensors group themselves in clusters, with a responsible cluster head. In some cases this cluster head is a more powerful node, enabling different schemes. In other cases, this role rotates among nodes to distribute the effort by the data gathering sensors.

On various proposals sensor nodes aggregate and combine data received from other nodes before forwarding it to the next sensor. This occurs when the computation is deemed less energy consuming than simple forwarding and capitalizes from the correlation of data between sensors. However, there are cases (e.g. video cameras) for which aggregation is not an option.

Depending on the objective of the application using the network, sensors may (i) provide continuous information (e.g. room temperature level), (ii) send information only when a threshold is met (temperatures above some value dependent on local weather can indicate a fire), (iii) or act on demand where the base station ask for data (query the current temperature on the pool).

In the next sections we will discuss routing and middleware bearing this diversity in mind.

3 Routing

This section bases its contents on the survey by Kemal Akkaya and Mohamed Younis [3], which addresses several of the routing protocols used in sensor networks. Here, we will mention their categorization and provide brief descriptions of protocols for each of the proposed classes.

As mentioned in section 2 sensor networks data paths aim to transport data from multiple points (the sensors in the area) to a single point (the base station that handles the information) [1]. The fact that the collected information is closely coupled can be exploited to minimize the transport of redundant data. The main objective of the protocols discussed next is to minimize the overall energy consumption and thus maximize the lifetime of the sensor network, while still guaranteeing that sensor information is forwarded through the network to the base station.

The survey in [3] groups the protocols as follows:

- **Data-centric** – the focus of these protocols is answering data queries from the base station. This precludes the necessity of address schemes for the sensors, enabling the approach of flooding the query to the interesting area and have nodes gather and combine the answer to be provided to the base station. To issue the query, some form of attribute naming must be defined for the query language.
- **Hierarchical** – the protocols of this category aim to reduce the flooding and the delay of data aggregation by defining a hierarchy in the routing path.
- **Location-Based** – these routing protocols use the location information to more directly query the sensors of the area of interest of the base station.
- **Network Flow and QoS Aware** – the protocols characterized in this group use QoS (Quality of Service) parameters for the cost function in the links and/or use flow values to route the data.

These categories are not disjoint in the sense that some of the protocols fall in more than one of the divisions. The authors of [3] provide a table with the classification of the protocols studied and their characteristics.

In the next sub-sections we will briefly describe some examples for each of the categories.

[1] Note that There can be more than one base station, nonetheless the flow paradigm is the same, as usually the data is transferred to the base stations on demand and not multicast.

3.1 Data-Centric

- **Direct Diffusion** – [4] defines attribute value pairs to query the needed information from the sensor network. Based on these pairs it is possible to broadcast our *interest* on the information, which is flooded throughout the network. Messages of this kind enable the inference of a gradient, defined by the data rate, duration and expiration time of the interest message. Sending this interest through specific nodes reinforces the path, thus enabling the definition of the trail. The interests are compared with received data from sensors to devise which nodes to transmit to. In Direct Diffusion nodes perform data aggregation, i.e. they combine the data received with their own collected measurements.
- **Rumour Routing** – in [5] when an event is generated/perceived by the sensors, they send a long lived packet that other sensors can cache. This serves to populate a local table leading back to the node that sensed the event. When a query is issued the sensors can use their local table to direct it to the correct node.
- **COUGAR** – [6] defines a new query layer[2] that enables query based declarative languages using in-network-data. COUGAR requires synchronization. Leader nodes are defined to aggregate data and communicate with the base station. This approach treats the sensor network as a distributed database.

3.2 Hierarchical

- **LEACH** – Low-Energy Adaptive Clustering Hierarchy [7] defines clusters based on the energy of the received signal. It defines cluster heads that are used as routers to communicate with the base station. These cluster heads are elected in a nearly random way, based on the desired cluster head percentage over the total number of nodes. The election algorithm enforces some rotation to distribute the "burden" by all sensors. In LEACH each sensor must be able to directly "talk" to the cluster head and the cluster head must also be at one hop from the base station.
- **PEGASIS** – Power-efficient GAthering in Sensor Information Systems [8] defines paths as chains from each sensor to the base station. It includes the definition of a leader node, which communicates directly with the base station. The data gathered along the chain is aggregated to reduce traffic. Hierarchical PEGASIS [9] defines tree like hierarchy chains that reduce bottleneck problems and delays inherent in PEGASIS. It also defines schemes for simultaneous transmission (using for example CDMA).

3.3 Location-Based

- **GAF** – Geographic Adaptive Fidelity (GAF) presented in [10] tries to preserve energy by turning off some sensors located in the same region. The location information is used to define sections where nodes are grouped. Each

[2] As such COUGAR can be viewed also as a middleware for sensors, which is addressed specifically in section 4.

node on the same area is considered at the same path cost and thus can be put in sleep mode. The authors define three states for each node: discovery (finding neighbours in the same area), active (participating in routing) or sleep.

3.4 QoS Aware

- **Maximum lifetime energy routing and data gathering** – These approaches try to maximize the lifetime of the sensor network. [11] uses the remaining sensor energy and the required transmission energy at each link as input to the cost function for the link, then uses Belman-Ford shortest path to define the path. [12] uses data-gathering heuristics. It defines sensor network lifetime as the number of periodic readings from sensors until the first one dies. It elaborates on an algorithm to produce schedules that define data aggregation tree based paths for each of the periods. The algorithm defines a schedule that maximizes the lifetime of the network.
- **QoS Routing** – SPEED [13] uses geographic information and *"packets' speed"* to infer the delay in end-to-end communication. Packet's ACKs are used to estimate the delay between neighbouring nodes. [14] uses the sensor's energy, the transmission energy needed, error rate and other QoS parameters to define the cost function. It uses class-based queuing to allow for different types of traffic.

4 Middleware

There is a current trend on trying to define a middle layer between the application that uses the sensor network and the sensor network itself. This layer would manage the network, getting the information from it and defining its operation using as input the application's requirements and the reliability attached to those requirements.

The rationale behind the approach is that the usage and functioning of the sensor network is highly mandated by the application. The required number of nodes, the geographic coverage, threshold reports, continuous monitoring, mobility, and so on, are defined by the goals of the application utilizing the network. Consequently, the natural solution would be to customize network management (what nodes are needed, can data be aggregated, what is the degree of reliability needed, etc) for each application. The middleware approach tries to provide a common layer that allows this customization. The application can define its necessities (through the middleware's application program interface, API) and the middleware will control the sensors to optimize data deliverance. The key contribution of this approach is a common interface for applications regardless of its objective — customization is taken care of by the middleware. In the next subsections we will briefly address related proposals.

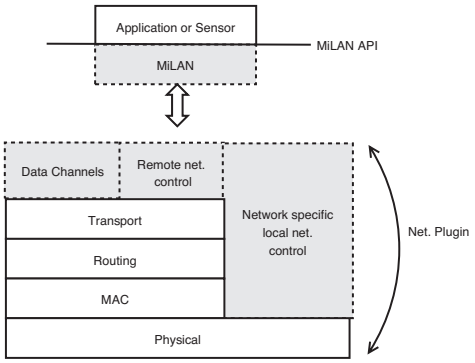


Fig. 2. MiLAN network plug-in architecture example based on [15]. Gray boxes correspond to developed components of MiLAN.

4.1 MiLAN

Middleware Linking Applications and Networks (MiLAN) [15] tries to address the issues raised, by tackling what they devise as the features of sensor applications: distribution, dynamicity in the availability of sensors, constraint application QoS demands, resource limitation (bandwidth and energy) and cooperative applications. This last issue is related to different applications using the same network to achieve different objectives; as such they must cooperate or at least not *"step in each other toes"*. MiLAN tries to cope with different application requests using their QoS requirements as input. It also takes into account the network information (energy and bandwidth) and the system's information on the relevance/precedence of the different applications.

The authors propose a middleware that also dwells on the network stack, so as to take into account and control the network properties- The main objective is to vary the network parameters over time and maintain the QoS needs for the applications.

MiLAN uses as input: the variables that the application requires, the required QoS for each variable and the level of QoS that each sensor or group of sensors can provide for each variable. This information is defined using *"State-based Variable Requirements"* (for the QoS of each variable) and *"Sensor QoS"* graphs (defining which sensors are used to satisfy the QoS). These graphs enable the definition of sets of sensors to fulfil the requirements of the applications.

The MiLAN plugin in the network stack is responsible for using the information of the network to decide what sensors are to be used and for what purpose. It uses a service discovery protocol to query the network on the attributes of the nodes. It should be possible to influence the states of the sensors as well as the routing protocol used to get the most of the network and define a set of sensors on the network capable of evaluating the needed information.

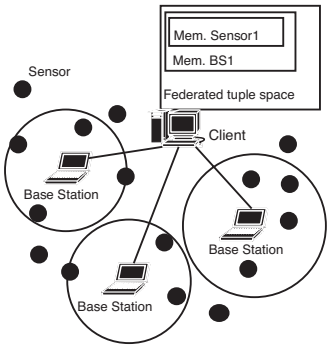


Fig. 3. TinyLime architecture example based on [16]. Denoting managers (clients), base stations and sensors. The federated tuple space in the client is also portrayed as a shared memory.

The final set of available sensors is defined by the intersection of the set given by the network plugin and the one from the upper layer given the QoS requirements of the application. This set is used to choose the sensors such as to maximize the time that the information can be given. Nonetheless, MiLAN is able to incorporate rules to sacrifice quality in order to extend the lifetime.

4.2 TinyLIME

[16] is based on a middleware model for Mobile Ad-hoc NETworks (MANETs) named LIME (Linda in a Mobile Environment). LIME [17] itself is based on LINDA. LINDA [18] uses a shared memory model to represent data. It defines tuples (typed fields) to hold the data structures. The coordination between processes is based on reading and writing on these tuples on the shared memory. It defines basic operations for: adding tuples (**out**), removing (**in**) and reading (**rd**). The operations have blocking and non blocking variants. The removal and reading can use patterns (defined also as tuples) to query for the data.

LIME breaks the tuple space on several spaces in order to decouple the shared memory from space and time. LIME is intended for mobile environment and as such the nodes are not available all the time. The *break* allows nodes to have each one a tuple space that is synchronized when the nodes are accessible. As such, the shared memory is being reshaped according to nodes' connectivity, creating a federated tuple space. The operations are extended to include a location parameter, which enables to indicate which tuple-space to query/write. LIME also adds reactions that enable an effect when a tuple matching a pattern is found in the tuple space.

In TinyLIME, the authors consider that it is not feasible to know all sensor locations, that multi-hop communications to the base station enforce a great burden on the sensors forwarding the data and that it is not feasible to assume that all sensors can communicate with the base station. As such they define a

model where the sensors only communicate with a one hop mobile base station that recalls sensor information. Managers of the data are clients of these mobile base stations. They define the model as a mixture of the flexibility of MANETs (as it uses its routing capacities between the managers and base stations) and the sensor network capabilities (sensing the environment). In this model sensors are different from other nodes, as they are only visible when there is a base station on their range (see fig. 3). In that case, their tuple-space is visible to base station, being part of it. However remove or modify operations are not possible in sensor data as they are considered read-only. TinyLIME introduces conditions to the reactions operations and a freshness parameter to these reactions. The last addition enables having a frequency for refresh to the queried reaction.

As can be seen TinyLime is more dedicated to data retrieval issues, providing a middleware that hides the details of retrieving data and using the simple operations of LINDA.

4.3 Energy Efficient Resource Allocation

[19] dictates as its primary goals to define an architecture to provide (a) a common standardized system to applications with diverse objectives, (b) an environment capable of supporting and coordinating multiple applications and (c) a means to use the sensor network resources efficiently and adaptively. As in MiLAN (see section 4.1) it intends to optimize the computation, communication and sensing (CCS) energy spendings. For such, it enables nodes to sleep (thus saving energy) and it defines guarding nodes that are responsible for detecting a target phenomenon and selectively wake up the sleeping nodes. The coordination of these active nodes to sense the phenomenon, aggregate data and route the decision to the base station is a key point.

As in MiLAN it also aims to use the application knowledge, but it tries to reach a balance between the specificity of the application and the generality needed by the middleware to cope with very different applications. It also tries to balance the different needs of each application to reach the greater common benefit possible. For this it suggests the use of adaptive fidelity algorithms and inter-application coordination.

It uses a cluster based approach where each cluster is considered a basic unit of the middleware, functioning as distributed software. The clusters are considered dynamic as the phenomenon to be sensed is. The nodes' resources also impose dynamicity on the cluster, with nodes leaving and joining the cluster to optimize energy. As such, on-the-fly self-configuring distributed clustering mechanisms are addressed. Periodic information regarding nodes CCS capabilities and energy should flow to the cluster head through efficient mechanisms.

Aggregation and comparison of data is also considered mandatory, and for this purpose distributed and energy-aware protocols should be used. Inter-cluster coordination is also of importance with cluster heads seen as another source of information from other clusters.

The proposal defines a three-phase heuristic to cope with the problems of distributing sensor tasks (see [20] for further details).

5 Conclusion

In this article, we introduced some of the problems in sensor networks. We presented some of the current proposals that address routing and middleware.

The routing protocols described try to minimize energy consumption while maximizing information delivery. Data-centric ones focus on broadcasting queries and defining methods for the sensors to forward the query to the appropriate nodes. Some approaches rely more deeply on defining a hierarchy for the data paths. As such they define clusters or hierarchy levels and rotate the leader nodes so to distribute the load. Others rely on the location information to selectively turn on/off sensors based on their proximity. We also described protocols that are more dedicated to QoS measures. They use network delay (and other QoS parameters) as part of the cost function for traversing links. None of these approaches can be categorized in one group, for example COUGAR is more prone to a data-centric approach, but also defines some hierarchy levels. One could say that data-centric and location based approaches are better suited for monitoring event based phenomenon whereas QoS aware protocols are more appropriate for real time monitoring.

Middleware solutions try to provide a common framework to all applications while trying to customize the sensor network to fit the applications needs. MiLAN and Energy Efficient Resource Allocation are similar in that they provide an API for input of QoS needs by the applications. The first expects to know which sensors are intended to yield the measured parameters and which quality they provide for that result. The second strives for a simple and general API with some focus on cluster definition in order to optimize CCS energy consumption. TinyLime provides a shared memory model approach where the given API enables the configuration and retrieval of sensor information. The focus is on information gathering based on a single approach regardless of the underlying network mobility and state.

The common view seems to be that there is no *"one size fits all"* solution – sensor networks are in essence application-specific – but middleware appears as a reasonable option.

There are two set of issues, which remain on our agenda: (1) security issues, which are either related to the vulnerabilities of the infrastructured world and the ad-hoc networking paradigm or specific to sensor networks; (2) data gathering, in its own is yet another source of ongoing research.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirc, E.: Wireless sensor networks: a survey. In: Computer Networks 38, Elsevier (2002) 393–422
2. Akyildiz, I.F., Kasimoglu, I.H.: Wireless sensor and actor networks: research challenges. In: Ad Hoc Networks 2, Elsevier (2004) 351–367
3. Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. In: Ad Hoc Networks 3, Elsevier (2003) 325–349

4. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'00). (2000)
5. Braginsky, D., Estrin, D.: Rumor routing algorithm for sensor networks. In: Proceedings of the First Workshop on Sensor Networks and Applications (WSNA). (2002)
6. Yao, Y., Gehrke, J.E.: The cougar approach to in-network query processing in sensor networks. In: Sigmod Record 31(3), Sigmod (2002)
7. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless sensor networks. In: Proceeding of the Hawaii International Conference System Sciences. (2000)
8. Lindsey, S., Raghavendra, C.: Pegasis: power efficient gathering in sensor information systems. In: Proceedings of the IEEE Aerospace Conference. (2002)
9. Lindsey, S., Raghavendra, C., Sivalingam, K.: Data gathering in sensor networks using the energy*delay metric. In: Proceedings of the IPDPS Workshop on Issues in Wireless Networks and Mobile Computing. (2001)
10. Xu, Y., Heidemann, J., Estrin, D.: Geography-informed energy conservation for ad hoc routing. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'01). (2001)
11. Chang, J.H., Tassiulas, L.: Maximum lifetime routing in wireless sensor networks. In: Proceedings of the Advanced Telecommunications and Information Distribution Research Program (ATIRP 2000). (2000)
12. Kalpakis, K., Dasgupta, K., Namjoshi, P.: Maximum lifetime data gathering and aggregation in wireless sensor networks. In: Proceedings of IEEE International Conference on Networking. (2002)
13. et al, T.H.: Speed: a stateless protocol for real-time communication in sensor networks. In: Proceedings of International Conference on Distributed Computing Systems. (2003)
14. Akkaya, K., Younis, M.: An energy-aware qos routing protocol for wireless sensor networks. In: Proceedings of the IEEE Workshop on Mobile and Wireless Networks. (2003)
15. Heinzelman, W.B., Murphy, A.L., Carvalho, H.S., Perillo, M.A.: Middleware to support sensor network applications. In: IEEE Network, IEEE (2004)
16. Curino, C., Giani, M., Giorgetta, M., Giusti, A., Murphy, A.L., Picco, G.P.: Tinylime: Bridging mobile and sensor networks through middleware. Third IEEE International Conference on Pervasive Computing and Communications (PerCom'05) (2005) 61–72
17. Murphy, A.L., Picco, G.P., Roman, G.C.: Lime: A middleware for physical and logical mobility. In: Proc. of the 21st Int. Conf. on Distributed Computing Systems (ICDCS). (2001) 524–533
18. Gelernter, D.: Generative communication in linda. ACM Computing Surveys 7(1) (1985) 80–112
19. Yu, Y., Krishnamachari, B., Prasanna, V.: Issues in designing middleware for wireless sensor networks. In: IEEE Network 18(1), IEEE (2004) 15–21
20. Yu, Y., Prasanna, V.: Energy-balanced task allocation for collaborative processing in wireless sensor networks. In: Algorithmic Solutions for Wireless, Mobile, Ad Hoc and Sensor Networks 10(1-2), Springer (2005) 115–131

Building a distributed system for dynamic information search, organization and classification for educational purposes

Joaquim Fernando Silva[1] , Francisco José Restivo[2]

Faculty of Engineering of University Of Oporto
[1]joaquim.silva.pt,[2] fjr@feup.pt

Abstract. Our focus is on a Web-based Knowledge Portal for educational purposes. We consider a distributed system for information retrieval and document collection, which will enable different forms of knowledge construction. Also accessing information from multiple information systems and integrating them into a knowledge portal are key issues in developing our system. Some tools are considered to be used namely ontologies, concept maps and software agents to deal with semantic issues.

Previous related initiatives are described and tools for dealing with semantics are also present. A comprehensive section on syntactic and semantic interoperability is described in detail.

A very high-level architecture is drafted along with ideas to deal with its implementation

Introduction

Currently we are designing a scaffold for learning with semantic interoperability, with two issues. One aspect is considering web search for information retrieval. Other is using reusability of learning contents available in autonomous and heterogeneous information systems.

Despite the technological aspects, conceptual and pedagogical issues arise when we develop a framework for being used by learners. All the design and system architecture are based on the social constructivist theory that sees learning as a participation in social processes of knowledge construction.

This knowledge portal is intended to be a reference place for students of an undergraduate or graduated course where all members (tutors, teachers, content producers) will have an active role in achieving good educational results. It is intended as a reinforcement in their study of a real subject, where they can revise

subjects, understand their basics, other fields of interest some how related and even go deep in the subjects of their course.

We have been studying Knowledge Management Systems (KMS) and B2B application integration to draw our framework. Both refer semantic, which is intended to be the core issue of our system. To deal with it we intend to use some available technologies, namely software agents, ontologies and Concept maps.

One might say that our approach of reusability tends to follow a method-oriented B2B application integration, but for us the trading partners are schools or education institutions. A common set of objects, that for us are content usable in disciplines, invites reusability and significantly reduces the need of redundant methods and applications. By defining methods these can be shared and integrated, in inter application trough distributed objects.

In this article we first focus KMS, B2B, semantic and syntactic interoperability and then analyze some architecture platforms with semantic support and some integration in data or logic level. We also present some considerations about semantic desktop. Finally a comprehensive framework is proposed to provide semantic interoperability among different systems web supported. Finally future work is referenced, namely building the system, evaluating its effectiveness and usefulness.

KMS

Knowledge management systems aren't a new idea. In fact in 1997 University of California, Berkeley [15] have already implemented a KIE (Knowledge Integration Environment) which is a learning environment that uses Internet to help middle and high school students develop an integrated understanding of science and a critical eye toward the complex resources found on the Web. The KIE combines network resources and software with sound pedagogical principles to improve science learning. KIE networking tools allow students to use scientific evidence in activities that foster knowledge integration. [4]

This project continued in the Web-based Inquiry Science Environment (WISE), which is a free on-line science learning environment for students in grades. [16]

Knowledge management systems (KMS) are also emerging applications that require mechanisms for dealing more explicitly with the meaning and use of the data with semantics.

B2B

After having studied some aspects related to data-oriented B2B application integration data movement, transforming tools and technologies we thought in applying these same concepts into a web portal which will provide a working place for students to deal with their school matters.

In an B2B application integration may occur at the data level or logic level. Also integration can occur in the B2B oriented application interface. In this layer we take into account the usefulness of APIs (Application Programming Interfaces) because we will be able to communicate to applications that use a proprietary interface with a standard interface. APIs developers may invoke the services of the entities evolved in order to obtain some value of them. For example the Google search engine provides its APIs for search and manipulates information on the web. In this particular case applications connect remotely the APIs threw the SOAP protocol, a mechanism based on XML format, which uses HTTP. [17]

Before implementing an application method-oriented one must decompose it to its scenarios or types. In a B2B application these can be rules, logic, data and objects.

Rules are defined as a set of conditions and are used to control the flow of information between partners.

Logic differs from rules in that way it is simply a sequence of instructions in a program. There are three classes of logic: sequential processing, selection and iteration. Sequential processing is related to a series of steps in data processing, while selection is the decision making dynamic within the program. Iteration can be seen as the repetition of a series of steps.

Data is sharing information between trading applications, computers and humans.

Objects are simply data and business services bound as objects. In fact they are bundles of data encapsulated inside an object and surrounded by methods that act upon that data. [24]

Process B2B integration can be defined as the ability to define a common business process model that defines the sequence, hierarchy, events, execution logic, and information movement between systems residing in multiple organizations.

For creating a portal with a web browser interface one must first design the portal application, including the users interface and application behaviour, as well as determine which information contained within the back systems. The portal application needs a traditional analysis-and-design life cycle , as well as a local database. Also a portal architecture is based in web clients, web servers, database servers, back-end applications as application servers. [24]

An architecture diagram (fig. 1) of a general portal, with its evident business logic layer, is called a three layer which enhances scability. Also we can see different types of usages made from different users. The technology involved to integrate different systems can be XML, RDF [22] or even OWL based. [23] In fact these three types distinguish different levels of semantic.

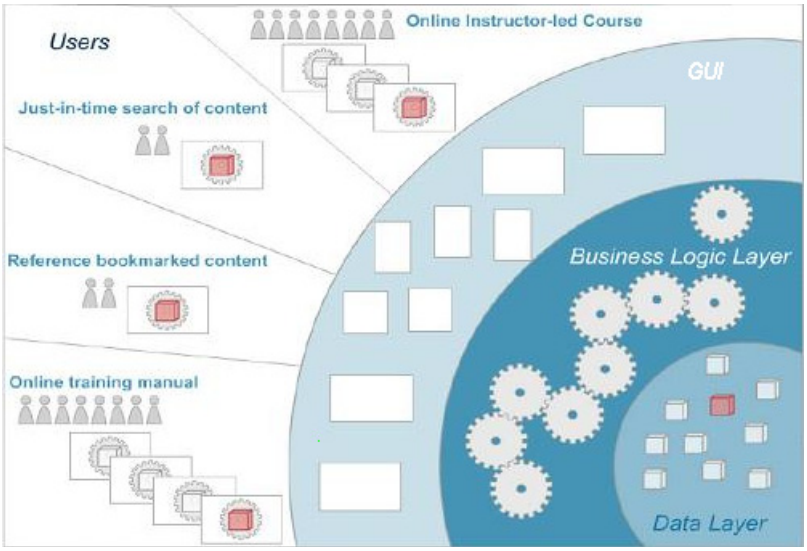


Fig. 1. Architecture diagram

XML provides a surface syntax for structured documents, but imposes no semantic constraints on the meaning of these documents and its associated XML Schema. A XML Schema is considered a language for restricting the structure of XML documents.

RDF [22] is a relation model for objects (”resources”) and relations between them. Provides simple semantics for this model, and these models can be represented in an XML syntax. In this case RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes.

Going a step further there’s OWL [23] which adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjoint ness), cardinality (e.g. ”exactly one”), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

Syntactic and semantic interoperability

Integration with interoperability brings a semantic problem as well as syntactic problem among heterogeneous and distributed information systems.

According to [2] semantic interoperability is the knowledge-level interoperability that provides cooperation between businesses. With the ability to bridge semantic conflicts, which arise from differences in implicit meanings, perspectives, and assumptions, we create a semantically compatible information environment based on the agreed concepts between different business entities.

Syntactic interoperability is the application-level that allows multiple software components to cooperate even though their implementation languages, interfaces, and execution platforms are different. [16]

Some available emerging standards such as XML and Web Services based on SOAP (*Simple Object Access Protocol*), UDDI (*Universal, Description, Discovery, and Integration*), and WSDL (*Web Service Description Language*), can resolve many application-level interoperability problems and give a technology solution. But semantic interoperability assumes its solutions in semiotic, linguistic, philosophical, or social environments.

Integration in Data level requires managing the differences in metadata and application semantics. Semantic conflicts might be resolved at the data level, or achieve interoperability resolving schema-level conflicts, that is, structural differences.

The design of a semantically interoperable system environment should provide the capability of detecting and resolving incompatibilities in data semantics and structures, as well as a standard query language for accessing information on a global basis. At the same time, it should involve minimal or no changes to existing systems to preserve the local autonomy of the participating systems. The environment must be flexible enough to permit adding or removing individual systems from the integrated structure without major modifications. [16]

Previous research in semantic interoperability can be categorized into three broad areas: mapping-based, intermediary-based, and query-oriented approaches.

The mapping-based approach attempts to construct mappings between semantically related information sources. It is usually accomplished by constructing a global schema and by establishing mappings between the global schema and the participating local. Mappings are not limited to schema components (i.e., entity classes, relationships, and attributes), but may be established between domains and schema. The drawback of the global schema approach is that it is not designed to be independent of particular schemas and applications. Explicit representations of semantics of information sources can help resolve the problems associated with interoperability when constructing mappings between them.

Another promising approach is the intermediary-based approach, which depends on the use of intermediary mechanisms (e.g., mediators, agents, ontologies, etc.). These intermediaries may have domain-specific knowledge, mapping knowledge, or rules specifically developed for coordinating various autonomous information sources. In most cases, such intermediaries use ontologies to share standardized vocabulary or protocols to communicate with each other. The advantage of using ontologies is its ability to capture the tacit knowledge within a certain domain in great detail in order to provide a rich conceptualization of data objects and their relationships. Its knowledge is domain-specific, but independent of particular schemas and applications. Even though such an approach may be theoretically valid, the inherent complexities of the knowledge domain bring enormous difficulties to

develop and maintain ontology in autonomous, dynamic, and heterogeneous databases. Therefore, this approach is typically applied only to a restricted application domain, which limits its general applicability in practice.

The third approach, query-oriented approach, is based on interoperable languages, most of which are either declarative logic-based languages or extended SQL. They are capable of formulating queries spanning several databases. In order to resolve semantic conflicts over data structure and data semantics, it is desirable to have high-order expressions that can range over both data and metadata. One of the main drawbacks of this approach is that it places too heavy a burden on users by requiring them to understand each of the underlying local databases. This approach typically requires users to engage in the detection and resolution of semantic conflicts, since it provides little or no support for identifying semantic. Consequently, users are also responsible for semantic conflict resolution.

These research approaches classified into these three categories may not be mutually exclusive. For example, the intermediary-based approach may not necessarily be achieved only through intermediaries. Some approaches based on intermediaries also rely on mapping knowledge established between a common ontology and local schemas. It is also often the case that mapping and intermediaries are involved in query-oriented approaches.

In B2B application integration, semantic conflict analysis can occur at the data level and at the schema level. Data-level conflicts are differences in data domains caused by the multiple representations and interpretations of similar data. Examples of data-level conflicts are data-value conflicts, data representation conflicts, data-unit conflicts and data precision conflicts. Data-level conflicts can be further classified into two different levels depending on the granularity of the information unit (IU). Semantic conflicts can occur at the level of objects' properties and their values (attributes as IU) or at the level of the objects themselves (entities as IU), but not necessarily try to resolve structural differences.

Schema-level conflicts are characterized by differences in logical structures with inconsistencies in metadata of the same application domain. Examples can be found in: naming conflicts, entity-identifier conflicts, schema-isomorphism conflicts, generalization conflicts, aggregation conflicts and schematic discrepancies.

Naming conflicts arise when labels of schema elements (i.e., entity classes, relationships, and attributes) are somewhat arbitrarily assigned by different database designers.

Entity-identifier conflicts are often caused by assigning different identifiers (primary keys) to the same concept in different databases.

Schema-isomorphism conflicts occur when the same concept is described by a dissimilar set of attributes, that is, the same concept is represented by a number of different attributes and, therefore, the sets of entities are not set operation-compatible. Generalization conflicts result from different design choices for modelling related entity classes.

Aggregation conflicts arise when an aggregation is used in one database to identify a set of entities in another database. Therefore, the properties and their values in one

database may aggregate corresponding Properties and values of the set of entities of another database schematic discrepancies can occur when the logical structure of a set of attributes and their values belonging to an entity class in one database are organized to form a different structure in another database. The pure schema-level approach, without data-level interoperability, however, may result in achieving interoperability between different schemas that may be semantically different but structurally similar. It is, therefore, desirable to achieve interoperability at both levels. [24]

Some architecture platforms with semantic support

We tried to analyse some architecture platforms which integrate knowledge and have information visualization methods, as the one we intend to build. We divided our analysis in two kinds of approaches:

The first considers *visualizing knowledge and information* for fostering learning and instruction.

They are mainly based on concept mapping technology. Information is conceived as a knowledge resource and is represented in the map. The map is functioning as a personal repository that has been constructed for facilitating visual search and access to knowledge elements and associated resources.

The later is a *knowledge-oriented information organization.*

They may serve as a developmental aid for course designers or as an information basis for students engaged in self-regulated learning in a resource-based learning environment. Concept maps functioning as organizational tools may also be used as navigational aids for fostering knowledge, providing facilities for the visual search of documents in broad information repositories, for example, the World Wide Web, digital libraries, or hypermedia environments. [5]

Both scenarios can be interested to cope in our system. The first scenario can be seen when the learner uses the portal for learning around a particular concept. He has multiple sources to understand it. On the second scenario one might consider the tutor to navigate throw the concept map to populate it with new repositories to be used by the learner. Or even the learner can navigate throw the concept map to visualize the entire subject.

In a first approach Tergan [6] recently brought a new perspective of using concept maps in educational scenarios. The potential of digital concept maps for supporting processes of individual knowledge management is analyzed. The author suggests digital concept mapping as a visual-spatial strategy for supporting externalized cognition in resource-based learning and problem solving scenarios. In fact many

authors dealing with cognitive demands inherent in a variety of educational, social, and workplace scenarios, refer explicitly to concept maps as a means for bridging the gap between knowledge visualization and information visualization.

Also Canãs [5] outline a conceptual knowledge represented in a Concept map, which may be linked with content knowledge and information resources coded as text, images, sound clips, or videos accessible in personal or public repositories. In CmapTools, the use of concept maps has been extended beyond knowledge representation to serve as a browsing interface to a domain of knowledge and associated information. The authors outline special features of the approach for integrating, making accessible, and using knowledge and information.

Other important platform, by Neumann [8] called ParIS, which is a learning environment that aims at fostering the development of competencies for self-regulated learning and media competencies as central components of scientific literacy. In ParIS, students solve everyday authentic problems by using Mind Mapping, a visual-spatial strategy to assist planning, gathering, generating, organizing, and using knowledge and knowledge resources. The presented instructional design approach transforms ideas of supporting resource-based learning by helping students visualize their knowledge and relate it to information associated with it.

On the second approach, Coffey [11] describes a learning environment organizer (LEO) that provides students and instructors with information and knowledge visualization capabilities. LEO serves as a meta-cognitive tool for course designers and an advanced organizer for students. It is an extension of the CmapTools developed by Cañas and associates. LEO helps to visualize and plan a course organization by using a concept map. The concept map itself is used as knowledge based visualization of the structure of course components and provides interactive access to the materials. It presents an interesting approach, doable to follow, by integrating both fields of research knowledge and information visualization.

Also Fiedler [12] develop the tool "Weblog authoring", which enables the user to represent information spontaneously and to maintain it in personal repositories, as well as to generate a social network and collective information filtering and routing. Weblog authoring supports the construction of a personal repository of information, as well as the ability to engage in shared dialogue about artefacts, with the possibility and the benefits of using concept mapping to make sense of the Weblog representations.

Even Lee [13] focuses on design rationales and implementations of an alternative Web search environment called "VisSearch". It has some advantages, particularly with regard to cognitive processes, in dealing with ill structured, open-ended research questions, as compared to conventional Web-search environments. The VisSearch environment facilitates information searching in dealing with such problematic search questions by means of visualizing the knowledge and associated Web resources of both the user and other users looking for useful Web based information on the same or similar topics. VisSearch employs a single, reusable concept map-like knowledge network, called search-graph for a variety of purposes, for example, visualizing Web search results, the history of Web search engine hits of a variety of iterative Web searches of different users, as well as user comments to Web sites and search queries.

The search-graph provides interactive access to all Web resources linked with the elements in the graph.

Going a step further we analysed Frank [14] which gives a strictly application-oriented approach. It focuses on the visualization of knowledge and information management activities underlying the development of the Management Information System (MIS) at DaimlerChrysler. The MIS is for the leaders of the department of research and technology, the central department for technical innovations and the management of technology. It is used not only as a tool with a controlling function, but as a general homogenous information and dialogue platform of high actuality and flexibility, serving as a knowledge and information space. The aim for developing the system was to match the user's needs, processes and visions as closely as possible. The authors show how complex processes and problem solutions in the development and maintenance of a MIS may be visualized and used for facilitating dialogue and for working with a large number of content elements, highly complex information structures and large knowledge networks.

This contribution opens up a perspective of how visualizations may be used on a large-scale basis for knowledge and information visualization in the application context.

Semantic Desktop

Semantic desktop for personal information management and collaboration intends to enable the integration of desktop applications, with the use of ontologies, metadata annotations and semantic web protocols on desktop computers with an integrated personal information management focusing information distribution and collaboration on the Web. This issue is not new. The Semantic Web effort (W3C-SW) [21] provides standards and technologies for the definition and exchange of metadata and ontologies.

Use open-source software with reusability and create, on top of existing sophisticated system, an open personal information management system and collaborative infrastructure based on Semantic Web. *Collaboration, acquisition and dissemination infrastructures* like Wikis and Blogs are providing the foundation for joint collaborative knowledge creation and are essentially simplified knowledge acquisition tools.

Social Software maps the social connections between different people into the technical infrastructure. Online Social Networking enables collaboration relationships as first class citizens, and allows exploiting these relationships for automated information distribution and classification.

P2P and Grid computing, especially in combination with the Semantic Web field, develops technology to interconnect large communities without centralized infrastructures for data and computation sharing, which is necessary to build heterogeneous, multi-organizational collaboration networks.

Several systems have been created already to explore this field, e.g., the Haystack [18] system at MIT, the Gnowsis [19] system at DFKI, or the Chandler [20] system by the OSA foundation. Each of these systems only addresses some parts of the picture.

Proposal of architecture for a Web Portal on a distributed platform of documents

Our framework for the Portal is a multitier architecture which supports various levels of user activities in managing the semantically interoperable environment. Bearing in mind that the Knowledge Portal is based for information finding, classification and even organizing the retrieval of courseware material by users. We are presenting its system architecture as well as its technologic architecture.

System Architecture

The Knowledge Portal comprises three layers: *Information Access* from the user's perspective, *Information Processing* features of the portal and the *Grounding Technologies*. [9]

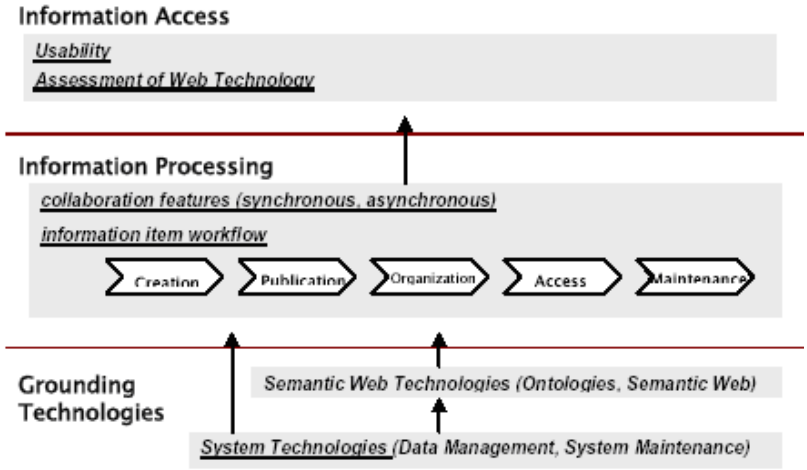


Fig. 2. Layer Architecture

Information Access layer comprises the usability and Assessment of Web Technology. In other words this layer is the front-end of our Knowledge portal for a virtual community, with semantic capabilities.

In detail we are building a knowledge portal in a collaborative perspective with different classes of users. In a brief discussion there are learners, visitors, tutors and administrators. Each member of a class can add new members, which will inherit its privileges.

Some considerations have been taken to the navigation structure. Also we must consider some usage and navigation rules in this layer.

In Information processing layer we are mainly concerned with collaboration features (synchronous, asynchronous) and information item workflow.

Some issues related to document management systems from creation, Publication, organization, access and destruction / maintenance have not yet been considered, but for us collaboration features enabling virtual groups of different domain of interest, such us different courses, must be taken in consideration with synchronous, such as forums and asynchronous features.

The layer grounding Technologies we should consider two aspects. In one hand System Technologies, such as Data Management and System Maintenance. In other hand Semantic Web Technologies namely Ontologies and Topic Maps.

For the former we have not yet decided which Data Management and Data Storage we should consider. But we bear in mind that the success depends upon the System Maintenance and System Administration. Especially the ontologies applied in the system as well as different levels of access with password-protection.

For the later we should consider ontologies based in RDF [22] or OWL [23] with agents to interact with portal members.

This layer is the core of the system, in an innovative perspective, because ontologies are the central components of the portal. Ontology provides term definitions of the domain of interest and it can be applied in different ways to enable Semantic Web enhanced functionalities. Two types of ontologies should be considered, domain ontologies and application ontologies. Also for a knowledge portal based semantic web there must exist some tools for Ontology Management and Editing capabilities using an ontology editor like PROTÉGÉ , OntoEdit, or an editor facility integrated in the portal.

We intend to use a dynamic ontology, tracking the changes with consistency.

The KAON (Karlsruhe Ontology and Semantic Web Framework) Framework disposes an ontology tool. KAON offers abstractions for ontologies and text corpora, an ontology editor and application framework, inferencing, and persistence mechanisms, etc. [10] .

Furthermore Semantic Web Services will add a new level of functionality to allow automatic content search, content publication, import and export documents in a distributed environment.

Technological Architecture

Agents are to some degree knowledgeable about the subject matter being learned and are also programmed to have some knowledge about how to find and filter information from available resources in the Internet.

Our portal will have some agents with precise roles. The *Pedagogical Agent* will interact to the user helping in his path. The *Search Agent* will help the tutor to find related documents to attach to the Portal. The *Ontology Agent* to maintain the Ontologies coherent during the dynamic changes made by tutors. Finally the Monitor Agent which will track user actions, activity levels, to facilitate the collaboration process in accordance by urging users to thoroughly discuss concepts, try to initiate discussions and encourage users to reach common ground when negotiations fail.

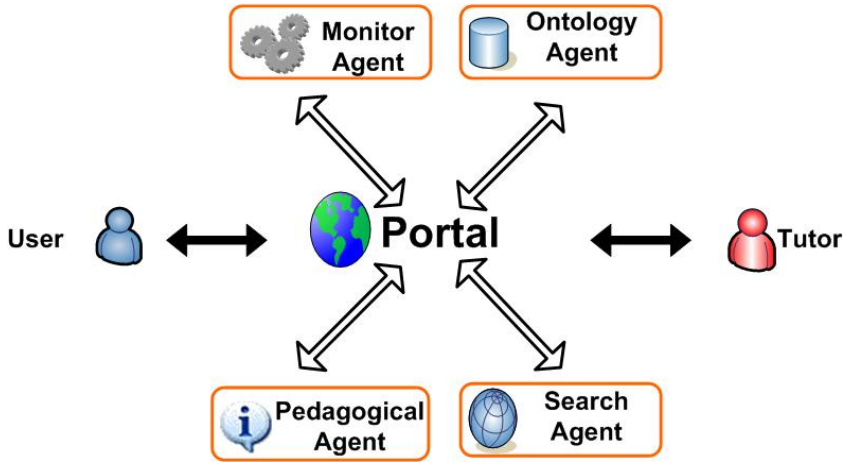


Fig. 3. System Architecture

Ontology Evolution

The Knowledge Portal as presented in this paper must not only have a proper ontology that reflects what the user is interested in, but also future interests which will change together with the teaching/learning subject itself. Therefore, the ontology and the topics represented therein need to be updated. One must deal with several requirements incorporated in such updates:

Modifying the ontology: The ontology must remain consistent at all times. We intend to use the evolution functionalities of the KAON API, which ensure that changes to the ontology will not corrupt it.

Introducing new concepts: First recognizing that a new concept (e. g. a new topic) has appeared in the course material available in the network or on the Web, then inserting this concept into the right place of the taxonomy, and finally linking it via further relations to other concepts.

Implementation Details

The technical implementation of the evolution component uses the functionalities of the KAON API to guarantee the coherence of the ontology. In the KAON API, a special care has been taken to make sure changes to the ontology reflect user changes while preserving the logical coherence of the ontology level.

The evolution component integrates the use of TextToOnto, a KAON component designed to help users in creating ontologies out of texts.

Overall, the Knowledge Portal is expected to indicate how a Semantic Web based approach increases the support of retrieval and management of remote (learning) resources, by providing tools for discovering and organizing them.

Conclusions and next steps

Our purpose his to achieve a portal with the right approach both in terms of technology and cognitive perspective. By analysing the problem approach presented here it seems an interesting approach, up to some extent original. Our main question is if we are in danger of the resulting Portal will not reach critical mass and thus will not be able to penetrate the user space wide enough to result in mass adoption. How to organize information in a user perspective and knowledge or matter perspective is a question to future resolution. Maybe using trees to connect and relate information should be a good idea.

One think is certain, we must use an ontological approach with software agents to interact with learners if we want to exchange information and enable automated processing of information items. With ontologies becoming an integral part of many academic applications, support for ontology evolution and versioning is important.

Next step should be a deep analysis to the framework presented, maybe in a single study, designing and deploying pedagogical agents in a distributed collaborative learning environment and finally have some results to give some test support to the Portal.

References

1. Linn, M. C.: Designing the Knowledge Integration Environment. International Journal of Science Education, 22(8) 781-796. (2000)

2. Park, Jinsoo, Ram, Sudha, : Information systems interoperability: What lies beneath?. ACM Transactions on Information Systems TOIS, v.22 n.4, p.595-632, (2004)
3. Vernadat, Françoes B,:Enterprise Modeling and Integration, principles and applications. London: Chapman & Hall (1996)
4. Bell, Philip, Davis, Elizabeth A. and Linn, Marcia C.: The Knowledge Integration Environment: Theory and Design. Education in Mathematics, Science, and Technology Graduate School of Education University of California, Berkeley (2000)
5. Keller, Tania, Tergan Sigmar-Olaf: Visualizing Knowledge and Information: An Introduction. Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
6. Tergan, Sigmar-Olaf: Digital concept maps for managing knowledge and information. Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
7. Cañas, Alberto, Carff, Roger, Hill, Greg, Carvalho, Marco, Arguedas, Marco: Concept maps: Integrating knowledge and information visualization., Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
8. Neumann, Anja, Wolfgang, Gräber1, Tergan, Sigmar-Olaf: *ParIS – Visualizing Ideas and Information in a Resource-Based Learning Scenario*. Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
9. Holger, Lausen, Stollberg ,Michael, Hernández, Rubén Lara, Ding Ying, Han Sung-Kook, Fensel Dieter :*Semantic Web Portals – State of the Art Survey*. DERI (Digital Enterprise Research Institute), Technical Report (2004)
10. KAON the Karlsruhe Ontology and Semantic Web Framework – Developer's Guide. http://kaon.semanticweb.org/Members/rvo/ KAON_Dev_Guide.pdf, (2003)
11. Coffey, John W, :LEO: A Concept Map Based Course Visualization Tool for Instructors and Students".Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
12. Fiedler, Sebastian, Sharma ,Priya: Navigating Personal information Repositories with Weblog Authoring and Concept Mapping. Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
13. Lee, Young-Jin: Facilitating Web Search with Visualization and Data Mining Techniques. Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
14. Hans, Frank, Jürgen, and Drosdol, Johannes: Information and Knowledge Visualization in Development and Use of a Management Information System (MIS) for DaimlerChrysler, A Visualized Dialogue and Participation Process". Knowledge and Information Visualization, LNCS Springer-Verlag Berlin Heidelberg (2005)
15. Web-based Integrated Science Environment (WISE) Project available at: http://kie.berkeley.edu/KIE.html (2005)
16. March, S.T., Hevner A., Ram S.: An Agenda for information technology research in heterogeneous and distributed environment. Inform. Syst. Res. 11, 4, 327-341. (2000)
17. Google Web APIs reference available at: http://www.google.com/apis/reference.html (2005)
18 . Haystack available at: http://haystack.lcs.mit.edu/system (2005)
19. Gnowsis available at: http://www.gnowsis.org/system (2005)
20. Chandler available at: http://www.osafoundation.org/ (2005)
21. W3C-SW available at: http://www.w3c.org/sw (2005)
22. RDF Available at: http://www.w3.org/RDF (2005)
23. OWL Web Ontology Language Guide, Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, W3C Recommendation, http://www.w3.org/TR/2004/REC-owl-guide-20040210/ , (2004)
24. Linthicum David S.: B2B Application Integration e-Business-Enable Your Enterprise. Addison-Wesley Information Technology Series, pp 75-104. (2004)