Actas das 3as Jornadas de Informática da Universidade de Évora

Lígia Ferreira
Vasco Pedro

u évora
ESCOLA DE CIÊNCIAS E TECNOLOGIA

Évora, Portugal
21–22 de Fevereiro de 2013

Actas das 3as Jornadas de Informática da Universidade de Évora — JIUE2013
Escola de Ciências e Tecnologia
Universidade de Évora
2013

ISBN: 978-989-97060-7-1

http://www.di.uevora.pt/jiue2013/

Prefácio

As Jornadas de Informática da Universidade de Évora são uma iniciativa do Departamento de Informática, já na sua terceira edição, destinada a divulgar e promover os trabalhos na área da Informática realizados na Universidade ou contando com a participação dos seus elementos. Nestes trabalhos inclui-se, fundamentalmente, a investigação, básica ou aplicada, efectuada no âmbito de projectos ou no contexto de teses de mestrado e de doutoramento.

As Jornadas oferecem, aos alunos de pós-graduação da Universidade, um fórum amigável e alargado onde discutir o seu trabalho. O ambiente e a audiência, diferentes dos com que contactam habitualmente, permitem-lhes aumentar a confiança no resultado dos seus esforços e ganhar experiência na sua apresentação pública. Para os alunos da Licenciatura em Engenharia Informática, é a oportunidade de um primeiro contacto com a investigação realizada no Departamento e uma hipótese de encararem a vida académica sob uma perspectiva de futuro.

Em linha com os objectivos enunciados, os oradores convidados desta edição são, pela primeira vez, membros do Departamento de Informática.

Depois de 2010 e 2011, a terceira edição das Jornadas acontece em 2013 devido à mudança de calendário destinada a integrar a apresentação dos trabalhos realizados no âmbito da disciplina de Introdução à Investigação, constante do plano curricular do Programa de Doutoramento em Informática. Nela, os alunos têm a dupla responsabilidade de serem autores de artigos, que serão submetidos a um processo de avaliação semelhante ao das conferências científicas, e de participarem no processo de avaliação.

Outra estreia, nesta edição, foi a apresentação de um artigo feita em directo da África do Sul.

As presentes actas integram os artigos escolhidos de entre todos os enviados para inclusão nas Jornadas e os de Introdução à Investigação. A comissão organizadora agradece a todos os seus autores, que manifestaram o interesse em participar nas Jornadas enviando-os e apresentado-os, e aos seus colegas do Departamento que colaboraram no processo de selecção. Agradecimentos especiais são devidos ao José Saias e ao Miguel Barão, por aceitarem o convite para apresentar o seu trabalho. Agradecemos, ainda, a ajuda e inúmeras sugestões e perguntas do José Saias.

Lígia Ferreira
Vasco Pedro

Évora, 22 de Fevereiro de 2013

Comissão de programa

Responsáveis

Lígia Ferreira
Vasco Pedro

Membros

Carlos Caldeira
Francisco Coelho
Irene Rodrigues
José Saias
Luís Rato
Miguel Barão
Paulo Quaresma
Salvador Abreu
Teresa Gonçalves
Vitor Nogueira

Workshop PhDI
Comissão de programa

Responsável

Salvador Abreu

Membros

Albertina Ferreira
Ana Paula Silva
Arlindo Silva
Carlos Caldeira
David Mendes
Dora Melo
Francisco Coelho
Francisco Guimarães
Irene Rodrigues
Joel Costa
José Duarte
José Saias
João Laranjinho
Luís Arriaga da Cunha
Luís Rato
Lígia Ferreira
Miguel Barão
Mohammad Moinul Hoque
Paulo Quaresma
Pedro Fialho
Shib Sankar Bhowmick
Sérgio Cardoso
Teresa Gonçalves
Vasco Pedro
Vitor Nogueira

JIUE2013

Oradores convidados

5a-feira, 21 de Fevereiro de 2013

Sessão 1

Sessão 2

Sessão 3

Sessão 4

6ª-feira, 22 de Fevereiro de 2013

Sessão 5

Sessão 6

Análise de Sentimentos sobre conteúdos da Web atual

José Saias

Departamento de Informática - ECT
Universidade de Évora, Portugal
jsaias@uevora.pt

Resumo Alargado

Com a utilização da Web, e em particular o uso crescente de blogs e redes sociais, foram surgindo grandes coleções de documentos onde as pessoas manifestam as suas ideias sobre os mais variados assuntos. Existe interesse na procura de opiniões expressas textualmente nesses documentos, que, por existirem em grande volume e em formas diversas, obrigam ao uso de técnicas automáticas.
A Análise de Sentimentos (AS) lida com a busca de opinião em texto, procurando manifestações de sentimento (positivo, negativo ou neutro) sobre entidades. O desafio inicia com a deteção de menções ou referências a entidades (pessoas, marcas, instituições). Uma menção pode ser objetiva, ou pode carregar alguma subjetividade com sentido pejorativo ou favorável.
O léxico de sentimentos é um recurso muito usado em AS, que consiste num dicionário com a indicação da polaridade que cada vocábulo pode carregar, num ou mais contextos. Daqui resulta que palavras como *bom* têm associado um sentimento positivo, enquanto *mau* tem a polaridade inversa. Numa abordagem simples, a AS pode usar esta informação para verificar a presença ou não de fontes de polaridade junto à entidade alvo. A distância desses termos à entidade, em número de palavras, é por vezes considerada na análise de opinião.
Para além de identificarem o tipo de sentimento, alguns sistemas indicam ainda um valor numérico para a intensidade do mesmo, estando os valores acima de zero associados à classe positivo e os valores abaixo de zero associados à classe negativo. A presença de alguns advérbios de intensidade (como *muito*, *pouco*) pode atenuar ou aumentar esse valor.
Modelos linguísticos com n-gramas podem também aplicar-se em AS, no sentido de encontrar sequências de palavras determinantes para a orientação do sentimento. Dependendo do tipo de texto e do domínio, há trabalhos onde os unigramas são a melhor escolha e trabalhos onde os melhores resultados são conseguidos com bigramas e trigramas.
Com maior análise linguística, as dependências entre os elementos do discurso podem indicar com maior precisão se a entidade é ou não abrangida pelo tipo de sentimento associado a alguns termos polarizados. A análise morfossintática do texto, complementada com o uso de dicionários de sinónimos e outros recursos semânticos permitem maior abrangência e correção na análise.
Algumas abordagens recorrem a técnicas de Aprendizagem Automática supervisionada da correspondência entre um conjunto de atributos e a polaridade do sentimento, aqui vista como uma possível classe (positivo, negativo, neutro) num exercício de classificação. O processo requer um treino, com um conjunto de exemplos classificados, cujos textos são descritos por um conjunto de atributos, usualmente representados num vetor. Cada atributo é relativo a um aspeto a considerar no texto, como a presença de uma palavra ou a sua frequência, ou quaisquer aspetos sobre n-gramas, morfologia ou função sintática. A lista de atributos em estudo pode misturar várias abordagens, desde a análise superficial de texto às orientadas à semântica.
As dificuldades num processo de AS decorrem da ambiguidade da linguagem natural, tal como acontece com outras tarefas de PLN. Ironia e anáforas são difíceis de tratar. O jargão e as expressões idiomáticas específicas de um tema ou grupo dificultam a conceção de uma técnica de aplicação global. Por outro lado, a escrita que se encontra nas redes sociais é pouco cuidada, misturando frases mal estruturadas com abreviaturas e símbolos, para além dos simples erros ortográficos. Este aspeto torna menos eficaz uma abordagem focada na análise linguística convencional, onde a correção do texto é importante para a classificação correta da função sintática.
As redes sociais favorecem ainda a identificação do autor do texto opinativo. Em grande escala, pode ajudar na identificação de tendências por faixa etária ou zona geográfica.

Dynamical Systems, Control and Applications

Miguel Barão
Departamento de Informática
Universidade de Évora

Abstract

Artificial Intelligence is deeply divided into subareas embracing different problems, techniques and sometimes having different scientific communities. Among these areas, control theory dedicates itself to the study of dynamical systems, construction of models and design of control systems. The theoretical tools obtained are usually sufficiently generic so that they can be applied in problems coming from many different fields such as robotics, electronics, mechanics, chemistry, medicine, economy, and others. In this talk, a few applications from past and ongoing research projects are presented to illustrate the power of the theory in practical problems.

Using Finite State Machines with Simple Learning Techniques for Classifying Questions in Question Answering System

[1]Mohammad Moinul Hoque, [2]Teresa Goncalves and, [2]Paulo Quaresma

[1, 2,] Department of Informatica, University of Evora
{d10861, tcg , pq} @ uevora.pt

Abstract. Question classification plays a significant part in Question Answering system. In order to obtain a classifier, we present in this paper a pragmatic approach that utilizes simple sentence structures observed and learned from the sentence patterns, trains a set of Finite State Machines (FSM) based on keywords appearing in the sentences and uses the trained FSMs to classify various questions to their relevant classes. Although, questions can be placed using various syntactic structures and keywords, we have carefully observed that this variation is within a small finite limit and can be traced down using a limited number of FSMs and a simple semantic understanding instead of using complex semantic analysis. We have used WordNet semantic meaning of various keywords to extend the FSMs capability to accept a wide variety of wording used in the questions. Various kinds of questions written in English language and belonging to diverse classes from the CLEF Question Answering track are used for the training purpose and a separate set of questions from the same track is used for analyzing the FSMs competence to map the questions to one of the recognizable classes. With the use of learning strategies and application of simple voting functions along with training the weights for the keywords appearing in the questions, we have managed to achieve a classification accuracy as high as 95%. The system was trained by placing questions in various orders to see if the system built up from those orders have any subtle impact on the accuracy rate. The usability of this approach lies in its simplicity and yet it performs well to cope up with various sentence patterns.

Keywords: Finite State Machine, question classification, machine learning, keyword weights

1 Introduction

Classifying a question to its appropriate class in an important subtask and plays a substantial role in the Question Answering (QA) systems. It can provide some useful clues for identifying potential answers in large collections of texts. The goal of this current work is to develop a classifier using Finite State Machines (FSM) to classify a set of questions into their relevant classes. Various techniques have already been tried by the community either to classify a question to its relevant class or to a finer subclass of a specific class. Results of the error analysis acquired from an open domain QA system demonstrates that more or less 36.4% of the errors were generated due to the wrong classification of questions [1]. So, this issue can be highlighted as a subject of interest and has arisen the aim of developing more accurate question classifiers [2]. Usually the answers generated from the classified questions have to be exact in nature and the size of the answer has to be within a restricted size [3-4] which greatly emphasizes the need of an accurate question classifier. Techniques involving support vector machines [5-6] showed a fare accuracy rate of over 90% in classifying

questions to their finer classes instead of diverse super classes. Li and Roth [7] investigated a variety of feature combinations using their Sparse Network of Winnows (snow) algorithm [8]. The Decision Tree (DT) algorithm [9] was also used for question classification with fair amount of accuracy rate. It is a method for approximating discrete valued target function where the learned function is presented in a tree which classifies instances. Naïve Bayes [9] method was also used in the question classification task with limited accuracy rate. In another work [10], where a function-based question classification technique is proposed, the authors of that paper claimed to have achieved as high as 86% precision levels for some classes of questions. Some attempts have been made to develop a language independent question classifier [13] with not a mentionable success rate.

This work focuses on the questions posed only in English language and uses questions from the Question Answering (QA) track of the Conference and Labs of the Evaluation Forum (CLEF) [14]. It classifies the questions into 5 major classes namely Factoid (FA), Definition (DE), Reason/Purpose (RP), Procedure (PR) and Opinion (OP) Class. CLEF QA track have some diverse types of questions and we required to fit each of the questions into any of the above mentioned classes. Factoid class of questions are mainly fact oriented questions, asking for the name of a person, a location, some numerical quantity, the day on which something happened such as 'What percentage of people in Bangladesh relies on medical insurance for health care?', 'What is the price of an air-conditioning system?' etc. Definition questions such as 'What/Who is XYZ?' asks for the meaning of something or important information about someone or an organization. 'What is avian influenza?', 'Define SME', 'What is the meaning of Bluetooth signal?' are some examples of the definition class questions. Reason/Purpose questions ask for the reasons/goals for something happening. 'Why was Ziaur Karim sentenced to death?' and 'What were the objectives of the National meeting?' are the example questions of this class. Procedural questions ask for a set of actions which is an accepted way of doing something. Such as: 'How do you calculate the monthly gross salary in your office?' Opinion questions ask for the opinions, feelings, ideas about people, topics or events. An example question of this type may be like 'What did the Academic Council think about the syllabus of informatics department?' A question is either mapped to only one class or may be classified as 'other'.

The next section of the paper describes the procedure used to create the states and transitions in the FSMs involving a simple learning mechanism and the section 3 presents the data set for the experimental verification of the procedures and outcome of the experiments followed by a section covering a discussion about the future works.

2 Classification using Finite State Machines (FSM) with learning strategy

From a large number of questions derived from the gold standard set of QA track of CLEF 2008 - 2010 and observing them manually, we came to a conclusion that it may be possible to classify a set a questions using a set of FSMs and the FSMs can be automatically built and adjusted according to the questions in the training set and later on can be used to classify the questions appearing in the test set. Initially we start off with some elementary states for each of the FSMs beginning with different headwords. The headwords are usually What, Why, How, When, Where etc. Questions that do not begin with a known headword

can be restructured to a suitable form. For example, 'In which country was the Vasco da Gama born?' can be changed to 'What country was the Vasco da Gama born?'. Similarly, 'What does SME stand for?' can be reformatted to 'What is SME?' and so on. The initial preprocessing module performs this question restructuring step. It also converts the keywords into its present tense and singular form to make sure that the keywords 'thought' and thinks are treated similarly. It also reduces the number of keywords in the set. Each FSM is represented with a directed graph and may have more than one state for each question class. Those states are called final states. Rests of the states are called 'undefined' (UN).

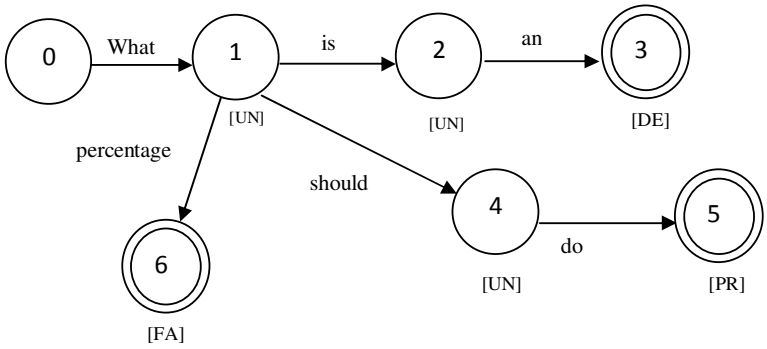


Figure 1. An FSM accepting questions Q1 and Q2.

An FSM can have many intermediate and undefined states as well as transitions between them. The inputs to a FSM are keyword tokens extracted from the questions. An example FSM beginning with the headword 'What' and accepting the questions Q1: 'What is an SME?' and Q2: 'What percentage of people relies on TV for news?' is depicted in the figure 1.

2.1 Learning new states and transitions in the FSM using keywords

FSMs continue to build up the states and transitions as it encounters more new question instances. Each of the questions in the training set is tokenized removing a few stop words and the keywords are then isolated from each of the questions to form a keyword structure (KS). Every keyword in the KS has a weight in context with the other keywords appearing in the question. In order to calculate the relevant weights of the n number of keywords, a Keyword Frequency Matrix (KFM) of n x n dimension is created first and the frequency of every keyword appearing before the every other ones is stored in the matrix. This KFM is prebuilt from all the question instances of the training set. Table 1 shows a dummy KFM with some sample frequency values.

Table 1: A sample 5x5 Keyword frequency matrix (KFM)

Before \ After	*What*	*The*	*Is*	*Meaning …*	*Think*
What	5	80	90	16	8
The	1	5	120	16	8
Is	2	71	6	12	6
Meaning	0	0	1	0	0
Think	0	3	2	0	0

When we are using a sentence to build or train up the FSMs, a subset of the KFM is created using only those keywords which are appearing in the question sentence and the weights of each keyword Z in the question sentence is calculated in context with other keywords appearing in that sentence using the formula followed. The formula sums up all the frequency values where the keyword Z appears after each of the other keywords in the sentence and subtracts from it the sum of the frequency values where Z appears before each of the other keywords in that question sentence. There are many keywords in various question sentences which appear more frequently than some other rarely used keywords. In order to make sure that such keywords do not receive highest weights all the time compared to the other significant but less frequently used keywords, we divide the weight value with the sum of the frequency value of Z where Z appears before each of the other keywords in the question sentence. In case of a negative weight, the weight is set to 0.0.

Probable Weight (Z) =

$$\frac{\sum_{j \in row} KFM(j)(index\ of\ Z\ in\ column) - \sum_{j \in col} KFM(index\ of\ Z\ in\ row)(j)}{\sum_{j \in row} KFM(j)(index\ of\ Z\ in\ column)}$$

Finally, the keyword structure (KS) is built from the question sentence and it comprises of the keywords along with their weights. An FSM is selected based on the headword appearing in the question sentence and it is built using the *Algorithm1.*
The algorithm detects the keyword boundary from the KS which is the position of the keyword having the highest weight value. If there are multiple keywords having the same highest weights, the position of the first keyword with the highest weight value is marked as the keyword boundary position. The FSM does not take any keyword input beyond this boundary position to build up on its own. When creating a transition to a state for an input keyword, synonyms of the keyword if there are any are also derived with the help of WordNet [11] and are added as inputs to that transition to extend the machines capability.

Major steps of Algorithm1:

For every keyword K_i in the KS of a question sentence
  Mark the keyword boundary which is the first position of the highest weighted keyword
End for
For every keyword K_i within the keyword boundary position in the KS
  Try to go through the FSM states using K_i as input to the FSM starting from the state S_0
    If a valid state S_j can be reached using a transition path with K_i as input
      Continue to repeat the above step with the next K_i from the state S_j
    Else
      If for the input K_i , no transition path can be found from state S_f
        If K_i and K_{i-1} were same
          Create a loop transition in that state S_f
        Else
          Create a new state and add a transition from the current state S_f to that new state
          Set the input of the transition to K_i and also synsets(Ki) using WORDNET
        End if
      If K_i appears at the keyword boundary
        Set the class of the state S_f according to the class of the question.

Else
Set the class of the state S_f to 'UNDEFINED'
End for

2.2 Voting function for a state

Different ordering of the similar kind of sentences belonging to different classes can lead to the development of an FSM with wrong classification states. For example, the question, 'What is the aim of the raid spectrum policy?' may be classified as a factoid question in one training set where as there may be 5 more questions of similar pattern that are classified as Reason/Purpose question in another training set. In this case, we propose a simple voting algorithm where every question terminating at a final state with a keyword appearing at the keyword boundary will vote for its class in its favor. The class of the state of the FSM will finally be determined according to the class that gets the maximum vote. Voting function ensures that the FSM does not label a state to a wrong class because of the different ordering of the questions appearing in the training set.

Major steps of the Voting Algorithm:
For every FSM_i in the FSM set
For Every Question Q_i in the question set
For Every Keyword K_i in the Q_i appearing at the keyword boundary and terminating at state S_f in FSM_i
Cast a vote for that state S_f in favor of the class that Q_i belongs to
End for
End for
Update each of the states of the FSM_i to the class that gets the maximum vote
End for

3 Experimental verification

For the experimental purpose, we took questions from CLEF question answering track for the year 2008-2010. A total number of 800 questions of various classes were selected for the evaluation purpose. Around 400 questions from various years were selected for the training purpose and a test set was created with the rest of the questions. The system was trained with a training set and a test set was used to test the capability of the system. Effectiveness of the classification was calculated in terms of precision and recall and the accuracy was calculated from the confusion matrix [12].

In order to make sure that the system does not get biased with specific patterns, we have trained and tested the system in various ways to see if any subtle changes occur in the case of precision and recall. We also trained the system with 50% questions from one year mixed up with the 50% question from another year to cope up with the variations used in question wording and syntactic structure. We also changed the question order to see if the FSMs built from different order cause any considerable error or not. Keyword frequency matrix was trained using a dataset and it continued to update itself with the introduction of new questions from the training set. The data set and the result is presented in table 2.

Throughout the training process, voting function was kept activated. Results are depicted in graphs as seen in the figure 2.

Table 2: Data set for the question classification using FSMs

Year	No. of Questions for training	No. of Questions for testing	Accuracy
2009	250	250	94.1777%
2010	80	90	95.1278%
2008	40	50	98.7190%
Mixed Set of questions	350	450	93.2311%

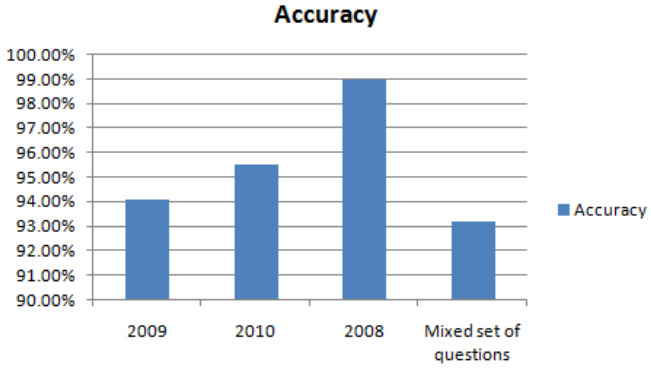


Figure 2. Accuracy of the system plotted on Table 2.

Precision and Recall for each class is calculated and is shown in Table 3. From the data, we can see that most of the questions were correctly classified by the FSMs, because it could find correct patterns for the questions belonging to specific classes. Wrong classifications were made in some cases where almost similar pattern existed in questions belonging to two different classes. Fortunately, our voting function took the feedback from the questions and responded accordingly to reduce the classification error by a margin.

Table 3: Precision and Recall for every class of questions

Question Class	**2009**	**2010**	**2008**	**Mixed**
Definition (Precision)	1.0	1.0	1.0	0.981
Definition (Recall)	0.938	0.966	1.0	0.921
Factoid (Precision)	0.899	0.903	1.0	0.967
Factoid (Recall)	0.955	1.0	1.0	0.911
Opinion(Precision)	0.0	1.0	0.0	1.0
Opinion (Recall)	0.0	1.0	0.0	1.0
Procedure(Precision)	0.977	1.0	0.0	0.965
Procedure(Recall)	0.957	0.939	0.0	0.991
Reason/Purpose (Precision)	0.959	1.0	0.0	0.978
Reason/Purpose (Recall)	0.967	1.0	0.0	0.988

Because of the inaccurate calculation of weights for some keywords in context with the others also contributed to the errors, though most of the time, the weight calculation function provided near correct assumption.

In order to check the building procedure of the FSM during the training steps using the training data, we were concerned about the question ordering. We have created an n x n question index matrix with each of the questions in the training set having an index number in that question matrix. We have randomly selected a question index and started to train the system from there on. The next question selected for the training was the question that was most similar (in terms of similar words) to the previously selected one. We did 4 runs and in every run, we have selected a question randomly and made sure that the same question does not get selected twice. We did the same for the most dissimilar ones and did 4 runs as well.
The average of the runs is listed in table 4. We can observe from the average run that, no significant change in accuracy, precision and recall are noticed with the change occurring in the question order.

Table 4: Precision and Recall for every class of questions with change in question order

Question Class	Change of order (Most similar ones) (Overall Accuracy : 93.2%)	Change of order (Most dissimilar ones) (Overall Accuracy: 94.1%)
Definition(Precision)	0.942	0.962
Definition (Recall)	0.943	0.943
Factoid(Precision)	0.913	0.921
Factoid(Recall)	0.891	0.890
Opinion(Precision)	1.0	1.0
Opinion (Recall)	1.0	1.0
Procedure(Precision)	0.911	0.925
Procedure (Recall)	0.986	0.962
Reason/Purpose(Precision)	0.912	0.911
Reason/Purpose (Recall)	0.912	0.921

4 Discussion and future works

In this current work we have tried to take a practical yet simpler approach towards the question classification problem. The approach came into existence when we realized that most of the time, we don't need to go through all the words and their semantic meanings in detail to map the questions to different classes. We thought it may be useful to give the machine this kind syntactic knowledge and a little semantic understanding to some extent to make it capable of classifying questions to its various classes. Instead of deriving handcrafted rules by watching each of the questions manually, we tried to establish a formalism through the Finite State Machines where the syntactic structure of the sentences could be learnt gradually with the example instances The result that we achieved shows that, the approach can be handy and may cope with various types of syntactic structure used for creating question sentences.

Because of not using deep semantic meaning analysis, our system failed to classify some of the questions to their corresponding classes. Lack of a proper recognizable structure was responsible for the failure in those cases. The system also made some wrong classifications when some very similar structures belonging to two different classes of questions came into

existence and this can be observed from the result that we achieved from the Recall parameter measurement of the each of the classes. The voting function we used rescued us to some extent to handle such situations. The result that we achieved encourages us to carry on with this approach further to improve it and use it in the other problem domain such as identifying question focus or may be in classifying questions to their finer classes.

References

1. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu., Performance issues and error analysis in an open domain question answering system., ACM Trans. Inf. Syst., 21(2):133–154
2. Zhang and W. Sun Lee., Question classification using support vector machines., In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 26–32, Toronto, Canada. ACM Press., 2003
3. Peters, M. Braschler, J. Gonzalo, and M. Kluck., Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, Rome, Italy, CLEF 2002
4. Voorhees., Overview of the TREC 2001 question answering track., In Proceedings of the 10th Text Retrieval Conference (TREC01), NIST, Gaithersburg, pages 157–165
5. Dell Zhang and Wee Sun Lee., Question Classification using Support Vector Machines., In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003.
6. K. Hacioglu and W. Ward., Question classification with support vector machines and error correcting codes., In Proceedings of NAACL/HLT-2003, Edmonton, Alberta, Canada, pages 28–30
7. X. Li and D. Roth., Learning question classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)., 2002.
8. A. Carlson, C. Cumby, J. Rosen, and D. Roth, The snow learning architecture., Technical Report UIUCDCS-R99-2101., University of Illinois at Urbana-Champaign, 1999
9. Mitchell, Tom M., 2nd edition., Machine Learning. McGraw-Hill, New York.
10. Fan Bu, Xingwei Zhu, Yu Hao and Xiaoyan Zhu., Function-based question classification., In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing., MIT, Massachusetts, USA, pages 1119–1128, 9-11 October 2010.
11. http://wordnet.princeton.edu/, accessed on January 22, 2013.
12. Kohavi R., and Provost F., 1998. On Applied Research in Machine Learning. In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Columbia University, volume 30, New York.
13. Thamar Solorio, Manuel P´erez, Manuel Montes-y-G´omez, Luis Villasenor-Pineda and Aurelio L´opez., A Language Independent Method for Question Classification., Proceedings of the 20th international conference on Computational Linguistics, Article No. 1374.
14. Question Answering for machine reading evaluation (QA4MRE) track of CLEF, http://celct.fbk.eu/ResPubliQA/index.php?page=Pages/pastCampaigns.php, accessed on February 14, 2013.

BCLaaS: implementação de uma base de conhecimento linguístico *as-a-service*

Mário Mourão[1] and José Saias[2]

[1] Cortex Intelligence - Évora, Portugal
mario.mourao@cortex-intelligence.com
[2] Departamento de Informática - ECT
Universidade de Évora, Portugal
jsaias@uevora.pt

Resumo Na área de Processamento de Linguagem Natural existem operações muito frequentes, independentemente do maior ou menor grau de análise linguística praticada. Um caso muito comum é a consulta da lista de sinónimos de um termo. Num ambiente com várias aplicações deste género, como Sistemas de Pergunta-Resposta, Análise de Sentimentos e outros, a manutenção destes recursos de apoio linguístico junto de cada aplicação torna-se pouco eficaz. Para cada ajuste numa coleção de sinónimos, por exemplo, seria necessário gerir o processo de atualização dos recursos individuais instalados junto das aplicações.
Este trabalho descreve a conceção de uma base de conhecimento linguístico *as-a-service*, considerando aspetos de armazenamento, comunicação e gestão de conteúdo, que permitam uma solução evolutiva e eficiente.

Keywords: Rede semântica, NoSQL, Grafos, *SaaS*

1 Introdução

A cada dia, pessoas e processos geram novos dados que se avolumam e propiciam o surgimento de novos serviços de procura de informação. Estes serviços aplicam técnicas complexas e semanticamente ricas na análise e relacionamento de dados, sejam estruturados, como quantitativos ou categóricos, ou não estruturados como uma publicação textual num fórum. Para tal, é comum a existência de bases de conhecimento, para o apoio na interpretação dos dados. A título de exemplo, uma tabela de sinónimos pode ajudar a captar o significado relevante de um termo. Numa pesquisa de notícias sobre compras de veículos, saber que *comprar* é sinónimo de *adquirir*, permitirá a identificação de mais casos de negócio, o que é fundamental neste serviço.
Na área de Processamento de Linguagem Natural (PLN), o uso destes serviços é cada vez mais comum, para recuperação de documentos, tradução automática e outras aplicações. Ao invés de cada aplicação ter o seu próprio repositório de conhecimento linguístico, este pode ser fornecido como um serviço autónomo que as aplicações consultam, evitando redundância e facilitando a atualização do conteúdo, com repercussão imediata em todas as aplicações cliente.
A centralização do conteúdo poderá trazer, por outro lado, eventuais limitações de desempenho, manifestadas por exemplo no tempo de resposta quando vários clientes geram múltiplas operações de leitura e escrita. Pretende-se que a base de conhecimento assente num serviço eficiente e escalável, cujo repositório tenha capacidade de evoluir sem prejuízo da consistência. Neste artigo, descrevemos uma solução para um serviço deste género, estudada no âmbito de um projeto entre a Cortex Intelligence e o Departamento de Informática da Universidade de Évora.

2 Trabalho Relacionado

O SentiLex-PT é um léxico de sentimento [7] com lemas e formas flexionadas, em Português. Cada entrada do léxico tem indicação de polaridade do sentimento e informação sobre o alvo dessa manifestação de sentimento, para adjetivos, nomes, verbos e expressões idiomáticas. Este recurso é

disponibilizado em ficheiro de texto CSV[3], deixando a cada aplicação cliente a responsabilidade de processar esse ficheiro e representar os dados num formato funcional próprio.
George Miller liderou o projeto WordNet [1], na Universidade de Princeton. Neste recurso, para o Inglês, cada conceito tem associado um *synset*, que aglomera um conjunto de sinónimos. Existem mais de 100.000 *synsets* e entre eles existem relações, incluindo sinonímia, hiperonímia e meronímia. Esta rede de conceitos pode ser consultada via Web[4] ou obtida para fins académicos ou comerciais, em formato de *scripts* Prolog, ficheiros de texto ou XML. A WordNet.PT [2] é uma base de dados de conhecimento lexical do Português desenvolvida no Centro de Linguística da Universidade de Lisboa, e que surge na sequência da WordNet de Princeton. Um conceito corresponde a um nó da rede, que pode ser representado por várias expressões lexicais, e cujo significado poderá ser deduzido pela posição relativa na rede, de acordo com as relações existentes. Existe uma interface Web para consultas aos serviço[5].
Os três casos referidos são bases de conhecimento com reconhecido valor, que funcionam como coleção de dados, que cada aplicação usará à sua maneira.
O WorldCat[6] é uma rede internacional de bibliotecas que dispõe de uma vasta base de conhecimento sobre dados bibliográficos e institucionais, cujo conteúdo evolui diariamente. Para facilitar a partilha da informação e a implementação de diversas aplicações junto dos parceiros desta rede, o acesso à base de conhecimento é normalizado, através de um serviço REST[7]. Na area da linguística, há vários exemplos da utilização de REST em serviços de dicionários[8] e *thesaurus*[9].
O sistema Wolfram|Alpha[10] tem uma abrangente base de conhecimento formada pela aplicação de múltiplos algoritmos sobre diferentes fontes de dados. Permite encontrar resposta a questões, não pela via de pesquisa na Web, mas através de cálculos dinâmicos sobre a base de conhecimento. Para colocar este serviço ao dispor da comunidade, a base de conhecimento é consultada com uma API baseada em REST, que uniformiza e facilita a integração das funcionalidades em aplicações móveis ou Web.

3 Solução Proposta

Esta secção enumera alguns aspetos e opções tomadas para os três pontos fulcrais do serviço: armazenamento, comunicação e gestão de conteúdo.

3.1 Armazenamento dos dados

No trabalho de Saias e Quaresma [5], a escolha automática do resultado para perguntas em língua natural é baseada numa rede semântica. As respostas candidatas, previamente extraídas pelo sistema, são validadas e ordenadas em função da afinidade semântica, entre o conjunto de hipóteses, e entre cada uma e elementos da pergunta. As técnicas empregues no sistema baseiam-se na ativação semântica ao longo das relações da base de conhecimento. Este repositório inclui conceitos associados a diversos domínios e relações como *hiperónimo*, *merónimo*, *sinónimo*, *antónimo*, *instânciaDe* [3]. Em trabalho posterior, sobre interfaces em língua natural [6,4] e análise de sentimentos, outros aspetos sobre o léxico (flexão de vocábulos) ou a semântica (novas relações) foram gradualmente acrescentados a esta base de conhecimento.
Uma vez que o serviço pode ser integrado em várias aplicações, com natureza distinta, e que poderão fazer acessos simultâneos ao sistema, a solução de armazenamento emerge como fator crítico

[3] CSV: *comma-separated values*, é um formato onde há um conjunto de valores em cada linha, separados por vírgula ou outro separador textual como ";" ou ":".
[4] http://wordnetweb.princeton.edu/perl/webwn
[5] http://www.clul.ul.pt/wn/
[6] http://www.oclc.org/uk/en/worldcat/
[7] REST significa *Representational State Transfer*, um protocolo de comunicação cliente-servidor sobre HTTP, alternativo a *Web Services*
[8] Merriam-Webster's Medical dic: http://www.dictionaryapi.com/products/api-medical-dictionary.htm
[9] Big Huge Thesaurus: http://words.bighugelabs.com/api.php
[10] http://products.wolframalpha.com/docs/WolframAlpha-API-Reference.pdf

para o serviço. Uma aplicação pode estar interessada apenas em sinónimos, outra apenas em conjugações de verbos, e outra poderá consultar polaridades de sentimento para alguns vocábulos. Assim, um dos aspetos a considerar é a ocorrência de acessos simultâneos a zonas diferentes do repositório. Depois, pela intenção de tornar o recurso multilingue, espera-se um crescimento substancial no repositório. E este crescimento não é necessariamente uniforme e organizado. Pode haver oportunidade de aumentar informação sobre variações lexicais, definir relações para traduções entre idiomas, adicionar uma entrada num catálogo de entidades mencionadas, ou outras. Assim, foi escolhida uma base de dados (BD) NoSQL[11] baseada em grafos, o Neo4J[12] . O modelo conceptual desta base de conhecimento, que é constituída de conceitos e relações, mapeia na própria estrutura de grafo da BD [8], com nós, propriedades e relações. Desta afinidade, espera-se uma vantagem na gestão dos conteúdos, e também no desempendo. A Figura 1 mostra a interface nativa do Neo4J para visualização da BD. Em Neo4J, as relações entre os nós podem ter um determinado tipo, tal como interessa neste serviço. Na figura, observamos relações do tipo *hiperónimo* e *sinónimo*. Tanto os nós como as relações da BD podem ter propriedades, com um nome e um valor. É através das propriedades Neo4J que se representa a maioria dos dados, como os vocábulos de um certo domínio do conhecimento, ou metainformação para identificar o idioma ou o contexto.
Para encontrar rapidamente o nó de base para um qualquer processo de análise da base de conhecimento, o Neo4J dispõe de um sistema de indexação flexível, adaptável a cada caso e baseado na tecnologia *Apache Lucene*[13]. A título de exemplo, há vantagem em indexar as palavras em Português num índice, usando outro para a indexação desses nós noutro idioma.

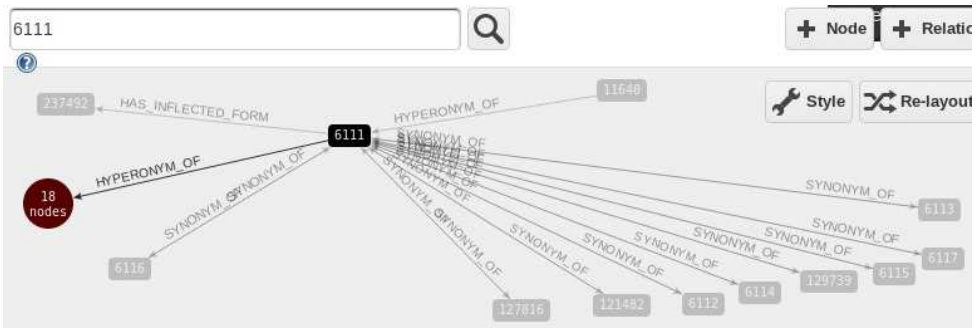


Figura 1. Interface Neo4j para consulta do repositório

3.2 API do serviço

Tendo em mente o futuro uso do sistema em diversas aplicações, possivelmente em ambiente *cloud*, era imperativa a adoção de um protocolo de comunicação universalmente aceite, que não dificultasse o processo de integração, independente da tecnologia da aplicação cliente. Assim, foi estabelecido que o protocolo do serviço seria REST, com formato de dados à escolha do cliente, entre JSON[14] e XML. Por omissão, o formato de envio dos dados é JSON, por ser mais compacto, como ilustrado na Figura 2 para um teste de equivalência semântica entre *adquirir* e *comprar*.
A maioria dos acessos ao serviço são consultas. Outros acessos destinam-se à gestão do conteúdo, com operações para editar os conceitos, relações, ou metainformação. A interface do serviço prevê uma separação destes perfis, um de consulta e outro de gestão de conteúdo.

[11] NoSQL: *not only SQL*. É uma classe de sistemas de gestão de bases de dados que não se baseiam no modelo relacional, nem usam SQL.
[12] Neo4j é uma tecnologia BD em grafos, *open-source* e de alto desempenho. http://www.neo4j.org/
[13] http://lucene.apache.org/
[14] *JavaScript Object Notation*: formato compacto para representação de dados. http://www.json.org/

Pela observação dos pedidos ao serviço, cedo se notou que as aplicações que requerem alguma análise linguística repetem operações. Por exemplo, em Sistemas de Pergunta-Resposta, a verificação da relação de sinónimo entre *t1* e *t2* pode ser necessária N vezes, apenas para o tratamento de uma pergunta. Como a comunicação até ao serviço atravessa a rede, o uso de cache tornou-se necessário. Isto poderia ser gerido pela aplicação cliente, decidindo quando realizar uma consulta ao serviço e quando usar alguma indicação prévia, que teria de manter localmente. A multiplicação desta necessidade, nos ambientes heterogéneos das diferentes aplicações cliente, levou à implementação de um sistema de cache no próprio serviço. Assim, a gestão da cache é transparente para a aplicação cliente, ficando disponível um conjunto de operações: ativar, desativar, ajustar tamanho máximo e o tempo de validade (*lease time*) das entradas.
As entradas em cache ficam em memória, do lado da aplicação cliente, representadas em objetos integrantes da própria API cliente do serviço, e ajudam a minimizar a comunicação com o repositório.

```
$ curl -HAccept:application/xml "http://localhost:8080/synonyms?t1=adquirir&t2=comprar"
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<BooleanValue>true</BooleanValue>

$ curl "http://localhost:8080/synonyms?t1=adquirir&t2=comprar"
{"value":true}
```

Figura 2. Resposta do serviço a um pedido de verificação de sinónimos: formatos XML e JSON

3.3 Aplicação Web para análise e edição do conteúdo

À medida que as aplicações cliente usavam o serviço, foram detetados alguns erros (devido à construção automática de parte do repositório) ou elementos em falta, como o simples estabelecimento de uma relação de sinonímia entre dois termos. Estas situações justificavam a validação, e eventual atualização, de alguns segmentos do repositório, por um humano, possivelmente um linguista. A interface da Figura 1 é demasiado técnica para o nível de abstração necessário a esta análise. Como tal, foi implementada uma aplicação Web para consultas e atualizações à base de conhecimento. Na Figura 3 podemos ver informação sobre um conceito da rede semântica, que, em Português, é *aluno*. Na zona central encontramos um círculo com vários segmentos de cor diferente. Cada segmento representa um tipo de relação. Em cada segmento há um conjunto de fatias, que representam as ligações individuais entre conceitos. No caso, o segmento de sinónimos está ativo. Esta forma amigável de visualização de informações em grafo foi implementada com uma versão modificada da biblioteca neovigator[15]. No canto superior direito podemos ver as principais propriedades do nó: um identificador numérico e a designação em Português. Depois surgem os tipos de relação, as direcionadas para o nó e as que partem daquele nó para outros. Quando é adicionada uma tradução para Inglês, o nó recebe uma nova propriedade com o nome CONCEPT_NAME_EN. Na interface adiciona-se uma tradução para determinado idioma. O modo como essa tradução é representada é escondido pela camada de armazenamento do sistema. A aplicação Web lida diretamente com os conceitos e relações, que aqui são a "lógica de negócio" da base de conhecimento, procurando ser independente da solução particular de armazenamento. Desta forma, o acesso aos dados realiza-se também através da interface REST. A Figura 4 ilustra uma operação em que se define que *aluno* é sinónimo de *discente* (em algum contexto e para a língua base).
Complementando a inserção manual de conteúdo no repositório, existe a possibilidade de importar dados de uma coleção. Uma das aplicações cliente, sobre análise de sentimentos, dispõe de um mecanismo de anotação de frases, para criação de regras para a extração de opinião. A Figura 5 tem um destes casos, onde é marcada a entidade *Universidade de Évora*, um verbo e o adjetivo *lucrativo*.

[15] https://github.com/maxdemarzi/neovigator

Daqui resultam dois elementos passíveis de importação para a base de conhecimento: a entidade e a polaridade positiva associada ao adjetivo. Este adjetivo não existe no recurso SentiLex-PT, pelo que ter esta informação na base de conhecimento será uma importante mais-valia para aquela aplicação cliente.

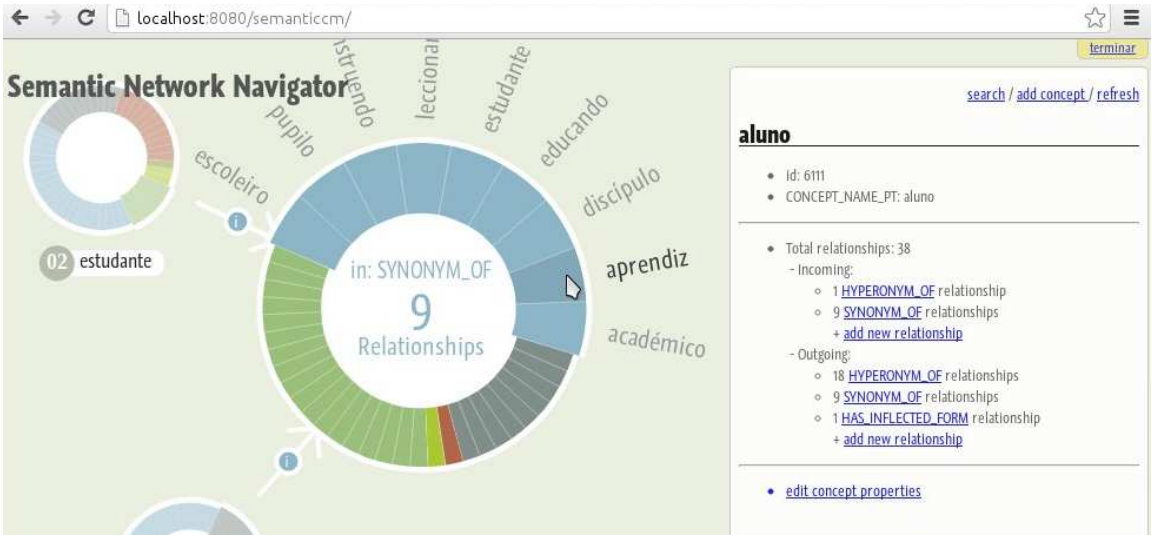


Figura 3. Visualização: resumo e relações de sinonímia para *aluno*

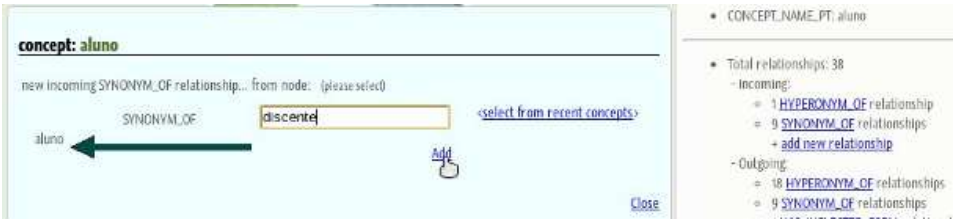


Figura 4. Edição do conteúdo: novo sinónimo

4 Conclusões

Este artigo relata um trabalho de transformação de um recurso estático convencional numa base de conhecimento *as-a-service*. Foram realçados aspetos cruciais para a implementação do serviço, envolvendo o armazenamento, a comunicação e a gestão de conteúdo. Ao evitar a distribuição de cópias do recurso estático convencional (como sucede com alguns dos exemplos referidos na secção 2), simplificamos a manutenção e coerência da base de conhecimento, que se pretende evolutiva e capaz de oferecer o mesmo grau de atualização a todas as aplicações cliente.
Em termos de trabalho futuro, existe o aspeto técnico, de monitorização da adequação desta arquitetura ao aumento do repositório e do número de pedidos a tratar, e uma vertente semântica, que incide sobre o conteúdo. A estrutura do serviço poderá evoluir, se o desempenho o determinar,

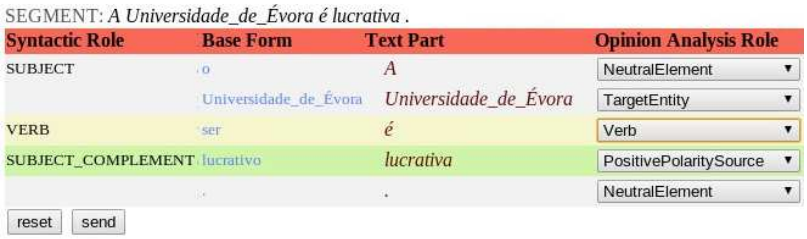


Figura 5. Evolução: importar dados da anotação em aplicações terceiras

designadamente pela introdução de replicação no repositório, o que permitiria algum balanceamento dos pedidos. O modelo de cache implementado no serviço reduz os acessos à BD, permitindo baixos tempos de resposta. Com exceção de cada primeiro pedido, o tempo de acesso à resposta local, em cache, é igual ou inferior ao verificado quando a base de conhecimento era mantida, local e integralmente, na aplicação cliente. Enquanto nesta abordagem o espaço de pesquisa era relativo a toda a base de conhecimento, a pesquisa na cache do serviço incide apenas sobre o subconjunto dos dados relevante para a aplicação cliente.
Paralelamente com a arquitetura do serviço, o conteúdo da base de conhecimento pode também evoluir. É aí que reside o valor deste serviço. A introdução de novas relações, novos conceitos ou traduções, são exemplos do trabalho contínuo de acompanhamento necessário a este recurso.

Agradecimento

Este trabalho enquadra-se numa investigação parcialmente financiada pelo programa QREN/PO Alentejo, no âmbito do projeto ALENT-07-0202-FEDER-018599.

Referências

1. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.
2. Palmira Marrafa, Raquel Amaro, Rui Pedro Chaves, Susana Lourosa, Catarina Martins, and Sara Mendes. Wordnet.pt – uma rede léxico-conceptual do português on-line. In *XXI Encontro da Associação Portuguesa de Linguística*, Porto, Portugal, Setembro 2005.
3. José Saias. *Contextualização e Activação Semântica na Selecção de Resultados em Sistemas de Pergunta-Resposta*. PhD thesis, Universidade de Évora, 2010.
4. José Saias, P. Quaresma, P. Salgueiro, and T. Santos. Binli: An ontology-based natural language interface for multidimensional data analysis. *Intelligent Information Management*, 4(5):225–230, September 2012.
5. José Saias and Paulo Quaresma. Semantic networks and spreading activation process for qa improvement on text answers. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology - STIL2011*, Cuiabá, Mato Grosso, Brasil, 2011. ISSN: 2175-6201.
6. José Saias and Paulo Quaresma. Di@ue in clef2012: question answering approach to the multiple choice qa4mre challenge. In *Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, Rome, Italy, September 2012. ISBN 978-88-904810-3-1.
7. Mário J. Silva, Paula Carvalho, and Luís Sarmento. Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, Lecture Notes in Computer Science (LNCS), pages 218–228. Springer-Verlag, 2012.
8. Jim Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, SPLASH '12, pages 217–218, New York, NY, USA, 2012. ACM.

POS-Tagging usando Pesquisa Local

João Laranjinho
joao.laranjinho@gmail.com
Irene Rodrigues
ipr@di.uevora.pt
Lígia Ferreira
lsf@di.uevora.pt

Universidade de Évora

Abstract. This paper presents a system of part-of-speech tagging, domain independent, for Portuguese and English grammatical tagging.
The system uses morpho-syntactic information that comes from a local dictionary to complement the information obtained using dictionaries avaible on the network as the Priberam and LookWAYup.
The tagger is based on an evaluation function, its parameters are optimized using a training corpus.
In the optimization of evaluation function parameters, some techniques are used such as local search to reduce the search space.
In the evaluation system, we use two texts from Reuters corpora: Testa (during training) and Testb (in testing phase).

1 Introdução

Os sistemas de part-of-speech tagging classificam gramaticalmente átomos de um texto.

As formas das palavras são frequentemente ambíguas no part-of-speech tagging. Numa expressão, essas ambiguidades normalmente são resolvidas pelo contexto das palavras.

Os sistema de part-of-speech tagging dividem-se em dois grupos: baseados em regras e estocásticos.

Para o inglês, alguns dos sistemas actuais conseguem um valor que ronda os 96-97% de precisão.

Neste artigo apresentamos um sistema automático independente do domínio para etiquetação gramatical de texto.

Este artigo vem no seguimento de um anterior trabalho de marcação de nomes próprios [1] que usa técnicas de pesquisa local.

Na etiquetação é usada informação morfo-sintáctica de dicionários que se encontram na WEB.

Para reduzir o espaço de pesquisa foram usadas algumas técnicas de pesquisa local.

O desempenho de um sistema de part-of-speech tagging pode ser medido com diversão métricas que representam o desempenho em valores numéricos.

As três métricas que normalmente são utilizadas para avaliar o desempenho são as seguintes: Abrangência (Recall), Precisão (Precision) e Medida-F (F-Measure).

- A Abrangência mede a relação entre o número de resultados correctos e o número de resultados existentes. A fórmula da Abrangência é a seguinte:
  Abrangência $= \frac{Resultados\ Correctos\ \cap\ Resultados\ Existentes}{Resultados\ Existentes}$
- A Precisão mede a relação entre o número de resultados correctos e o número de resultados obtidos. A fórmula da Precisão é a seguinte:
  Presisão $= \frac{Resultados\ Correctos\ \cap\ Resultados\ Obtidos}{Resultados\ Obtidos}$
- A Medida-F é uma métrica harmónica de Precisão (P) e Abrangência (A). A fórmula da *Medida-F* é a seguinte:
  Medida-F $= 2^* \frac{P*A}{P+A}$

2 Arquitectura do Etiquetador

O etiquetador contém duas etapas: optimização e etiquetação. Na figura 1 são apresentados os módulos de optimização e na figura 2 são apresentados os módulos de etiquetação.

2.1 Optimização

A etapa de optimização contém os seguintes módulos: pré-processamento, análise lexical, pesquisa local e saída.

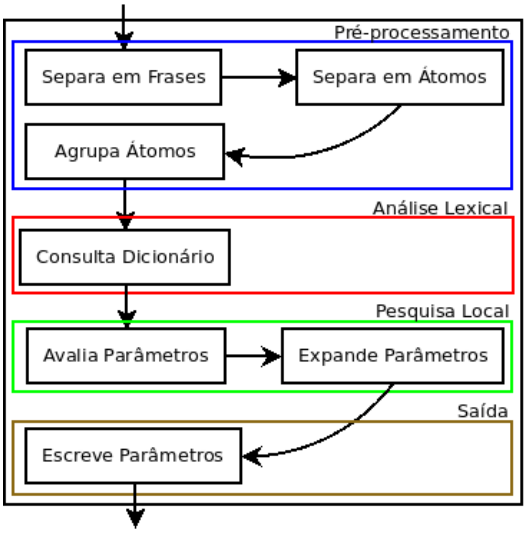


Fig. 1. Arquitectura do Optimizador

No pré-processamento separa-se o texto em frases e as frases em átomos. As frases são constituídas por átomos e os átomos por sequências de caracteres. Ainda no pré-processamento os átomos são agrupados em tripos para serem usados na função de avaliação.

Na análise lexical consulta-se em dicionários *on-line* a informação morfo-sintática das palavras que não se encontram no dicionário local, guardando-se essa informação no dicionário local.

Na pesquisa local geram-se conjuntos de parâmetros iniciais, que posteriormente serão avaliados. Quando um conjunto de parâmetros contém outros conjuntos vizinhos com valor de avaliação superior, expandem-se os vizinhos e em seguida avaliam-se. A avaliação termina quando não são encontrados mais vizinhos com valor de avaliação superior ou um critério de paragem ter sido alcançado.

Finalmente na saída transcreve-se para um ficheiro o conjunto de parâmetros que obteve o valor mais alto de avaliação.

2.2 Etiquetação

A etapa de etiquetação contém os seguintes módulos: pré-processamento, análise lexical, avaliação e saída.

No pré-processamento separa-se o texto em frases e as frases em átomos. As frases são constituídas por átomos e os átomos por sequências de caracteres. Ainda no pré-processamento os átomos são agrupados em tripos para serem usados na função de avaliação.

Na análise lexical consulta-se em dicionários *on-line* a informação morfo-sintática das palavras que não se encontram no dicionário local, guardando-se essa informação no dicionário local.

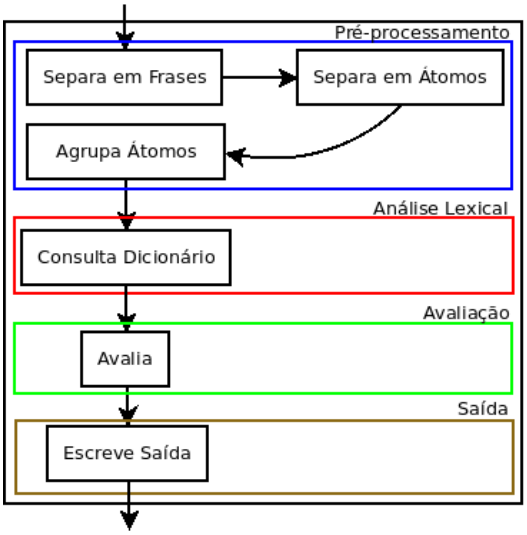


Fig. 2. Arquitectura do Etiquetador

Na avaliação são atribuídas classes gramaticais a cada átomo através de uma função que usa os parâmetros apurados na fase de optimização.

Finalmente na saída escreve-se num ficheiro para cada átomo a categoria correspondente.

3 Corpus

Nos testes com o etiquetador usamos os seguintes 2 ficheiro do corpus da Reuters: testa (na fase de treino) e testb (na fase de teste).

Para conhecermos um pouco melhor o corpus fizemos testes com: átomos ambíguos, átomos sem contraditórios e todos os átomos. A distribuição dos átomos encontra-se na tabela 1

	testa	testb
Todos	51.360	46.435
Ambíguos	24.144	–
Sem Contraditório	45.184	–

Table 1. Distribuição dos átomos no Corpus

Os átomos ambíguos são átomos que se encontram no dicionário com entrada em mais que uma classe gramatical. Os átomos contraditórios são átomos com iguais características no dicionário que ocorrem no corpus com diferentes classificações.

4 Função de Avaliação

Na função de avaliação estuda-se o impacto das classes gramaticais na etiquetação gramatical de texto. O estudo inclui informação sobre: átomo anterior, átomo em análise e átomo seguinte.

A função de avaliação usada é a seguinte:

$F\_CLASSE(A, A-1, A+1) = \sum_{Ci} Pi * Ci(A) + Pi' * Ci(A-1) + Pi'' * Ci(A+1)$

Na função de avaliação *A-1*, *A* e *A+1*, representam átomo anterior, átomo em análise e átomo seguinte, e os *Pi*s são parâmetros multiplicativos das classes gramaticais *Ci*s. Os *Pi*s podem tomar

valores no intervalo -150 a 150 e os *Ci*s podem tomar o valor 0 (átomo é da classe) ou 1 (átomo não é da classe).

As classes gramaticais consideradas são classes simplificadas do *package* NTLK [1] que podem ser vistas na tabela 2.

No etiquetador é usada uma função de avaliação para cada classe gramatical na qual são apurados os parâmetros para a classe em questão.

Tag	Meaning	Examples
ADJ	adjective	new, good, high, special, big, local
ADV	adverb	really, already, still, early, now
CNJ	conjunction	and, or, but, if, while, although
DET	determiner	the, a, some, most, every, no
EX	existential	there, there's
FW	foreign word	dolce, ersatz, esprit, quo, maitre
MOD	modal verb	will, can, would, may, must, should
N	noun	year, home, costs, time, education
NP	proper noun	Alison, Africa, April, Washington
NUM	number	twenty-four, fourth, 1991, 14:24
PRO	pronoun	he, their, her, its, my, I, us
P	preposition	on, of, at, with, by, into, under
TO	the word to	to
UH	interjection	ah, bang, ha, whee, hmpf, oops
V	verb	is, has, get, do, make, see, run
VD	past tense	said, took, told, made, asked
VG	present participle	making, going, playing, working
VN	past participle	given, taken, begun, sung
WH	wh determiner	who, which, when, what, where, how

Table 2. Conjunto de tags simplificado do NLTK

5 Avaliação

Num dos testes fizemos 3 experiências nas quais apuramos os parâmetros usando a função de avaliação para cada classe gramatical com as seguintes informações do ficheiro testa: átomos ambíguos, átomos sem contraditórios e todos os átomos. Posteriormente com os parâmetros encontrados foi feita etiquetação do ficheiro testb, os resultados encontram-se na tabela 3.

Num outro teste fizemos também outras 3 experiências na quais etiquetamos o ficheiro testb usando os parâmetros encontrados para cada uma das classes gramaticais com as seguintes informações do ficheiro testa: átomos ambíguos, átomos sem contraditórios e todos os átomos. Neste teste escolhemos para cada átomo a classe gramatical que obteve o valor mais alto de avaliação. Os resultados alcançados com os átomos ambíguos, átomos sem contraditórios e todos os átomos, foram respectivamente 0.8348, 0.8522 e 0.8598.

Finalmente num outro teste fizemos 3 experiências nas quais marcamos: adjectivos, substantivos e nomes próprios. No ficheiro de treino testa estas classes são aquelas que têm maiores frequências de átomos e onde a etiquetação teve menor desempenho. Nesse teste antes de etiquetarmos uma determinada classe, verificamos quais os erros de etiquetar outras classes dado etiquetar a classe pretendida. A etiquetação foi feita sucessivamente das classes que obtiveram menor erro associado para as classes com maior erro durante 3 iterações. Os ganhos na etiquetação das classes adjectivo, substantivo e nome próprio, foram respectivamente 0.0746, 0.0673 e 0.0708.

[1] http://nltk.org/

	Ambíguos			Sem contraditórios			Todos		
CAT	PREC	COB	MED-F	PREC	COB	MED-F	PREC	COB	MED-F
ADJ	0,6899	0,7434	0,7157	0,6878	0,7934	0,7368	0,6909	0,8217	0,7507
ADV	0,7661	0,2161	0,3371	0,8461	0,6358	0,7260	0,8761	0,6424	0,7413
CONJ	0,9793	0,5569	0,7100	0,9922	0,9935	0,9928	0,9961	0,9908	0,9934
DET	0,9846	0,9825	0,9836	0,9774	0,9872	0,9823	0,9847	0,9882	0,9865
EX	0,8889	0,9412	0,9143	0,9655	0,8235	0,8889	0,8857	0,9118	0,8986
FW	1,0	0,0000	0,0000	1,0	0,0000	0,0000	1,0	0,0000	0,0000
MOD	0,9431	0,9888	0,9654	0,9462	0,9851	0,9653	0,9336	0,9963	0,9639
N	0,8019	0,8356	0,8184	0,7602	0,8750	0,8136	0,7713	0,8688	0,8172
NP	0,8023	0,5702	0,6666	0,9407	0,6032	0,7351	0,8891	0,6615	0,7586
NUM	0,9792	0,9890	0,9840	0,9816	0,9990	0,9902	0,9798	0,9863	0,9830
PRO	0,9965	0,9567	0,9762	0,9900	0,9867	0,9883	0,9955	0,9767	0,9860
P	0,9260	0,9835	0,9539	0,9284	0,9773	0,9522	0,9265	0,9766	0,9509
TO	1,0	0,4315	0,6029	1,0	0,9963	0,9982	1,0	0,9988	0,9994
UH	0,5556	0,7143	0,6250	0,8333	0,7143	0,7692	0,8000	0,5714	0,6667
V	0,9136	0,7305	0,8118	0,8828	0,7848	0,8309	0,8994	0,8001	0,8469
VD	0,8795	0,8546	0,8669	0,8769	0,9141	0,8951	0,8879	0,9182	0,9028
VG	0,8488	0,6033	0,7053	0,8229	0,9215	0,8694	0,8550	0,9256	0,8889
VN	0,8318	0,7136	0,7682	0,8411	0,7760	0,8072	0,8739	0,7206	0,7899
WH	0,9464	0,8689	0,9060	0,9579	0,9705	0,9642	0,9581	0,9738	0,9659
SYM	0,9854	0,0690	0,1289	0,9924	0,9990	0,9957	0,9929	0,9973	0,9951

Table 3. Etiquetação isolada usando no treino átomos ambíguos, átomos sem contraditórios e todos os átomos

		MARCAÇÃO		
		ADJ	N	NP
% DE ERRO	ADJ	–	0,0285	0,0160
	ADV	0,0424	0,0030	0,0037
	CONJ	0,0	0,0	0,0017
	DET	0,0030	0,0007	0,0075
	EX	0,0	0,0	0,0
	FW	0,0	0,0	0,0020
	MOD	0,0	0,0019	0,0018
	N	0,0811	–	0,0335
	NP	0,1436	0,1327	–
	NUM	0,0089	0,0064	0,0292
	PRO	0,0	0,0	0,0023
	P	0,0050	0,0004	0,0040
	TO	0,0	0,0	0,0017
	UH	0,0	0,0	0,0006
	V	0,0026	0,0440	0,0021
	VD	0,0073	0,0037	0,0002
	VG	0,0083	0,0064	0,0005
	VN	0,0069	0,0010	0,0020
	WH	0,0	0,0	0,0003
	SYM	0,0	0,0	0,0018

Table 4. Percentagem de erros na marcação de adjectivos, substantivos e nomes próprios

6 Conclusão e Trabalho Futuro

Na experiência em que usamos os parâmetros que foram apurados com todos os átomos do ficheiro testa conseguimos melhores resultados. No entanto, a diferença em relação à experiência na qual usamos os átomos sem contraditórios não foi significativa (não chegou a 0.01 de erro). Já na experiência na qual usamos apenas átomos ambíguos existiu uma perda de cerca de 0.02.

Antes de etiquetar uma determinada classe deve-se verificar quais os erros de etiquetar outras classes dado etiquetar a classe pretendida. A etiquetação deve ser feita das classes que obtiveram menor erro para as classes que obtiveram maior erro durante várias iterações até não existir mais ganhos no desempenho do sistema.

Como trabalho futuro pensamos fazer estudos nos quais:

- retirar os contraditório com menores frequências;
- retirar átomos não ambíguos;
- marcar sucessivamente as classes que obtêm maior percentagem de erro até não existir perda no desempenho;
- etiquetar nomes próprios antes de etiquetar todas as outras classes gramaticais;
- adicionar informação de dois ou mais átomos anteriores e de dois ou mais átomos seguintes ao átomo em análise;

References

1. João Laranjinho, Irene Rodrigues, and Lígia Ferreira. Marcação de nomes próprios usando técnicas de pesquisa local e recorrendo a fontes de conhecimento na internet. In *JIUE2011*, 2011.

A Review on Cooperative Question-Answering Systems

Dora Melo[1], Irene Pimenta Rodrigues[2], and Vitor Beires Nogueira[2]

[1] Iscac, Instituto Politécnico de Coimbra e CENTRIA, Portugal
dmelo@iscac.pt,
[2] Universidade de Évora e CENTRIA, Portugal
{ipr,vbn}@di.uevora.pt

Abstract. The Question-Answering (QA) systems fall in the study area of Information Retrieval (IR) and Natural Language Processing (NLP). Given a set of documents, a QA system tries to obtain the correct answer to the questions posed in Natural Language (NL). Normally, the QA systems comprise three main components: question classification, information retrieval and answer extraction. Question classification plays a major role in QA systems since it classifies questions according to the type in their entities. The techniques of information retrieval are used to obtain and to extract relevant answers in the knowledge domain. Finally, the answer extraction component is an emerging topic in the QA systems. This module basically classifies and validates the candidate answers. In this paper we present an overview of the QA systems, focusing on mature work that is related to cooperative systems and that has got as knowledge domain the Semantic Web (SW). Moreover, we also present our proposal of a cooperative QA for the SW.

Keywords: Question-Answering Systems, Information Retrieval, Information Extraction, Natural Language Processing

1 Introduction

The QA systems try to find answers that are accurate and concise to questions stated in NL, posed by the user in their own terminology [16]. These systems belong to the Computer Science (CS) area and are directly related to the studies in IR and NLP, and fit within the building systems that automatically answer to questions raised by users in NL.

To find an answer to a question, a QA system can resort not only to structured databases but also to sets of documents in NL. The research domain can vary between small sets of documents locally stored, to enterprise internal documents, to networks of news reports, and even to the internet. The main goal of the QA systems is to provide accurate answers to the question posed by users, by consulting its knowledge base.

Research in this area deals with a wide range of question types, including: facts, lists, definitions, hypothetical, semantically limited, language-independent questions (*cross-lingual questions*). Prior knowledge of the type of expected answer help QA systems to extract accurate and correct answers from the collections of documents that make up their knowledge domain.

The first QA systems were developed in 1960's and were essentially NL interfaces for intelligent systems built for specific domains. The advance of the internet reintroduced the need for research techniques pleasing to the user that reduce information overload, posing new challenges for research in automating question answering.

The amount of information on the internet has increased exponentially over the years, with content covering almost any subject. As a result, when users look for certain information, get a little confused with the vast amount of information returned by search engines. Virtually any type of information is available in the internet in one way or another, having billions of web pages available on the internet. Managing such quantities of information is not a simple task. Search engines such as Google and Yahoo, return links along with fragments of text of all documents in response to the request made by users, and that will allow them to navigate the content through a long list of results to look for the answer wanted.

The development of QA systems emerged as an attempt to solve this problem of information overload. QA systems can be classified into two categories according to their domains: closed and open domain. QA systems with closed domain deal with questions based on a specific domain (eg, medicine, music, etc.). These can be seen as simpler systems, since the techniques of NLP can explore specific areas of knowledge, often formalized in ontologies. The specific area of a QA system involves the intensified use of NLP, formalized through the construction of an ontology of the considered domain. Domains may refer to contexts where a limited type of questions are accepted. Open domain QA systems handle questions about anything, and can only rely on general ontologies and world knowledge. Usually there is more information available from which to extract answers.

The remainder of the paper is structured as follows. In Section 2, we present the proposals in the field of QA system that we consider more relevant for our work, namely the ones targeting cooperation and the semantic web. In Section 3, we present the architecture of a typical QA system. In Section 4, we introduce some characteristics about the methodologies that are used more often in QA. Section 5, enumerates several challenges inherent to the development of these systems. In Section 6, we present some current research topics. In Section 7 we present a general view of our proposal for a cooperative QA systems for the SW developed under the PhD in Informatics. Finally, in Section 8, we establish the final conclusions.

2 State of the Art

The most important QA application areas are information extraction from the entire web, online databases, and inquiries on individual websites. Current QA [1] systems use text documents as their underlying knowledge source and combine various NLP techniques to search for an answer to an user question. In order to provide users with accurate answers, QA systems need to go beyond lexical-syntactic analysis to semantic analysis and processing of texts and knowledge resources. Moreover, QA systems equipped with reasoning capabilities can derive more adequate answers by resorting to knowledge representation and reasoning systems like Description Logic and ontologies. A survey on ontology-based QA is presented in [21]. A study on the usability of NL Interfaces and NL query languages, over ontology-based knowledge, for the end-users is presented in [18]. To that end, the authors present four interfaces that enable different search languages and they present a comparative study of their use. They conclude that users have a clear preference for queries expressed in NL and a small set of expressions composed with some keywords or some formal structures.

Several conferences and workshops have been focusing in aspects of search in QA systems. Starting in 1999, the Text REtrieval Conference (TREC)[3] has invested in a trajectory involving QA systems having as main goal the evaluation of systems answering to factual questions using a set of documents from the TREC corpus. A significant number of systems presented in this evaluations were able to successfully combine IR and NLP techniques. In [2], the authors present a review of QA systems and they compare three main approaches to QA systems based in NLP, in IR and in questions modelling, emphasizing their main differences and the application context that is more adequate to each system.

Cooperative QA is an automated QA in which the system, taking as the starting point an input query, tries to establish a controlled dialogue with its user, i.e, the system collaborate automatically with users to find the information that they are seeking. These systems provide users with additional information, intermediate answers, qualified answers, or alternative queries. One form of cooperative behaviour involves providing associated information that is relevant to a query. Relaxation generalizing a query to capture neighbouring information is a means to obtain possibly relevant information. A cooperative answering system described in [12] uses relaxation to identify automatically new queries that are related to the original query. A study on adapting machine learning techniques defined for information extraction tasks to the slightly different task of answer extraction in QA systems is presented in [17]. The authors identified the specificities of

[3] http://trec.nist.gov/

the systems and also tested and compared three algorithms, assuming an increasing abstraction of NL texts. In [7], a semantic representation formalism dedicated to cooperative QA system is presented, this system is based in conceptual and lexical structures and represents homogeneously web texts, NL questions and related answers. This author also presents and analyses some of the prerequisites in order to build cooperative answers depending on the resources, the knowledge and the process. In order to enhance cooperative QA systems, in [23] a set of techniques to improve these systems is presented and the potential impact of their use is discussed.

A cooperative answer [10,13] to a NL question is an indirect answer that is more useful to the user then a direct and literal answer. A cooperative answer may explain a failure that has occurred during the results computation and/or suggest related questions in order to continue with the search. When the system can obtain some results, a cooperative answer can supply additional information that was not explicitly required by the user. Cooperative answers fit into the context of QA systems and they were originally motivated by the wish to approximate system user dialogue from a human dialogue. The cooperative answer processing is preferable to usual techniques of answer extraction focusing on the users since: first it humanizes the system; second it enables the use of adapted vocabulary; and finally it allows the introduction of non-solicited information that may interest the user.

There are some examples of works that try to build answers, instead of merely extract and retrieve. In [28], the authors propose a model for a QA system where the system, departing from the user question, tries to establish a controlled dialogue with the user. In the dialogue, the system has its main goal to identify the user question and to suggest new question related with the user initial question. The dialogue controller is based on the concept structure in the knowledge base, in the domain constraints and in conditioning specific rules. In [15] a system prototype is presented, this system returns cooperative answers, corrects missing concepts, it intends to meet the user needs and it uses the database semantic information in order to formulate coherent and informative answers. The main characteristics of lexical strategies that were developed by humans intellect in order to answer questions are presented in [8]. This author also presents how this strategies can be reproduced in the construction of QA systems in particular in intelligent cooperative QA systems. A answer search method to find answers that are in a neighbour of an answer to the user initial question is presented by [14], this method can be used to process answers that can satisfy the user needs and claims.

Advanced reasoning techniques that are used in QA systems raise new challenges to researchers since answers are not just extracted directly from the text or from structured databases, building an answer can evolve several reasoning forms with the goal of generate explained and justified answers. The integrated knowledge representation and reasoning mechanisms enable the systems, for instance, to anticipate an answer to questions that may raise and to solve cases where the answer can not be found in the knowledge base. These systems should identify and explain false assumptions and others conflict types that might be found in a question.

In [26], an approach to cooperative QA systems is presented, using databases as the domain knowledge source. In [6], the author proposes a logic based model used to generate intentional and precise answers in a cooperative QA system. This author in [5] presents an approach to draw logic based QA systems, WEBCOOP, these systems integrates knowledge representation and reasoning techniques in order to generate cooperative answers to NL questions posed on the web. PowerAqua [20] is a multi-ontologies based QA system that given a NL question returns answers that are computed using relevant resources distributed in SW.

3 The Architecture of a Question-Answering System

The typical architecture of a QA system comprise three main and distinct phases: question classification; information retrieval and document processing; and information extraction.

The question classification is the first phase and consists of classifying questions according to a defined type, generates the kind of the expected answer, extracts keywords and reformulates the questions into multiple questions semantically equivalent. Reformulate a question into a number

of questions with similar meaning is also known as question expansion and provides the basis for increasing and improving the performance of information retrieval mechanism.

The information retrieval phase is very important for QA systems. If it is not found in any document a correct answer, the continuity of the process in searching for an answer is finished. The fragments precision and classification that are candidates for the answer may also affect system performance, during the information recovery phase.

The extraction of the answer is the final phase of the QA systems, and states the difference between what is considered a QA system and the usual meaning given to a text retrieval system. The answer extraction technology becomes an influential and decisive factor in the QA system to achieve the final results. Thus, the answer extraction technology is also considered a necessary and important module for the QA systems.

4 Methodologies used

The QA systems are directly dependent on a good search in corpus - without documents containing the answer, there is very little that the QA systems can do. So it makes sense that larger sets of documents generally provide better performance on QA systems, unless the domain of the question is orthogonal to the set of documents. The concept of data redundancy in massive collections of documents, such as internet, i.e. small fragments of information that are susceptible to be formulated in many different ways, in different contexts and documents [19], leads to two benefits: by having the right information and appear in many forms, the burden done in QA systems to perform complex techniques of NLP in order to understand the text is smaller; the correct answers can be filtered out of false positives, taking into account that a correct answer may appear more often in documents that incorrect answers.

Most of the QA systems use NL text documents as domain of knowledge. The techniques of NLP are used both for the processing the questions as well as to index or process the text corpus where answers are extracted.

An increasing number of QA systems use the internet as its corpus of text and knowledge. However, many of these tools do not produce a pleasurable, cooperative and informative answer to the user, which in turn employ superficial methods (techniques based on correspondence between words, models, etc.) to produce a list of documents containing the probable answer.

In current QA systems [1], typically, the questions classifier determines the type of question and the type of the expected answer. After the question is analysed, the system normally uses several modules that apply techniques increasingly complex, in a gradually reduced amount of text. Retrieving documents uses search engines to identify the documents or paragraphs in documents collections that are susceptible to contain the correct answer. Subsequently, a filter select small fragments of text that contain strings of the same type than the expected answer. For instance, if the question is "Who invented Penicillin?", the filter returns the text that contains names of people. Finally, the answer extraction search for more information or tracks in the text that determines whether a candidate answer can really answer the question.

5 Challenges of Developing Question-Answering Systems

The development of QA systems have released several challenges motivated, mostly, by the exponential increase of the information available, the advance in technology and by the demands and requirements of users. Wherefore, it is now possible to enumerate a collection of problems that continue to have full attention among researchers and were initially identified by a group of researchers and presented in [9]:

Question classes - Different types of questions require the use of different strategies to find the answer. Question classes are arranged hierarchically in taxonomies.

Question processing - The same information can be expressed in various ways. A semantic model of question understanding and processing would recognize equivalent questions, regardless of how they are presented. This model would enable the translation of complex questions into a series of simpler questions, would identify ambiguities and treat them in context or by interactive clarification.

Context - Questions are usually asked within a context and answers are provided within that specific context. The context can be used to clarify a question, resolve ambiguities or keep track of an investigation performed through a series of questions. For instance, the question, "Why did Joe Biden visit Iraq in January 2010?" might be asking why Vice President Biden visited and not President Obama, why he went to Iraq and not Portugal or some other country, why he went in January 2010 and not before or after, or what Biden was hoping to accomplish with his visit. If the question is one of a series of related questions, the previous questions and their answers might guide the system on the intentions of the user.

Data sources - Before a question can be answered, it must be known what knowledge sources are available and relevant. If the answer to a question is not present in the data sources, no matter how well the question processing, information retrieval and answer extraction is performed, a correct result will not be obtained.

Answer extraction - Answer extraction depends on the complexity of the question, on the answer type provided by question processing, on the actual data where the answer is searched, on the search method and on the question focus and context.

Answer formulation - The result of a QA system should be presented in a way as natural as possible. For example, when the question classification indicates that the answer type is a name (of a person, organization, etc.), a quantity (size, distance, etc.) or a date, the extraction of a single datum is sufficient. For other cases, the presentation of the answer may require the use of fusion techniques that combine the partial answers from multiple documents.

Real time question answering - There is need for developing QA systems that are capable of extracting answers from large data sets in several seconds, regardless of the complexity of the question, the size and heterogeneity of the data sources or the ambiguity of the question.

Cross-lingual - The ability to answer a question posed in one language using an answer corpus in another language (or even several). This allows users to consult information that they cannot use directly.

Interactive - It is often the case that the information needed is not well captured by a QA system; the question processing part may fail to classify properly the question; or the information needed for extracting and generating the answer is not easily retrieved. In such cases, the questioner might want not only to reformulate the question, but to have a dialogue with the system.

Advanced reasoning - More sophisticated users expect answers that are outside the scope of written texts or structured databases. To upgrade a QA system with such capabilities, it would be necessary to integrate reasoning components operating on a variety of knowledge bases, encoding world knowledge and common-sense reasoning mechanisms, as well as knowledge specific to a variety of domains.

Information clustering - Information clustering for QA systems is a new trend that is originated to increase the accuracy of question answering systems through search space reduction [27].

User profile - The user profile captures data about the user, comprising context data, domain of interest, reasoning schemes frequently used by the user, common information established within different dialogues between the system and the user. The profile may be represented as a predefined template, where each template slot represents a different profile feature.

6 Current Research Topics

In recent years, the QA systems evolved to incorporate additional domains of knowledge [22,4]. For instance, the QA systems have been developed to automatically answer to questions of geospatial

and temporal context, questions of terminology and definitions, biographical questions, cross-lingual questions, and questions about audio content, images or even video. The current research topics of QA system include:

- Cooperation and clarification of questions and/or answers
- Answers reuse
- Knowledge representation and reasoning
- Social media analysis
- Sentiment analysis

7 Cooperative Question-Answering System for Semantic Web

The wide range of challenges related to the development of QA systems, presented above; the growing need for intelligent search engines able to satisfy the demands of many different kinds of Internet users; the need for cooperation and interaction between users and systems; the need to produce, by the system, accurate answers, informative and expressed as closest as possible to NL; were reasons enough to make the decision to proceed with the arduous task: the development of a cooperative QA system for the SW [25,24].

The proposed cooperative QA system receives NL questions and is able to produce a cooperative answer, also expressed in NL, obtained from knowledge base. When the system can not decide the correct path to obtain the answer, it starts a controlled clarifying dialogue with the user. The system includes deep parsing, makes use of ontologies, OWL2 descriptions and other web resources such as WordNet [11] and DBpedia [3].

Our goal is to provide a system that is independent of prior knowledge of the semantic resources by the user and is able to provide a cooperative, direct, accurate and informed answer to questions posed in NL. To this purpose, the architecture of the proposed system is enriched with a Discourse Controller (DC). The DC is invoked after transforming the NL question into its semantic representation and controls all the steps until the end, i.e. until the system can return an answer to the user: from the phase of question classification, passing through the phase of information retrieval, until the phase of answer processing. That is, the DC tries to make sense of the initial question by: analysing the question and the type expected answer; analysing the ontology structure and the structured information available on the web (such as DBpedia); and use the correspondence of similarity between strings and generic lexical resources (such as WordNet), with the objective to provide a clear and informative answer.

The DC deals with the set of discourse entities: verifies the question presupposition, to decide the sources of knowledge to be used; decides when the answer has been achieved or iterates using new sources of knowledge. The decision of when to relax a question in order to justify the answer, when to clarify a question and how to clarify it, is also taken in this module. Thus, the DC represents the intentions and beliefs of the system and the user, the structure of discourse and the context of the question, includes implicit context (such as spatial and temporal knowledge), entities and information useful for the semantic interpretation (like discourse entities used for anaphora resolution, on finding what an instance of an expression is referring to), that allow to add the ability to deal with multiple answers and provide justified answers.

The QA systems strongly depend on reasoning, fact that led us to choose the Logic Programming, specifically Prolog, for their development. Furthermore, there is a vast amount of libraries and extensions for handling and questioning OWL2 ontologies, as well as incorporate the notions of context in the process of reasoning.

8 Conclusion

The main objective of the QA systems is to provide accurate answers to questions posed by users, rather than returning lists of complete documents or fragments of documents that are closer of the expected answer, as with most IR systems. In this paper we presented an overview of the

QA systems, focusing on mature work that is related to cooperative systems and that has got as knowledge domain the SW, highlighting aspects of typical architecture of a QA system, some features on methodologies that are used more often in QA systems, development challenges of QA systems and some current research topics. Finally, we also presented our proposal of a cooperative QA for the SW.

References

1. Allam, A., Haggag, M.: The Question Answering Systems: A Survey. International Journal of Research and Reviews in Information Sciences 2(3), 211–221 (2012)
2. Andrenucci, A., Sneiders, E.: Automated question answering: Review of the main approaches. In: ICITA (1). pp. 514–519. IEEE Computer Society (2005)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J.: Dbpedia: A nucleus for a web of open data. The Semantic Web 4825(Springer), 722–735 (2007)
4. Azzam, S., Humphreys, K.: New Directions in Question Answering. Information Retrieval 9(3), 383–386 (Jun 2006)
5. Benamara, F.: Cooperative question answering in restricted domains: the WEBCOOP experiment. In: Proceedings of the Workshop Question Answering in Restricted Domains, within ACL (2004)
6. Benamara, F.: Generating intensional answers in intelligent question answering systems. Natural Language Generation (2), 11–20 (2004)
7. Benamara, F.: A semantic representation formalism for cooperative question answering systems. In: Proceeding of Knowledge Base Computer Systems (KBCS) (2008)
8. Benamara, F., Saint-Dizier, P.: Lexicalisation strategies in cooperative question-answering systems. In: Proceedings of the 20th international conference on Computational Linguistics. p. 1179. No. Cruse 1986 in COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
9. Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.Y., Maiorano, S., Miller, G., Others: Issues, tasks and program structures to roadmap research in question & answering (Q&A). Document Understanding Conferences Roadmapping Documents pp. 1–35 (2001)
10. Corella, F., Lewison, K.: A brief overview of cooperative answering. Journal of Intelligent Information Systems 1(2), 123–157 (Oct 2009)
11. Fellbaum, C.: WordNet: An electronic lexical database. The MIT press (1998)
12. Gaasterland, T.: Cooperative answering through controlled query relaxation. IEEE Expert: Intelligent Systems and Their Applications 12(5), 48–59 (Sep 1997)
13. Gaasterland, T., Godfrey, P., Minker, J.: An overview of cooperative answering. Journal of Intelligent Information Systems 1(2), 123–157 (1992)
14. Gaasterland, T., Godfrey, P.: Relaxation as a platform for cooperative answering. Journal of Intelligent Information 1(3), 293–321 (1992)
15. Gaasterland, T., Godfrey, P., Minker, J., Novik, L.: A cooperative answering system. In: Logic Programming and Automated Reasoning. pp. 478–480. No. X, Springer (1992)
16. Hirschman, L., Gaizauskas, R.: Natural language question answering: The view from here. Natural Language Engineering 7(4), 275–300 (2001)
17. Jousse, F., Tellier, I., Tommasi, M., Marty, P.: Learning to extract answers in question answering: Experimental studies. In: CORIA. p. 85 (2005)
18. Kaufmann, E., Bernstein, A.: Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics: Science, Services and Agents on the World Wide Web 8(4), 377–393 (Nov 2010)
19. Lin, J.: The Web as a resource for question answering: Perspectives and challenges. In: Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 2120–2127. No. Lrec, Citeseer (2002)
20. Lopez, V., Motta, E.: Poweraqua: Fishing the semantic web. Semantic Web: Research and Applications (2006)
21. Lopez, V., Uren, V., Sabou, M., Motta, E.: Is question answering fit for the semantic web?: a survey. Semantic Web? Interoperability, Usability, Applicability 2(2), 125–155 (September 2011)
22. Maybury, M.: New directions in question answering. Elements pp. 533–558 (2004)
23. Mcguinness, D.L.: Question Answering on the Semantic Web. IEEE Intelligent Systems pp. 6–9 (2004)
24. Melo, D., Rodrigues, I.P., Nogueira, V.B.: Puzzle out the semantic web search. International Journal of Computational Linguistics and Applications 3(1), 91–106 (June 2012)

25. Melo, D., Rodrigues, I.P., Nogueira, V.B.: Work out the semantic web search: The cooperative way. Adv. Artificial Intellegence 2012 (2012)
26. Minker, J.: An overview of cooperative answering in databases. Flexible Query Answering Systems pp. 282–285 (1998)
27. Perera, R.: Ipedagogy: Question answering system based on web information clustering. 2012 IEEE Fourth International Conference on Technology for Education 0, 245–246 (2012)
28. de Sena, G.J., Furtado, A.L.: Towards a cooperative question-answering model. Flexible Query Answering Systems 1495, 354–365 (1998)

Discovery of Disambiguation Rules for the POS Problem Using Genetic Algorithms

Ana Paula Silva[1] and Irene Rodrigues[2]

[1] Escola Superior de Tecnologia do Instituto Politécnico de Castelo Branco
{dorian,arlindo}@ipcb.pt
[2] Universidade de Évora
ipr@uevora.pt

Abstract. In this work, we modeled the part-of-speech tagging problem as a combinatorial optimization problem, which we solve using a genetic algorithm. The search for the best combinatorial solution is guided by a set of disambiguation rules that we first discovered using a classification algorithm, that also includes a genetic algorithm. Using rules to disambiguate the tagging, we were able to generalize the context information present on the training tables adopted by approaches based on probabilistic data. We were also able to incorporate other type of information that helps to identify a word's grammatical class. The results obtained on two different corpora are amongst the best ones published.

Keywords: Part-of-speech Tagging, Disambiguation Rules, Evolutionary Algorithms, Natural Language Processing

1 Introduction

The automatic part-of-speech tagging is the process of automatically assigning to the words of a text a part-of-speech (POS) tag. The words of a language are grouped into grammatical categories that represent the function that they might have in a sentence. These grammatical classes (or categories) are usually called part-of-speech. However, in most languages, there are a large number of words that can be used in different ways, thus having more than one possible part-of-speech. To choose the right tag for a particular word, a POS tagger must consider the surrounding words' part-of-speeches. The neighboring words could also have more than one possible way to be tagged. This means that, in order to solve the problem, we need a method to disambiguate a word's possible tags set.

Traditionally, there are two groups of methods used to tackle this task. The first group is based on statistical data concerning the different context possibilities for a word [1–6], while the second group is based on rules, normally designed by human experts, that capture the language properties [7–9]. Most current taggers are based on statistical models, defined on a set of parameters, whose values are extracted from texts marked manually. The aim of such models is to assign to each word in a sentence the most likely part-of-speech, according to its context, i.e., according to the lexical categories of the words that surround it. In order to do this, statistics on the number of occurrences of different contexts, for each word part-of-speech assignment possibilities, are collected.

Other type of taggers are rule-based systems, that apply language rules to improve the tagging's accuracy. The first approaches in this category were based on rules designed by human linguistic experts. There are also attempts to automatically deduce those rules, with perhaps the most successful one being the tagger proposed by Brill [7]. This system automatic extracts rules from a training corpus, and applies them in a iterative way, in order to improve the tagging of the text. The rules presented in [7] are called transformation rules and are driven toward error correction.They allow to consider not only the tags that precede one particular word, like the traditional probabilistic taggers, but also the tags of the words that follow it.

More recently, several evolutionary approaches have been proposed to solve the tagging problem. These approaches can also be divided by the type of information used to solve the problem, statistical information [2–6], and rule-based information [8]. Shortly, in the former, an evolutionary

algorithm is used to assign the most likely tag to each word of a sentence, based on a context table, that basically has the same information that is used in the traditional probabilistic approaches. On the other hand, the later are inspired by [7]. In this case a genetic algorithm (GA) is used to evolve a set of transformations rules, that will be used to tag a text in much the same way as the tagger proposed by Brill.

In this work, we modeled the part-of-speech problem as a combinatorial optimization problem and we investigate the possibility of using a classification algorithm to evolve a set of disambiguation rules, that we then use as an heuristic to guide the search for the best tags combination. These rules contemplate, not only context information, but also some information about the words' morphology. They are not oriented toward error correction, like in [7, 8], instead they are a form of classification rules, which try to generalize the context information that is used by probabilistic taggers. The discovery of the disambiguation rules was done by a classification algorithm based on a covering approach that integrates a genetic algorithm (GA) to perform the search for the best rules. For each rule found a quality value was saved. The classification problem was divided into n different problems, with n the number of part-of-speech tags that were pre-established for the experimental work. The selection of the predictive attributes took into consideration, not only the context information, but also some aspects about the words' internal structure. The tagging itself was performed by another genetic algorithm (witch we called GA-Tagger). This algorithm searches for the best combination of tags for the words in a sentence, guided by the disambiguation rules found earlier. Therefore, our system is composed by two steps. First, a set of disambiguation rules are discovered by a classification algorithm, and than a GA-Tagger is used to tag the words of a sentence, using the rules found in the first step.

The rest of the paper is organized as follows: Section 2 describes the classification algorithm used to discover the disambiguation rules. In section 3 we present the GA-Tagger and the results achieved. Finally, Section 4 draws the main conclusions of this work.

2 Classification Algorithm for Disambiguation Rules Discovery

In this section we describe the use of a classification algorithm, based on a covering approach, to discover a set of disambiguation rules, that will be used as an heuristic to solve the part-of-speech tagging problem. We chose to use a genetic algorithm to perform the search of the best rule for each iteration of the covering algorithm. The motivation for using a GA in this task, is that genetic algorithms are robust, adaptive search methods that perform a global search in the space of candidate solutions.

2.1 The Covering Algorithm

The main steps of the implemented covering algorithm are presented in algorithm 1. This algorithm was executed for each tag of a pre-defined tag set. As we can see, the genetic algorithm is invoked as many times as necessary to cover all the positive examples of the training set, evolving a rule in each run. After each execution, the rule returned by the genetic algorithm is stored, along with its quality value. This value is determined during the search process. The training set is then updated by removing all the positive examples that were covered by the returned rule.

In the next sections we will describe the main steps of the genetic algorithm implemented. We begin by selecting the predictive attributes to be used in the antecedents' rules, since they will determine the individuals' representation. Then we will present the genetic operators, selection scheme, and the fitness function used.

2.2 Attribute Selection

Our aim is to discover a set of rules that take into consideration not only context information but also information about the words' morphology. For the context, we decided to consider the same information that was used in [7, 8]. Thus, we consider six attributes: the lexical category of the

Algorithm 1 Covering Algorithm. sp and sn represent the sets of positive and negative examples. ps and gm give the population size and the maximum number of generations.

Require: sp, sn, ps, gm
Ensure: set_of_rules
  while $sp \neq \varnothing$ **do**
    $best\_rule \leftarrow GeneticAlgorithm(sp, sn, ps, gm)$
    $sp \leftarrow RemoveExamples(sp, best\_rule)$
    $set\_of\_rules \leftarrow Add(set\_of\_rules, best\_rule)$
  end while

third, second and first word to the left and the lexical category of the third, second and first word to the right. For the words' morphology information we decided to include the following attributes:

- The word is capitalized
- The word is the first word of the sentence
- The word ends with ***ed*** or ***ing*** or ***es*** or ***ould*** or ***'s*** or ***s***
- The word has numbers or '.' and numbers

The possible values for each of the first six attributes are the values of the corpus tag set from which the classification algorithm will extract the rules. This set will depend on the annotated corpus used, since the set of labels will vary for different corpora. The last nine attributes are boolean, and so the possible values are simply **True** and **False**.

2.3 Representation

Genetic algorithms for rule discovery can be divided into two dominant approaches, based on how the rules are encoded in the population of individuals. In the Michigan approach, each individual encodes a single rule, while in the Pittsburgh approach each individual encodes a set of prediction rules. The choice between these two approaches depends strongly on the type of rules we want to find, which in turn is related to the type of data mining task we are interested to solve. In our work, we are interested in finding a set of rules that will be used, not as a classifier, but as an heuristic to solve the combinatorial optimization problem that can be formulated from the part-of-speech tagging problem. In this sense, the Pittsburgh's approach seems to be more appropriate. However, there is an important question to consider when we adopt this type of representation, and that concerns the size of the individuals. We could adopt a traditional fixed length representation, or we could adopt a non standard variable length representation. In the first case, the problem is to define which size to consider, since we usually don't know how many rules are necessary for a certain classification task. On the other hand, in the non standard variable length representation, there is a difficult problem to deal with, which concerns the control of the individuals' length. Individuals tend to grow as the evolutionary algorithm generations increase, making it progressively slower - the well known bloat problem.

Since we will have a very large training set, and therefore the algorithm will be very time consuming, we have decided to adopt the Michigan's approach, so that we don't have to deal with the bloat problem. However, we didn't consider all the population as a set of rules representing a solution to the classification problem. Instead, we adopted a covering algorithm approach, i.e., we run the genetic algorithm as many times as necessary to cover all the positive examples of the training set, evolving a rule in each run.

In our approach each individual represents a rule of the form **IF** ***Antecedent*** **THEN** ***Consequent***, where *Antecedent* consists of a conjunction of predictive attributes and *Consequent* is the predicted class.

To encode each of the first six attributes we used $1 + k$ bits. The first bit indicates whether the attribute should or should not be considered, and the following k bits encode an index to a table with as many entries as the number of elements of the tag set adopted. If the value, d, encoded by the k bits exceeds the table's size, m, we use as index the remainder of the integer division of

d by m. The extra bit for each attribute allow us to ignore it, as in the previous representation when all the bits are 1. The remaining attributes were encoded in a similar way by 2×9 bits. In each pair of bits, the first one indicates if the attribute should, or should not, be ignored, and the second represents the logical value of the respective boolean attribute. In short, each individual is composed by $k \times 6 + 2 \times 9$ bits.

As we said before, we divided the classification problem into n different problems. The object to classify in each problem, is one of the n tags belonging to the tag set being considered, and the possible classes belong to a set of discrete values with only two elements: **Yes** and **No**. Since we are only interested in positive rules, we didn't need to encode the rule's consequent.

2.4 Initial Population

Half the individuals of the initial population were randomly generated and the other half were obtained by randomly choosing examples from the set of positive examples. These examples were first converted to the adopted binary representation and then added to the population.

2.5 Fitness Function

The formula used to evaluate a rule, and therefore to set its quality, is expressed in equation 1. This formula penalizes a individual representing a rule that ignores the first six attributes (equation 2), which are related with the word's context, forcing it to assume a more desirable form. The others are evaluated by the well known F_β-measure (equation 3). The F_β-measure can be interpreted as a weighted average of precision (equation 4) and recall (equation 5). We used $\beta = 0.09$, which means we put more emphasis on precision than recall.

$$Q(X) = \begin{cases} F_\beta(X) & \text{If context(X) = True} \\ -1 & \text{Otherwise} \end{cases} \tag{1}$$

$$context(X) = \begin{cases} True & \text{If } X \text{ tests at least one of the first six attributes} \\ False & \text{Otherwise} \end{cases} \tag{2}$$

$$F_\beta(X) = (1 + \beta^2) \times \frac{precision(X) \times recall(X)}{\beta^2 \times precision(X) + recall(X)} \tag{3}$$

$$precision(X) = \frac{TP}{TP + FP} \tag{4}$$

$$recall(X) = \frac{TP}{TP + FN} \tag{5}$$

where:

- TP - True Positives = number of instances covered by the rule that are correctly classified, i.e., its class matches the training target class;
- FP - False Positives = number of instances covered by the rule that are wrongly classified, i.e., its class differs from the training target class;
- FN - False Negatives = number of instances not covered by the rule, whose class matches the training target class.

2.6 Genetic Operators and Selection

Since our representation is a typical binary representation, we didn't need to use special operators. We used a traditional two point crossover and binary mutation as genetic operators. In the two point crossover operator, two crossover points were randomly selected, and the inner segments of each parent were switched, thus producing two offsprings. The mutation operator used was the standard binary mutation: if the gene has the allele 1, it mutates to 0, and vice versa. We used

a mutation probability of 0.01 and a 0.75 crossover probability. These values were empirically determined.

For the selection scheme, we used a tournament selection of size two with $k = 0.8$. We also used elitism, preserving the best individual of each generation by replacing the worst individual of the new population by the best of the old one (see algorithm **??**).

2.7 Pre-processing Routines - Data Extraction

We used the Brown corpus to create the training sets that we provided as input to the evolutionary algorithm. For each word of the corpus, we collected the values for every attribute included in the rule's antecedent, creating a specific training example. Then, for each tag of the tag set, we built a training set composed by positive and negative examples of the tag. In this process we decided to use only the examples determined by ambiguous words. The algorithm used to define each of the training sets was the following: for each example e_i of the set of examples, with word w and tag t, if w is an ambiguous word, with $\mathbb{S}$ the set of all its possible tags, then put e_i in the set of positive examples of tag t, and put e_i in the set of negative examples of all the tags in $\mathbb{S}$, except t. It is worth noting that each example has associated the number of times it occurs in the training corpus.

2.8 Experimental Results

We developed our system in Python and used the resources available on the NLTK (Natural Language Toolkit) package in our experiences. As we said before, tagged corpora use many different conventions for tagging words. This means that the tag sets vary from corpus to corpus. To extract the disambiguation rules from a set of annotated texts, we need to run our algorithm for each of the tags belonging to the tag set. However, if we want to test the resulting rules in a different corpus, we will not be able to measure the performance of our tagger, since the corpus tag set may be different. To avoid this, we decided to use the *simplify_ tags=True* option of the *tagged_ sentence* module of NLTK corpus readers. When this option is set to *True*, NLTK converts the respective tag set of the corpus used to a uniform simplified tag set, composed by 20 tags. This simplified tag set establishes the set of classes we use in our algorithm. We ran the covering algorithm for each one of these classes and built, for each one, the respective sets of positive and negative examples.

We processed 90% of the Brown corpus in order to extract the training examples, and, for each word found, we built the corresponding instance. The total number of examples extracted from the corpus equaled 929286. We used 6 subsets of this set (with different cardinality) to conduct our experiments. We used sets of size: $3E4$, $4E4$, $5E4$, $6E4$, $7E4$ and $8E4$. The examples were taken from the beginning of the corpus. For each subset adopted, we built the sets of positive and negative examples for each ambiguous tag, using the process described in the previous section.

The genetic algorithm was run with a population size of 200 individuals for a maximum of 80 generations. These values were established after some preliminary experiments. We run the algorithm for each of the defined sets of training examples. The best tagging was achieved with the rules learned from the sets of positive and negative examples defined from the first $8E4$ training examples. The rules set has a total of 4440 rules.

3 GA-Tagger

Our GA-Tagger was designed to receive as input a sentence, S, made of n words; a dictionary, $\mathbb{D}$; and a set of sets of disambiguation rules, $\mathbb{R}_i$, with $i \in \mathbb{T}$, the tag set adopted. As output it should return S, but now with each word, w_i, associated with the proper tag $t_i \in \mathbb{T}$.

Assuming we know the set of possible tags, $\mathbb{W}_i$, for each word w_i of S, the part-of-speech tagging problem can be seen as a search problem with $\mathbb{W}_0 \times \mathbb{W}_1 \times ... \times \mathbb{W}_n$ as its state space. This means that we should be able to solve it by applying a search algorithm designed to solve the intrinsic combinatorial optimization problem. In this work we investigate the possibility of applying

a genetic algorithm to search for the best combination of tags for the words of a sentence, using as an heuristic the sets of disambiguation rules $\mathbb{R}_i$. The main aspects of the genetic algorithm implemented will be presented in the following sections.

3.1 Representation

An individual is represented by a chromosome made of a sequence of genes. The number of genes in a chromosome equals the number of words in the input sentence. Each gene proposes a candidate tag for the word in the homologous position.

The individual phenotype is the set of all pairs $\langle x_i, t_i \rangle$ determined by each gene, g_i, and its corresponding word w_i. t_i is the tag proposed by g_i for the word w_i and x_i is a 15-tuple with the values of the disambiguation rules' attributes. When there is no gene (no corresponding word) in one of the positions contemplated in the context, we adopted an extra tag named 'None'. This can happen with the first three and last three genes of the individual. We adopted a symbolic representation, i.e. the possible alleles of a gene are the tags of the tag set adopted for the corpus in which the experiences will be executed. However, the allowed alleles of a gene are only the ones that correspond to the possible tags of the word the gene represents.

The initial population is generated by choosing randomly, for each gene g_i, one of the values presented in $\mathbb{W}_i$. The input dictionary gives the possible tags for each word of the input sentence. However, If some word, w is not in the dictionary, the algorithm chooses randomly one of the tags whose rules set has a rule which covers the instance defined by the 15-tuple determined by w. If none of the rules cover the 15-tuple, the algorithm chooses randomly one of the tags belonging to $\mathbb{T}$.

3.2 Genetic Operators and Selection

We used a typical one point crossover with a 0.8 probability. The mutation operator randomly chooses another allele from the set of possible alleles for the particular gene and was applied with a 0.05 probability. Again, if the word is unknown, the sets of rules will be used to determine which ones include a rule that covers the 15-tuple, and one of the possibilities will be randomly chosen and assigned to the corresponding gene.

We adopted a tournament selection of size two with $k = 0.7$ and also used elitism, replacing the worst individual of each new population with the best of the old one. All the values were empirically determined in a small set of preliminary experiments.

3.3 Fitness Function

The performance of an individual with phenotype $\mathbb{P}$ is measured by the sum of the quality values of the pairs $\langle x_i, t_i \rangle \in \mathbb{P}$. These quality values are obtained from the disambiguation rules by applying equation 6. The $Quality(z)$ function gives the quality value associated with rule z, which was computed by the classification algorithm.

$$F(\langle x_i, t_i \rangle) = \begin{cases} Quality(r_k) & \text{if } r_k \in \mathbb{D}_{t_i} \text{ and } r_k \text{ covers } x_i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The fitness of an individual with phenotype $\mathbb{P}$ is given by equation 7:

$$Fitness(\mathbb{P}) = \sum_{i=1}^{n} F(\langle x_i, t_i \rangle) \tag{7}$$

Table 1. Results achieved by the GA-Tagger for the test set defined from the Brown corpus.

Population	Generations	Average	Best	Standard Deviation
50	10	0.9672170	0.9675561	$1.9200E-4$
	20	0.9672968	0.9674231	$1.1707E-4$
100	10	0.9672591	0.9675561	$1.4097E-4$
	20	0.9672835	0.9675117	$1.0978E-4$

3.4 Experimental Results

We tested our GA-Tagger with different population sizes and number of generations, on a test set of the Brown corpus (different from the one used to learn the disambiguation rules). We ran the algorithm 20 times with a population of 50 and 100 individuals during 10 and 20 generations. The test set was composed by 22562 words. The results achieved are shown in table 1.

Although there are no significant differences between the average results achieved with the different parameters' combinations, the smallest standard deviation was achieved with a population of 100 individuals during 20 generations. Nevertheless, we can conclude from the experiments that the GA-tagger usually finds a solution very quickly. We also tested the GA-tagger on a test set of the WSJ corpus that is available in the NLTK software package. The test set was made of 100676 words. We ran the algorithm with 50 individuals during 10 generations, and the best output had an accuracy of 96.66%. The results achieved show that there are no significant differences on the accuracy obtained by the tagger on the two test sets used. At this point, it is important to recall that the disambiguation rules used on the tagger were extracted from a subset of the Brown corpus. This bring us to the conclusion that the rules learned on step one are generic enough to be used on different corpora of the same language, and are not domain dependent.

Table 2 presents the results achieved by the GA-tagger on the two corpora used in the experiments performed, along with the results achieved by the approaches that are more alike to the one presented here.To better understand the table information, we would like to point out that the results presented in [8] were achieved for the set used to perform the training, that is why we used 'none' for this field of the table. Also, since we learned the disambiguation rules from the Brown corpus, we didn't presented any value for the size of the training set for the WSJ corpus.

Table 2. Results achieved by the GA-Tagger on two different corpus, the Brown corpus and the WSJ corpus, along with the results achieved by the approaches more similar to the one presented here

Corpus	Tagger	Training set	Test set	Best
Brown	GA-Tagger	80000	22562	96.76
	[2]	185000	2500	95.4
	[6]	165276	17303	96.67
	[6]	165276	17303	96.75
WSJ	GA-Tagger	none	100676	96.66
	[8]	600000	none	89.8
	[7]	600000	150000	97.2
	[6]	554923	2544	96.63

4 Conclusions and Future Work

We described a new approach to the part-of-speach tagging problem that defines it as a combinatorial optimization problem. The results achieved are comparable to the best ones found in the area bibliography (see table 2). Although there are other approaches to this problem that use, in some way, evolutionary algorithms, as far as we know this is the first attempt that uses these algorithms

to solve all aspects of the task. In our approach to the problem, we used an evolutionary algorithm to find a set of disambiguation rules and then used those rules as an heuristic to guide the search for the best combination of tags for the words of a sentence, using another evolutionary algorithm to perform the search. We suggest a new way to collect and represent relevant information to solve the part-of-speech tagging problem. The use of disambiguation rules allows to generalize the information usually stored in training tables by the probabilistic approaches. This generalization is reflected in a reduction of the information volume needed to solve the problem, and also in a less domain dependent tagger. Also, the flexible format of the rules used, easily adapts to the inclusion of new aspects, that might be useful to solve the tagging problem. Moreover, the information is presented in a way that could be easily interpreted by a human observer. All these aspects were achieved without loosing the statistical information, represented here in the form of the rules' quality value. Another important contribution of this work is the formalization of the part-of-speech tagging problem as a combinatorial optimization problem, allowing the use of a global search algorithm like genetic algorithms to solve it.

We tested our approach on two different corpora: in a test set of the corpus used to discover the disambiguation rules, and on a different corpus. The results obtained are among the best ones published for the corpora used in the experiments. Also, there were no significant differences between the results achieved in the subset belonging to the same corpus from which we defined the training set, used to discover the rules, and the results obtained on the sentences of the other corpus. This confirms our expectations concerning the domain independence of the obtained rules.

Although we consider our results very promising, we are aware of the necessity of test our approach with a larger tag set, and to apply it to more corpora. We intend to test the tagger on other languages, as well. We also think that this approach could be applied to other natural language processing tasks, like noun phrase chunking and named-entity recognition.

References

1. Brants, T.: Tnt: a statistical part-of-speech tagger. In: Proceedings of the sixth conference on Applied natural language processing. ANLC '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 224–231
2. Araujo, L.: Part-of-speech tagging with evolutionary algorithms. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Volume 2276 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2002) 187–203
3. Araujo, L.: Symbiosis of evolutionary techniques and statistical natural language processing. Evolutionary Computation, IEEE Transactions on **8**(1) (feb. 2004) 14 – 27
4. Araujo, L.: How evolutionary algorithms are applied to statistical natural language processing. Artificial Intelligence Review **28**(4) (2007) 275–303
5. Araujo, L., Luque, G., Alba, E.: Metaheuristics for natural language tagging. In: Genetic and Evolutionary Computation - GECCO 2004, Genetic and Evolutionary Computation Conference. Volume 3102 of Lecture Notes in Computer Science., Springer (2004) 889–900
6. Alba, E., Luque, G., Araujo, L.: Natural language tagging with genetic algorithms. Information Processing Letters **100**(5) (2006) 173–182
7. Brill, E.: Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput. Linguist. **21** (December 1995) 543–565
8. Wilson, G., Heywood, M.: Use of a genetic algorithm in brill's transformation-based part-of-speech tagger. In: Proceedings of the 2005 conference on Genetic and evolutionary computation. GECCO '05, New York, NY, USA, ACM (2005) 2067–2073
9. Nogueira Dos Santos, C., Milidiú, R.L., Rentería, R.P.: Portuguese part-of-speech tagging using entropy guided transformation learning. In: Proceedings of the 8th international conference on Computational Processing of the Portuguese Language. PROPOR '08, Berlin, Heidelberg, Springer-Verlag (2008) 143–152

A Swarm Intelligence Approach to SVM Training*

Arlindo Silva[1] and Teresa Gonçalves[2]

[1] Escola Superior de Tecnologia do Instituto Politécnico de Castelo Branco
arlindo@ipcb.pt
[2] Universidade de Évora
tcg@uevora.pt

Abstract. In this paper we outline a new swarm intelligence based approach to the problem of training support vector machines with non positive definite kernels. Past approaches using particle swarm optimizers have been shown to compare poorly with other evolutionary computation based methods. In this paper, we describe a new heterogeneous particle swarm optimizer, specifically tailored for the training of support vector machines. We present experimental results of the comparison of this algorithm with traditional support vector machine training algorithms and recent evolutionary approaches. The comparison is made both with positive definite and non positive definite kernels. The new algorithm is shown to be competitive with all approaches in terms of final classification accuracy and the fastest of the evolutionary computation based algorithms. The results also suggest that the evolutionary and swarm intelligence optimizers can achieve better classification results that the traditional methods when non positive definite kernels are used.

1 Introduction

Support vector machines (SVMs) are the best known representative of a very successful group of data analysis techniques, called kernel methods [1]. SVMs classify new data by comparing it with a learned hyper-plane that maximizes a margin between data points of different classes [2]. The remarkable success with which these methods have been applied to many areas is the result of their specific properties: low computational cost; robustness, with solid theoretical bases in statistical learning; and generality, since the choice of an appropriate kernel function allows the algorithm to learn non-linear decision functions and deal with non-vectorial and even heterogeneous data.

Despite the success of this approach, the application of a SVM to a new problem still presents a number of difficulties. A kernel function must be chosen and its parameters optimized. A real parameter C must also be chosen to balance error and capacity in the SVM. If a new kernel function has to be developed, care must be taken to ensure the kernel is positive semi definite (PSD), since training relies on quadratic programming based techniques, where a unimodal concave function is optimized. These issues have been addressed using both analytic techniques and heuristic search algorithms. Recently, several approaches based on evolutionary methods, such as genetic algorithms (GA), genetic programming (GP) and particle swarm optimization algorithms (PSO) have been proposed. These algorithms are advantageous when search and/or optimization is done in complex, non-vectorial spaces, or when the function to optimize is multi-modal, with many local optima in alternative to a single solution. In this paper, we deal with evolutionary approaches to the training of support vector machines.

2 Support Vector Machines

In their most common formulation [1–3], support vector machines are classification mechanisms, which, given a training set $T = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \{\pm 1\}$, assuming n

* A substantially enlarged and revised version of this paper, with the title "Training Support Vector Machines with an Heterogeneous Particle Swarm Optimizer", was accepted for oral presentation at ICANNGA'13, 11th International Conference on Adaptive and Natural Computing Algorithms. It will also be published in the conference proceedings in a Springer Lecture Notes in Computer Science - LNCS - volume.

exemples with m real attributes, learn a hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, with $\mathbf{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$, which completely separates the example labels as -1 from the ones labeled as $+1$. Using this hyperplane, a new instance $\mathbf{x}$ is classified using $f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$.

The maximization of the distance, or margin, between the discriminating hyperplane and the closest examples, is a characteristic of large margin methods, of which support vector machines are an instance. This maximization reduces the so-called structural risk, which is related to the quality of the decision function. Support vector machines therefore try to minimize the structural risk, which comprises not only the empirical risk, but also a measure of the classifier's quality.

Support vector machines are extended to the non-linear case by implicitly mapping the data to a secondary space with higher dimensionality - the feature space - where a linear separation is possible. This mapping is achieved by using a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ instead of the product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. The most common kernel function is the radial basis function, but many others are used. If the kernel is positive definite, the optimization problem is a quadratic problem with a concave optimization function and several algorithms are available to solve it efficiently, with the most popular implementations being mySVM and LIBSVM. If, however, the kernel is not positive definite, the problem is not guaranteed to possess a single optimum and these algorithms can become trapped in local optima, failing to find the best solution.

Haasdonk, however, has proposed an alternative interpretation for SVMs based on non positive definite kernels, which accounts for their good experimental results [4]. In this interpretation, SVMs are optimal hyperplane classifiers, not by margin maximization, but by minimization of distances between convex hulls in pseudo-Euclidean spaces. Both this work and the one by [5] conclude that traditional methods, e.g. LIBSVM, can converge to a stationary point of the non concave optimization problem that results from the use of indefinite kernels. Obviously, there is no guarantee that this point is the global optimum, which leads us to the interest in using heuristic global optimization methods, like the evolutionary algorithms we discuss here. An in-depth discussion of the relevance of learning with non PSD kernels can be found in [6], while the training of SVM using these kernels is thoroughly discussed in [7, 5, 4].

3 The Scouting Predator-Prey Optimiser

Particle swarm optimizers [8, 9] can be an obvious answer to the problem of optimizing SVMs with non positive definite kernels, since they are population based global optimization algorithms with successful application to hard optimization problems with many optima [10]. We recently proposed a new heterogenous particle swarm algorithm, called scouting predator-prey optimizer (SPPO), which showed good performance, even on hard optimization problems [11]. We also proposed the use of scout particles to improve a swarm optimizer by using problem specific knowledge. Here, we will describe a version of the SPPO specifically tailored to the training of SVMs.

In particle swarm optimization, each member is represented by three m-size vectors, assuming an optimization problem $f(\mathbf{x})$ in $\mathbb{R}^m$. For each particle i we have a $\mathbf{x}_i$ vector that represents the current position in the search space, a $\mathbf{p}_i$ vector storing the best position found so far and a third vector $\mathbf{v}_i$ corresponding to the particle's velocity. For each iteration t of the algorithm, the current position $\mathbf{x}_i$ of every particle i is evaluated by computing $f(\mathbf{x}_i)$. Assuming a maximization problem, $\mathbf{x}_i$ is saved in $\mathbf{p}_i$ if $f(\mathbf{x}_i) > f(\mathbf{p}_i)$, i.e. if $\mathbf{x}_i$ is the best solution found by the particle so far. The velocity vector $\mathbf{v}_i$ is then computed and used to update the particle position.

One of the limitations of the standard particle swarm algorithm is its inability to introduce diversity in the swarm after it has converged to a local optimum. Since there is no mechanism similar to a mutation operator, and changes in $\mathbf{x}_i$ are dependent on differences between the particles' positions, as the swarm clusters around a promising area in the search space, so does velocity decreases and particles converge to the optimum. This is the desirable behavior if the optimum is global, but, if it is a local optimum, there is no way to increase velocities again and allow the swarm to escape to a new optimum. We use a predator-prey effect in SPPO to alleviate this problem. The predator particle's velocity $\mathbf{v}_p$ is updated, oscillating between the best particle's best position and the best particle's current position. This update rule makes the predator effectively chase the best particle in the search space.

Since the predator chases the best particle, the perturbation in the swarm is more likely when all the particles are very near, i.e. during the exploitation phase, and becomes almost inexistent when the particles are far apart. This mechanism allows for particles to escape and find new optima far from the current attractor even in the last phases of exploitation.

Scout particles, or scouts, are a subset of the swarm that implement exploration strategies different from the one used by the main swarm. They can be used to introduce improvements to the global algorithm, e.g. a local search sub-algorithm, or to implement problem dependent mechanisms to better adapt the algorithm to a specific problem. In this work, we will use two scout particles to tailor the SPPO to the specific problem of training SVMs. The first scout is a local search particle which, from previous work [11], we know can be used to increase the convergence speed without compromising the final results. For this scout we choose the best particle at each iteration and perform a random mutation on one of its dimensions j.

The second scout particle uses specific problem knowledge to accelerate the training process. Since we know that in the final solution only the few α_i corresponding to support vectors will be different from 0, and that, from these, most will be C, we will, at every iteration move the scout particle, in a random dimension, to an extreme of the search space, with an 80% probability of that extreme being 0 and 20% of being C. This scout will consequently explore the border of the search space, where we know the solution should be in a large majority of dimensions.

4 Experimental Results

In the first set of experiments, we tested three evolutionary approaches, including the best previous ES based approach [12], a constricted PSO and the SPPO algorithm previously described, against the two most popular quadratic programming based methods, mySVM and LIBSVM, in a set of 10 benchmark problems, 3 of which are synthetically generated and the remaining 7 are real world benchmark problems. Table 1 lists the datasets' names, source, number of attributes n and number of instances m. e_d is the error for a classifier that always returns the most frequent class. The kernel used in all experiences was the radial basis function and its parameter σ was previously optimized for the mySVM method using a grid search (See Table 1). We used $C = 1$ for all datasets.

Table 1. Dataset and kernel parameters.

Dataset	Source	n	m	e_d	σ	σ_E	d
Checkerboard	Synthetic	1000	2	48.60	100	0.92	6.54
Spirals	Synthetic	500	2	50.00	100	0.12	3.84
Threenorm	Synthetic	500	2	50.00	1	61.60	9.38
Credit	UCI MLR	690	14	44.49	0.001	564.62	0.65
Diabetes	UCI MLR	768	8	34.90	0.1	220.51	4.87
Ionosphere	UCI MLR	351	34	35.90	0.1	2.44	7.48
Liver	UCI MLR	345	6	42.03	1	61.59	6.90
Lupus	StatLib	87	3	40.23	0.1	241.63	7.42
Musk	UCI MLR	476	166	43.49	0.1	63.12	6.93
Sonar	UCI MLR	208	60	46.63	0.1	61.63	6.90

The evolutionary algorithms were run for 500 iterations with 20 individuals/particles, except the SPPO, which only used 18 particles to compensate for the extra evaluations of the scout particles. Evaluation is done using 20-fold cross-validation. Experiences were run using RapidMiner software [13] with additional operators.

Table 2 presents average error rates and respective standard deviations for all pairs algorithm/dataset. We found that the best evolutionary approaches performed as well as the classical methods, both in terms of accuracy (error percentage) and robustness (standard deviation). Only the simple PSO performed poorer, which is compatible with the findings in [12]. Since the classical

approaches are significantly faster than the evolutionary ones, there's no particular reason to prefer the later for the training of SVMs with PSD kernels. These results are still useful to demonstrate some debilities of the standard PSO and to demonstrate that the SPPO is the first competitive swarm intelligence based approach to the training of support vector machines.

Table 2. Experimental results (error percentage) using the RBF kernel.

Dataset	MySVM	LIBSVM	ES	PSO	SPPO
Checkerboard	5.6 (2.6)	5.6 (3.2)	5.5 (3.0)	8.1 (4.3)	5.7 (3.4)
Spirals	0.4 (1.2)	0.2 (0.9)	0.6 (1.4)	2.2 (2.0)	0.6 (1.9)
Threenorm	15.0 (6.3)	14.8 (4.9)	14.6 (6.5)	14.2 (5.0)	14.8 (8.1)
Credit	14.5 (6.5)	14.5 (4.5)	14.5 (6.5)	13.8 (6.4)	14.4 (5.7)
Diabetes	22.4 (5.0)	22.5 (4.8)	23.7 (5.5)	29.8 (5.6)	23.3 (7.4)
Ionosphere	6.8 (5.9)	6.2 (5.6)	7.1 (5.7)	24.2 (9.1)	7.1 (5.7)
Liver	29.9 (11.9)	28.4 (10.5)	29.3 (11.9)	31.7 (10.4)	28.7 (11.1)
Lupus	26.0 (22.4)	24.8 (22.8)	25.0 (22.4)	26.0 (24.5)	25.0 (16.6)
Musk	7.8 (5.2)	7.5 (4.5)	9.9 (6.0)	11.1 (6.9)	8.6 (7.4)
Sonar	14.4 (7.6)	16.0 (11.8)	14.4 (9.9)	15.3 (10.2)	14.3 (13.1)

In the second set of experiments, we investigate how the best evolutionary algorithms compare with one of the standard approaches when training SVMs with a non PSD kernel, i.e., in a multimodal optimization problem. We use the same datasets, but the RBF is substituted by an Epanechnikov kernel, which can be proved to be indefinite. C was set to 1 and kernel parameters are presented in Table 1. Since we observed in the convergence graphs of the previous experiences that the evolutionary algorithms converged a lot sooner than the allotted 500 generations, we reduced the iteration limit to 100 (150 for the synthetic problems). In Table 3 we present the classification error, and, for the evolutionary approaches, the average best value found for the objective function, $f(\alpha^*)$.

This second set of results allows us to draw several conclusions. First, all algorithms were able to learn with the non-PSD kernel. In fact, for two of the datasets, Lupus and Sonar, the best overall results were obtained using the Epanechnikov kernel, in both cases using the SPPO algorithm. Second, with the lower iteration limit, there is a large difference, both in classification accuracy and best $f(\alpha^*)$, between the evolutionary approaches. This leads us to conclude that the SPPO needs significantly less function evaluations to achieve similar (or superior) classification accuracy, when compared with the best ES based approach. And, finally, we can see four datasets for which the SPPO performs better than the MySVM algorithm (results are similar for the rest). Since, from the previous experiments, we know that the algorithms are able to obtain identical results when using the same parameters in the concave optimization problems, these data apparently illustrate situations where, for a multimodal problem, the quadratic based approach fails to converge to the global optimum. This result confirms our proposition that evolutionary approaches can be useful tools in the training of SVMs when using non PSD kernels.

5 Conclusions

In this paper we proposed the first known particle swarm based approach to the problem of training SVMs with non PSD kernels. Our algorithm is a specially tailored version of the scouting predator prey algorithm [11], an heterogeneous particle swarm optimizer. To improve the algorithm performance in this particular problem, we embedded two scout particles in the algorithm, one to perform a local search and another to take advantage of specific problem knowledge. We compared our algorithm with the best known evolutionary approach to this task, as well as with two popular classical SVM training algorithms, using both a PSD and a non PSD kernel. The experimental results supported the assertions made in [7], since the evolutionary approaches, most specifically

Table 3. Experimental results (error percentage) using the Epanechnikov kernel.

Dataset	MySVM	ES ($f(\alpha^*)$)	ES	SPPO ($f(\alpha^*)$)	SPPO
Checkerboard	6.5 (4.6)	-304.1 (37.6)	8.2 (4.0)	60.9 (12.3)	7.5 (4.9)
Spirals	11.0 (6.5)	163.2 (3.2)	19.2 (7.8)	188.5 (2.5)	**7.8 (5.6)**
Threenorm	14.0 (7.5)	31.0 (11.9)	15.0 (6.9)	132.5 (4.7)	14.4 (6.6)
Credit	14.4 (5.5)	254.8 (6.2)	14.2 (6.5)	299.6 (5.5)	13.6 (4.4)
Diabetes	24.8 (6.2)	-62.4 (197.2)	29.4 (6.6)	297.5 (7.8)	25.2 (8.1)
Ionosphere	26.9 (10.6)	79.8 (3.9)	24.0 (9.3)	99.8 (1.6)	**16.1 (9.0)**
Liver	40.8 (3.7)	175.1 (5.5)	35.9 (12.1)	224.1 (4.7)	**35.4 (10.6)**
Lupus	27.8 (15.8)	48.6 (1.9)	24.0 (15.9)	58.4 (1.3)	**21.5 (18.8)**
Musk	9.6 (5.7)	104.5 (2.4)	11.8 (7.1)	118.4 (2.1)	9.5 (5.5)
Sonar	12.4 (11.4)	175.1 (5.5)	12.8 (10.8)	224.1 (4.7)	11.9 (9.6)

the SPPO, were able to outperform the classical method on several benchmark problems, when training the SVMs with the non PSD kernel. Regarding the evolutionary approaches, the results show that the SPPO can achieve significantly better values for the optimization function, with corresponding similar or better classification accuracy, than the ES based approach, for the same number of function evaluations.

References

1. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge Univ. Press, Cambridge (2004)
2. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press (2000)
3. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2** (1998) 121–167
4. Haasdonk, B.: Feature space interpretation of svms with indefinite kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 482–492
5. Lin, H.T., Lin, C.J.: A study on sigmoid kernel for svm and the training of non-psd kernels by smo-type methods. Technical report, National Taiwan University, Taipei, Department of Computer Science and Information Engineering (2003)
6. Ong, C.S., Mary, X., Canu, S., Smola, A.J.: Learning with non-positive kernels. In: Proceedings of the twenty-first international conference on Machine learning. ICML '04, New York, NY, USA, ACM (2004)
7. Mierswa, I., Morik, K.: About the non-convex optimization problem induced by non-positive semidefinite kernel learning. Advances in Data Analysis and Classification **2** (2008) 241–258
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks. Volume 4. (1995) 1942 –1948 vol.4
9. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intelligence **1** (2007) 33–57
10. Poli, R.: Analysis of the publications on the applications of particle swarm optimisation. J. Artif. Evol. App. **2008** (January 2008) 4:1–4:10
11. Silva, A., Neves, A., Gonçalves, T.: An heterogeneous particle swarm optimizer with predator and scout particles. In: Autonomous and Intelligent Systems. Volume 7326 of Lecture Notes in Computer Science. Springer (2012) 200–208
12. Mierswa, I.: Evolutionary learning with kernels: a generic solution for large margin problems. In: GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, New York, NY, USA, ACM (2006) 1553–1560
13. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (KDD-06). (2006)

Information Discovery over Annotated Content (Survey)

Sérgio M. Esteves Cardoso

Polytechnic Institute of Santarem, School of Management and Technology
sergio.cardoso@esg.ipsantarem.pt

Salvador Abreu

University of Évora
spa@di.uevora.pt

Abstract. *Tags* are metadata generated by users to annotate *web* content. Apparently they have great potential as a tool for *web* content discovery, but this use is hindered by challenges like *tag space* ambiguity, the lack of structure and the influence of certain linguistic or semantic aspects.

This work presents an overview on the use of annotated content over social bookmarking services for information discovery. It is also included a short literature review, with the aim of identifying the most relevant inputs and the state of the art in this field.

Keywords: Social bookmarking, tags, content discovery, facets

1 Introduction

There are numerous second-generation Internet services (Web 2.0) that allow users to annotate content in the form of tags reflecting their own individual interests. Many of these services, such as *Del.icio.us, Flickr* or *Digg*, have great acceptance and mobilize a significant number of users, and therefore have attracted research attention. Authors as [1] assume that in terms of volume, the information gathered, will have a scale comparable to the www itself, if it continues to grow at the pace it has grown over this last few years.

Terms (*tags*) are chosen freely by each user, according to individual criteria, with the objective to organize access and share web content that is available over numerous formats (*documents, images, video, etc.*). The result of cataloguing specific web resources to such a large users scale, produces implicit classification under several dimensions (*e.g. facets*), which suggests its potential as an element of information discovery (*based on notions of "wisdom of crowds"*).

Many online services offer tag clouds to allow users to visualize the existing *tag space*. Often, those clouds are purely based on frequency measures and highlight the most frequently used terms, looking to highlight the importance of those before others. This is normally just one way of presenting independent (*e.g. with no clear relation between them*) search terms to the user.

The use of tags by itself presents some specific challenges, such as those resulting from the language used in the classification, some ambiguity as to the meaning and significance of the terms actually used, as well as the absence of a multidimensional classification space that contributes to define a research context in which a user can interact repeatedly over the same dataset.

2 Background and Related Work

2.1 Annotated Content and Social Tagging

Web collaborative spaces provide the tools that allow users to discover web content and annotate it using tags of their own, which then can be used to organize, share, browse, and later search content. Annotated content has a great potential since a large number of users categorize resources, as well as produce annotations over the web. However, it is also clear that annotated content is disorganized and inaccurate. A single user ends up using several approaches which ultimately lead to contradictory classification [2].

These authors explain the different nature and objectives of the tags used. Linguistic and semantic influences are also present as synonyms and homonyms, lexical forms, errors and other common expressions [3] are used as tags to annotate web content. The search and information discovery over web resources using annotated content becomes more challenging due to the presence of these effects.

It has been possible to demonstrate that the fact that a large number of users annotate the same resource will protrude the most relevant dimensions of the classification. On the other hand, one could argue that some annotated resources remain under imprecise classification just because there aren't sufficient users annotating it [4].

Authors like [5] have done work showing that clusters formation of equivalent terms *(e.g. Tag Clusters)* allows to diminish some language and semantic problems and thus simplifies the operation of these systems. Clustering is done by referencing tags on the resources and the relationships between these. As more closely related the terms are, the more prone they are to grouping and uniform representation.

Another relevant facet is the fact that the tags used to categorize a web resource are often found in the title (16%) and/or on the body of the document (50%), which exposes them to common indexation mechanisms and pre-existing search methods [1]. This conclusion is mainly based on the analysis of textual content and ignores content annotation potential over different media such as images, video, audio clips, and its relevance to information discovery over "*non-textual*" supports.

Moreover, the fact that there is an apparent relation between the tag used and the web content itself, according to the work mentioned above, led us to speculate that, for a significant portion of annotated resources, the terms used are effectively appropriate and therefore useful for information discovery.

This aspect has been addressed by authors like [6]. In their studies, they specify different types of tags and different motivations for content annotation - which ultimately is reflected in the actual tag used on a particular web resource annotation *(tag: "thingsToRead")*. They also show the influence of social or linguistic contexts. It is clear that in these latter cases, content discovery will need to resort to strong normative references, such as ontologies, classification schemes or complementary analyses aimed to limit the impact of those effects.

Ontological dimensions arise in the work of authors like [7], since they address the underlying data model with formal concerns, seeking concept representation by exploration of their relational semantics. This effort is oriented towards removing or reducing *tag space* ambiguity and to highlight the most significant terms.

Another line of work is based on the fact that people with common interests *(e.g. like photographers)* would easily contribute heavily to define a set of tags that would become broadly used *(as a result of their individual annotations)*, making them to be more meaningful and appropriate, as they share the same context and background, compared to other freely formed communities.

Has [1] also states in its conclusions, the developments to be introduced in the platforms will validate this possibility. Those techniques can be already observed in some websites *(including del.ico.us)* suggesting annotation terms for a given web resource, as a strategy to limit the classification diversity and therefore contain the impact of heterogeneous annotations.

2.2 SNA - Social Network Analysis

A social network is often presented as a set of entities *(people, organizations or other)* that are linked together as they establish relationships [8], variable in nature and in motivation. The analysis on such networks has allowed the development of analytical research over those structures. Also common is the belief that this form of analysis focuses primary on existing relations and complementary on the entities themselves.

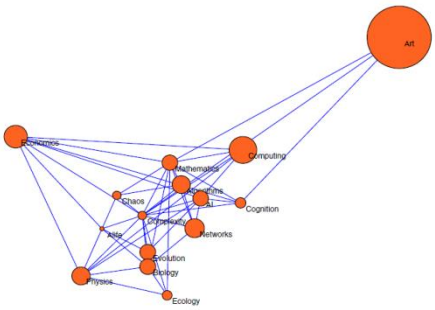
The extraction of information from web collaborative platform uses existing interfaces that are invoked via *RSS* or *JSON* techniques, enables one to accumulate contents on tags, users and web resources. Afterwards, one has to improve data quality and address the representativeness of the information obtained *(e.g. remove all entries with only one or two symbols)*. Some authors perform these tasks under criteria that allow them to isolate key aspects of their research, such as information about a specific domain, the presence of a particular note or tag, among other examples.

Recent studies [3] have used representative datasets *(e.g. with sufficient size and variety)* obtained by direct extraction of users, sets of tags and web resources, and later subjecting them to techniques of *SNA - Social Network Analysis* [9]. The information obtained is then modeled as tripartite graphs and then combined in the form of a network. Methods for calculating centrality measures for each of the nodes in these networks are later applied. The network is subjected to quantitative analysis in software as *Pajek* or *Gephi*. The network analysis introduces tags as terms significant to web resources classification and a relational dimension is extracted from the annotated content. These contributions are particularly significant because they allow the emergence of relational and structural elements that did not exist prior to the method application.

Pajek (http://pajek.imfm.si/doku.php) is an open-source framework of SNA, installable on Windows and other platforms, which allows the representation *(e.g. visualization)* of large scale networks, as well as perform the calculation of important centrality measures [3] such as Degree, Betweenness and Eigenvector.

The Degree centrality measure quantifies the importance of a node based on the number of adjacent nodes. Betweenness measures the number of times a node belongs to a shortest route between any other two nodes *(also called geodesic distance)*. The Eigenvector measurement does not assume that all nodes have the same significance. Its calculation is influenced by the notion that the contribution of central nodes is higher compared with other more peripheral ones [3].

In addition to the analytical part, it is possible to show more meaningful network segments according to how the nodes appear *(much / little)* related and the type of links actually established. I include a representative picture shown in [10] for this type of networks:



The analysis exposes categories *(also called facets)*, which allow the usage of the network over search interfaces, as a relational structure of concepts that underlay in the web resources themselves. At his work [8] summarized SNA main features:

- Generates mathematically regular structures, predictably and efficiently, showing the existent relations;
- Complex elements, yet to be specified, may exist along others;
- It is possible to add new objects and to establish relations with pre-existing others;

2.3 Search Interfaces

The search interfaces operate over substantial datasets and are designed to deal with user interaction, maintaining a search context on the behalf of the user, and displaying relevant and organized content.

Flamenco *(Flexible Access information using metadata in Novel Combination - Flamenco Framework-http://flamenco.berkeley.edu)* is one of such search interfaces which targets primarily to allow users to move through a large information space in a flexible way [11], [12]. An important property is that the categories *(the ones that have resulted from SNA techniques)* are displayed to the user in order to direct it to possible search terms and therefore organizes subsequent results.

Another important aspect relates to the fact that all subsequent searches result in the presentation of new search paths as the revelation of new search categories occurs - which necessarily exist in the data - once these are built over the information space. If stated in another way: a user can scroll through numerous dimensions that relate to a given resource and his path leads to new aspects that he would not necessarily have envisioned before.

3 Conclusions and Future Work

The discussed research results suggest that many of the challenges on using tags and other annotated content in information discovery can be greatly overcome by the usage of the techniques presented in this work. The use of tags and other annotated content precedes the outline of relevant classification on web resources.

Several authors stress the need to automatically build more formal reference systems. The purpose is to use annotated content in a well-ordered manner. Frequently this line of work leads to further investigate on *Ontologies*, *taxonomies* and *folksonomies*, as they are representation and classification systems – formally or informally established.

We would like to highlight some aspects to be kept under consideration:

- The fact that these techniques are primarily applied offline datasets, after extraction from de web platforms themselves and improved to eliminate senseless data. Would it be possible to extend these techniques in such a way so that they could be applied over real online datasets? In a way that it becomes independent from where it came?
- The origin of the data *(datasets are predominantly acquired in Del.ico.us, in numerous studies).* Other platforms that allow annotated content and can provide data, after applying the same techniques can support the same conclusions?
- Linguistic aspects impact on the ability to use annotated content for information discovery? The network established by association of the *user*, *resource* and *tag* provides the same type of responses, regardless of the etymological or linguistic origin of the tags actually used?
- It is possible to use recommender systems or other decision support systems with the purpose of supporting the user in his web resources tagging, with advantage on the following information discovery?

Extending these later aspects will presumably lead to develop studies in order to stabilize answers on important facets such as:

- Improve extraction techniques and data reorganization;
- Reduce tags ambiguity, by introducing classification systems;
- Establish nodes relationship formal representation;
- Introduce web search interfaces use for exploration and results evaluation purposes probably in the form of a usable prototype.

This work remains under development and will make a path of greater differentiation from the contributions already made. This path will identify an area of actual research based on a more independent perspective yet to be fully established.

4 References

1. [1] P. Heymann, G. Koutrika, and H. Garcia-molina, “Can Social Bookmarking Improve Web Search ? Categories and Subject Descriptors,” WSDM, pp. 195–205, 2008.
2. [2] S. A. Golder and B. A. Huberman, “The Structure of Collaborative Tagging Systems,” J. Inform. Science, vol. 2, no. 32, pp. 198–208, 2005.
3. [3] W. E. I. Wei and S. Ram, “Using a Network Analysis Approach for Organizing Social Bookmarking Tags and EnablingWeb Content Discovery,” ACM Trans. Manage. Inf. Syst. Article 15, vol. 3, no. 3, p. 16 pages, 2012.
4. [4] G. Begelman, P. Keller, and F. Smadja, “Automated Tag Clustering : Improving search and exploration in the tag space,” Proceedings of the Collaborative Web Tagging Workshop at the International Conference on theWorld Wide Web, pp. 22–26, 2006.
5. [5] L. Boratto, S. Carta, and E. Vargiu, “RATC: A robust automated tag clustering technique,” Proceedings of the International Conference on E-Commerce andWeb Technologies, pp. 324–335, 2009.
6. [6] M. Gupta, R. Li, Z. Yin, and J. Han, “Survey on social tagging techniques,” ACM SIGKDD Explorations Newsletter, vol. 12, no. 1, p. 58, Nov. 2010.
7. [7] P. Mika, “Ontologies are us: A unified model of social networks and semantics,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, no. 1, pp. 5–15, Mar. 2007.
8. [8] W. E. I. Wei, “Utilizing Social Bookmarking Tag Space for Web Content Discovery: A Social Network Analysis Approach,” University of Arizona, 2010.
9. [9] S. Wasserman and K. Faust, Social Network Analysis - Methods and Applications. Cambridge, UK: University Press, 1994, p. 867.
10. [10] H. Halpin, V. Robu, and H. Shepherd, “The Complex Dynamics of Collaborative Tagging,” www, pp. 211–220, 2007.
11. [11] M. A. Hearst, “Design Recommendations for Hierarchical Faceted Search Interfaces,” Proceedings of the ACMSIGIR Workshop on Faceted Search, 2006.
12. [12] M. A. Hearst, “UIs for Faceted Navigation Recent Advances and Remaining Open Problems,” Proceedings of theWorkshop on Human-Computer Interaction and Information Retrieval (HCIR), 2008.

Manuscritos portugueses do século XVIII: uma ontologia do parentesco para extracção semiautomática de relações

Lígia Gaspar Duarte

Universidade de Évora, Centro Interdisciplinar de História, Culturas e Sociedades – CIDEHUS, Palácio do Vimioso, 7002-554 Évora
duarteligia@live.com.pt

Abstract. No presente trabalho pretende-se expor o propósito de um projecto em fase inicial, consiste em desenvolver a representação e extracção semiautomática de relações de parentesco, a partir de manuscritos portugueses do séc. XVIII, o que significa ter em conta as convenções histórico-linguísticas inerentes.

Das relações sociais enunciadas no *corpus* Gazetas Manuscritas da Biblioteca Pública de Évora (1729-1754), a de parentesco assume uma expressão política e social incontornável, como princípio organizador da sociedade. As inúmeras referências que pontuam a narração permitem estabelecer uma tipologia de fórmulas que expressam os paradigmas vigentes dos laços consanguíneos e artificiais. A partir destas desenha-se um possível percurso para a elaboração de um modelo conceptual para a construção de uma ontologia, aplicação e avaliação.

Keywords: Ontologias; Linguagem Natural; Manuscritos Modernos; Relações de Parentesco

1 Introdução

Na confluência da antropologia histórica, da historiografia do período moderno e da computação (via inteligência artificial) encontra-se o presente trabalho da representação do parentesco no Portugal do Antigo Regime (séc. XVIII). Pretende ser uma plataforma alargada que contribua para o estudo das relações sociais. Enquadra-se no programa de doutoramento em Ciências da Informação e Documentação da Universidade de Évora, edição de 2012/2015.

A possibilidade de extracção semiautomática de relações em linguagem natural emerge como uma ferramenta incontornável na edição de fontes históricas. Neste contexto, a importância da fonte documental *Gazetas Manuscritas da Biblioteca Pública de Évora (1729-1754)*, enquanto fonte de referência para estudos afins à corte e sociedade no reinado de D. João V, dá forma a uma questão partilhada por vários projectos ligados à História, como área de investigação; corresponde às possibilidades de representar o conhecimento de um dado conjunto documental, de maneira a extrair os vários níveis de informação explícita e implícita. Ora, a documentação histórica abarca as mais diversas tipologias documentais, revestindo-se de uma complexidade que dificilmente pode ser generalizada. A estrutura das *Gazetas Manuscritas* apresenta níveis susceptíveis de representação muito diversificados. Todavia, uma das que encontra eco nos grupos de trabalho e investigação afins à temática prende-se com a representação de relações e de diferentes tipos de redes sociais, inferidas da documentação. Das relações sociais enunciadas na fonte, a de parentesco assume uma expressão política e social incontornável, como princípio organizador da sociedade.

A questão central do presente trabalho deriva desta preocupação que serve quer investigadores das Humanidades e Ciências Sociais quer os utilizadores/leitores dos arquivos e bibliotecas patrimoniais.

Neste texto faz-se uma síntese do projecto, procura-se aqui sublinhar apenas os aspectos fundamentais do percurso inicial. Optou-se por restringir o estado da arte ao núcleo onde confluem as

várias disciplinas envolvidas. No domínio a representar, mencionam-se os conceitos e contextos de relevo de forma a enunciar os principais problemas. Na metodologia sintetizam-se etapas, deixando em aberto várias possibilidades técnicas, por ora ainda não definidas.

2 Estado da Arte

As diferentes áreas do saber que integram a problemática em análise são diversas. Entre estas, a computação tornou-se, há já algumas décadas, um dos parceiros mais cobiçados, nomeadamente pela linguística e antropologia, no desenvolvimento de novas tecnologias, com implicações na criação de novos métodos de trabalho, bem como na própria formulação dos problemas de estudo. No entanto, apesar da visibilidade destes trabalhos, o despertar de outras disciplinas deu-se muito lentamente, como é no caso da história. É na viragem do século e mais propriamente na primeira década de 2000 que se assiste a um conjugar de esforços assinalável. Designações de campos de trabalho como *History & Computing*, *Humanities Computing* ou a actual *Digital Humanities* reflectem a consciência de esforços conjugados.

No relatório *Past, present and future of historical information science* [1] de 2006 da Royal Netherlands Academy of Arts and Sciences apontam-se percursos de investigação em aberto, entre os quais o recurso às ontologias, enquanto instrumento de recuperação da informação.

A utilização de ontologias em qualquer domínio do conhecimento é vista com algumas reservas por parte da comunidade da Inteligência Artificial, muito embora se multipliquem os projectos que reúnem esforços para a sua aplicação em âmbitos que se desviam de uma lógica racional linear do conhecimento [2]. Os estudos para contornar os obstáculos nas áreas das Humanidades e Ciências Sociais demonstram que as ferramentas disponíveis podem ser reconduzidas para as necessidades específicas deste vasto campo de estudo [3].

A história dispõe já de alguns trabalhos para a representação de conhecimento em fontes documentais, que se debruçam sobre conceitos-chave comuns. Neste contexto, nomeiam-se os contributos de Nagypál [4] e Ide e Woolner [5] quanto às possibilidades de integração dos factores "tempo" e "espaço" na construção do modelo conceptual. Os dois últimos autores recorrem nomeadamente à criação de ontologias independentes, de modo a poderem representar as constantes alterações de factos, ao longo de um determinado período de tempo, solução encontrada pela comunidade da web semântica para modificar, corrigir e desenvolver versões de determinadas ontologias. Foi neste procedimento comum que se encontrou uma ponte para a utilização de ontologias de forma dinâmica, respondendo à problemática da representação "tempo-espaço".

Das experiências realizadas, o projecto *Henry III fine rolls*, desenvolvido entre os National Archives -UK, o King´s College (London) e Christ Church University em Canterbury, prefigura um dos empreendimentos mais significativos [6]. Ao procurar trabalhar a representação de documentação medieval do ponto de vista das questões dos historiadores, elabora uma ontologia que através da marcação do texto permite a identificação de relações entre os indivíduos e as entidades mencionadas nos documentos [7].

Tudo isto está em língua inglesa e centrado maioritariamente em documentação medieval. Pretende-se neste projecto desenvolver ontologias aplicadas a fontes portuguesas e para a época Moderna (século XVIII).

3 O parentesco histórico como objecto de representação

Como sublinha Héritier [8], a familiaridade do termo parentesco induz à acepção errónea de que qualquer pessoa domina o conceito. Este, como área interdisciplinar de estudo encontra-se claramente divorciado da enganosa familiaridade do termo e a sua análise em vários períodos cronológicos implica a utilização de ferramentas que permitam aferir os contextos particulares a que se remetem.

No Antigo Regime, a concepção de *família* encontra-se dividida entre a ideia de coabitação e de consanguinidade, incluindo todos os residentes, nomeadamente os serviçais. O *parentesco* apre-

senta uma conotação restrita aos laços de sangue, muito embora inclua os de afinidade. Esta noção abrangente de parentesco é, aliás, distinta da de *parentela* que respeita a todos os consanguíneos e afins de um único indivíduo, por via paterna e materna. Por outro lado, a *linhagem* circunscreve-se aos descendentes de um mesmo ancestral, normalmente por via paterna. E, enquanto a linhagem tem uma forte componente da ideia de "raça", "sangue", "honra" e "nome", o conceito de *casa* encontra referentes no de linhagem, e conjuga a noção de geração e de família [9], isto é, os laços de afinidade e de sangue que coabitam com outros sob a autoridade de um mesmo chefe de família, herdeiro de um património imaterial e material que se pretende perpetuar. Os aspectos apontados não passam, no entanto, de um esboço tímido e frágil face à complexidade das questões abordadas.

Ainda relativamente à terminologia, sublinha-se que, apesar de se inscrever num mapa terminológico ocidental, apresenta uma dinâmica muito distinta. É o caso da ordem de nascimento dos filhos legítimos que corresponde à hierarquia estabelecida pelo direito de primogenitura e varonia, na sucessão e correspondente sistema de herança (filho sucessor, herdeiro; filho segundo). Da mesma maneira, quanto aos filhos ilegítimos também existe diferenciação terminológica, conforme a sua origem (filho natural, adulterino, bastardo, sacrílego).

À sobreposição de conceitos de parentesco, para expressar e/ou reforçar determinadas relações, precede o entendimento de uma realidade social que atribui uma dimensão jurídica e moral às relações enunciadas e, através desta dimensão, delimita os comportamentos lícitos e ilícitos e circunscreve as acções dos indivíduos mediante o controlo da igreja e do Estado Moderno. Além da terminologia do parentesco é necessário representar o contexto em que esta emerge, daí a complexidade da análise.

De que forma é que são expressas as relações de parentesco nas *Gazetas Manuscritas*? Sugere-se como exemplo, a análise da formalização das relações num dos registos de 1731 [10].

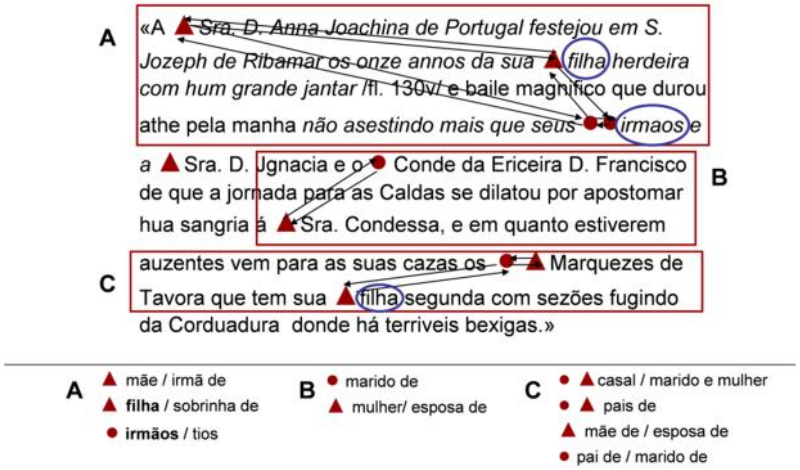


Fig. 1. Excerto de uma "gazeta" manuscrita de 1731. Biblioteca Pública de Évora, Códice CIV/1-5d, fls. 130-131.

Destaca-se desde logo o escasso número de termos de parentesco em comparação com o número de relações expressas de forma implícita, facto possível através do princípio da reciprocidade de relações (cada termo inclui dois sentidos na ligação). Da mesma forma que chama a atenção para outro tipo de expressão: a identificação de titulares nobiliárquicos e respectivas consortes.

Na expressão das relações identificam-se duas vertentes distintas: as implícitas e as explícitas. As primeiras definem-se pela ausência de termos específicos de parentesco, apesar de revelarem formas que implicam alguns graus/formas de relações deste género; as segundas definem-se pela presença estruturante de termos de parentesco, em conjugação ou não com outros elementos.

A terminologia é a pedra chave na expressão de relação de parentesco. Esta pode ser directa (termos específicos: mãe, avô, irmão, filho), ou associada (palavras relativas a ac-

ções/acontecimentos em grupos de parentesco: casar, adoptar, nascer, baptismo, herança). Em ambos os casos, a articulação destes termos com as diversas formas de identificação de indivíduos intervenientes (cargos, estatutos, estado, títulos, ofícios, naturalidade, entre outros) resultam em inúmeras fórmulas que expressam relações que extravasam o domínio do parentesco, já por si complexo.

A profusão de vocabulário a trabalhar estabelece, no entanto, uma plataforma de correspondência para outros aspectos da sociedade de Antigo Regime, permitindo por isso a construção de várias pontes para outro tipo de relações sociais. Aliás, a ligação entre relações de parentesco e redes sociais estende-se para além dos limites das considerações óbvias, revelando-se um processo complexo, com múltiplas conjugações. Uma delas é o estabelecimento de laços de afinidade, como o casamento, que origina uma multiplicidade de relacionamentos distantes que ultrapassam naturalmente a esfera familiar alargada, e assumem um outro tipo de papéis que se fundem no domínio social. As frequentes relações de interdependência entre o parentesco, as redes de influência e consequentes redes de poder, são inalienáveis das respectivas repercussões sociais, e vice-versa. Ao contextualizarmos estas noções às realidades socioculturais do Antigo Regime em que a hereditariedade, a parentela, e os privilégios políticos e sociais são indissociáveis, cada item referido parece encontrar, neste período, uma das suas expressões mais exuberantes.

O âmbito alargado da representação do parentesco pode fomentar assim as possibilidades de reutilização da ontologia a criar.

4 Metodologia

Conforme a literatura referente à metodologia para a construção de ontologias [11], elaborou-se um plano faseado que procura integrar as várias áreas disciplinares implicadas e respectivos métodos de trabalho, como se pode verificar na figura 2.

O plano faseado da metodologia seguida consiste em três vertentes essenciais: a compilação terminológica do parentesco, através do levantamento das expressões de relação de parentesco para o intervalo de 1729-1740; a elaboração de um mapa conceptual para as gazetas manuscritas que descreva os processos desencadeados pelas relações entre os conceitos envolvidos; e a construção e implementação de uma ontologia que se pretende avaliar através da respectiva aplicação a amostras de *corpora* distintos (manuscritos/impressos).

Destaca-se, sobretudo, a segunda etapa, uma vez que prefigura o esqueleto robusto que fornecerá os elementos necessários para o ensaio da extracção semiautomática de relações de parentesco. Não obstante, a terceira etapa é o núcleo estruturante de todo o projecto. Os recentes trabalhos sobre métodos de extracção semiautomática de relações em texto apontam vários percursos possíveis [12]. E, neste sentido, a reflexão acerca do caminho mais adequado será crucial na concretização dos objectivos do projecto. Apesar da vertente da execução manual dominar grande parte do trabalho, sobretudo na construção da ontologia, desenha-se um possível recurso a formas semiautomáticas para a anotação do texto.

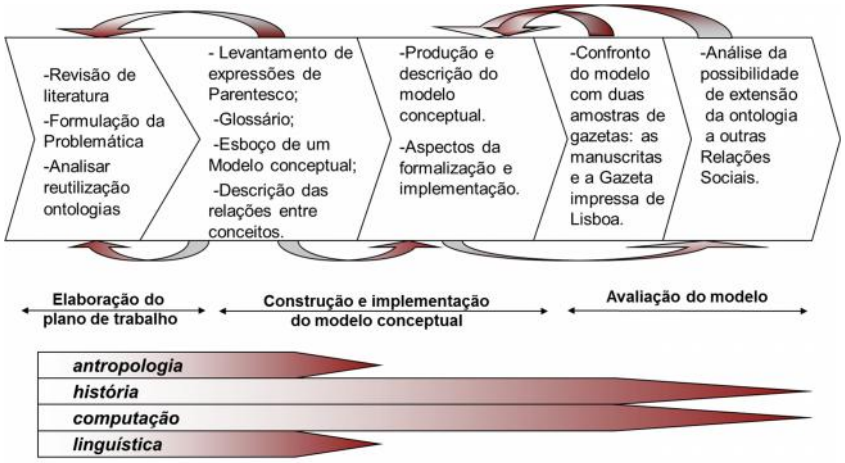


Fig. 2. Esquema genérico das etapas metodológicas e transversalidade disciplinar diferenciada.

5 Conclusões

Da apresentação inicial do problema de estudo, a uma síntese do estado da arte, passando pela reflexão sobre o vocabulário e os contextos a representar procurou-se sublinhar a pertinência do âmbito de trabalho, dado a inexistência de referências para o caso português.

Referências

[1] Boonstra, O., Breure, L., Doorn, P.: Past, present and future of historical information science. Amesterdam: DANS - Data Archiving and Networked Services / The Royal Netherlands Academy of Arts and Sciences. (2006)

[2] Stevens, R., Aranguren, M. E., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A.: Using OWL to model biological knowledge. International Journal of Human-Computer Studies [on-line]. 65(7): (2007) 583-594 http://dx.doi.org/10.1016/j.ijhcs. 2007.03.006

[3] Kahlert, R.; Sullivan, J.: *Toward a Knowledge Representation Corpus of Historical Events* [on-line]. (2006) http://clio-knows.sourceforge.net/corpus.pdf

[4] Nagypál, G.: History Ontology Building: the tecnical view. Proceedings of the XVI international conference of the Association for History and Computing. Amsterdam. Royal Netherlands Academy of Arts and Sciences. (2005) 207-214 www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/**2005**1064.pdf

[5] Ide, N.; Woolner, D.: Historical Ontologies. In: Ahmad, K., Brewster, C., Stevenson, M. - Text, Speech and Language Technology. Words and Intelligence II - Essays in Honor of Yorick Wilks 36. Netherlands. Springer. (2007)
http://www.cs.vassar.edu/~ide/papers/festschrift.pdf

[6] Ciuala, A.; Spencer, P.; Vieira, J. M.: Expressing Complex Associations in Medieval Historical Documents: the Henry III Fine Rolls Project. Literary and Linguistic Computing. 23(3): (2008) 311-325 http://llc.oxfordjournals.org/content/23/3/311.abstract

[7] Ciuala, A.; Vieira, J. M.: Implementing an RDF/OWL Ontology on Henry the III Fine Rolls. *OWLED* 258 (2007) 1-10
http://www.webont.org/owled/2007/PapersPDF/submission_6.pdf

[8] Héritier, F.: Parentesco. Enciclopédia Einaudi – *Parentesco*. 20. Imprensa Nacional-Casa da Moeda. (1990) 28

[9] Bluteau, R.: Vocabulário Portuguez e Latino. II. Coimbra. (1713-1728) 174

[10] Diário de 1731. Biblioteca Pública de Évora. Códice CIV/1-5d, fls. 130-130v

[11] ABECKER, A.; HITZLER, P.; GRIMM, S.: Knowledge Representation and Ontologies. Logic, Ontologies and Semantic Web Languages. In: STUDER, R.; GRIMM, S.; ABECKER, A. (Eds.) - SemanticWeb Services. Concepts, Technologies, and Applications. Berlin Heidelberg New York, Springer. (2007) 80

[12] Choi, M.; Harksoo, K.: Social Relation Extraction from Texts Using a Support-vector-machine-based Dependency Trigram Kernel. Information Processing and Management. 49 (1): (2013) 303–311

Components for Spoken Dialogue Systems: a brief survey

Pedro Fialho[1,2,⋆], Paulo Quaresma[1], and Luísa Coheur[2]

[1] Universidade de Évora, Portugal
[2] L2F/INESC-ID Lisboa, Portugal

Abstract. Natural human communication with machines has a major role in widespread usage of computational resources in everyday life. Namely, the capability of understanding natural language, for fluent human-machine dialogues, may be achievable by a series of parsing and knowledge matching operations, ideally featured in systems with some sensorial human-like resemblance. These features/components are spread over distinct fields of study, from which a selection was made (mostly based on available resources/information), briefly covered in this paper.

Keywords: speech, natural language processing, dialogue systems

1 Introduction

Interacting with machines by using human language is a task from the field of Dialogue Systems (DS), where domain dependent Question Answering (QA) is combined with out of domain knowledge handling. The core efficiency of this task relies on textual Natural Language Understanding (NLU) and Natural Language Generation (NLG), which may involve simple pattern matching or more advanced techniques from the fields of Natural Language Processing (NLP) and/or Machine Learning (ML), as briefly described in section 2.

Spoken Dialogue Systems[6] (SDS) extend DS with speech processing technologies, essentially for Automatic Speech Recognition (ASR) and Text to Speech Synthesis (TTS), covered in section 3. SDS will be further mentioned when any of these capabilities is required, for distinction from text based DS.

Speech enables a more human-like machine interface and is best complemented by a virtual character and/or environment, requiring technologies from the field of Computer Graphics (CG). These virtual environments enable visual complements for natural language (NL) outputs, through the use of graphical assets representing objects involved in the dialogue, thus allowing a more unambiguous and obvious description of knowledge already stated in NL. The virtual character represents the user interface, with all user input being directed to it, accounting for a friendly and inviting environment that triggers curiosity and enforces system adoption by users unfamiliar with SDS and assorted technologies.

Due to increasing popularity as a discussion topic, DS are being referenced by multiple names[3] (often not compatible) for enforcement of usage or application specific details. Key terms in these names, describing representative features and concepts, include:

- multimodal, the usage of multiple input and/or output types of representing information (such as touch, text, graphics or speech), eventually combined;
- agent, an autonomous entity underneath a system's interface, capable of understanding and generating human language. Mainly inherited from the Artificial Intelligence (AI) field's concept[20] of the same name;
- assistant, an agent providing support on specific domains; as opposed to the usage of chat or conversational related terms, which declare a capability or intention for phatic talk[11];
- avatar (or any term resembling humans or the existence of a human body), an agent represented visually in a virtual environment;

⋆ Corresponding author. Email: prpfialho@gmail.com

[3] http://www.chatbots.org/synonyms/

Currently, SDS are best known from commercial applications for handheld devices, that provide assistance on daily tasks (domain dependent[23] or not[4]) while the QA task became popular after a computer defeated humans in a knowledge based television contest[4]. Limited SDS (essentially QA systems with few out of domain topics) can be found on a variety of public spaces (such as museums[19]) and for training purposes[18]. Their usage on automotive environments has also been explored[1] and is an active discussion topic[10], since voice interaction with in-car devices and controllers has shown to reduce distractions while driving[22].

Recent advances in DS can be tracked in the conference of the Special Interest Group on Discourse and Dialogue (SIGDIAL), the Discourse & Dialogue journal and several smaller venues such as the Young Researchers Roundtable on Spoken Dialogue Systems and the workshop on Knowledge and Reasoning in Practical Dialogue Systems. Dialogue based NL handling is also (at least partially) covered in the Language Resources and Evaluation conference (LREC) and the conferences of the International Speech Communication Association (INTERSPEECH) and the Association for Computational Linguistics (ACL and EACL), though these are mostly focused on components typically used for interface and feedback in DS.

2 Handling core knowledge

Text interfaces in DS have specific NLU problems, such as abbreviations and misspelled words, not present in speech interfaces. However, human inputs (text or speech) always end up converted to a textual form, for further NLU that allows information retrieval from knowledge bases (necessarily built in textual form) with reduced ambiguity.

NL parsers can be used for low level text analysis, allowing proper segmentation and classification of linguistic components. Parser usage in DS has been explored[8], although not being actively discussed. However, current robustness and multilingual support by such parsers suggests that their usage can be an asset to DS. Therefore, below are described some of these parsers and generic NLU and NLG techniques, from knowledge authoring and reasoning platforms known as chatbots.

2.1 Parsers

NL parsers can be generated from grammars (handmade using application specific formalisms), with tools such as DepPattern[5] or AGFL[6]. Although suitable for small scale domains, the analysis of unrestricted NL usages requires tools offering more robustness and coverage, due to the inherent complexity and extent of NL. The below described parsers use rules, ML or both to increase global performance, representing valuable approaches to the state of the art in NLP, although many others could be mentioned[9].

FREELING[15] is a NLP toolkit and library providing language identification, grapheme to phoneme[7] conversion (G2P) and lexical, morphological, syntactic and semantic information for text inputs (of varying length) in 9 languages, being bundled with varying feature coverage across them, as summarized in [15]. These features are gathered from isolated modules, where some may: a) be chained; b) use ML, rules or a combination of both; c) be configured with regular expressions or application specific notations. Feature coverage is richer for English and Spanish, although relevant/essential features (for SDS) are available for all languages, such as part of speech tagging (POS) or named entity detection.

TREETAGGER[21] features a ML approach to POS, by statistically inference on hand annotated corpora. Also with ML, the Berkeley[8] and Stanford[9] parsers are able to infer probabilistic context free grammars from partial or limited lexical information[16, 7].

[4] http://www.apple.com/ios/siri/
[5] http://gramatica.usc.es/pln/tools/deppattern.html
[6] http://www.agfl.cs.ru.nl/
[7] Text and speech units, further discussed in section 3.
[8] http://code.google.com/p/berkeleyparser/
[9] http://nlp.stanford.edu/software/lex-parser.shtml

2.2 Chatbots

The type of conversation usually assumed for a DS has a mainly QA purpose which is not realistic after observation of interactions from real users[14] (unaware of the underlying domain), who will eventually try to mimic human conversation through small talk. A chatbot is a type of agent (usually developed in community/hobby projects) unfocused on a particular or useful domain and therefore underestimated and out of the scope of organizations.

In most chatbots, domain and agent profile knowledge is hand authored in an implementation specific scripting language, usually reusing syntax and design concepts from well known knowledge representation formalisms and markup languages. Content authoring comprises a complex task, requiring familiarity with NL usage and dialogue specific phenomena (such as question anticipation and syntactic variations) due to unpredictability in real user's NL usage, particularly in public spaces and web based deployments.

A scripting language is designed or adapted for chatbot knowledge representation, dialogue flow control and/or pattern matching, intended to ease the authoring process[10]. However, a common problem with chatbot knowledge bases is their size and verbosity[11], particularly when scripting inherits features from computational languages, not intended or designed for humans and thus negatively affecting readability and knowledge management.

Chatbot evaluation is usually subjective due to absence of a definition for correctness in answers, having the Loebner Prize[17] for state of the art ranking, where human judges select who is human after a series of written conversations[12]. NLU platforms feasible for the chatbot task include:

- AIML[13], a rule description language based on XML, which has a wide community of users and developers (supplying engines for all major programming languages) as its main strength.
- CHATSCRIPT[14], a rule based language and engine, featuring: a) linguistic and domain dependent constraints; b) dynamic fact triples/tables; c) dialogue specific operators (such as discourse control variables and rejoinders); d) embeddable C-like programming. Script organization is inspired by the Scone project[3] although authoring is made in a newly created language, aimed at improving readability by humans (for comparison, a CHATSCRIPT chatbot won the Loebner bronze with ten times less rules than a previously winner AIML chatbot);

Chatbot related problems and approaches can represent (and summarize) the main module of NLU and NLG in a DS. Therefore, a chatbot engine is a domain independent NLU and NLG platform/framework (from which results a core DS - the chatbot), using a specific knowledge description language, having text as the only form of input and output and (usually) forcing conversation to happen in turns.

3 Speech processing

Rich interfaces for SDS require human-like behaviors and multimedia features. One of the richer features in SDS is the ability to input and output information reliably and exclusively through speech. Speech is the most natural and innate form of language expression, and its interpretation by computers can be seen as a data mining problem where an input utterance is matched or generated on a collection/model of examples, following some ML scheme.

Speech interfaces lack the issues commonly seen in text inputs, for obvious lack of abbreviations in spoken language (although unpredicted neologisms may appear) and due to ASR design as a classification problem (recognized words must exist in the language model). However, automatic

[10] Authors may not be aware of computational formalisms/languages.
[11] `http://gamasutra.com/blogs/BruceWilcox/20120104/9179/`
[12] The gold prize was never won, therefore current state of the art chatbots only compete for bronze.
[13] `http://www.alicebot.org/aiml.html`
[14] `http://sourceforge.net/projects/chatscript/`

speech transcriptions may contain mismatched words (due to phonetic similarity) and incomplete or incorrectly structured sentences, usually due to signal capturing issues.

A conceptual relation between speech and text can be understood from their underlying knowledge representation units, phonemes and graphemes respectively. Letters, numbers and other symbols used to imply/convey meaning in text, alone (as in Chinese) or combined, are called graphemes and may have a correspondent, and eventually non unique, sound or effect (such as pauses and intonation changes) when spoken. Each sound in a spoken language is represented in textual form with an element from a phonetic alphabet - the phoneme - consisting of one or more symbols (such as letters, the tilde or the at sign) which may describe speech production features such as nasality.

3.1 Automatic Speech Recognition

As with any data mining problem, more data means better matching/recognition performance, therefore speech has been collected and studied by companies and research institutes for generation and combination of computational models that represent acoustic/phonetic and language/linguistic features.

ASR technology details are not widely published (only macro components/organization), since fine tunning of ML schemes strongly depends on sensitivity and knowledge of a spoken language's sounds (as text is for NLU/NLG), which comprises a valuable/secret asset.

AUDIMUS[13] is a speech recognition engine using Hidden Markov Models (HMM) - among other ML techniques, out of the scope of this paper - for phoneme classification/detection on acoustic features extracted from a speech/audio signal. Phonemes are mapped to their corresponding graphemes (the reverse of G2P) and matched against a language model describing combinations/usages of a language's words (compiled into an n-gram model), which may be domain restricted, based on a generic/large vocabulary (usually from broadcast news corpora) or a combination of both. It has been successfully applied for ASR in English, Portuguese (varieties included) and Spanish[12].

AUDIMUS is also a toolkit and library, providing speech/audio related features such as speaker identification and jingle detection, which are particularly useful in SDS designed for real time usage, where speech is streamed/input constantly (as opposed to push-to-talk interfaces). AUDIMUS is developed by the Spoken Language Systems Laboratory of the INESC-ID research institute (L2F/INESC-ID); other organizations with well known success/results in the ASR task include Nuance (which integrates Loquendo), Microsoft and the Carnegie Mellon University (CMU) with its SPHINX project.

3.2 Text To Speech synthesis

TTS is one of the key technologies for human-like SDS, allowing an agent to express an answer/decision in audible form, eventually discarding the need for textual/visual feedback.

Common TTS technologies include unit selection synthesis[2], where features from utterances in a speech database are combined to form a new speech signal, and HMM based synthesis[24] where a ML model is calculated from fluently read speech recordings.

Evaluation of TTS is essentially subjective, covering features such as the naturalness of prosody (detecting issues such as the presence of glitches or incorrect intonations) and the intelligibility of synthesized speech (for incorrectly spelled words and pitch variations). The Blizzard challenge is the main TTS competition, ranking state of the art approaches.

Typical speech tools for TTS include support for model generation from newly supplied resources. Synthesized speech's audio quality strongly depends on the original resources, but also on the underlying technology. For instance, concatenative synthesis with monophones (a phonetic set comprised of one phoneme per time/state) is worse than with diphones (a phonetic set whose elements combine/encode the current and previous state's phonemes), which allows a softer transition between basic sounds. Well known TTS approaches/organizations include L2F/INESC-ID's DIXI, the University of Edinburgh's FESTIVAL, DFKI's (the German Research Center for Artificial Intelligence) MARY and Microsoft.

4 Conclusion and uncovered topics

The components that may form a SDS are spread across current, highly relevant and understudy knowledge fields. Apart from the many references left out of this paper, there are also fields and technologies/approaches not covered here. These include SDS hubs/platforms (such as OLYMPUS[15] and GALATEA[16]), that allow inter component communication on a specific protocol, aimed at the ultimate task of joining multiple components and therefore build a SDS. An also major task/component left out is graphical agent design and manipulation, which involves game development tools such as UNITY[17] or CRYENGINE[18], with the latter being better in covered CG technology (therefore allowing better quality graphics - near photo realistic - in real time) and worse in ease of usage. Particularly relevant usages of a fully featured SDS are in tutoring and entertainment (computer, eventually serious, games - also applicable for educational purposes) such as E-TUTOR[5] or FAÇADE[19], with the latter being known by its exclusively natural language based narrative control.

References

1. Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Ivana Kruijff-Korbayová, Andreas Korthauer, Manfred Pinkal, Michael Pitz, Peter Poller, and Jan Schehl. Natural and intuitive multimodal dialogue for in-car applications: The sammie system. In *Proceedings of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva del Garda, Italy*, pages 612–616, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.
2. Robert A. J. Clark, Korin Richmond, and Simon King. Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49(4):317–330, 2007.
3. Scott Fahlman. Using scone's multiple-context mechanism to emulate human-like reasoning. In *AAAI Fall Symposium Series*, 2011.
4. David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010.
5. Pedro Fialho, Sérgio Curto, Luísa Coheur, and Ricardo Ribeiro. E-tutor: a tutoring agent. In *International Conference on Computational Processing of Portuguese (Propor 2012)*, volume Demo Session, April 2012.
6. K. Jokinen and M. McTear. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2010.
7. Dan Klein and Chris Manning. Accurate unlexicalized parsing. In *Proceedings of ACL*, 2003.
8. Gergely Kovásznai. Algorithmic improvements in natural language parsing within dialogue systems: Priority patterns and wildcards. In *Proceedings of the International Conference on Applied Informatics*, pages 129–138, 2004.
9. Jonathan K. Kummerfeld, David Hall, James R. Curran, and Dan Klein. Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of EMNLP*, Jeju Island, South Korea, July 2012.
10. Andrew L. Kun, editor. *International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '12, Portsmouth, NH, USA - October 17 - 19, 2012*. ACM, 2012.
11. David Marsh. Book Review : Small Talk – Analysing Phatic Discourse Klaus P. Schneider Hitzeroth: Marburg, 1988, 352 pp. *RELC Journal*, 20(2):88–89, December 1989.
12. H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. P. Neto. The l2f broadcast news speech recognition system. In *Proceedings of Fala2010*, Vigo, Spain, 2010.
13. Hugo Meinedo, Diamantino Caseiro, João Neto, and Isabel Trancoso. Audimus.media: a broadcast news speech recognition system for the european portuguese language. In *Proceedings of the 6th international conference on Computational processing of the Portuguese language*, PROPOR'03, pages 9–17, Berlin, Heidelberg, 2003. Springer-Verlag.

[15] `http://wiki.speech.cs.cmu.edu/olympus/index.php/Olympus`
[16] `http://hil.t.u-tokyo.ac.jp/~galatea/`
[17] `http://unity3d.com/`
[18] `http://www.mycryengine.com/`
[19] `http://www.interactivestory.net/`

14. Pedro Mota, Luísa Coheur, Sérgio Curto, and Pedro Fialho. Natural language understanding: From laboratory predictions to real interactions. In *15th International Conference on Text, Speech and Dialogue (TSD)*, volume 7499 of *Lecture Notes in Artificial Intelligence*. Springer, September 2012.
15. Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.
16. Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, pages 404–411, Rochester, New York, April 2007. Association for Computational Linguistics.
17. David M. W. Powers. The total turing test and the loebner prize. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, pages 279–280, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
18. Albert Rizzo, Thomas D. Parsons, John Galen Buckwalter, Belinda Lange, and Patrick G. Kenny. A new generation of intelligent virtual patients for clinical training. *American Behavioral Scientist*, 2012.
19. Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.
20. S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2010.
21. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.
22. Patrick Tchankue, Janet Wesson, and Dieter Vogts. Are mobile in-car communication systems feasible?: a usability study. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '12, pages 262–269, New York, NY, USA, 2012. ACM.
23. Markku Turunen, Jaakko Hakulinen, Olov Ståhl, Björn Gambäck, Preben Hansen, Mari C. Rodríguez Gancedo, Raúl Santos de la Cámara, Cameron Smith, Daniel Charlton, and Marc Cavazza. Multimodal and mobile conversational health and fitness companions. *Comput. Speech Lang.*, 25(2):192–209, April 2011.
24. Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black, and Keiichi Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6)*, August 2007.

OGCP - A new ontology for clinical practice knowledge representation and a proposal for automated population

David Mendes** Irene Pimenta Rodrigues* and Carlos Fernandes Baeta**

* Universidade de Évora; ** Hospital José Maria Grande
{dmendes,ipr}@uevora.pt
http://www.uevora.pt

Abstract. We introduce the **OGCP** - *Ontology for General Clinical Practice*. The *Ontology for General Medical Science* ***(OGMS)*** *complemented with the Computer-Based Patient Record Ontology* ***(CPR)*** *is based on several upper ontologies which may have formal ontological relations according to the OBO Foundry principles. These ontologies accordant to the underlying* Ontological Realism *render a structure with reasoning capabilities that reach further than those possible with logical formalisms alone. We propose to extend carefully the OGMS taking into account the diverse ontological relations found in the recently proposed Basic Formal Ontology V2, FMA and SNOMED-CT as foundational ontologies in order to extract axioms for ontology enrichment from natural language text. With these cautions in mind, using careful instantiation we improve largely the reasoning capabilities over the resulting OWL knowledge base. Most of the clinical practice knowledge is currently recorded in text format, namely in the semi-structured* ***SOAP*** *(Subjective, Objective, Analysis, Plan) framework format. We extend the OGMS with the CPR structure into an Ontology for General Clinical Practice* ***(OGCP)*** *for the generation of adequate ontologically rich axioms from the SOAP text segments.*

Keywords: *OGMS, CPR, OGCP, Ontological Realism, SOAP, Clinical Practice Knowledge, OWL*

1 Introduction

Our main intention is to be able to enrich automatically an extension of the OGMS [19] capturing knowledge from clinical reports text. Ontologies formalize the meaning of terms in a vocabulary and provide a mechanism to integrate knowledge from different sources through semantic annotation of data. Interoperability of ontological resources is required to automatically analyze data across different data repositories and to enable automatic reasoning for knowledge discovery. One milestone has been the development and establishment of ontologies in the biomedical research community with the goal of integrating knowledge from different scientific resources and domains. In recent years, more emphasis has been put on the standardization, formalization and interoperability of the data resources and ontologies that characterize them. However, the proliferation of domain-specific ontologies has resulted in an urgent need to develop an approach to bridging the increasing gaps between them. It has now become necessary to automatically resolve inconsistencies across these resources to facilitate automated reasoning, formulation of complex queries across a variety of data resources, testing of hypotheses against the current body of knowledge and translational research.

Barry Smith and Werner Ceusters make an extraordinary argument in favor of the coordinated evolution for scientific ontologies through the application of Ontological Realism [28]. The pragmatic good results of providing a controlled system for ontological annotation are shown in [3] by illustrating the GO [1] success.

In our work we do our best to overcome the different issues identified by the several experts in [4]. In order to maximize the reasoning capabilities based in our extended OGMS ontology, different considerations in the referred work by Brochhausen et al. were taken into good account.

* The current work is funded by the 'Bento de Jesus Caraça' scholarship.

We complemented the **OGMS** ontology with the **CPR** into what we call the **OGCP** that is intended to be a more supportive structure for representation of clinical practice while, at the same time, embodies a formal medical theory of disease and healthcare.

2 Previous Work

We base our work in [18] for what matters about the foundational principles of structuring meaningful knowledge representation as a framework for clinical reasoning. Although focusing our previous work [15] mainly over the CPR ontology [6], we are now targeting the OGMS because it is more promising as suitable for representation of a disease theory, model and corresponding reasoning capabilities.

3 Methodology and approach

In the following sub-sections the current line of work is illustrated. The division provides a structuring approach to introduce and understand the different issues.

3.1 The problem with current Clinical Medicine ontologies

Regarding ontologies in the sub-domain of *Clinical Medicine*[1] there are some issues lacking thorough study. These can be stated as current problems for the effectiveness of using ontologies as knowledge support for clinical reasoning. Problems found in current ontologies and enumerated in literature [9] that lead to reasoning hurdles are:

- Lack of adequate modularization [20] to achieve the minimum amount of implicit differentiation among primitive concepts.
- Inadequate clear separation of digital entities from the reality they represent.
- Inability to avoid the *knowledge acquisition bottleneck* [29] in order to speed start any automatic enrichment.

3.2 Ontological Realism and Relations

Ontological Realism is a methodology to avoid mistakes that cannot be detected by logical formalisms alone [5]. We still want to highlight the reasoning power that formal ontological relations provide to a carefully crafted ontology given the higher semantic level that these relations comprise [28]. The formalization of *Ontological Relations* has been advocated for many years and it succeeded in the development of *"relations that obtain between entities in reality, independently of our ways of gaining knowledge about such entities"* [25].

3.3 OGCP as suitable support for Clinical Practice Knowledge

OGCP includes very general terms about entities involved in the healthcare practice domain that are used across medical disciplines, including: 'disease', 'disorder', 'disease course', 'diagnosis', 'patient', and 'healthcare provider'. OGCP uses the Basic Formal Ontology (BFO) as an upper-level ontology. The scope of OGCP is restricted to humans, but many terms can be applied to a variety of organisms. OGCP provides a formal theory of disease that can be further elaborated by specific disease ontologies. This theory is implemented using OWL-DL and is available in OWL. So far we are elaborating in Chardiovascular related healthcare.

[1] The study of disease by direct examination of the living patient

3.4 Ontological relations to leverage

SNOMED CT as the primary terminology aggregation In [27] is well illustrated that "SNOMED CT and the OBO Foundry differ considerably in their general approach". Nevertheless, a general trend towards more formal rigor and cross-domain interoperability can be seen in both and we argue that this trend should be accepted by all similar initiatives in the future. Our effort will take advantage of the breadth of coverage of SNOMED CT in our domain of interest. SNOMED - CT is a comprehensive terminological framework for clinical documentation and reporting. Comprised of about half a million concepts: Clinical findings, procedures, body structures, organisms, substances, pharmaceutical products, specimen, quantitative measures, and clinical situations. It has an underlying description logic ($\mathcal{EL}$ family). $\mathcal{EL}$ family has shown to be suitable for medical terminology and subsequently, $\mathcal{ELHR}+$ is the performance target of many modern classifiers including those based in consequence driven reasoning capable of classifying SNOMED CT in practical and acceptable processing times as shown by Kazakov in his IJCAI "best paper award winner" [13] with recent proposed extension for concurrent processing [14,23] that benefits of current advances in *BigData* cluster processing. As we show in [15] we use the SNOMED CT and FMA as Golden Standard Ontologies for terminological alignment.

3.5 Ontology learning from text

Accepting as evidence the fact that most of the clinical information is maintained in text form, we shifted our focus from the extraction based in semantic models in the EHRs like the HL7 v3 RIM [16] to the automated acquisition from clinical reports. This introduced us the problem of acquiring the knowledge necessary for learning ontologies known as the *"Knowledge Acquisition Bottleneck"*. This challenging [29] issue remains one of the main barriers for automated acquisition and we tried to circumvent it by using a progressive tutored learning approach. Our approach is to depart from semi-structured text and use the semi-automated translation tasks to generate a controlled *domain specific vocabulary* on which further acquisition tasks build upon [15] minimizing ambiguity and redundancy for better reasoning capabilities. When trying to instantiate individuals (populate) formal heavyweight ontologies like the OGCP we do not normally intend to enrich the ontology but instead turn them from theoretical models of the domain into reasonable knowledge bases.

3.6 SOAP reports as NLP base

We are acquiring from what can be called a semi-structured repository of clinical practice information: the personal SOAP[2] framework reports. When trying to apply the principles of well defined formal ontologies depicted in [26] and trying to avoid the errors mentioned in [5] we decided to get our hands dirty with a simple approach to the representation of disease and diagnostic as illustrated in [24].

In our system there is a clear support for text divided by 4 pre-defined subsections curiously acronymed SOAP after Subjective, Objective, Analysis and Plan. For any particular encounter (actually for any Clinical Episode) the text for any of these may be collected in the form of text suitable for processing into the Ontology and the appropriate suggested points in the OGCP time frame are the following:

Where Subjective notes are those kind of signs normally expressed by the patient as well as soft validated symptoms. Objective findings are all kinds of observations and results from quantifiable exams. Processing and populating the Ontology with the Analysis record labeled as Diagnosis the "Clinical Picture" is completed and is only lacking the Plan being instanced to render the therapeutic response associated with this encounter.

[2] *Subjective, Objective, Action and Plan*

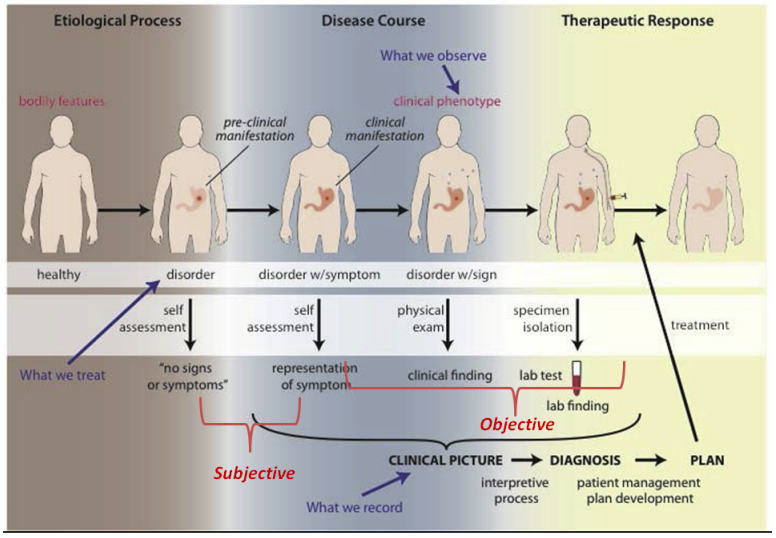


Fig. 1. SOAP Points Insertion

4 Automated acquisition from Clinical Episodes Text

4.1 The generic situation

As reviewed in [29] the state-of-the-Art for acquisition from Clinical Text has enjoyed strong developments in recent years. Here we are delving into the more generic possibility of extracting from free text present in most interfaces used by clinicians. Going from clinical episodes free text that is usually presented in a human friendly format to one adequate for computer processing involves a fair amount of text processing to handle situations like:

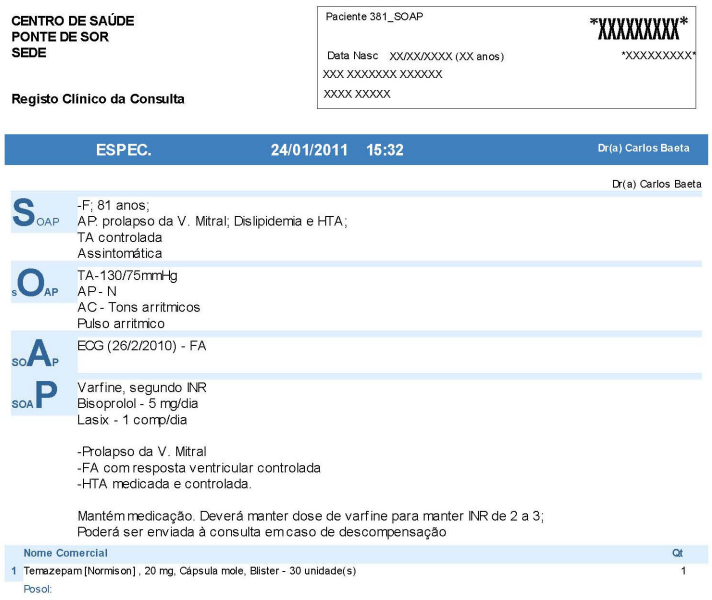
- Reports aggregate information from different clinical episodes that are not uniquely identified or not even individually dated
- The clinician is only identified by his/her name if any identification is made at all
- The information conveyed in free text is intended only to be understandable by fellow practitioners or even by the clinician himself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community
- Text is profoundly intermixed with decorative elements for better legibility, normally in PDF or HTML files
- The time spanning of the processes depicted in natural language are difficult to represent formally
- The clinicians natural language is other than English without concepts defined in the foundational thesaurus like SNOMED CT or FMA for instance that don't exist in that particular language

4.2 Turning the Multilingual problem into an advantage

We can take advantage of the fact that we have to translate from jargon to English to customize the Google translator toolkit[3] with our own Translation Memories and Glossaries. Let us introduce some demonstrative examples taken from a sample document gently provided by Dr. Carlos Baeta and properly de-identified:

We will, in the process of using the Google toolkit, create Translation Memories with the identified personal acronyms like:

[3] https://translate.google.com/toolkit


CENTRO DE SAÚDE
PONTE DE SOR
SEDE

Registo Clínico da Consulta

Paciente 381_SOAP *XXXXXXXXX*
Data Nasc XX/XX/XXXX (XX anos) *XXXXXXXXX*
XXX XXXXXXX XXXXXX
XXXX XXXXX

ESPEC. 24/01/2011 15:32 Dr(a) Carlos Baeta

Dr(a) Carlos Baeta

S OAP
-F; 81 anos;
AP: prolapso da V. Mitral; Dislipidemia e HTA;
TA controlada
Assintomática

SO AP
TA-130/75mmHg
AP - N
AC - Tons arritmicos
Pulso arritmico

SOA P
ECG (26/2/2010) - FA

SOA P
Varfine, segundo INR
Bisoprolol - 5 mg/dia
Lasix - 1 comp/dia

-Prolapso da V. Mitral
-FA com resposta ventricular controlada
-HTA medicada e controlada.

Mantém medicação. Deverá manter dose de varfine para manter INR de 2 a 3;
Poderá ser enviada à consulta em caso de descompensação

	Nome Comercial	Qt
1	Temazepam [Normison] , 20 mg, Cápsula mole, Blister - 30 unidade(s)	1

Posol:

Fig. 2. SOAP Report Sample

– AP (Antecedentes Pessoais) into Personal History
– HTA (Hiper Tensão Arterial) into High Blood Pressure
– FA (Fibrilhação Auricular) into Atrial Fibrilation
– V. Mitral (Válvula Mitral) into Mitral Valve

Some which are acronyms that can be given the suitable translated concept like:

– ECG (Electro Cardio Grama) into Electro Cardio Gram

or those that are even English acronyms:

– INR (International Normalized Ratio) into International Normalized Ratio

Included in this sample are notorious some more complex problems that are not related to the translation itself but with some other problems like the time spanning of concepts like “1 comp/dia” which is adequately translated to “1 tablet per day” using the defined Translation Memory but has to be posteriorly well defined as time delimited occurring process this kind of problems.

SOAP Report This report depicts a clinical encounter in a semi-structured way. As seen previously in the figure in this section we find sections that can be associated with

Symptoms, the subjective section S where we extract directly to `ogcp:symptom-recording`.
Signs, the objective section O that are `ogcp:sign-recording` that we take as generator for `ogcp:clinical-findings`.
Actions, the analysis section A which are the ogcp:clinical-investigation-act whose outputs can be `ogcp:clinical-artifact` to investigate things that can be `ogcp:isConsequenceOf` any of `ogcp:physiological-process` or `ogcp:pathological-process`

and finally

Plan, the plan section P where the therapeutic acts can be extracted with all the timing, posology and prescriptions registered in a particular clinical encounter.

4.3 Services to annotate clinical concepts in free text

Apart from being able to provide "our own" Web Services for various tasks given the availability of downloading several types of terminologies like MeSH or SNOMED CT CORE and generally the UMLS Metathesaurus, currently there are a miriad of WS at reach that can be configured to be connected to our CP-ESB as service providers. Among those we think that are worth mentioned here the BIOPortal [4], OntoCAT[5] and UTS[6].

All these provide Web Services that offer specific tasks for Biomedicine terminology. Carefully chosen endpoints provide features that range from simple term lookups to complete semantic concept acquisition. Normally all these offerings are available at no cost upon registering and access granting.

CP-ESB The advantages of using a Software component as important as the ESB in the current SOA world of application composition is beyond the scope of this work. An important source of information to get up-to-date might be the Wikipedia page [7]. In our case the articulation of the providers and the Ontologies are well defined and are not handled point-to-point but always through the ESB routing and intercepting capabilities. All the core features are already implemented to enable plug-and-play ability for interchanged modules as stated above.

Building over the suggested infrastructure the systems are rather composed as opposed to monolithicaly built and so manifest high capabilities of plug-and-play configuration allowing for interchangeable providers (as Web Services), Reference Ontologies (Feeders), and target ontologies. The Java best-practices for pragmatic development include a number of Patterns as in JEE [8] compiled in http://java.sun.com/blueprints/corej2eepatterns/Patterns/ or the pragmatic approaches developed in such successful projects as OSGi or Spring[9].

The flowchart that depicts graphically the acquisition from the source texts in Portuguese to the creation of the appropriate OGCP instance is:

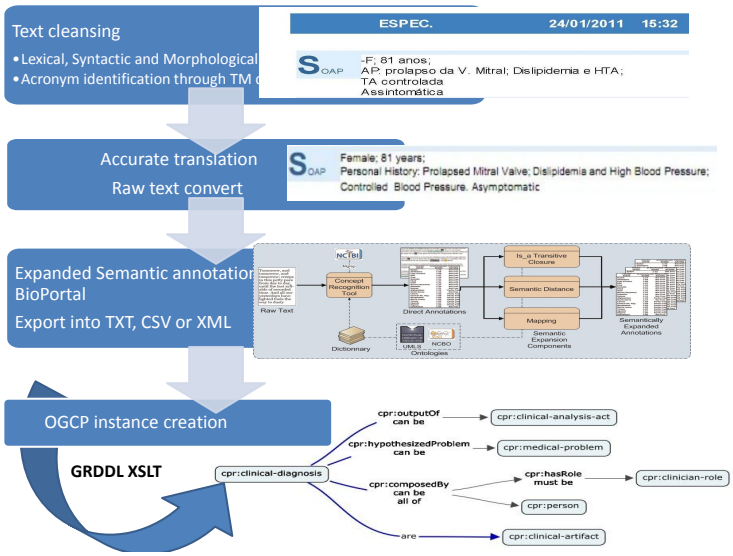


Fig. 3. Acquisition Flowchart

[4] http://bioportal.bioontology.org
[5] http://www.ontocat.org
[6] https://uts.nlm.nih.gov/home.html
[7] http://en.wikipedia.org/wiki/Enterprise_service_bus
[8] Java Enterprise Edition
[9] http://www.springsource.com/

5 Conclusions

We present our contribution for automatically developing a formal way to reason about clinical practice. So we propose an **OGMS** extension using the adequate ***ontological realism*** approaches and incorporating the **CPR** and its upper level ontologies as framework for an $\mathcal{EL}$ reasoning workhorse. We present our efforts for knowledge base population from semi-structured clinical text reports and discuss the underlying problem of automatic instance creation from them into the proposed knowledge representation structure.

Acknowledgments

The authors want to show their appreciation to Instituto de Investigação e Formação Avançada (IIFA) *for the support available under the* Bento de Jesus Caraça scholarship..

References

1. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M. & Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet, 25, 25-29, (2000) 1
2. BFO 2012 Basic Formal Ontology 2.0
URL: http://code.google.com/p/bfo. Last accessed: 18/12/2012.
3. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform, 67-79 (2008) 1
4. Brochhausen, M.; Burgun, A.; Ceusters, W.; Hasman, A.; Leong, T.; Musen, M.; Oliveira, J.; Peleg, M.; Rector, A.; Schulz, S. & others Discussion of" Biomedical Ontologies: Toward Scientific Debate" Methods of information in medicine, 50, 217, (2011) 1
5. Ceusters, W., Smith, B.; Kumar, A. and Dhaen, C.: Mistakes in medical ontologies: where do they come from and how can they be detected? Stud Health Technol Inform., (2004) 2, 3
6. CPR (2009) - Computer-based Patient Record Ontology
URL: http://code.google.com/p/cpr-ontology. Last accessed: 18/12/2012. 2
7. Ghazvinian, A.; Noy, N.; Musen, M. & others How orthogonal are the OBO Foundry ontologies J Biomed Semantics, 2, S2, 2011
8. Grenon, P.; Smith, B. & Goldberg, L. Biodynamic Ontology: Applying BFO in the Biomedical Domain Stud. Health Technol. Inform, IOS Press, 20-38, (2004)
9. Hoehndorf, R.; Dumontier, M.; Oellrich, A.; Rebholz-Schuhmann, D.; Schofield, P. & Gkoutos, G. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning PloS one, Public Library of Science,6, e22006, 2011 2
10. Hoehndorf, R.; Dumontier, M. & Gkoutos, G. V. Towards quantitative measures in applied ontology CoRR, (2012)
11. IFOMIS 2004 Basic Formal Ontology
URL: http://www.ifomis.org/bfo. Last accessed: 18/12/2012.
12. Karlsson, D.; Berzell, M. & Schulz, S. Information Models and Ontologies for Representing the Electronic Health Record ICBO, 153-157, (2011)
13. Kazakov, Y. Boutilier, C. (Ed.) Consequence-Driven Reasoning for Horn SHIQ Ontologies. IJCAI, 2040-2045, (2009) 3
14. Kazakov, Y.; Krtzsch, M. & Simanck, F. Concurrent classification of EL ontologies Proceedings of the 10th international conference on The semantic web - Volume Part I, Springer-Verlag, 305-320, (2011) 3
15. Mendes, D. & Rodrigues, I. P. Advances to Semantic Interoperability through CPR Ontology extracting from SOAP framework reports electronic Journal of Health Informatics, (2012) 2, 3
16. Mendes, D. & Rodrigues, I. P. A Semantic Web pragmatic approach to develop Clinical Ontologies, and thus Semantic Interoperability, based in HL7 v2.xml messaging Information Systems and Technologies for Enhancing Health and Social Care, IGI Global, (2012) 3
17. Noy, N. F.; Shah, N. H.; Whetzel, P. L.; Dai, B.; Dorf, M.; Griffith, N.; Jonquet, C.; Rubin, D. L.; Storey, M.-A.; Chute, C. G. & Musen, M. a. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37, W170-3, (2009)

18. Ogbuji, C. A Framework Ontology for Computer-Based Patient Record Systems Proceedings of the ICBO: International Conference on Biomedical Ontology, 217-223, (2011) 2
19. OGMS (2010) - Ontology for General Medical Science
URL: http://code.google.com/p/ogms. Last accessed: 18/12/2012. 1
20. Rector, A. L. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL Proceedings of the 2nd international conference on Knowledge capture, ACM, 121-128, (2003) 2
21. RO (2012) - Relation Ontology
URL: http://obofoundry.org/ro . Last accessed: 18/12/2012.
22. Sarkar, I. & others Biomedical informatics and translational medicine J Transl Med, 8, 22, (2010)
23. Simancik, F.; Kazakov, Y. & Horrocks, I. Walsh, T. (Ed.) Consequence-Based Reasoning beyond Horn Ontologies. IJCAI, IJCAI/AAAI, 1093-1098, (2011) 3
24. Scheuermann, R. H.; Ceusters, W. & Smith, B. Toward an Ontological Treatment of Disease and Diagnosis 2009 AMIA Summit on Translational Bioinformatics, 116-120, (2009) 3
25. Smith, B.; Ceusters, W.; Klagges, B.; Köhler, J.; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A. L. & Rosse, C. Relations in biomedical ontologies. Genome Biol, Institute for Formal Ontology and Medical Information Science, Saarland University, 6, R46, (2005) 2
26. Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L. J.; Eilbeck, K.; Ireland, A.; Mungall, C. J.; Leontis, N.; Rocca-Serra, P.; Ruttenberg, A.; Sansone, S.-A.; Scheuermann, R. H.; Shah, N.; Whetzel, P. L. & Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration Nature biotechnology, 25, 1251-5, (2007) 3
27. Smith, B. & Brochhausen, M. Putting Biomedical Ontologies to Work Methods Inf Med, 49, (2010) 3
28. Smith, B., Ceusters, W.: Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied ontology, 5, 139-188, (2010) 1, 2
29. Wong, W.; Liu, W. & Bennamoun, M. Ontology learning from text: A look back and into the future ACM Comput. Surv., ACM, 44, 20:1-20:36, (2012) 2, 3, 4

Afetação de Unidades Térmicas Considerando as Emissões Poluentes

R. Laia[1], H.M.I. Pousinho[1], R. Melício[1],V.M.F. Mendes[2]

[1] University of Évora, Évora, Portugal, rui.j.laia@gmail.com , hpousinho@gmail.com ruimelicio@uevora.pt
[2] Instituto Superior of Engenharia de Lisboa, Lisbon, Portugal, vfmendes@deea.isel.pt

Resumo. Neste artigo é proposta uma metodologia de otimização que considera não só os benefícios económicos, mas também as emissões antropogénicas para resolver do problema de afetação de unidades térmicas em empresas produtoras de energia tomadoras de preços com poucas unidades. A metodologia é suportada por um problema de programação matemática biobjetivo. A aplicação computacional desenvolvida para o problema recorre ao uso do método das somas ponderadas, da programação dinâmica e da programação não linear. É apresentado um caso de estudo para demonstrar a eficácia da aplicação.

Palavras-Chave: Afetação de unidades térmicas, unidades térmicas, emissões, programação dinâmica, soluções ótimas-Pareto.

1 Introdução

Políticas ambiciosas e medidas concretas foram propostas para encorajar a mitigação das emissões antropogénicas dos gases com efeito de estufa (GEE) e assegurar uma sustentabilidade ambiental global. Uma parte significativa das emissões de GEE na atmosfera resulta da actividade de centrais térmicas. A gestão racional destas centrais em mercado liberalizado competitivo força as empresas produtoras de energia (GENCO- Generation Companies) a otimizarem a afetação de unidades térmicas.

Com a evolução tecnológica dos recursos computacionais foi possível modelizar com maior detalhe o problema da afetação de unidades, permitindo incluir novas restrições não apenas relacionadas com custos operacionais e limites técnicos, mas também, entre outras restrições, a de reserva girante, de tempos mínimos de operação e de paragem, de taxas de variação de potência e de limites de potência. Ainda acresce que em ambiente de mercado da energia elétrica, novo paradigma, existem diversos mecanismos, como por exemplo, o contrato bilateral, adequado às empresas produtoras de energia, que tem de ser modelizado. O contrato bilateral é usado para limitar a volatilidade dos preços, sendo estabelecido entre um produtor de energia e um consumidor, designando o preço acordado e a quantidade de energia que será fornecida ao longo de estádios futuros de entrega [1]. Pelo que, é necessário modelizar os mecanismos de mercado e esta comunicação constitui uma contribuição no que respeita a empresas tomadoras de preços com poucas unidades térmicas cuja existência advém do paradigma vigente.

2 Estado da Arte

No espólio de literatura sobre o estado da arte são relatados vários trabalhos de investigação que recorrem a diversas técnicas de otimização para resolver o problema de afetação de unidades e o de despacho económico.

A lista de prioridades foi um dos primeiros métodos usados para resolver o problema de afetação de unidades [2], devido à sua fácil implementação e necessidade de poucos recursos computacionais. No entanto, este método não garante uma solução ótima global, podendo conduzir a maiores custos com a operação das unidades [2].

No âmbito dos métodos clássicos, a programação dinâmica (PD) foi usada para resolver o problema, permitindo obter uma solução ótima global. Contudo, a "maldição da

dimensionalidade", inerente à PD, obriga a uma excessiva quantidade de memória e de tempo para processamento o que é uma limitação relevante. Embora a relaxação lagrangeana [3] ultrapasse esta limitação, nem sempre está assegurada uma solução exequível. Pelo que, existe a necessidade de recorrer ao uso de metodologias com base em processos heurísticos. Os métodos ditos de inteligência artificial baseados em algoritmos genéticos [4], algoritmos evolucionários [5] e "simulating annealing" [6] têm como maior limitação a baixa probabilidade de obter uma solução próxima do ótimo global, em particular quando o número de unidades térmicas é reduzido. Pelo que, existe a necessidade de investigar métodos apropriados ao problema, visto que, é relevante para as GENCOs tomadoras de preços com poucas unidades.

Na literatura e no âmbito das preocupações ambientais são abordados principalmente problemas de despacho económico [7], determinando apenas a potência de cada unidade. No domínio da afetação de unidades térmicas cumpre determinar em cada hora do horizonte temporal: i) as unidades térmicas em serviço; ii) a potência elétrica de cada unidade em serviço. Tradicionalmente, o problema de afetação de unidades só considera a minimização do objetivo económico consumo de combustível, mas atualmente têm de considerar quer consumo de combustível, quer o nível de emissão poluente resultante [8], atendendo às políticas ambientais. As consequências da afetação de unidades são significativas, já que uma melhor operação permite obter não apenas uma redução no consumo de combustível [9], mas vai ao encontro de preocupações ambientais. A problemática das emissões na afetação de unidades térmicas [10] não foi ainda tão aprofundada como para o caso do despacho económico. Pelo que, constitui um tema em aberto para o qual ainda são esperadas contribuições.

3 Formulação do Problema

O problema de afetação de unidades térmicas em estudo determina a sequência de funcionamento de cada unidade térmica i em cada período t num número de períodos T constituintes do horizonte temporal, otimizando critérios de desempenho envolvendo custos, emissões e parcelas relacionadas com a modelização de objetivos provenientes do ambiente de mercado, satisfazendo a um conjunto de restrições quer técnicas quer económicas que são impostas na operação das unidades.

3.1 Função Objetivo

A formulação inicial para o problema em estudo considera os objetivos conflituantes associados respetivamente ao custo total de combustível requerido na operação das unidades térmicas e ao nível das emissões poluentes, sendo o vetor biobjetivo dado por:

$$\{\sum_{t=1}^{T}\sum_{i=1}^{I} C_{it}(u_{it}, p_{it}), \sum_{t=1}^{T}\sum_{i=1}^{I} \omega E_{it}(u_{it}, p_{it})\}. \quad \textbf{(1)}$$

I é o número de unidades térmicas; T é o número de períodos no horizonte temporal; $C_{it}(u_{it}, p_{it})$ e $E_i(u_{it}, p_i)$ são respetivamente as funções que determinam o custo de combustível e o nível de emissões; u_{it} e p_{it} são respectivamente as variáveis de decisão referentes ao estado da unidade (on/off) e à potência elétrica. Sendo expectável que o custo de combustível e as emissões sejam objetivos conflituantes [11], i.e., é expectável não ser possível obter uma única solução ótima que simultaneamente minimize os dois objectivos. Um dos métodos usados para obter o conjunto das melhores soluções de compromisso, conhecidas como ótimas de Pareto, para problemas de programação matemática biobjetivo, recorre à soma ponderada das funções objetivo, dada por:

$$(1-\lambda)\sum_{t=1}^{T}\sum_{i=1}^{I} C_{it}(u_{it}, p_{it}) + \lambda\sum_{t=1}^{T}\sum_{i=1}^{I} \omega E_{it}(u_{it}, p_{it}). \quad \textbf{(2)}$$

ω é o preço associado à penalização das emissões e λ o coeficiente de ponderação que determina a combinação convexa em (2) entre o custo de combustível e o das emissões, sendo $0 \le \lambda \le 1$. A GENCO pode vender energia por licitação no mercado e ou por contrato bilateral. Pelo que, um termo que reflita esta venda deve ser considerado na função objetivo (2). A função objetivo é então dada por:

$$(1-\lambda)\sum_{t=1}^{T}\sum_{i=1}^{I} C_{it}(u_{it}, p_{it}) + \lambda\sum_{t=1}^{T}\sum_{i=1}^{I} \omega E_{it}(u_{it}, p_{it}) - \sum_{t=1}^{T}\sum_{i=1}^{I} \pi_t(p_{it} - d_t). \quad \textbf{(3)}$$

π_t é o preço da energia elétrica no período t e d_t é a energia contratada com acordos bilaterais que deve ser fornecida no período t. Esta função objetivo permite admitir como possível adquirir energia no mercado caso a produção não seja suficiente para satisfazer os acordos. A função objetivo (3) pode ser interpretada como a aplicação do método das somas ponderadas ao vetor biobjetivo dado por:

$$\{\sum_{t=1}^{T}\sum_{i=1}^{I} C_{it}(u_{it}, p_{it}) - \pi_t(p_{it} - d_t), \sum_{t=1}^{T}\sum_{i=1}^{I} \omega E_{it}(u_{it}, p_{it}) - \pi_t(p_{it} - d_t)\}. \quad \textbf{(4)}$$

As funções que determinam o custo de combustível e as emissões de cada unidade térmica em funcionamento dependem da potência elétrica dessa unidade e podem ser modelizadas como funções quadráticas, dadas respetivamente por:

$$C_{itop}(p_{it}) = a_{it} + b_{it}p_{it} + 1/2c_{it}p_{it}^{2}. \quad \textbf{(5)}$$

$$E_{itop}(p_{it}) = \alpha_{it} + \beta_{it}p_{it} + 1/2\gamma_{it}p_{it}^{2}. \quad \textbf{(6)}$$

a_i, b_i, c_i e α_i, β_i, γ_i são, respetivamente, os coeficientes de custo e das emissões da unidade térmica i. O custo associado com o combustível requerido no arranque das unidades térmicas depende do número de períodos x_t em que a unidade esteve parada antes da colocação em serviço. Este custo de arranque é dado por:

$$SU_{it} = uccool_{i0}(1 - e^{-\frac{5\,x_t}{utcool_i}}). \quad \textbf{(7)}$$

$uccool_{i0}$ é o custo de arranque a frio e $utcool_i$ é a constante de tempo de arrefecimento.

3.2 Restrições

O problema de otimização está sujeito a um conjunto de restrições devido a condições de operação, por exemplo:

a) Disponibilidade de operação de unidades térmicas:

$$\sum_{i \in I} u_{it} \le NMV_t. \quad \textbf{(8)}$$

b) Restrição do balanço de potência:

$$\sum_{i \in I} p_{it} \ge D_t. \quad \textbf{(9)}$$

c) Restrição da reserva girante:

$$\sum_{i \in J \subset I} \overline{p}_{it} \ge D_t + R_t. \quad \textbf{(10)}$$

d) Restrições de taxa de variação da potência:

$$DR_i \le p_{it+1} - p_{it} \le UR_i. \quad \textbf{(11)}$$

e) Restrições de capacidade do gerador:

$$u_{it}\underline{p}_{it} \le p_{it} \le u_{it}\overline{p}_{it}. \quad \textbf{(12)}$$

f) Tempo mínimo de serviço:

A unidade está em serviço no mínimo durante esse tempo antes ser parada.

g) Tempo mínimo de paragem:

A unidade está parada no mínimo durante esse tempo antes de entrar em serviço.

h) Restrição de satisfação de acordo bilateral no período *t*:

$$\sum_{i\in I} p_{it} - d_t \ge 0. \quad \textbf{(13)}$$

É possível considerar mais restrições que em geral implicam maior tempo de processamento. Por exemplo, é possível haver acordos bilaterais com diferente flexibilidade, i.e., em apenas alguns períodos é obrigatório satisfazer a energia contratada, sendo que em outros períodos poderá haver a possibilidade da não satisfação por cláusula contratual a modelizar.

4 Caso de Estudo

O caso de estudo consiste numa afetação de três unidades, tendo um horizonte temporal de 168 horas. Os dados usados para a simulação estão disponíveis em [12]. A simulação foi realizada num computador de 2 GB de memória RAM e processador de 1.9 GHz. A linguagem de programação escolhida VBA teve em consideração o fato de ter elevada disponibilização e custo nulo para os utentes do Microsoft Excel. Foram utilizados métodos de PD e programação não linear para a determinação das soluções exequíveis. As unidades são codificadas por um autómato finito, cujo diagrama de transição da Máquina de estados para uma unidade com 4 horas de tempo de mínimo de serviço, estados {1, 2, 3, 4}, e de paragem, estados {-1,-2,-3,-4} é apresentado na Fig. 1.

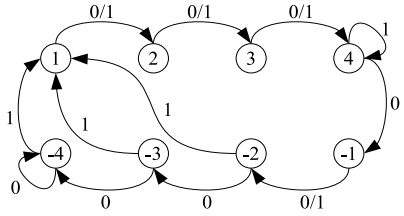


Fig. 1. Diagrama de transição da Máquina de Estados.

Com base nos resultados obtidos na simulação nomeadamente os estados, o número total de estados é de 1120, e as potências elétricas de cada unidade em cada hora ao longo do horizonte temporal foi efetuado um processamento extra para determinar grandezas globais que são relevantes para o suporte à gestão da afetação das unidades.

A energia contratada e a fornecida ao longo do horizonte temporal para alguns valores de λ é apresentada na Tabela 1.

Tabela 1. Energia contratada e fornecida

λ	Custo total (Eur)	Energia contratada (MWh)	Energia fornecida (MWh)
0.0	4117780	119096	138084
0.4	3923691	119096	119135
1.0	3408946	119096	119096

A maior diferença entre a energia fornecida e a contratada é obtida para $\lambda = 0$, visto que, o custos associado com as emissões não é considerado. As potências elétricas por cada unidade respetivamente para $\lambda = 0$ e $\lambda = 1$ são apresentadas nas Fig. 2 e Fig. 3.

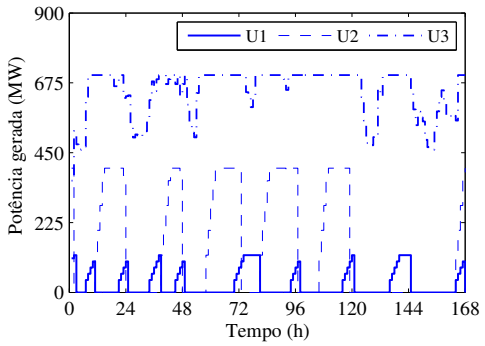


Fig. 2. Potência elétrica por unidade, $\lambda = 0$.

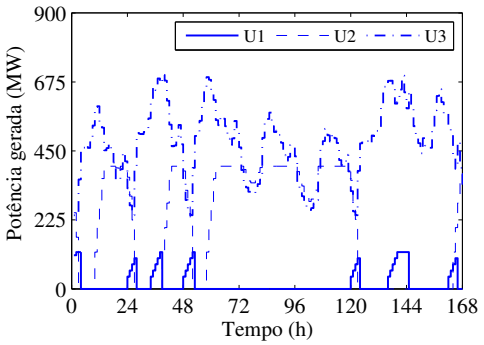


Fig. 3. Potência elétrica por unidade, $\lambda = 1$.

Na Fig. 2, são mostradas as sequências de operações para cada unidade térmica, que apresentam diferenças devido aos custos de arranque. Ambas as unidades U1 e U2 estão paradas quando o preço da energia é baixo, enquanto que U3 está em serviço para evitar o custo de arranque caso a unidade fosse parada. A comparação da Fig. 2 com a Fig. 3 mostra que a potência elétrica para cada unidade é diferente: a afetação das unidades térmicas U2 e U3 são diferentes devido às funções de custos de combustível e das emissões. A unidade U1 entra ao serviço para menores valores de potência contratada tanto para $\lambda = 0$ como para $\lambda = 1$. A curva de compromisso entre o custo total de combustível requerido na conversão e o custo das emissões é apresentada na Fig. 4.

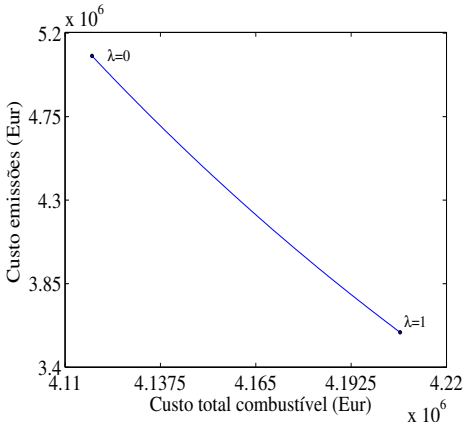


Fig. 4. Frente de Pareto.

A Fig. 4 ilustra que as funções objetivo são conflituantes devido à necessidade de compromisso entre unidades mais económicas, mas com maiores níveis de emissões, e unidades menos económicas, mas com menores níveis de emissões. As primeiras são mais favoráveis quando a consideração da função de custo de combustíveis e mais relevante na solução de compromisso. Contrariamente, às segundas que são mais favoráveis quando a consideração das emissões é mais relevante.

6 Conclusões

No âmbito de ambiente de mercado de energia elétrica é descrita uma metodologia baseada em otimização biobjetivo para a afetação de unidades térmicas, permitindo o suporte às decisões de gestão em pequenos GENCO's que são tomadores de preços. A modelização descrita considera não só os benefícios económicos que derivam da venda de energia elétrica, mas também as emissões antropogénicas associadas com os combustíveis fosseis requeridos. Na modelização são consideradas restrições de tempos mínimos de serviço e de paragem codificadas por um autómato finito. A resolução da afetação de unidades térmicas é resolvida com recurso ao uso do método das somas ponderadas, à PD e à programação não linear. Os resultados numéricos demonstram que a

metodologia proposta é adequada, sendo ilustrada a sequência de operações de afetação das unidades e a curva de compromisso que permite suportar a tomada de decisão de redução das emissões tendo como contrapartida um aumento no custo de combustíveis requeridos na operação. A aplicação desenvolvida oferece ao decisor um suporte para tomar opções em ambiente de mercado no contexto de decisão sobre contratos bilaterais e licitação, utilizando a linguagem de programação VBA Microsoft Excel, que é de elevada disponibilização.

Referências

1. Heredia, F.J., Rider, M.J., Corchero, C.: A stochastic programming model for the optimal electricity market bid problem with bilateral contracts for thermal and combined cycle units. Annals of Oper. Res., 193, pp. 107--127 (2012)
2. Senjyu, T., Shimabukuro, K., Uezato, K., Funabashi, T.: A fast technique for unit commitment problem by extended priority list. IEEE Trans. Power Syst., 18, pp. 882--888 (2003)
3. Zhuang, F., Galiana, F.D.: Towards a more rigorous and practical unit commitment by Lagrangian relaxation. IEEE Trans. Power Apparatus Syst., 102, pp. 1218--1225 (1983)
4. Kazarlis, S.A., Bakirtzis, A.G., Petridis, V.: A genetic algorithm solution to the unit commitment problem. IEEE Trans. Power Syst., 11, pp. 83--92 (1996)
5. Dhillon, J.S., Kothari, D.P.: Economic-emission load dispatch using binary successive approximation-based evolutionary search. IET Gener. Trans. Distrib., 3, pp. 1--16 (2009)
6. Wong, S.Y.W.: An enhanced simulated annealing approach to unit commitment. Int. J. Electr. Power Energy Syst., 20, pp. 359--368 (1998)
7. Abido, M.A.: Multiobjective particle swarm optimization for environmental/economic dispatch problem. Elect. Power Syst. Res., 79, pp. 1105--1113 (2009)
8. Chandrasekaran, K., Hemamalini, S., Simon, S.P., Padhy, N.P.: Thermal unit commitment using binary/real coded artificial bee colony algorithm. Electr. Power Syst. Res, 84, pp. 109--119 (2012)
9. Wood, A.J., Wollenberg. B.F.: Power Generation, Operation and Control. New York: Wiley, (1996)
10. Gjengedal, T.: Emission constrained unit-commitment (ECUC). IEEE Trans. Energy Convers., 11, pp. 132--138 (1996)
11. Kumar, N., Parmar, K.P.S., Dahiya, S.: Optimal solution of combined economic emission load dispatch using genetic algorithm. Int. J. Computer Appl., 48, pp. 15--20 (2012)
12. Laia RJR. Afectação de unidades térmicas considerando as emissões poluentes, Master Thesis (in Portuguese); Lisbon, Portugal: ISEL/Area Dep. Eng. Sist. Pot. Autom. (2011).

A Simulation for Acceptance of Two-level Converters in Wind Energy Systems

M. Seixas[1,2], R. Melício[1], V.M.F. Mendes[2]

[1] University of Évora, Évora, Portugal, mafalda.seixas@gmail.com, ruimelicio@uevora.pt
[2] Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal, vfmendes@deea.isel.pt

Abstract. This paper focuses on the use of the computational tool Matlab/Simulink to simulate an energy conversion between a variable frequency into and a constant frequency system using a two-level power converter. The simulation of this converter shows an acceptable behavior, justifying the use of a two-level power converter in wind energy systems.

Keywords: Variable speed wind energy conversion, two-level power converter, modeling and simulation.

1 Introduction

The demand for energy, the shortage of fossil fuels and the need for carbon footprint reduction have resulted in a global awareness of the importance of energy savings and energy efficiency [1] and programs on the Demand-side Management have been developed in order to assist consumers on energy usage. Also, renewable energy sources coming from wind and solar energy sources are attractive to go into exploitation, considering not only large scale systems, but also micro and mini scale conversion systems [2], Disperse Generation (DG), that can be owned by consumer. But electric energy coming from renewable energy sources typically due to the electric frequency does not meet the requirement to be directly injected into the electric energy system. Power electronic converters conveniently assist on the conversion of this electric energy into feasible one in order to inject into the electric energy system.

The exploitation of DG renewable energy source cannot disregard the economic benefit coming from a convenient choice of the power electronic convert. The reduced complexity and cost of two-level power converters relatively to multi-level power converters are significant advantages in favor of two-level power converters [3]. Therefore a focus on research and supporting simulation of more economical hardware implementation for two-level power converter use in DG renewable energy source is a highlight item.

This paper is concerned with mathematical modeling and simulation of AC/DC/AC two-level power converter, converting the energy of a variable frequency source in injected energy into the grid with constant frequency. Pulse modulation by space vector modulation associated with sliding mode is used for controlling the converter. This study relies on use of the computational mathematical tools Matlab/Simulink, which allows a friendly simulation for capturing the behavior of the system.

2 Modeling

The two-level power converter scheme considering the conversion of electric energy between variable frequency and constant frequency energy sources is shown in Fig. 1.

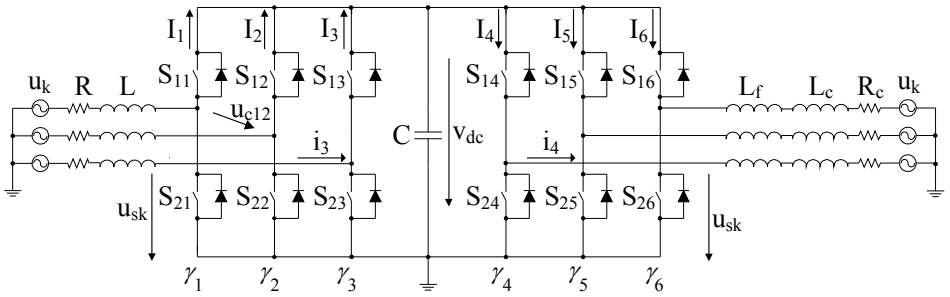


Fig. 1. Two-level power converter scheme.

Fig.1 shows from the left to the right: the electric generator describing the electric part AC of the DG renewable energy source, with the output connected to the rectifier; the capacitor bank in the CC part connected between the rectifier and the inverter; the series inductor filter connected between the inverter and the electric grid; the electric grid is modeled by an equivalent electric generator.

The state equation for the generator output current, applying Kirchhoff laws to the circuit, is given by:

$$\frac{di_k}{dt} = \frac{1}{L}(u_k - u_{sk} - R\ i_k) \qquad k \in \{1,2,3\}\ . \tag{1}$$

L and R are respectively the generator inductance and resistance, u_k is the electromotive force, i_k is the generator output current, and u_{sk} is the generator output voltage or the input voltage for the converter. The two-level converter is an AC/DC/AC converter, the index i with $i \in \{1,2\}$ identifies the upper and the down IGBT's in Fig. 1. The groups of two IGBT's linked to the same phase constitute a leg k of the converter. The AC/DC/AC converter has six unidirectional commanded IGBT's, S_{ik}, used as a rectifier, and has the same number of IGBT's used as an inverter. The following assumptions are considered for the modeling of the two-level converter: i) the IGBT's are ideal and unidirectional, and never subject to inverse voltage, due to the arrangement of anti-parallel diodes connection; ii) the diodes are ideal: in conduction, the voltage between terminals is null, and in blockage, the current is null; iii) the continuous voltage in the output of the rectifier is $v_{dc} > 0$; iv) each leg k of the converter has always one IGBT in conduction [4]. The index k with $k \in \{1,2,3\}$ identifies the leg for the rectifier and $k \in \{4,5,6\}$ identifies the leg for the inverter. The switching strategy on the two-level converter for the k leg must ensure that the switches S_{ik} are always in complementary states [5]. The switching variable γ_k is used to identify the state of the upper IGBTof the leg k, i.e., the valid conditions for the switching variable of each leg k [4] is given by:

$$\gamma_k = \begin{cases} 1, & (S_{1k} = 1\ and\ S_{2k} = 0) \\ 0, & (S_{2k} = 1\ and\ S_{1k} = 0) \end{cases} \qquad k \in \{1, \ldots, 6\}\ . \tag{2}$$

Each switching variable can be viewed as logic value on the conduction of the upper IGBT of the leg. The satisfaction of assumptions iv) is ensured by a restriction [4] on the leg k given by:

$$\sum_{i=1}^{2} S_{ik} = 1 \qquad k \in \{1, \ldots, 6\}\ . \tag{3}$$

Neglecting switch delays, on-state semiconductor voltage drops, auxiliary networks, supposing small dead times [5] and assuming a balanced three phase electrical generator, the leg output voltage u_{sk} as a function of γ_k is given by:

$$u_{sk} = \frac{1}{3}\left(2\gamma_k - \sum_{\substack{j=1 \\ j\neq k}}^{3}\gamma_j\right) v_{dc} \qquad k \in \{1,2,3\}. \tag{4}$$

The current on the capacitor bank is given by:

$$i_{dc} = \sum_{k=1}^{3}\gamma_k i_k - \sum_{k=4}^{6}\gamma_k i_k \,. \tag{5}$$

The voltage v_{dc} is modeled by the state equation [4] given by:

$$\frac{dv_{dc}}{dt} = \frac{1}{C}\left(\sum_{k=1}^{3}\gamma_k i_k - \sum_{k=4}^{6}\gamma_k i_k\right). \tag{6}$$

The output voltage for the inverter is given by (4) but with $k \in \{4,5,6\}$. The state equation of the injected current in the electrical grid is given by:

$$\frac{di_k}{dt} = \frac{1}{L_f + L_c}(u_{sk} - u_k - R_c\, i_k) \qquad k = \{4,5,6\}. \tag{7}$$

L_f is the filter inductance, L_c and R_c are respectively the electrical grid inductance and resistance, u_k is the voltage on the delivery point of the electrical grid, u_{sk} is the converter voltage and i_k is the converter output current, injected in the electrical grid.

3 Control Method

Pulse modulation by space vector modulation associated with sliding mode is used for controlling the converters. The sliding mode control is successful used, due to known characteristics of robustness, system order reduction, and propensity to control variable structure systems such as switching power converters [5]. The sliding mode control is important for controlling the converters, by guaranteeing the choice of the most appropriate space vector. Their aim is to let the system slide along a predefined sliding surface by changing the system structure. Particularly, they cannot switch at infinite frequency. Also, for a finite value of the switching frequency, an error $e_{\alpha\beta}$ will exist between the reference value and the control value [4]. In order to guarantee that the system slides along the sliding surface $S(e_{\alpha\beta},t)$, it's necessary that the state trajectory near the surfaces verifies the stability conditions [6], given by:

$$S(e_{\alpha\beta},t)\ \frac{dS(e_{\alpha\beta},t)}{dt} < 0\,. \tag{8}$$

In practice a small error $\varepsilon > 0$ for $S(e_{\alpha\beta},t)$ is allowed. Hence, a switching strategy has to be considered. This strategy is given by:

$$-\varepsilon < S(e_{\alpha\beta},t) < +\varepsilon\,. \tag{9}$$

Strategy (9) is implemented by hysteresis comparators.

The output voltage for each converter leg has two possible values $(0, v_{dc})$; therefore the three legs of the rectifier or of the inverter have eight possible combinations for the output voltage. The output space vectors in the $\alpha\beta$ coordinates for the rectifier of the two-level converter are shown in the Fig. 2.

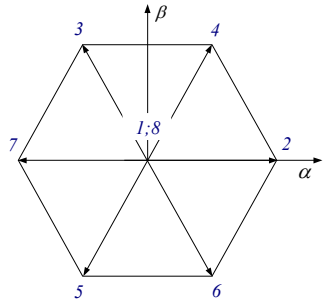


Fig. 2. Output space vectors for the rectifier of the two-level converter.

The comparison of e_α and e_β errors with the sliding surface $S(e_{\alpha\beta},t)$ enables to find at each instant the variables σ_α and σ_β, taking values -1, 0 or 1, if the error is below, within or above the sliding surface, respectively. The selection of each vector is processed in accordance with the values shown in Table 1.

Table 1. Vectors, according to the outputs of the hysteresis comparator

$\sigma_\beta \setminus \sigma_\alpha$	*-1*	*0*	*1*
-1	5	5;6	6
0	7	1;8	2
1	3	3;4	4

When the outcome has two possible vectors, it is chosen the one that introduces less variation in the states of IGBTs. The sliding mode control is a lower level of control as it is normally envisaged with the PI controller one for the rectifier and another one for the inverter.

4 Simulation

The simulation was carried out with for a variable speed wind energy generator given by voltage source data with a stair sequence of frequency ranging between 35 Hz and 65 Hz. Several case studies were performed with different time rates for the change in the frequency of the generator. Those results have a similar behavior. The results chosen to be reported were obtained by a variation over time on the frequency at every 0.1 seconds. The frequency variation used and the rectifier input voltages are respectively shown in Figure 3 and Figure 4.

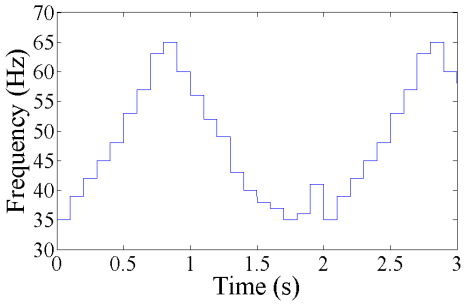


Fig. 3. Variable Frequency.

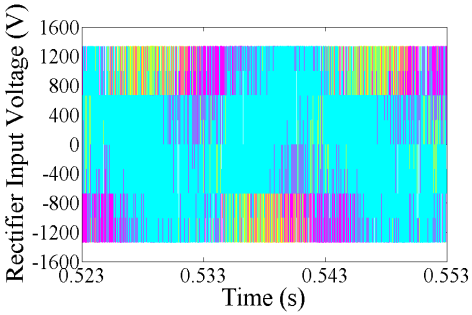


Fig. 4. Rectifier input voltages.

The capacitor bank reference and the v_{dc} voltages are shown in the Fig. 5, and the grid injected current is shown in the Fig. 6.

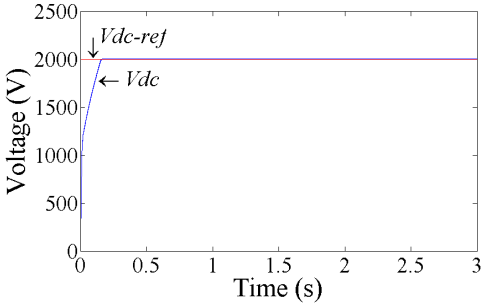


Fig. 5. Capacitor bank reference voltage and capacitor voltage.

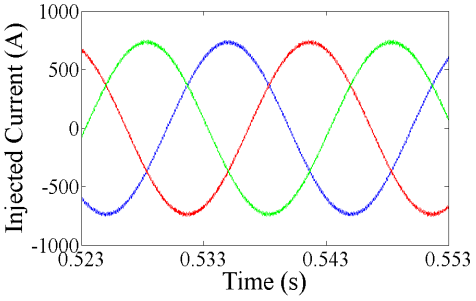


Fig. 6. Current injected into the grid.

The obtained results are consistent with what would be expected for this system.

5 Conclusion

This paper reports on the use of a computational tool to carry out a simulation which proposes a variable frequency power source system connected to an electric grid through a two-level power electronic converter.

Nowadays to investigate any real system is indispensable to use modeling computational tools able to friendly carried out with rigor a considerable processing set of computations, as for instance, to investigate beneficial arrangements or damaging impact into a system due to, although, undesirable but possible inputs. The friendly and confidence utilization for mathematical model implementation using computational tools is an important positive feature to support studies in order to obtain reliable results about the behavior of the systems, avowing undesirable performance that should be in due time prevented or supporting improvements in the system in what regards the cost of the components without deteriorating the desirable performance of the system.

A case study using Matlab/Simulink is illustrated and from the behavior given by the simulation results is admissible to support a favorable use for the applicability of two-level power electronic converter on DG renewable energy sources, leading to justified the utilization due the wellknown lower complexity, implying less failure rates, and cost relatively to multi-level power converters.

References

1. Popovic-Gerber J., Ferreira J.A.: Power electronics for sustainable energy future–quantifying the value of power electronics. In Proc. IEEE Energy Conv. Cong. And Exp. – ECCE, pp. 112–119, Atlanta, USA, (2010)
2. Fazeli M., Asher G.M., Klumpner C., Yao L., Bazargan M.: Khomfoi J.S., Tolbert L.M.: Novel integration of wind generator-energy storage systems within microgrids. IEEE Trans. on Smart Grid, vol. 3, pp. 728-737 (2012)
3. Khomfoi J.S., Tolbert L.M.: Multilevel power converters: Chapter 31, Power Electronics Handbook; Knoxville, U.S.A.: Tennessee Univ., Elsevier (2006)
4. Melício R.: Modelos dinâmicos de sistemas de conversão de energia eólica ligados à rede eléctrica: PhD Thesis, Covilhã, Portugal: UBI/FE/DEE, (in Portuguese) (2010)
5. Silva J.F., Rodrigues N., Costa J.: Space vector alpha-beta sliding mode current controllers for three-phase multilevel inverters. In Proc. IEEE 31st Annual Power Electronics Specialists Conference–PESC 2000, Galway, Ireland, (2000)
6. Melício R., Mendes V.M.F., Catalão J.P.S.: Power converter topologies for wind energy conversion systems: integrated modelling, control strategy and performance simulation. Renewable Energy, vol. 35, pp. 2165-2174 (2010)

A Wind Turbine Control Simulation

Carla Viveiros[1], R. Melicio[2], José Manuel Igreja[1], V.M.F. Mendes[1]

[1] Instituto Superior de Engenharia de Lisboa, 1959-007 Lisboa, Portugal
[2] Universidade de Évora, 7004-516 Évora, Portugal
{cviveiros, jigreja, vfmendes}@deea.isel.ipl.pt, ruimelicio@uevora.pt

Abstract. This paper deals with a wind turbine control simulation supported by the use of computational tools appropriate to simulate complex systems. High performance and reliability are required for wind turbines to be competitive within the energy market. A control design of a publicly available wind turbine benchmark model is proposed and simulations results by Matlab/Simulink are shown in order to prove the effectiveness of the design.

Keywords: Wind turbines, Matlab/Simulink, control.

1 Introduction

Evolution of technology has increased electric energy demand in order to sustain the society needs and ambient concerns with anthropogenic emissions are in nowadays a spotlight. Consequently, the exploitation of renewable energy sources became necessary, as fossil fuel resources are limited and an anthropogenic emission source. Among the renewable energy sources available today, wind power is the fastest growing one for several reasons: it is cheap, inexhaustible, clean, and climate friendly [1]. With the widespread use of wind turbines as renewable energy systems, supervision and control should be included in the system design to prevent performance degradation [2],[3]. In this paper a control simulation was performed on the benchmark model developed by [4]. The rest of the paper is organized as follows: Section 2 describes a wind turbine benchmark; Section 3 presents the mathematical modeling; Section 4 presents a case study of a wind turbine benchmark followed by simulation results; Section 5 concludes the paper.

2 Wind Turbine Description

Wind turbines are designed in such a way as to conveniently allow for electrical energy to be attained from conversion of wind kinetic energy. Wind kinetic energy is capture by the blades receiving a twist action force which causes the blades to rotate and deliver the mechanical energy to turn the speed shafts of an electric generator. The wind turbine can be analyzed on a benchmark block diagram with functional systems namely: the blade and pitch system, drive train system, generator and power converter system and the controller. A benchmark model was presented in [4] for a specific kind of wind turbine having three blade horizontal axis and equipped with a full power converter, the block diagram is show in Fig. 1.

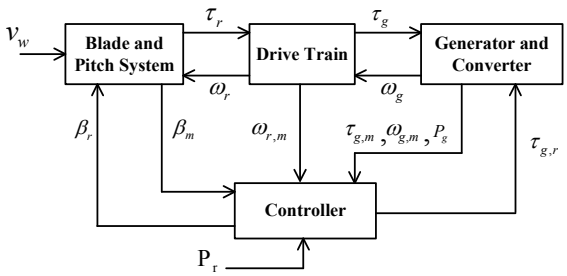


Fig. 1. Block diagram of the wind turbine benchmark [4].

Where:

$v_w\,[m/s]$	wind speed	$\tau_r\,[Nm]$	rotor torque
$\tau_g\,[Nm]$	generator torque	$\omega_r\,[rad/s]$	rotor speed
$\omega_g\,[rad/s]$	generator speed	$\beta\,[°]$	pitch angle
$P_g\,[W]$	generator power	$P_r\,[W]$	rated power

r , m subscripts designate respectively references or rotor and measurements values.

2.1 Blade and Pitch System

The blade and pitch system is the entry point to the wind turbine system. The wind provides input to the wind turbine and to the blade and pitch system. The blade and pitch system is made up of a hub and blades, together forming the rotor. The blades are designed to optimize the aerodynamic properties of the wind turbine. Furthermore the positions of the blades, also known as the pitch, corresponding to the angle formed by the blades to an imaginary horizontal axis plane, are kept perpendicular to the direction of the wind by the yaw system. Ideally, positioning the blades to the direction of the wind optimizes the electrical power converted from the wind in both low and high wind conditions. The rotor transmits the torque which serves as an input to the drive train system. The hub provides a connection from the blade and pitch system to the drive train system.

2.2 Drive Train System

The drive train system is made up of low-speed shaft, a high-speed shaft and gear system if needed. This system is used to transmit a torque and rotational force required by generator to obtain the electrical power. The principle of operation of the drive train is based on a speed transmission mechanism. First, if needed, a planetary gear system is used to boost the speed obtained by the low-speed shaft into a high speed shaft. The drive train system also contains a brake which is used to slow down the low-speed shaft in extreme high wind condition. The high-speed shaft transmits a torque and rotational speed to the generator at a level required to allow the convenient condition for the conversion into electric energy.

2.3 Generator and Power Converter System

This system utilizes the rotational energy transferred from the drive train system and converts the energy into electrical energy. The generator is used to achieve this purpose while the power converter provides the power output from the system.

2.4 Control System

The objective of the control system is to follow a reference power output, keeping the conversion at a conveniently optimal rating, but when this cannot be achieved, the controller should minimize the error. The pitch angle and the tip speed ratio, ratio between the speed of the blade tip and the wind speed, are an important value to conveniently achieve the objective of the control. The tip speed ratio is given by

$$\lambda(t) = \frac{\omega_r(t)R}{v_w(t)} \quad \textbf{(1)}$$

where R is the radius of the blades, $v_w(t)$ is the wind speed, and $\omega_r(t)$ the angular rotor speed. With a particular pitch angle, the optimal chose of the tip speed ratio allows a conversion at the maximum power permissible with that angle. One of the key challenges of the control system is due to the fact of the wind turbine is driven by the wind power which is an uncontrolled input and also acts as a disturbance. Four regions of operation of a wind turbine can be distinguished as shown in Fig. 2.

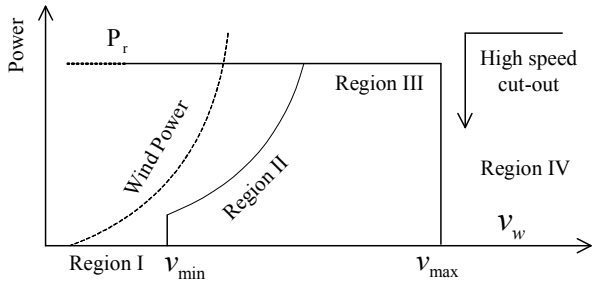


Fig. 2. Regions of power by wind speed [5],[6].

v_{min} and v_{max} are the cut-in and cut-out wind speeds, respectively. Region II corresponds to power optimization with values of wind speed such that conversion is possible at the global optimum rating in safety conditions. The control objective in this region is to capture all possible wind power with a pitch angle is equal 0 degrees, attaining global maximum power. Region III corresponds to a conversion at constant power due to the fact that the wind has more power than the one that is possible to convert, ensuring that the wind turbine works within its limits. The control objective in this region is to operate the wind turbine at the nominal power.

3 Wind Turbine Model

This section presents the mathematical model used and given in [4] for the different components of the wind turbine.

3.1 Blade and Pitch System Model

This model is a combination of the aerodynamic and pitch system model. The aerodynamics of the wind turbine is modeled in order to determine the torque acting on the blades. The aerodynamic torque is given by

$$\tau_r(t) = \sum_{j=1}^{j=3} \frac{\rho\pi R^3 C_p\left(\lambda(t), \beta_j(t)\right) v_w(t)^2}{6} \quad \textbf{(2)}$$

ρ is the air density, $C_p\left(\lambda(t),\beta_j(t)\right)$ is the power coefficient, which is a function of the pitch angle $\beta_j(t)$ (referring to blade j) and tip speed ratio. The pitch system consists of three actuators that use a hydraulic mechanism to rotate the blades in order to define different pitch angles. The pitch actuator can be modeled as a second order system with a time delay given by t_d. Hence, the pitch actuator model is given by

$$\ddot{\beta}(t) = -2\xi\omega_n(t)\dot{\beta}(t) - \omega_n^2\beta(t) + \omega_n^2\beta_r(t-t_d) \quad \textbf{(3)}$$

3.2 Drive Train Model

The drive train model consists of a low-speed shaft and a high-speed shaft having inertias J_r and J_g, and friction coefficients B_r and B_g. The shafts are interconnected by a transmission having gear ratio N_g, combined with torsion stiffness K_{dt}, and torsion damping B_{dt}. This result in a torsion angle $\theta_\Delta(t)$, and a torque applied to the generator $\tau_g(t)$, at a speed $\omega_g(t)$. The linear model for the drive train is given by:

$$J_r\dot{\omega}_r(t) = \tau_r(t) + \frac{B_{dt}}{N_g}\omega_g(t) - K_{dt}\theta_\Delta(t) - \left(B_{dt} + B_r\right)\omega_r(t) \quad \textbf{(4)}$$

$$J_g\dot{\omega}_g(t) = \frac{K_{dt}}{N_g}\theta_\Delta(t) + \frac{B_{dt}}{N_g}\omega_r(t) - \left(\frac{B_{dt}}{N_g^2} + B_g\right)\omega_g(t) - \tau_g(t) \quad \textbf{(5)}$$

$$\dot{\theta}_\Delta(t) = \omega_r(t) - \frac{1}{N_g}\omega_g(t) \quad \textbf{(6)}$$

3.3 Generator and Power Converter Model

The power converter dynamics is modeled by a first order system with time delay, t_{dg}, where α_{gc} is the inverse of the first order time constant. This model is given by:

$$\tau_g(t) = -\alpha_{gc}\tau_g(t) + \alpha_{gc}\tau_{g,r}(t-t_{dg}) \quad \textbf{(7)}$$

the power produced by the generator is given by:

$$P_g(t) = \eta_g(t)\omega_g(t)\tau_g(t) \quad \textbf{(8)}$$

where $\eta_g(t)$ denotes the efficiency of the generator.

3.4 Controller Model

The focus of the control design in this paper is on the normal operation region. Proportional integral discrete controller in Region III is given by:

$$\begin{aligned} \beta_r(k) &= \beta_r(k-1) + k_p e(k) + \left(k_i.T_s - k_p\right)e(k-1) \\ e(k) &= \omega_r(k) - \omega_{nom}(k) \end{aligned} \quad \textbf{(9)}$$

4 Simulation and Results

The simulation was performed using Matlab/Simulink environment. The wind turbine benchmark is linearized for a power set-point, P_r, of 4.8 MW and a wind speed, v_w, of 13 m/s. White noise is added to wind speed sequence in order to simulate an actual wind variation. The generator power and the reference power are shown in Fig. 3. The parameters of the wind turbine are given by: $R = 57.5$, $\rho = 1.225$, $\xi = 0.6$, $\omega_n = 11.11$, $\alpha_{gc} = 50$, $\eta_g = 0.98$.

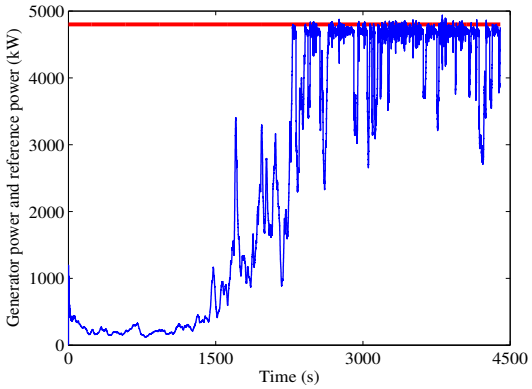


Fig. 3. Generator power and reference power.

Fig. 3 shows that the generator power follows the reference power between times 2200 s and 4500 s where the power set-point is linearized. The wind speed and the pitch angle are shown in Fig. 4.

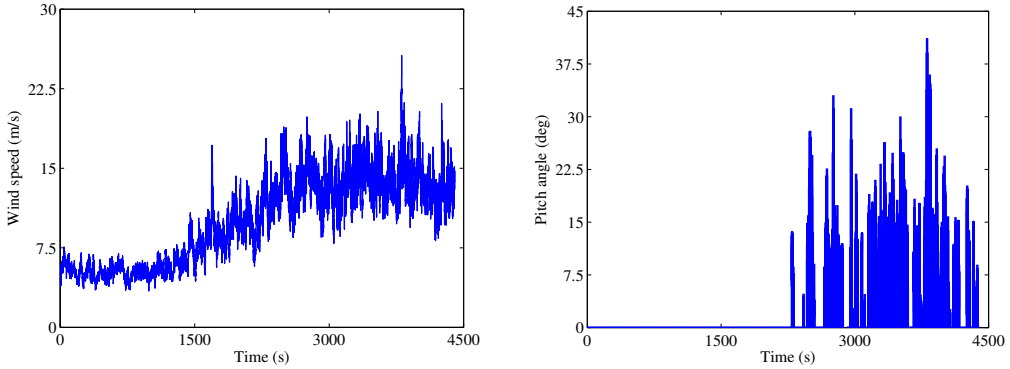


Fig. 4. Wind speed (left) and pitch angle (right).

Fig. 4 shows the wind speed and the blades adequate positions.

5 Conclusions

A wind turbine description, mathematical modeling and simulation results are presented for the control of a wind turbine in normal operation. The simulation was carried out in Matlab/Simulink which is a friendly tool, delivering information on adequacy of the control.

Further work will be held on the detection and diagnosis area, an important area in the early detection of faults and consequently preventing system failures.

References

1. Wind Energy News. Wind Energy is the World's Fastest Growing Energy Source, 2007. http://windenergynews.blogspot.com/2007/02/wind-energy-is-worlds-fastest-growing.html.
2. C. Sloth, T. Esbensen, J. Stoustrup, "Robust and Fault-Tolerant Linear Parameter-Varying Control of Wind Turbines", Mechatronics, vol 21, no. 4, pp. 645-659, 2011.
3. B. Boussaid, C. Aubrun, N. Abdelkrim, "Active Fault Tolerant approach for wind turbines", IEEE International Conference on Communications, Computing and Control Applications (CCCA), 2011 , pp. 1-6.
4. P. Odgaard, "Fault Tolerant Control of Wind Turbines - a Benchmark Model", in Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, pages 155-160, June 2009.
5. K. Johnson, L. Pao, M. Balas, and L. Fingersh, "Control of Variable-Speed Wind Turbine, Standard and Adaptive Techniques for Maximizing Energy Capture", IEEE Control Systems Magazine, vol. 26, pp. 70-81, Jun. 2006.
6. F. Bianchi, H. Battista, R. Mantz, "Wind Turbine Control Systems". Springer London, 2007.

First Look at ProbLog Implementation

Cindy Silva[1] and Salvador Abreu[2]

[1] m9293@alunos.uevora.pt
[2] spa@di.uevora.pt
Universidade de Évora

Abstract. Probabilistic logic programming languages have been a research topic for years, during which time usable interpreters and compilers have emerged for this niche paradigm. One such framework, called ProbLog, uses the YAP Prolog engine to drive pure logic queries while adding tools necessary to deal with probabilistic inference. By studying Problog's implementation, we wish to use it's theoretical concepts and apply them to GNU Prolog using constraint programming.

Keywords: Artificial Intelligence, Prolog, Probabilistic Logic

1 Introduction

ProbLog is a probabilistic logic programming language library built on top of and tightly integrated into the YAP Prolog runtime, thereby allowing probabilistic inference to be used in Prolog programs. This extension applies an algorithm which combines iterative search with binary decision diagrams (acyclic graph that represents boolean functions). This approach emerged from real biological applications that has big amounts of information about genes, proteins, tissues, organisms, etc. The biological data can be represented by graphs, where the weight of each edge is the probability of the relationship between the two nodes [3].

CLP(FD) (Constraint Logic Programming over Finite Domains) is a tool designed to solve constraints, which is translated on a set of rules that describes how the variables on the domain are related to others variables. A special case of constraints are the disjunctive contraints which can provide an improvement, in some cases.

This article intends to discover a way to built ProbLog into the Gnu Prolog engine, and so it presents the likelihood tools to built it. On section 2 we explain the ProbLog architecture and how ProbLog is processed. Section 3 shows how to represent a ProbLog program and on section 4 we explain how ProbLog solves the Binary Decision Diagrams (BDDs) and how it calculates the path probability. Finaly on section 6 we explain how CLP(FD) works and what are disjunctive constraints.

This work is supported by the FCT grant PTDC/EIA-EIA/100897/2008.

2 ProbLog Architecture

The ProbLog library is implemented as a YAP Prolog module and resides within the YAP source tree. A ProbLog program is composed by a set of facts labled with probabilities, and a set of predicates that make up the program's background knowledge. Once loaded, these facts and their associated background knowledge can be fed to the ProbLog library by calling one of its top-level predicates (for example: problog_exact/3, problog_montecarlo/3).

In [3], A. Kimmig et al. explain the module implementation: facts labeled with probabilities are expanded into Prolog statements via the term_expansion/2 mechanism which stores the facts with the logarithm of the probability.
After tranforming ProbLog into Prolog statements, proofs are translated into a binary decision diagram (BDD) by a command-line utility written in *C* that calls *SimpleCUDD* and that in turn writes the BDD. Then the BDD is processed by an external library, CUDD, whose sole purpose is to process and solve BDDs. After that the results will be retrieved by *SimpleCUDD* and then written in a temporary file that ProbLog will receive and use to present the results to the user. Figure 1 is a diagram explaining how ProbLog processes queries, starting with program compilation all the way to returning the result to the user.

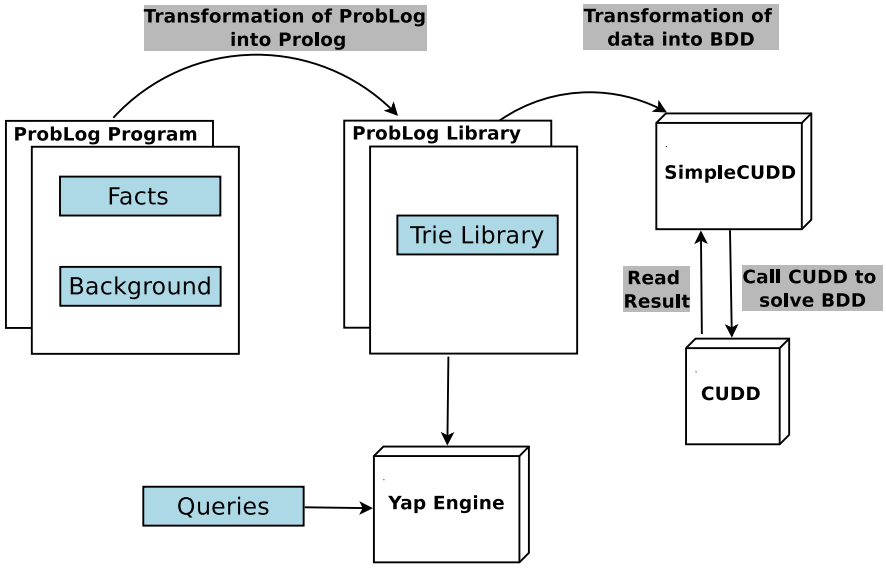


Fig. 1. ProbLog Architecture

3 Representing ProbLog

ProbLog defines the probability distribution of logic programs, by specifying the probability p_1 for each clause c_i, being mutualy independent. Consider the graph shown in figure 2, this can be easily translated in ProbLog code has shown below:

```
% Probabilistic facts
0.7::edge(a,b).
0.8::edge(a,c).
0.6::edge(b,c).
0.9::edge(c,d).
0.5::edge(e,d).
0.8::edge(c,e).

% Background knowledge
path(X,Y) :- edge(X,Y).
path(X,Y) :- edge(X,Z), path(Z,Y).
```

In a probablistic graph, the importance of the connection can be measured by the probability that a path exists between two valid nodes. These kinds of queries are easily expressed in logic, we just need the predicate *path(N1,N2)* that verifies the existence of a path between N1 and N2. Unlike *Prolog* that only cares if the query succeds or fails, *Problog* has an interest in the likelihood that a query will succeed.

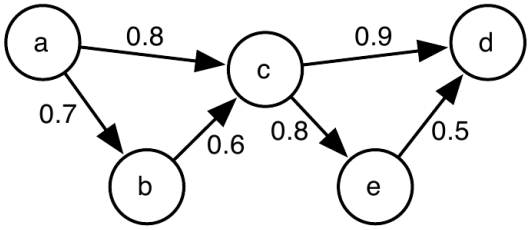


Fig. 2. Simple graph implemented in ProbLog for this test.

4 Deconstructing a ProbLog program

Consider the graph shown in figure 2. With this simple problem we intend to explain how probabilities are calculated in ProbLog, for that we will begin with a simple case and then progress to a complex case.

4.1 Edges

Imagine that we want to know the probability of the path between c and e. As you can see there is only one way to go form c to e, via edge ce. Despite being a small problem to solve, ProbLog still creates a BDD even though it has only one node.

The calculation of the probability of the path between c and e is given by:

$$Prob(path(c,e)) = edge(c,e)$$
$$Prob(path(c,e)) = 0.8$$

4.2 Path Finding

ProbLog builds a set of proofs that can be used to answer user queries. Each proof, as explained in [3], is the logical conjunction of all the *edges* in a given *path*, for example: *path(c,d)* can either be described by edge cd or by the conjunction of edges ce and ed. The union of all the possible proofs returns the probability that a given path is true.

The calculation of the path probability can be given by either one of the following equations since they are mathematically equivalent:

$$Prob(path(c,d)) = edge(c,d) + (edge(c,e) * edge(e,d) * (1.0 - edge(c,d)))$$
$$Prob(path(c,d)) = 0.9 + 0.8 * 0.5 * 0.1$$
$$Prob(path(c,d)) = 0.94$$

$$Prob(path(c,d)) = edge(c,e) * edge(e,d) + (edge(c,d) * (1.0 - edge(c,e) * edge(e,d)))$$
$$Prob(path(c,d)) = 0.8 * 0.5 + 0.9 * 0.6$$
$$Prob(path(c,d)) = 0.94$$

4.3 Eliminating Repeated Data

Now consider that we want to calculate the likelihood of *path(b,d)*. We can go from b to d by two paths, *bcd* or *bced*. Notice that both paths share edge *bc*, if we calculate the probability of *path(b,d)* we will count edge *bc* twice. In order to minimize the number of nodes when storing proofs, ProbLog will store edge *bc* only once. Figure 3 contains the BDD for the query *path(b,d)*.

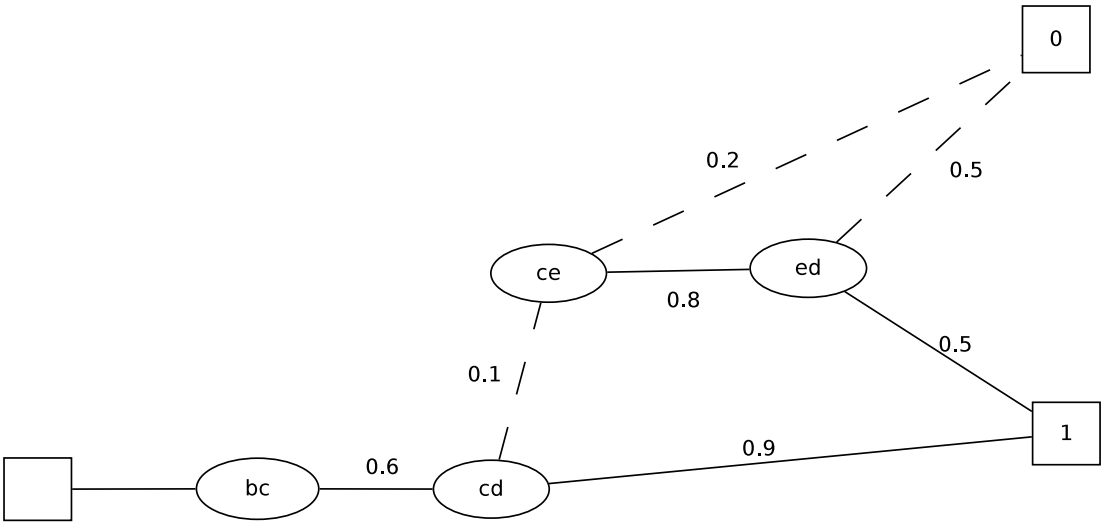


Fig. 3. BDD for path(b,d)

The probability of *path(b,d)* is given by:

$$Prob(path(b,d)) = edge(b,c) * (edge(c,d) + (edge(c,e) * edge(e,d) * (1.0 - edge(c,d))))$$
$$Prob(path(b,d)) = 0.6 * (0.9 + 0.8 * 0.5 * 0.1)$$
$$Prob(path(b,d)) = 0.564$$

5 Moving Away From CUDD

In order to avoid the use of external programs and libraries (which inevitably creates undesirable computational overhead), it is our objective to represent BDDs using *GNU Prolog's* constraint logic framework. By keeping data structures and BDD solving algorithms within *GNU Prolog's* runtime, we hope to avoid the type of interprocess comunication carried out between *ProbLog* and *CUDD*.

6 CLP(FD)

In fields such as computational logic and logic programming there is a class of problems known as *Constraint Satisfation Problems* (or CSP) whose states are limited or must obey certain restrictions. Thus *Constraint Logic Programming over Finite Domains* (or CLD(FD)) emerged as a means of tackling CSPs through the use of logic programming.

Codognet and Diaz [1] describes the implementation of a tool designed to solve constraints (also called CLP(FD)) which uses consistency and propagation techniques from CSPs. The idea behind this constraint solver is to have a single restriction X in r, where r represents an interval (for example $t1..t2$). This translates complex restrictions into other simpler ones. Each restriction consists of a set of propagation rules which describe how a variable's domain is related to all the other variables. Such as restriction can be represented by the following formula:

$$X \text{ in } r \tag{1}$$

Were X is a variable and r represents the range of values that variable can have. This range is a non-empty collection of natural numbers within a finite domain.

6.1 Disjunctive Constraints

The usual way to handle disjunctive constraints in CLP is to use the non-determinism of the underlying logical engine, which is very convenient from the programming point of view but is very inefficient because of the naive backtracking scheme of Prolog [1]. Intelligent backtracking can provide improvement in some cases however when the constraint network is strongly connected and constrained variables are all inter-linked to one-another the intelligent backtracking is useless.

The idea consists in defining a formula F from a formula $E \equiv c_1 \vee c_2 \vee \cdots \vee c_n$ so that there is no disjunction in F . Two cases are interesting:

(1) E $\Leftrightarrow$ F : the addition of F to the store suffices and no choice point is needed.
(2) E $\Rightarrow$ F : the addition of F to the store is not enough to ensure the correctness which will be then ensured by a choice point.

Simple Example

Consider the *"five houses"* problem, that involves five men living in five houses. Each man has a different profession, nacionality, favorite animal and favorite drink. The problem consists in assigning everyone and all attributes to their respective houses, and then identifying the house whose inhabitant's favourite animal is a zebra, for example. The facts are formulated as equality or inequality constraints between these variables. For instance a fact like "the Norwegian's house is next to the red one" means that it can be either on the left of the red house or on the right. This leads to a constraint of the form:

```
Norwegian = Red + 1 or Norwegian = Red - 1
```

It has to introduce a predicate plus_or_minus defined by:

```
plus_or_minus(X,Y,C):- X = Y-C.
plus_or_minus(X,Y,C):- X = Y+C.
```

Wich we can translate in CLP(FD) as:

```
plus_or_minus(X,Y,C):- X in dom(Y)-C : dom(Y)+C,
                       Y in dom(X)+C : dom(X)-C.
```

The predicate defined in clp(FD) removes the number of the red house, which means that the domain of Y doesn't have that value and will never create any choice point for the red house value. This behaviour corresponds to a constructive disjunction.

7 Conclusion

As future work we intend to implement a functional ProbLog-like program in Gnu Prolog Engine using CLP(FD) with disjunctive constraints. For that we have to find a way to represent BDDs using disjunctive constraints and without using any third-party libraries or programs such as CUDD/SimpleCUDD.

References

1. Philippe Codognet and Daniel Diaz. Compiling constraints in clp(fd). *J. Log. Program.*, 27(3):185–226, 1996.
2. Bernd Gutmann, Manfred Jaeger, and Luc De Raedt. Extending problog with continuous distributions. In *ILP*, pages 76–91, 2010.
3. Angelika Kimmig, Bart Demoen, Luc De Raedt, Vítor Santos Costa, and Ricardo Rocha. On the implementation of the probabilistic logic programming language problog. *TPLP*, 11(2-3):235–262, 2011.
4. Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *IJCAI*, pages 2462–2467, 2007.

Quadro interativo de baixo custo com interação através de dispositivos móveis

Rui Rebocho[1], Vitor Beires Nogueira[1], and António Eduardo Dias[2]

[1] Universidade de Évora, Portugal
ruirrebocho@gmail.com
vbn@di.uevora.pt
[2] Universidade Nova de Lisboa, Portugal
aed.fct@gmail.com

Abstract. Com a criação do Plano Tecnológico da Educação o Ministério da Educação Português tinha como objetivo dotar as escolas públicas de Internet de alta velocidade, dois alunos por computador, um videoprojector por sala e um quadro interativo por cada três salas. Todo este avultado investimento tinha como objetivo dotar as escolas públicas de equipamentos de última geração e que permitissem acompanhar os avanços tecnológicos na área da educação. No entanto, a rentabilização dos recursos disponibilizados foi um pouco descurada. Assim sendo, foram gastos alguns milhares de euros em equipamentos que continuam a ser subaproveitados e para os quais os educadores necessitavam de formação adequada que não veio a ser disponibilizada. A solução apresentada neste trabalho combina um quadro interativo de baixo custo proposto por Johnny Lee, com a utilização do Smoothboard (software minimalista para trabalhar com o quadro interativo proposto) e a sua integração com a plataforma Moodle (Learning Management System utilizado em muitas instituições de ensino). Este sistema possibilita uma interação quase universal com os equipamentos que os alunos diariamente transportam consigo (computadores portáteis, telemóveis e tablets).

1 Introdução e Motivação

Os Quadros Interativos Multimédia (QIM), são dispositivos que ligados a um computador e a um projetor de vídeo, permitem conceber uma outra forma de trabalhar no ensino os mais variados conteúdos. Podem ser também importantes ferramentas para a formação, apresentações, conferências, seminários, etc.

A informação e as aplicações que estão no computador, passam para o QIM através da ligação com o projetor, podendo ser trabalhadas e manipuladas no quadro, através de uma "caneta" (ou outro dispositivo) que funciona como um rato, possibilitando executar as aplicações, acrescentar notas/informações, fazer remoções, aceder à Internet, etc. Tudo o que acontece no QIM pode em seguida ser guardado, impresso, distribuído para os alunos através de email ou para uma página da Internet.

Através do Plano Tecnológico da Educação (PTE), o Ministério da Educação (ME) equipou as escolas com 5613 novos QIM. Estes equipamentos ficaram aquém do objetivo de 12.363 que previa em cada escola, 1/3 das salas equipadas com um QIM. Antes de efetuar este elevado investimento, inicialmente previsto de 9.000.000 de euros, mil euros por cada quadro interativo [1], o ME poderia ter feito uma pesquisa mais cuidadosa das opções disponíveis e quem sabe com menos gastos, equipar todas as salas de aulas com um QIM.

A proposta apresentada por Johnny Chung Lee [2], amplamente divulgada pela Internet, é uma das hipóteses que poderia ter sido considerada, devido aos seus baixos custos de implementação, aproximadamente 42 euros. Johnny Lee sugere a criação de um quadro interativo com a utilização do Wii Remote (comando da consola da Nintendo Wii) e de uma caneta que consiga emitir um sinal infravermelho. Colocando o Wii Remote a apontar para uma área de projeção ou monitor e "escrevendo" com a caneta nessa área temos um quadro interativo. Como o Wii Remote consegue controlar até quatro pontos podemos ter até quatro canetas.

Para este sistema funcionar teremos que ligar o comando ao computador por Bluetooth e em seguida utilizar um software de comunicação entre o comando e o computador. Depois da proposta

apresentado por Johnny Lee surgiram diversos softwares apresentados para este sistema: softwares personalizáveis pelo utilizador, estáticos, livres, comerciais, etc. A nossa escolha recaiu sobre o Smoothboard de Boon Jin [3] programa que desde 2009, data da sua primeira versão, tem sofrido constantes atualizações e que como podemos ver na sua nova versão, Smoothboard Air, tem a possibilidade de acesso ao sistema através dos vários dispositivos móveis disponíveis no mercado. As suas ferramentas e a possibilidade de personalização das mesmas tiveram grande peso na decisão.O registo do software tem o valor de $\$49,99$ o que acrescenta à nossa solução aproximadamente o valor de 38 euros, ou seja, a solução apresentada teria um custo total aproximado de 80 euros, 8% do custo previsto para a aquisição dos QIM pelo ME.

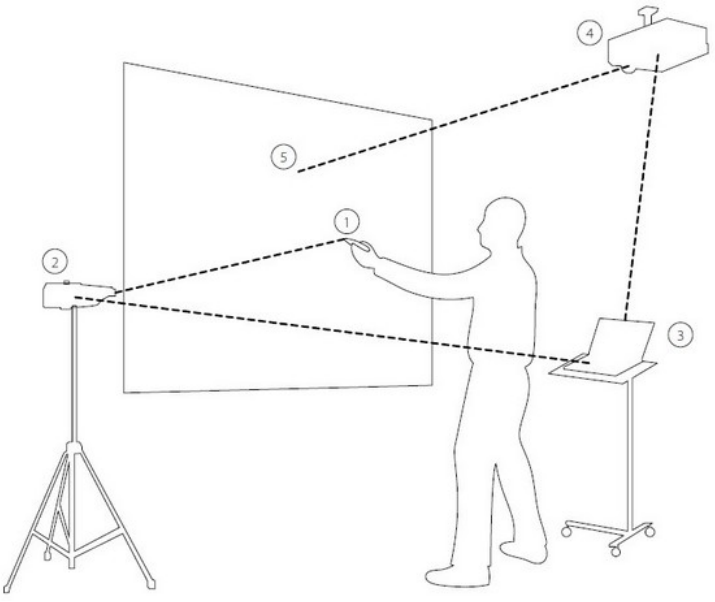
Para permitir uma maior interação por parte dos alunos com o QIM foi escolhido o Learning Management System (LMS) Moodle, uma vez que consideramos ser de fácil acesso para alunos e professores que já o utilizam nas suas atividades letivas. Assim sendo, foi criada uma aplicação que adicionada ao Smoothboard permite ao professor disponibilizar no Moodle os conteúdos lecionados na aula através do quadro interativo. É igualmente possível solicitar a participação dos alunos na aula através do Moodle com questões, inquéritos, etc.

Ao longo do trabalho apresentado neste artigo pretende-se dar a conhecer como criar um QIM de baixo custo, como personalizar a aplicação Smoothboard para a adequar a cada utilizador e ainda como foi criada a ligação entre este software e o Moodle.

2 Quadro Interativo

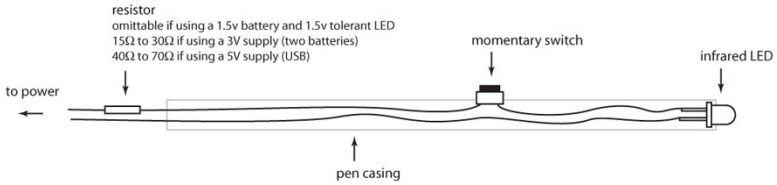
Um quadro interativo é uma superfície que pode reconhecer a escrita electronicamente e que necessita de um computador e de um videoprojector para funcionar, alguns deles permitindo interação com a imagem do computador projetada.

Fig. 1. Funcionamento do Wii Remote Whiteboard: (1) Caneta de infravermelhos; (2) Wii Remote; (3) Computador; (4) Videoprojector; (5) Superfície de projeção.(adaptado de Clinik)



O QIM proposto por Johnny Lee (Figura 1) [4], torna qualquer superfície de projeção num quadro interativo recorrendo à utilização de um videoprojector para projetar a imagem do computador e o Wii Remote e uma caneta de infravermelhos para controlar o computador.

Fig. 2. Esquema para construção de uma caneta de infravermelhos. (Autor Johnny Lee)

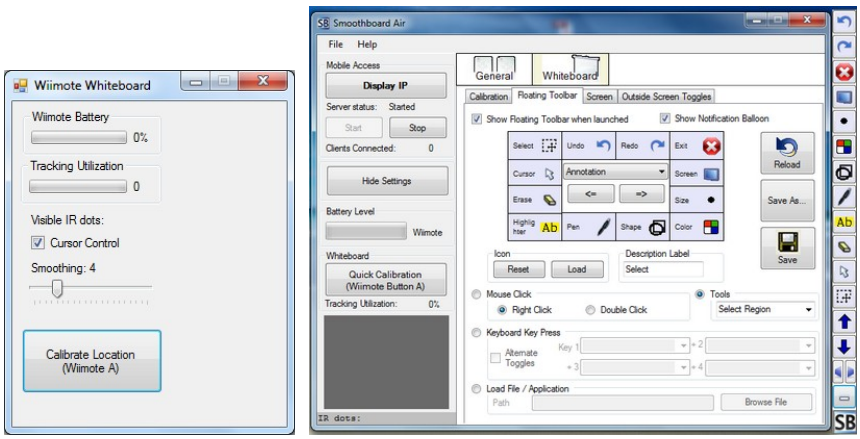


Para criar uma caneta de infravermelhos é necessário um LED[3] de infravermelhos, uma pilha, um botão que permita ativar e desativar o LED. Na Figura 2 podemos ver o esquema para uma caneta de infravermelhos proposto por Johnny Lee, aquando da apresentação do Wii Remote Whiteboard.

A escolha do LED é muito importante para o correto funcionamento do quadro interativo com recurso ao Wii Remote. Testes realizados por Julien Delmas [5] permitiram-lhe concluir que é possível utilizar a caneta a uma distância de mais de quatro metros sem problemas desde que se utilize a alimentação correta (pilha de $1,5V$), um LED com um comprimento de onda o mais próximo de $1000nm$ e um ângulo de propagação igual ou inferior a 30^o para que o Wii Remote obtenha a melhor qualidade de receção do sinal.

Da experiência adquirida, pelos testes feitos e por informações recolhidas [6], permite-nos considerar que a melhor localização do Wii Remote será na parte superior do videoprojector se este se encontrar junto do teto. Esta localização permite que o utilizador não se tenha que preocupar com a sua posição relativamente ao Wii Remote e elimina a busca da melhor posição do Wii Remote para que se possa obter uma boa área de calibração.

Em termos de software de configuração, vamos restringir a nossa descrição aos dois que consideramos mais relevantes (Figura 3: o Wiimote Whiteboard (versão 0.3) de Johnny Chung Lee e o Smoothboard Air with Duo de Boon Jin.



(a) Wiimote (b) Smoothboard Air with Duo

Fig. 3. Software para o Whiteboard

[3] Light Emitting Diodes/ Diodo Emissor de Luz

O Wiimote Whiteboard (Figura 3(a)) é bastante simplista em termos de funcionalidades: apresenta só a opção de calibração do QIM, de suavização dos movimentos e mostra ao utilizador o estado da bateria do seu Wii Remote e o *tracking utilization*[4]. O Smoothboard (Figura 3(b)) é o único que implementa o botão do lado direito do rato, suporta a utilização de vários Wii Remotes em simultâneo, suporte à utilização do PowerPoint. Mais, a versão denominada por Smoothboard Air permite colaborar com o QIM através de um dispositivo móvel, smarthphone, tablet ou qualquer outro dispositivo desde que possua um browser compatível com HTML5 e estejam ambos ligados na mesma rede. Adicionalmente, entre outras funcionalidades, disponibiliza informações sobre o estado da bateria e localização/área de calibração reconhecida pelo Wii Remote, denominado *tracking utilization*. Outra das suas grandes vantagens é a personalização da barra de ferramentas consoante as preferências do utilizador.

Clicando na opção "Mostrar configurações" podemos encontrar entre outras opções a opção de configuração da Barra Flutuante. Nesta área podemos personalizar o ícone, a etiqueta de descrição e em seguida o tipo de ação que queremos que o botão faça. As funções disponíveis são os botões do rato (direito e esquerdo), várias ferramentas (cursor, caneta, destacar, apagar, apagar tudo, sair do modo de anotações, *shape*, cor, tela de ferramentas, quadro branco, desfazer, refazer, selecionar uma área), pressionar tecla(s) do teclado e executar um ficheiro ou aplicação. Terminada a personalização da barra de ferramentas poderemos guardar essas alterações utilizando o comando *guardar* ou *guardar como*. As barras criadas ficam disponíveis para futuras utilizações ou para partilhar com outros utilizadores.

3 Integração entre o Smoothboard e o Moodle

Por forma a guardar o resultado do trabalho desenvolvido na aula, sessão ou apresentação com recurso ao QIM, sugere-se que o professor ou utilizador vá guardando o seu trabalho com recurso à opção do Smoothboard, captura de ecrã (*print screen*). No final da aula o utilizador pode disponibilizar toda a informação recolhida através de um script que foi criado para o efeito e adicionado ao Smoothboard através da personalização da barra flutuante do mesmo. O script tem como responsabilidade criar um ficheiro pdf com toda a informação recolhida. Em seguida apresenta ao utilizador uma autenticação à plataforma Moodle, de modo que o documento seja automaticamente enviado para a área pessoal do utilizador. Este mais tarde poderá disponibilizar esses documentos aos seus alunos.

A conversão das imagens em pdf fica a cargo do programa JPEGtoPDF de Jesse Yeager [7], escrito em VB .Net. Suporta a conversão de múltiplas imagens num único ficheiro pdf ou em vários, o redimensionamento das imagens, a alteração do seu posicionamento e os formatos de imagem: BMP, GIF, PNG, TIF, WMF, EMF, para além de JPG, JP2, J2K. A sua utilização neste script é totalmente transparente para o utilizador. O documento pdf é gerado pelo programa com o nome data e hora atual. O código utilizado faz a chamada do programa com os parâmetros, nome do pdf resultante e a pasta onde se encontram as imagens, exemplo:

```
jpegtopdf.exe "01-09-20121021.pdf" "C:\Boon Jin\Smoothboard\Snapshots\*"
```

Quando o script termina as suas tarefas encaminha o utilizador para uma autenticação no Moodle, de forma a permitir o envio do ficheiro gerado anteriormente para a sua pasta privada no Moodle. Esta ação é possível com o recurso aos WebServices(WS) do Moodle, que dão acesso a diversas mensagens de interação com o servidor. Essa troca de mensagens permite a gestão de utilizadores e cursos através de instrumentos externos, dando assim a possibilidade de acesso de outras ferramentas ao sistema e possibilitando a expansão e integração do Moodle com outras aplicações.

Com o abandono na versão 2.0 do WS `moodle_file_upload` [8] seria necessário recorrer a outro WS ou recorrer aos métodos alternativos disponibilizados, como é o caso do recurso à função PHP cURL para o envio do ficheiro. Foi utilizada como ponto de partida a sugestão apresentada na

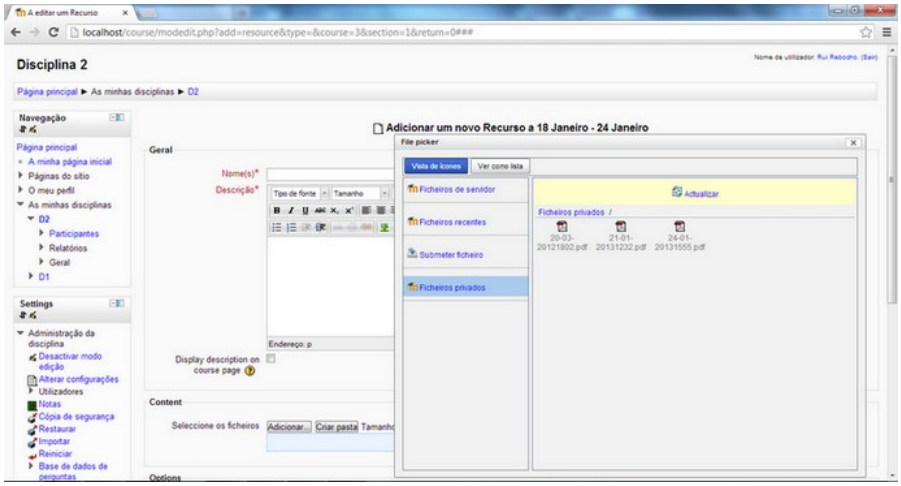
[4] Tracking utilization é o rácio entre a área do ecrã a calibrar e o campo de visão do Wii Remote.

área de desenvolvimentos dos manuais do Moodle, disponibilizados no seu site oficial, sobre a temática *Web Services file handling*. Esta implementação denominada `PHP_File_handling` por Jérôme Mouneyrac no site Github [9], necessita de um token válido na plataforma para conseguir utilizar o envio do ficheiro. Para gerar o token foi utilizado o WS disponibilizado no Moodle, `moodle_mobile_app`, que para além de outros serviços disponibiliza a criação de tokens para os utilizadores que pretendam interagir com o Moodle. Esta implementação utiliza um método HTTP POST para fazer o upload do ficheiro. Se os ficheiros forem enviados com sucesso estes ficarão disponíveis na área privada do utilizador e a informação enviada no formato JSON[5] a confirmar, em caso contrário, é enviada uma mensagem JSON a informar a falha.

```
$params = array('file_box' => "@".$localfilepath,
                'filepath' => $serverfilepath,
                'token' => $token);
$ch = curl_init();
curl_setopt($ch, CURLOPT_HEADER, 0);
curl_setopt($ch, CURLOPT_VERBOSE, 0);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_USERAGENT, "Mozilla/4.0 (compatible;)");
curl_setopt($ch, CURLOPT_URL, $domainname . '/webservice/upload.php');
curl_setopt($ch, CURLOPT_POST, true);
curl_setopt($ch, CURLOPT_POSTFIELDS, $params);
$response = curl_exec($ch);
```

O código apresentado é o responsável pelo envio do ficheiro para o Moodle, utiliza a função curl_setopt(resource ch, string option, mixed value) para definir as opções de uma sessão CURL identificada pelo parâmetro ch. O parâmetro option é a opção que se quer definir, e o value é o valor da opção dada por option. Todas as opções definidas são as necessárias para que o CURL consiga enviar o ficheiro. De referir que anteriormente tiveram que ser definidos o caminho onde o ficheiro se encontra localmente, o local onde irá ser armazenado no servidor e o token para que se consiga autenticar na plataforma Moodle.

Fig. 4. Partilha do recurso criado - Moodle



[5] JSON (JavaScript Object Notation) é uma estrutura de dados leve, utilizada para troca de informações e de fácil leitura e escrita por humanos.

Concluídos todos estes processos o utilizador ficará com a sua sessão disponível no Moodle, em seguida poderá partilhá-la com outros utilizadores. Na Figura 4 é apresentada uma forma de partilha do material criado na sessão através da inserção de um recurso na sua disciplina.

4 Conclusões e Trabalho Futuro

Para contornar o preço elevado dos quadros interativos existentes no mercado, aliada à necessidade de acompanhar a evolução tecnológica registada com o surgimento de novos métodos pedagógicos a solução *low cost* do quadro interativo Wii Remote, permite a implementação de um QIM de bom desempenho e que disponibiliza a maior parte das funcionalidades de um QIM tradicional. O seu sistema aberto permite todo o tipo de personalização necessária à sua adoção como um bom meio para promover a utilização da informática como dinamizador da aquisição e do desenvolvimento do conhecimento. A ligação com o Moodle também beneficia este sistema porque permite ao utilizador interligar duas ferramentas que atualmente já utiliza. O Moodle também permite que vários tipos de equipamentos acedam à informação e participem nas sessões com recurso ao QIM.

Do ponto de vista técnico, o posicionamento e a estabilidade do Wii Remote, aliada à qualidade do LED e à respetiva carga de alimentação da caneta, assumem-se como únicos fatores críticos para o sucesso do quadro interativo Wii Remote.

Na implementação do modelo proposto constataram-se exatamente estes problemas. A utilização de um LED de infravermelhos vulgar, igual ao utilizado nos comandos de televisão, permitiu-nos perceber que a caneta não conseguia realizar um traço contínuo. Os traços mais longos transformavam-se em tracejados. A utilização de um LED Vishay Tsal 6400 com um comprimento de onda de $940nm$ permitiu ultrapassar este problema do traço. De referir que, como este LED necessita de uma alimentação de $1,5V$, foi utilizada como fonte de alimentação uma pilha LR6 AA. No que à posição do Wii Remote diz respeito, chegou-se à conclusão que esta poderá variar. Sendo que, a que menos problemas de calibração oferece, é a colocação do Wii Remote junto ao projetor de vídeo, de preferência junto do teto.

Ultrapassados estes constrangimentos técnicos com recurso às soluções apresentadas anteriormente, tentar colocar o Wii Remote junto do videoprojector no teto, uma boa escolha de LED e da sua alimentação, podemos concluir que se trata de uma boa solução para a qualidade/preço que apresenta. E que a sua interligação com o Moodle, ferramenta que muitas escolas portuguesas utilizam torna ainda melhor esta solução. Outra das vantagens é a simplicidade de utilização do software Smoothboard, bastante intuitivo e personalizável às necessidades do utilizador. Devido à sua simplicidade, não necessidade da frequência de formação específica que é necessária para utilização dos QIM tradicionais e dos seus softwares, um diferente para cada tipo de quadro.

O trabalho futuro deste projeto passa essencialmente pelo aperfeiçoamento do revestimento da caneta e na divulgação junto da comunidade educativa, principalmente dos docentes, desta solução. O objetivo é permitir que mais pessoas conheçam este sistema e comecem a perceber as vantagens da utilização dos QIM e do Moodle nas suas atividades letivas. Os docentes e as escolas poderão assim, preparar-se para o futuro da educação e das suas práticas e os alunos poderão assistir às aulas, num ambiente mais motivante, interagir de outra forma com o professor, anotar e realçar tópicos do material apresentado.

References

1. Gabinete de Estatística e Planeamento da Educação, M.d.E.: Kit tecnológico. estudo de implementação. (2009)
2. : Johnny chung lee. `http://johnnylee.net/` Accessed: 24/01/2013.
3. : Smoothboard. `http://www.smoothboard.net/` Accessed: 24/01/2013.
4. : clinik. `http://clinik.net/wiimote/` Accessed: 24/01/2013.
5. : Julien delmas. `http://www.prtice.info/?Placement` Accessed: 24/01/2013.
6. Silva, F.V.d., Torres, J.M.: Avaliação da utilização em sala de aula de um quadro digital interactivo baseado no wiimote. (2009)

7. : Jpegtopdf. `http://www.compulsivecode.com/Project_ImageToPDF.aspx` Accessed: 24/01/2013.
8. : Web services files handling. `http://docs.moodle.org/dev/Web_services_files_handling` Accessed: 24/01/2013.
9. : Php-http-filehandling. `https://github.com/moodlehq/sample-ws-clients/tree/master/PHP-HTTP-filehandling` Accessed: 24/01/2013.

Serious Games: a First Look

José Duarte and Luís Rato

Universidade de Évora
m10401@alunos.uevora.pt, lmr@uevora.pt

Abstract. This article surveys serious games, aiming a better understanding of the concept. First, a historical review of games for learning is made. Then, the motivational elements behind this kind of games are explored. The inherent problems with the development is analyzed and discussed. Finally, the use of artificial intelligence algorithms in order to create adaptive serious games is examined.

Keywords: Serious Games, Learning, Motivation, Edutainment

1 Introduction

Games are part of our culture for a long time, and not just for entertaining purposes. Many researchers point out that playing is an important mean of socialization and learning mechanism for humans and for many species in the wild. Lions, dogs, cats and many other animals learn how to hunt by models and by playing[22].

No one likes to lose. So to play a game, you need to define strategies, manage resources, memorizing things, think better, learn rules and do that fast and accurate in order to win[11]. Because playing can be so rewarding, you can do that for many hours without rest. Ben Sawyer, co-founder of the *Serious Games Initiative* and *Games for Help Project*, defends that it is possible to learn by playing any game[19].

Many studies were made in order to understand how this features, present in games, could be used for serious purposes. This is how the concept of Serious Games was born. In this survey, we provide an overview of the concept, the motivational elements, the benefits and the difficulties associated to it. After that, we will debate what the future can bring in terms of adaptive serious games.

2 State of the Art

2.1 Historical review

The importance of games in the society is ancient. Games involve interactions and mental (or physical) challenges. Strategic, tactical and logical skills are regularly trained and tested by playing.

For example, the Mancala is maybe the oldest strategy game. It is a board game with probably more than 7000 years [18]. Its origin is uncertain, but the popularity of this game in Africa and Asia can be compared to the chess in the occidental world [10].

One of the main caracteristics of a game is "the fun". Thus, this motivational aspect is a key feature of any kind of game. In the 60s, many studies [1, 3, 6, 13] were made in order to determine how motivation affects learning and behavior.

It is in 1985 that a very detailed study about games for learning is published[5]. It was a study made for the National Council of Teachers of Mathematics in two different states of the United States. The objective was to determine how, and if, games could help students to learn math. The performance of more than 1600 students was evaluated following for three years in a row. Two variables were studied, the *instruction methods* and the *cognitive domain*[4]). The instructional methods were, *before*, *during* and *after* the subject been taught in classes. The cognitive categories were, *Knowledge*, *Comprehension*, *Application* and *Analysis*. The results in Table 1 shows us a

Table 1. Results of the effectiveness of mathematical instructional game (adapted from[5])

Cognitive Domain	Intructional method		
	After	During	Before
Knowledge	Positive	Positive/Neutral	Very positive
Comprehension	Positive	Neutral	Very positive
Application	Positive	Positive	Neutral
Analysis	Positive	Positive	-

clear effectiveness of instructional games in the learning of math.

Maybe inspired by this results, many started to study the influence of games in learning. In late 80s a new concept emerges, the Edutainment[21]. This name results form the words Education and Entertainment. The concept earned followers quickly. Many believed that edutainment could be "the savior of education because of their ability to simultaneously entertain and educate"[7]. The main target of this tools, that educates through entertainment, were preschool and young children[15]. Cute animals or TV characters were frequently used to provide fun while lessons of reading, math or science were presented. However, many edutainment projects failed, the games were tedious, not funny, and boring[22]. Many critics [16, 22, 23] pointed out problems like, poorly designed games, involving repetition and memorization, (ie. Drill-and-Practice learning) as the reason for that.

In 2002, America's Army is released, the official United States Army game. It is financed by the United States government and distributed for free. The game has been developed as a recruiting tool [15]. The teenager, or other civilian, have a chance to "play soldier". Intelligence, first aid, survival, marksman and leadership are just a few of the many skills trained and tested during this virtual recruiting simulation. Still in the same year, and maybe because of the huge success of America's Army, Woodrow Wilson Center for International Scholar in Washington D.C. had founded the *Serious Games Initiative* (http://www.seriousgames.org/), and the concept "Serious Games" became widespread [21].

2.2 What is a Serious Game?

Serious Games main objectives are to train, to inform, but above all, to provide learning by doing[11]. *Serious* is not the game, is the *message*. The game must be fun and engaging for the player. If the player is motivated to play, then it is ready to assimilate the message [15]. Paula Rego, et al.[17] defines Serious Games as "games that engage the user, and contribute to the achievement of a defined purpose other than pure entertainment (whether or not the user is consciously aware of it)". In our opinion the keyword is *engage*. But how to manage that? The answer is simple, you need to keep the player motivated.

2.3 Motivational Elements

There are two kinds of motivation, the intrinsic and the extrinsic. In the first one the person is self-motivated (internal) to do something, the second one is when a person does something just to get a (external) reward[6]. The objective of transmitting a serious message trough a game is to stimulate the intrinsic motivation. Converting "have to" in "want to" is not an easy task[11].

Taxonomy of Intrinsic Motivation
In 1987, Thomas Malone and Mark Lepper published "Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning"[14] as the result of many years of research. This taxonomy unveil the mechanisms of the endogenous motivation and guidelines for the design of intrinsically motivating environments.The taxonomy is divided in two groups, the individual motivations and the interpersonal ones. The first group is subdivided in *challenge*, *curiosity*, *control* and *fantasy*.

The second one in *cooperation*, *competition* and *social recognition*. Understanding how to apply this heuristic is a key element to create a appealing serious game [2, 5, 7, 12, 20, 22].

The Flow Theory
Another important motivational element is described in the studies of Mihaly Csikszentmihalyi, the *Flow* or *The Psychology of Optimal Experience*[8]. When one is really involved in a completely engaging process, sometimes feelings like losing track of time, not feeling hungry or tired are observed because all your attention and focus it is on the task. Somehow it is like your existence been temporarily suspended, that happens when you are in the *Flow*[9]. So, the *Flow* it is a mental state of full and deep immersion in a activity such as art, play or work. In Fig. 1 a graph of the mental state in terms of challenge ans skill levell is presented. Notice that the optimal point of experience happens for medium skill and challenge levels. On other hand, flow channel only appears for high skill and challenge levels. The concept of *Flow* has been applied in many fields as sports, teaching,

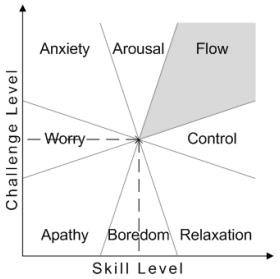


Fig. 1. The *Flow* or *The Psychology of Optimal Experience*[8]

business, leadership, among others.

A closer look at the motivational elements here described, it is possible to realize that the *Flow* is a deeper study of how the challenge of the taxonomy should be presented to the player.

3 Discussion

After a quick overview of the concept of serious games, several things need to be discussed. Is it possible to create games with so many and complex rules? The answer is yes. In fact, many of the features described in section 2.3 are currently present in many regular games (ie. games of pure entertainment). For example, looking at massive multiplayer games like Farmville, Lineage, World of Warcraft or Minecraft, all of them have many of the items described in the Taxonomy of Intrinsic Motivation as can be seen in Table 2.

Table 2. Presence of motivational elements in regular games

Individual motivations	Interpersonal motivations
Challenge - Quests, missions, goals; Curiosity - Mysterious gifts, engaging stories; Control - Player decisions can change the course of things; Fantasy - Emotion appealing, character customization;	Cooperation - shared a tasks, gifts; Competition - To finish a task first; Social recognition - Clan leaders, missions completion notifications;

On the other hand, the *Flow* it is a little harder to implement. Somehow you need to be able to evaluate the player preformance and, based on that, the game needs to adapt to the player

skills. Francesco Bellotti et al. [2] have made studies about adaptive serious games. They have implemented an experience engine module using genetic algorithms. By using artificial intelligence algorithms it is possible to automatically adjust the difficulty level or the order of the taught matters, providing a personalized experience.

4 Related work

Adaptive Experience Engine for Serious Games.
A well described article about adaptive serious games has been published in 2009. Bellotti et al. [2] have implemented an experience engine module for serious games based in Artificial Intelligence paradigms like genetic algorithms and reinforcement learning.A well described article about adaptive serious games has been published in 2009. Bellotti et al. [2] have implemented an experience engine module for a sandbox serious games based in Artificial Intelligence (AI) paradigms like genetic algorithms and reinforcement learning. In sandbox games the player can freely move along the environment, interacting with contextualized situation, building his own narrative by deciding which task, mission or quest wants to do next. Behind the scenes, the experience engine it is evaluating the player performance and, by using AI algorithms, selecting the right tasks to be present to the player.This way the player engagement is granted by showing the best challenges for the player skills (ie. the flow is granted).

Power of Research: a new online game to inspire the scientists of the future [1]
A new strategy browser game, named "Power of research" (http://www.powerofresearch.eu) was officially launched in 2011. It is a free multiplayer game where the player must assume the role of a doctor/researcher. The main objective is to inspire young Europeans to pursue scientific careers.

This game is been developed by two Austrian companies and is supported by the European Commission. The game received 617.000 Eur from this commission through the 7th Framework Programme for research.

Playing Surgery - A Laparoscopy Game for Surgeons on the Nintendo Wii[2]
In March 9, 2012, at the Google Tech Talks, a serious game for surgeons was presented. It is been developed by the University Medical Centers of Groningen and Leeuwarden and the game developer Grendel Games. The main objective is to practice basic motor skills needed to perform laparoscopies. It is visually attractive game which have an immersive storyline where the player is asked to solve exciting puzzles. In order to do that the player must use customized Nintendo Wii controls that imitate a surgeon's instruments. In contrast with the traditional simulators, where you just need to move pegs around, in this game you need to frequently do unexpected movements. The medical context has been removed from the game in order to provide extra motivation. The game has been tested by surgeons with very good results and reviews.

5 Conclusions and future work

After the survey for this paper it is possible conclude that serious games are highly motivational tools with great potential. It is a fairly new concept, still struggling for recognition and space in the scientific spectrum.
As benefits of serious games we can mention:

- High motivational;
- Easily adaptable to e/b-learning;
- Logfiles can be created for better follow up of the student/player;

[1] http://ec.europa.eu/research/index.cfm?pg=newsalert&lg=en&year=2011&na=na-230211
(Last access in 24-01-2013)
[2] http://www.youtube.com/watch?v=rpSvDvYvJGk
(Last access in 24-01-2013)

– Adjustable level of difficulty;
– Provide some kind of fun while learning.

As downsides:

– Multidisciplinary development teams are needed;
– It is hard to make good games;
– Many person are still suspicious about the effectiveness of this kind of games.

As line of thoughts and research to be done, we see the adaptability of the contents based of the player performance as good point to be study. Many Artificial intelligence concepts, as Intelligent Tutoring System, can linked to the serious games concept in order to provide a better learning experience.

The lack of research in this particular area makes it difficult but very interesting. For instance, *The First Workshop on AI for Serious Games*[3] took place in Stanford University in the end of last year.

Acknowledgements

José Duarte acknowledges Carlos Pampulim Caldeira, PhD for all the support on previous researchs. The authors want to acknowledge the reviewers for the tips and suggestions that helped to improve this paper.

References

[1] J.W. Atkinson. *An introduction to motivation.* Van Nostrand, 1964.
[2] F. Bellotti, R. Berta, A. De Gloria, and L. Primavera. Adaptive experience engine for serious games. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1(4):264–280, 2009.
[3] D.E. Berlyne et al. Curiosity and exploration. *Science (New York, NY)*, 153(731):25, 1966.
[4] B.S. Bloom, MD Engelhart, E.J. Furst, W.H. Hill, and D.R. Krathwohl. Taxonomy of educational objectives: Handbook i: Cognitive domain. *New York: David McKay*, 19:56, 1956.
[5] G.W. Bright, J.G. Harvey, and M.M. Wheeler. Learning and mathematics games. *Journal for Research in Mathematics Education. Monograph*, 1, 1985.
[6] J.S. Bruner. *Toward a theory of instruction*, volume 59. Belknap Press, 1966.
[7] D. Charsky. From edutainment to serious games: A change in the use of game characteristics. *Games and Culture*, 5(2):177–198, 2010.
[8] M. Csikszentmihalyi. *Flow: the psychology of optimal experience.* Harpercollins, 1990.
[9] M. Csikszentmihalyi. What makes a life worth living? http://www.ted.com/talks/lang/eng/mihaly_csikszentmihalyi_on_flow.html, 2004. Accessed: 14/01/2013.
[10] J. Erickson. *Sowing games.* Cambridge University Press, 1996.
[11] J. Gee, C. Aldrich, and H. Jenkins. Games in education. http://vc.ocde.us/archive/default.htm?v=tgame-1, 5 2005. Accessed: 14/01/2013.
[12] G. Gunter, R.F. Kenny, and E.H. Vick. A case for a formal design paradigm for serious games. *The Journal of the International Digital Media and Arts Association*, 3(1):93–105, 2006.
[13] J.M. Hunt. Intrinsic motivation and its role in psychological development. In *Nebraska symposium on motivation*, volume 13, pages 189–282. University of Nebraska Press Lincoln, 1965.
[14] T.W. Malone and M.R. Lepper. Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction*, 3:223–253, 1987.
[15] D.R. Michael and S. Chen. *Serious games: games that educate, train and inform.* Course Technology PTR, 2006.
[16] S. Papert. Does easy do it? children, games, and learning. *Game developer magazine*, 1998. http://www.papert.org/articles/Doeseasydoit.html.
[17] P. Rego, PM Moreira, and L.P. Reis. A survey on serious games for rehabilitation. In *5th DSIE'10 Doctoral Symposium in Informatics Engineering*, pages 267–278, 2010.

[3] https://www.cra.com/aiide/index.html
(Last access in 24-01-2013)

[18] L. Russ. *The Complete Mancala Games Book: How to Play the World's Oldest Board Games*. Marlowe & Company, 2000.

[19] B. Sawyer. Games everywhere the larger role for web platforms. http://www.youtube.com/watch?v=XPaCwjhZ2aY, 2011. Accessed: 14/01/2013.

[20] A.J. Stapleton and P.C. TAYLOR. Why videogames are cool & school sucks! In *Paper presented at the annual Australian Game Developers Conference (AGDC)*, volume 20, page 23, 2003.

[21] T. Susi, M. Johannesson, and P. Backlund. Serious games: An overview. *The American surgeon*, 73:10, 2007.

[22] R. Van Eck. Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE review*, 41(2):16, 2006.

[23] M. Warschauer and C. Meskill. *Technology and second language teaching*. Lawrence Erlbaum, 2000.

Um Serious Game para Reabilitação Cardíaca

Jorge Mota

Universidade de Évora

Resumo Os acidentes cardiovasculares são uma das principais causas de morte em todo o mundo. Quando uma pessoa sobrevive a um acidente destes, ou lhe é detetado indícios da doença, executa um plano de reabilitação cardíaca, o que nem sempre é tão bem sucedido quanto poderia ser devido nomeadamente ao custo do tratamento e à falta de meios físicos e humanos. Neste trabalho propõe-se uma solução para controlo da carga de esforço em reabilitação cardíaca, que se baseia na informação obtida dos sensores VitalJacket e Kinect, para determinar o nível de dificuldade de um *serious game*, de modo a elevar o nível de esforço do paciente a uma Zona Alvo de Treino (ZAT) prescrita clinicamente. Toda a configuração do plano de exercícios é ajustável *on-demand* e personalizável para cada paciente.

Keywords: Kinect, VitalJacket, reabilitação cardíaca, serious game

1 Introdução

Problemas cardiovasculares são uma das principais ameaças à vida do ser humano. Em Portugal, este tipo de doença é considerada a principal causa de morte [1].

Embora hoje em dia já existam métodos clínicos para tratar funcionamento do sistema cardiovascular destes pacientes, os métodos não são acessíveis a todos devido ao custo, falta de meios e elevada procura. Pelo fator risco, que neste caso é transversal a todos nós, a saúde é uma das área onde mais se tem evidenciado a utilização de tecnologia, nomeadamente sensores.

Este trabalho consiste numa implementação de um *serious game* que pretende motivar o paciente a executar determinado exercício físico com uma certa duração e esforço. O sensor utilizado neste projeto para avaliar os movimentos do paciente é o Microsoft Kinect e o sensor de ritmo cardíaco escolhido é o VitalJacket. Estes dois sensores são a base de informação que este jogo utiliza para controlar, através da dificuldade, a carga de esforço do paciente tendo tem conta também a Zona Alvo de Treino (objectivo) que foi prescrita clinicamente para o paciente. Para essa correção do esforço, em função de uma carga aplicada, é proposta uma solução matemática.

Tanto quanto é do conhecimento do autor, o sistema proposto neste trabalho é inédito.

A secção 2 apresenta os sensores utilizados. A secção 3 apresenta as diversas partes que constituem o jogo. A secção 4 detalha o modelo de controlo de carga proposto, enquanto a secção 5 conclui o trabalho.

2 Sensores

Integrar o comportamento do ritmo cardíaco do paciente no jogo, tem como meta recolher toda a informação necessária ao *software* para que seja possível o sistema perceber o estado atual do paciente e aplicar as devidas ações e correções de forma a que este atinja durante determinados momentos o ritmo cardíaco pretendido para o treino.

Para dar vida ao jogo, fazer com que o paciente se esqueça por instantes do seu problema de saúde e ao mesmo tempo incentivar o paciente a exercitar o corpo (obrigatório para alterar o ritmo cardíaco através do esforço) utilizou-se o sensor de movimentos Kinect. Como apenas o sensor de movimento não seria suficiente para o bom funcionamento do jogo, e uma vez que este é direcionado à reabilitação cardíaca, é também necessário analisar o comportamento cardíaco do paciente ao mesmo tempo que este executa determinado exercício, o que será conseguido através do VitalJacket.

2.1 Kinect

O Kinect é um produto da Microsoft criado com o contributo da Prime Sense que combina as características dos seus diferentes tipo de sensores para possibilitar às aplicações, que interagem com ela, reconhecer posições e movimentos do corpo humano em tempo real – o Kinect é normalmente apelidado de sensor de movimentos (Figura 1) [6]. Os diversos tipos de sensores que o constituem são:

- **Sensor de profundidade** – através de raios infravermelhos é obtida uma perceção de profundidade (traduzida na distância do objeto em relação ao sensor) e movimento do objeto;
- **Câmara** – devolve uma matriz de pixeis com a respetiva cor de cada um (RGB);
- **Microfone** – idealizado para principalmente reconhecer comandos de voz e possibilitar ao Kinect uma perceção mais concreta do ambiente na qual está inserida como por exemplo saber o número de pessoas a detetar (através do ruído).

Com o auxílio dos seus controladores e *software*, através da informação que dispõe permite detetar até 48 pontos que servem para representar o esqueleto humano.

Figura 1. Sensor de movimento Kinect

2.2 VitalJacket

Criado pela BioDevices, o VitalJacket é um dos primeiros produtos desenvolvidos que integram tecnologia de Monitorização Biológica no corpo humano através do conceito *wearable technology*. Este conceito significa fazer com que uma qualquer tecnologia faça parte do corpo humano sem que se dê pela sua presença (por exemplo relógios inteligentes)[2].

A BioDevices germinou com a criação do VitalJacket que foi desenvolvido na Universidade de Aveiro e aprovado no Hospital S. Sebastião. O VitalJacket, como o próprio nome indica, é um dispositivo, na forma de colete, capaz de medir e registar o comportamento cardíaco de quem o está a utilizar (Figura 2). Esta é uma alternativa viável face ao habitual eletrocardiograma, com a vantagem do VitalJacket ser menos dispendioso e permitir que sejam recolhidos os dados em pleno exercício físico sem que a pessoa tenha que estar ligada a uma máquina imóvel. Para além dessas funcionalidades o VitalJacket está ainda preparado para armazenar internamente os dados recolhidos pelos seus sensores, ou enviá-los em tempo real para outros dispositivos através de Bluetooth. A versão do dispositivo utilizado apenas possibilita a recolha da informação, de forma direta, do ritmo cardíaco.

Figura 2. VitalJacket

3 Jogo

O objetivo principal deste projeto, é ajudar o paciente durante todo o processo da sua reabilitação cardíaca, recorrendo não só ao auxílio do sensores Kinect e VitalJacket, mas também ao *software* desenvolvido. Tal como acontece nos processos de reabilitação cardíaca da medicina convencional, neste jogo cada paciente tem um programa personalizado em termos de exercícios a executar, carga horária de treino, ritmo cardíaco alvo entre outros.

Pode-se definir o ritmo cardíaco / frequência cardíaca como o número de batimentos do coração por unidade de tempo, normalmente por minuto (*bpm*). O ritmo cardíaco pode ser medido através de um eletrocardiograma, monitores cardíaco ou de forma manual.

Para permitir um controlo rigoroso da frequência cardíaca é necessário definir os três limites que serão aplicados ao paciente em determinado exercício: ZAT (Zona Alvo Treino). A ZAT Referência é o ideal a atingir e apesar de nem sempre ser equidistante da ZAT Mínima e Máxima, tem que se situar obrigatoriamente entre estas. A ZAT Mínima correspondente ao número mínimo de *bpm* que o paciente deve ter durante o exercício enquanto a ZAT Máxima corresponde ao máximo de *bpm* do paciente durante todo o exercício. As Zonas Alvo Treino são representadas por um número cuja unidade de medida é igualmente o *bpm*.

O paciente interage com o jogo através do Kinect, movimentando-se de forma a exercer exercício físico. Estão disponíveis vários tipos de jogo 2D e 3D com diferentes características (Figura 3). Quem decide que jogo o paciente irá jogar em determinada sessão, bem como as respetivas ZAT, será o profissional de saúde no Painel de Controlo do jogo, tendo em conta que jogos diferentes implicam movimentos físicos diferentes no paciente.

Toda a configuração de cada sessão de treino, é definida no Painel de Controlo do jogo. Ao paciente compete apenas fazer a autenticação no jogo, e seguir as recomendações prescritas previamente pelo profissional de saúde.

Através do VitalJacket o jogo tem conhecimento do ritmo cardíaco do paciente em tempo real. Como também são conhecidas as diferentes ZATs para determinado exercício/momento, o jogo sabe a cada instante se o paciente está ou não a atingir o ritmo cardíaco desejado. Dessa forma o jogo exigirá mais, ou menos esforço do paciente tendo em conta a sua prestação e o objetivo. A esse ajuste de dificuldade física, dá-se o nome de **carga**. Neste jogo, a carga é representada por um número decimal entre 0 (repouso absoluto) e 1 (esforço máximo). A carga modificará o estado do jogo alterando variáveis como por exemplo a velocidade, direção e distância dos objetos virtuais (que fazem parte do jogo), em relação à representação do paciente no jogo.

Figura 3. A posição do paciente no jogo 2D (imagem à esquerda) é representada pelo círculo branco. Para pontuar é necessário apanhar o objeto cor-de-rosa e escapar ao verde. Quando o paciente alcança o objeto cor-de-rosa, é instanciado um novo objeto no mapa. A representação gráfica do jogo equivale a uma vista aérea de uma sala na qual o paciente se move. No jogo 3D (imagem à direita) o paciente tem que intersetar e escapar aos objetos (dependendo do tipo), podendo utilizar os por exemplo os braços. Nos diversos modos de jogo, um aumento de carga reflete-se na velocidade, número e direção dos objetos.

3.1 Painel de controlo

Para permitir a criação de novos perfis (representação para o sistema de novos pacientes) e editar a parametrização de exercícios de cada um , criou-se um género de painel de controlo implementado em .Net (Figura 4). Foi escolhida esta tecnologia entre outras pelo sua excelente integração com o servidor de base de dados escolhido e sistema operativo alvo: Microsoft Windows.

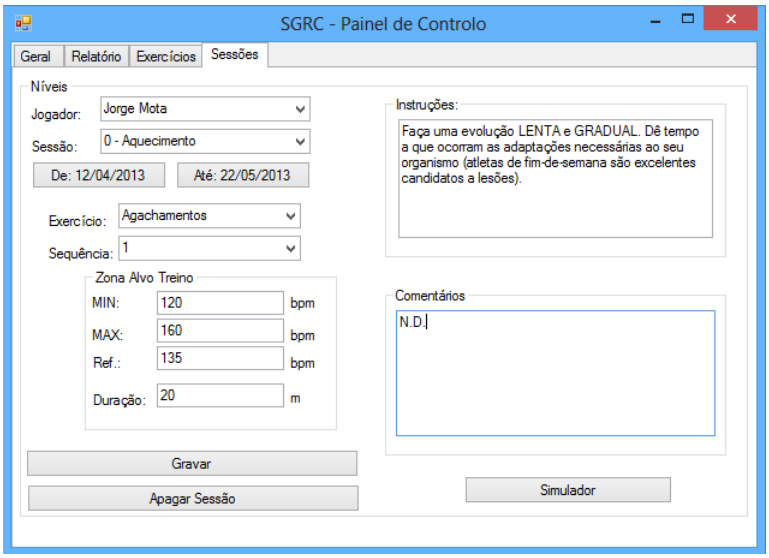


Figura 4. Painel de controlo especificando parte do exercício do paciente

3.2 Armazenamento de dados

Para armazenar todos os dados resultantes da parametrização do painel de controlo, registar os acessos ao jogo bem como registos efetuados durante o próprio jogo, foi criada uma instância de servidor Microsoft SQL Server e nela a respetiva base de dados (Figura 5). O servidor SQL Server encontra-se configurado pelo protocolo TCP/IP, o que neste caso permite uma ligação remota entre o jogo, painel de controlo e SQL Server. No mesmo servidor, encontra-se uma instância de SSRS (SQL Server Reporting Services) que é utilizada para suportar todos os relatórios criados (através de *queries* à base de dados e fórmulas matemáticas para tratamento dos respetivos dados) [3].

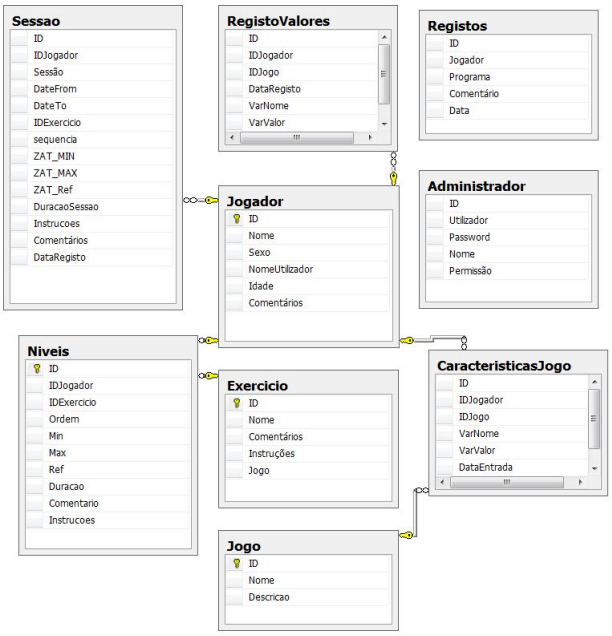


Figura 5. Diagrama da base de dados

3.3 Motor de jogo

O motor de jogo utilizado para desenvolver este trabalho é o Unity 3D (*C++*) e funciona como o *front-end* da aplicação. Este motor de jogo é o produto principal da empresa Unity Technologies e é particularmente forte na reconstituição de acontecimentos regidos pelas leis da física [4].

Este motor suporta, de forma nativa, as principais plataformas da atualidade: iOS, Android, Linux, Xbox, Windows. Outro fator que o distingue dos motores de jogo concorrentes é a facilidade de integração com outras *frameworks Open Source* como por exemplo o OpenNI. A versão do Unity utilizada é gratuita, e ao compilar o jogo permite ao utilizador configurar o seu desempenho gráfico, possibilitando dessa forma otimizar o jogo tendo em conta os recursos disponíveis na máquina [4].

A integração da Kinect no Unity foi feita não através do Microsoft Kinect SDK mas sim pelo *plugin* Zigfu em conjunto com o *framework* OpenNI.

O Zigfu funciona como um pacote de *drivers* que possibilita integrar a Kinect na maioria dos sistemas operativos ao mesmo tempo que simplifica a comunicação entre as aplicações e o sensor de movimentos, o que não aconteceria entre o Kinect SDK e o Unity.

Insistindo na mesma ideia seguida pela escolha do Unity como motor de jogo e do Zigfu como *plugin* de comunicação com a Kinect, seguiu-se a escolha natural do OpenNI, pois esta *framework* funciona com vários sistemas operativos. O OpenNI desempenha a função de abstrair o programador dos dados que são recebidos pelo sensor (neste caso por intermédio do Zigfu). É possível por exemplo reconhecer o número de pessoas que estão à frente do sensor, obter as relações angulares e métricas do corpo do paciente e, nas últimas revisões desta *framework*, identificar gestos do jogador [5].

Os dados fornecidos pelo OpenNI relativos à posição do paciente são disponibilizados da seguinte forma:

– **Pontos 3D** – cada ponto simboliza neste caso, o eixo de ligação entre dois segmentos/"ossos". É utilizado um sistema de coordenadas tridimensionais *(x,y,z)* sendo que x e y são a localização de determinado ponto numa tela 2D fictícia, equivalente às coordenadas de um determinado objeto numa fotografia sendo que o ponto *(0;0)* se situa no centro geométrico da mesma. A coordenada z é a distância entre esse ponto e o sensor;

– **Matriz rotação** – para determinar o ângulo que é feito entre cada ligação e o vector de coordenadas do sensor, é utilizada uma matriz. Esta informação é essencial para distinguir se por exemplo o paciente se encontra de frente ou de costas para o sensor [5];
Com esta informação é possível ser criada uma representação gráfica da posição do paciente (Figura 6).

Figura 6. Exemplo gráfico do *Skeletal Tracking*

Durante o desenvolvimento do jogo, foram criados diferentes ambientes físicos e multi-dimensionais (2D e 3D) com o objetivo de despertar diferentes reações físicas no paciente. No caso do jogo 2D é necessária uma abstração da coordenada y o que implicou ter que modificar a projeção das coordenadas x e z.

4 Modelo de controlo de carga

O estudo do metabolismo do ser humano é bastante complexo devido ao elevado número de variáveis que o constituem e influenciam. Para fazer com que o ritmo cardíaco do paciente x atinja a Zona Alvo Treino de referência em determinado exercício, o jogo aplica uma carga no esforço do paciente. Essa carga aplicada traduz-se na exigência física que o jogo exige ao paciente, para o elevar/descer/manter em determinado ritmo.

O comportamento do ritmo cardíaco x é modelado por uma equação diferencial de primeira ordem, onde o estado x representa o ritmo cardíaco e u é uma variável que define o nível de esforço ou dificuldade que o jogo apresenta ao paciente.

$$\frac{dx}{dt} = a(x - ZAT_{Min}) + bu \tag{1}$$

Nesta equação a é um número qualquer inferior a zero enquanto b é um fator qualquer a definir no jogo para amplificar a carga u. Sendo esta equação apenas um modelo, a e b servem para moldar o comportamento do ajuste da carga como por exemplo o tempo que o paciente deverá passar até que atinja a carga de referência. Como nas ciências da computação não é possível trabalhar equações com variáveis contínuas, recorreu-se à discretização:

$$\frac{dx}{dt} \approx \frac{x(t+\Delta t) - x(t)}{\Delta t} \Leftrightarrow \frac{x(t+\Delta t) - x(t)}{\Delta t} = a(x(t) - ZAT_{Min}) + bu(t) \tag{2}$$

$$\Leftrightarrow x(t+\Delta t) = x(t) + \Delta t a x(t) + \Delta t(-aZAT_{Min} + bu(t))$$

Obter a carga de referência (u_{ref}) implica que o ritmo cardíaco obtido seja igual ao ritmo cardíaco objetivo (x_{obj}).

$$\Rightarrow \frac{dx}{dt} = 0 \Leftrightarrow a(x_{obj} - ZAT_{Min}) + bu_{ref} = 0 \tag{3}$$

Resolvendo em ordem a u:

$$\Leftrightarrow u_{ref} = -\frac{a(x_{obj} - ZAT_{Min})}{b} \tag{4}$$

Com este conjunto de variáveis e relações podemos então definir a carga u como:

$$u = u_{ref} + K(x_{obj} - x) \tag{5}$$

A diferença entre ritmo cardíaco de referência (x_{obj}) e o ritmo cardíaco lido pelo VitalJacket (x), a somar com um erro de leitura que multiplica por uma constante K (positiva) que por sua vez soma com a carga de referência (u_{ref}) dá origem a uma nova carga u. K, tal como a e b nas equações anteriores, não está definido porque pode ser alterado para se obter por exemplo um comportamento diferente em u. Modificar K implica, por exemplo, obter um ajuste de carga mais ou menos sensível.

Por sua vez, u atua sobre a pessoa e é então obtido (lido) um novo ritmo cardíaco x (Figura 7).

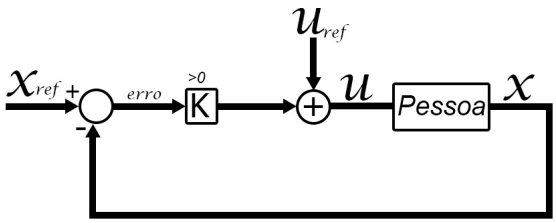


Figura 7.

Uma das possíveis simulações do comportamento da carga e ritmo cardíaco seguindo este algoritmo é o exemplo da Figura 8.

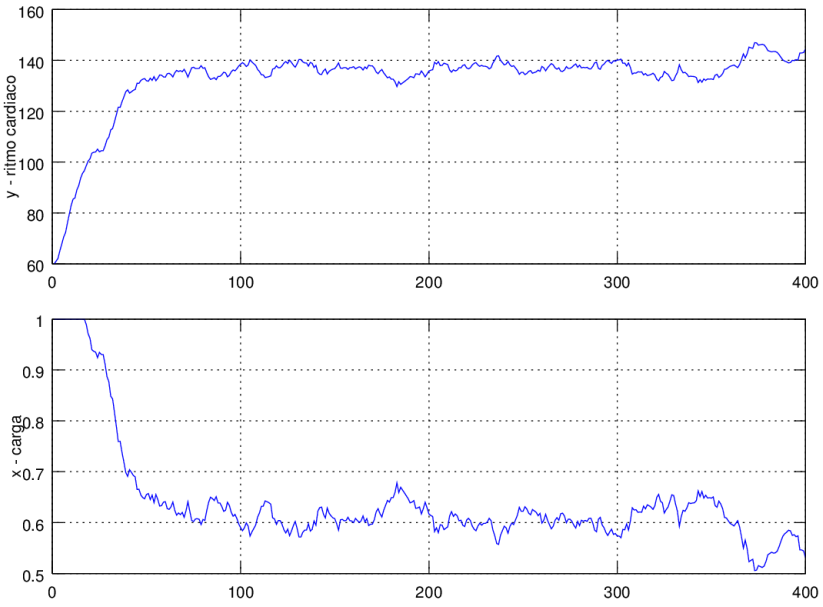


Figura 8. No gráfico de cima consegue-se ver que a pessoa começou no instante de tempo 0 com um ritmo cardíaco de 60 *bpm*. Ao longo do tempo foi aplicada uma carga (gráfico debaixo) que aproximou o ritmo cardíaco do paciente ao objetivo: 140 *bpm*. Note-se que como o metabolismo do paciente é algo irregular, aconteceram quebras e aumentos no ritmo, o que refletiu variações na carga.

5 Conclusão

Os sensores utilizados juntamente com a restante tecnologia utilizada revelam-se compatíveis, tornando possível o desenvolvimento de um sistema completo.

No futuro seria interessante que o sensor de movimento interpretasse e assinalasse em tempo real a correção à postura do paciente, rentabilizando assim o esforço durante todo o exercício ao mesmo tempo que prevenia o aparecimento de problemas físicos derivados a uma eventual postura errada durante a execução dos exercícios.

A ideia de se obter um sistema semi-autónomo está executada, pois apenas é necessária a parametrização dos valores referentes ao jogo/paciente pelo profissional de saúde, e mesmo esta pode ser feita à distância tal como a análise do registo de resultados do paciente.

O algoritmo de balanceamento de cargas, de facto conduz o paciente a atingir a ZAT de referência. Esse ritmo de referência é difícil de manter, pois a forma como a carga atua no jogo (velocidade dos objetos) não atua diretamente no jogador. O que faz com que a carga se altere constantemente.

Através da combinação destas tecnologias é possível explorarmos novas utilidades e paradigmas das aplicações informáticas direcionadas à saúde.

Referências

1. Ricardo Almendra: Geografia da Doença Cardiovascular em Portugal Continental: enfarte agudo do miocárdio - padrões e sazonalidade. (2012)
2. BioDevices: Studies. www.biodevices.pt/ visitado 06-2012.
3. Microsoft: Documentação sql server. http://msdn.microsoft.com/en-us/library/ee229559(v=sql.10).aspx visitado 09-2012.
4. Unity Technologies: Documentação. http://unity3d.com/company/support/documentation/ visitado 08-2012.
5. Borenstein, G.: 3D vision with Kinect, Processing, Arduino, and MakerBot. O'Reilly (2012)
6. Tim Carmody: How motion detection works in xbox kinect. www.wired.com/gadgetlab/2010/11/tonights-release-xbox-kinect-how-does-it-work visitado 11-2012.

Entre os dados qualitativos e a análise de redes

Albertina Ferreira[1] Carlos Caldeira[2] Fernanda Olival[3]

[1] Instituto Politécnico de Santarém; albertina.ferreira@esa.ipsantarem.pt
[2] Universidade de Évora; ccaldeira@di.uevora.pt
[3] Universidade de Évora; mfo@uevora.pt

Abstract. A preparação de dados de modo a adequá-los para serem utilizados em análise de redes é um aspeto fundamental nas bases de dados prosopográficas que envolvem o registo de relações. Mais complexo é ainda quando esses dados se reportam a sociedades pretéritas (séculos XVI a XVIII), como é o caso.
Neste ensaio referimos a importância da identificação de ocorrências anómalas, resultantes da introdução incorreta de dados, e mencionamos ainda as metodologias seguidas para identificar algumas dessas situações. Sugerimos também os procedimentos a seguir para colocar os dados num formato adequado à sua integração em software de análise de redes. Por último, apresentamos alguns dos resultados obtidos para a rede que analisámos.
O repositório de dados que utilizámos tem armazenada informação sobre eventos biográficos e relacionais, sendo o tratamento dos dados fundamental para o estudo das redes de relações entre os diversos atores sociais.

Keywords: base de dados prosopográfica, análise de redes, dados qualitativos

1 Introdução

O estudo da teoria de redes nas ciências físicas e sociais tem sido uma área pela qual os investigadores demonstram crescente interesse. Borgatti *et al.* [6] salientam que a teoria das redes tem possibilitado explicações para os mais diversos fenómenos sociais numa ampla diversidade de contextos.
Como atua uma rede? Como evolui? Podemos encontrar leis e derivar modelos que expliquem essa evolução [12]?
Estas questões são só um exemplo entre muitas outras que se podem colocar. Para que consigamos obter respostas para todas estas questões, há que percorrer um caminho: desde a obtenção de dados, passando pela sua manipulação e exploração, culminando com a análise efetiva da rede.
Uma rede social consiste num conjunto de atores e das relações que estes mantém entre si [10] [13]. Assim, à identificação desses atores está associada a identificação das suas relações. Estas podem incluir colaborações, amizades, laços comerciais, links, citações, fluxos de recursos, fluxo de informações, ou qualquer outra ligação que se possa determinar [13]. Neste estudo vamos referir apenas 2 das cerca de 500 relações guardadas na base de dados SPARES (Sistema Prosopográfico de Análise de Relações e Eventos Sociais).
A representação de redes sociais pode ser seguida e acompanhada em diversos textos como Andery *et al.* [2], sendo o estudo das redes, na teoria dos grafos, um dos pilares fundamentais da matemática discreta [9] [10].
A maneira mais frequente de apresentar visualmente redes sociais é utilizando grafos. Através deles é possível resolver múltiplos problemas, nomeadamente aqueles com que se deparam as redes sociais [1] [2]. Deste modo, qualquer rede, com qualquer tipo de relacionamento, pode ser transposta em grafos [2].
Um grafo é composto por nós, também designados por vértices, atores ou pontos, ligados por arestas, também denominadas relações [9] [10]. Os grafos podem ser classificados consoante o tipo de relação. Das diversas classificações encontradas referimos a apresentada por Abraham *et al.* [1] que consideram, entre outros, grafos não dirigidos (para representar exclusivamente relações simétricas) e grafos dirigidos (para representar essencialmente relações assimétricas).

Hanneman e Riddle [9] questionavam as posições estruturais nos diversos tipos de redes, nomeadamente porque é que um vértice tem uma posição privilegiada relativamente a outro vértice noutra posição. Nesta perspetiva estes autores abordaram diversas medidas de centralidade, das quais destacamos o grau. Também Borgatti e Halgin [5] referem esta medida.
A determinação do número de relações incidentes num determinado vértice representa o grau desse vértice. Nos grafos não dirigidos, os vértices diferem entre si apenas no que diz respeito ao número de relações em que estão envolvidos. Com grafos dirigidos pode ser importante fazer distinções ao nível do grau de entrada (quantas relações chegam a um vértice) e do grau de saída (quantas relações saem de um vértice) [5] [9]. Se um vértice recebe muitas relações é designado *prominent*, possuindo grande prestígio. Se de um vértice saem muitas relações, este é designado *influential* e possui um forte domínio sobre os outros vértices [9].
Embora os autores anteriormente citados considerem essencialmente redes a funcionar na atualidade, grande parte dos estudos que realizam poderão ser estendidos a outras épocas, bem como a outras sociedades. Neste trabalho abordamos o tratamento dos dados e referimos de que modo este se torna importante na análise dos diversos relacionamentos, entre os vários intervenientes nos processos de Familiaturas do Santo Ofício. Este estudo enquadra-se numa das tarefas propostas - *Developing SPARES: social network analysis* - do projeto aprovado e financiado pela FCT: PTDC/HIS-HIS/118227/2010 – CIDEHUS (Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora).
Os objetivos deste trabalho são: 1) analisar a importância da preparação dos dados qualitativos para serem tratados pelo software disponível para a análise de redes; além do mais há sempre erros, quando se trabalha com grandes números; 2) discutir a questão da escolha do software e da validação dos resultados numa equipa multidisciplinar.
Note-se que o facto de trabalharmos com dados reais e com uma equipa multidisciplinar é muito relevante, pois é sempre possível tentar garantir empiricamente o controlo dos resultados e até confrontar diferentes interpretações dos mesmos.

2 Procedimentos Metodológicos

Os dados a utilizar encontram-se distribuídos por três séculos, recaindo o nosso estudo sobre aproximadamente 107000 registos, os quais se encontram disponíveis na base de dados prosopográfica SPARES, desenvolvida no âmbito do projeto FCOMP-01-0124-FEDER-007360 – Inquirir da Honra: Comissários do Santo Oficio e das Ordens Militares em Portugal (1570 – 1773). Tratando-se de uma base de dados prosopográfica tem armazenada informação histórica de indivíduos que são considerados parte relevante da dinâmica social da época.
O Sistema SPARES é uma base de dados relacional desenvolvida de acordo com a Ecologia dos Dados [7] e, fisicamente construída no sistema de gestão de base de dados relacional MySQL. Este repositório de dados está alojado num servidor central com sistema operativo Linux, o qual pode ser acedido por ODBC (Open Database Connectivity) e, como tal, é utilizável por clientes muito diversificados como, por exemplo, sistemas Windows, Linux ou MacOS. Na Figura 1 apresenta-se o modelo de dados que suporta o SPARES.
Relativamente a trabalhos relacionados, em Portugal, apenas é do nosso conhecimento a aplicação Time Link, que pode ser acedida em `http://timelink.fl.uc.pt/`. As principais componentes desta aplicação são a base de dados e a interface gráfica de utilizador. Através deste repositório de dados é possivel reconstruir árvores genealógicas, sendo esta uma das funcionalidades mais atraentes para quem a consultar [8] .
Numa primeira fase automatizámos a extração dos dados a partir da base de dados prosopográfica SPARES, de modo a que estes possam ser manipulados por dois softwares de rede: PAJEK e GEPHI.
Inicialmente escolhemos o PAJEK para construirmos e analisarmos a nossa rede. Com efeito, esta aplicação consegue, por um lado, explorar e manipular redes de grande dimensão e, por outro, encontrar-se disponível gratuitamente, para uso não comercial. Pode ser acedida a partir de: `http://vlado.fmf.uni-lj.si/pub/networks/pajek/` [4] [11]. Embora através desta aplicação se

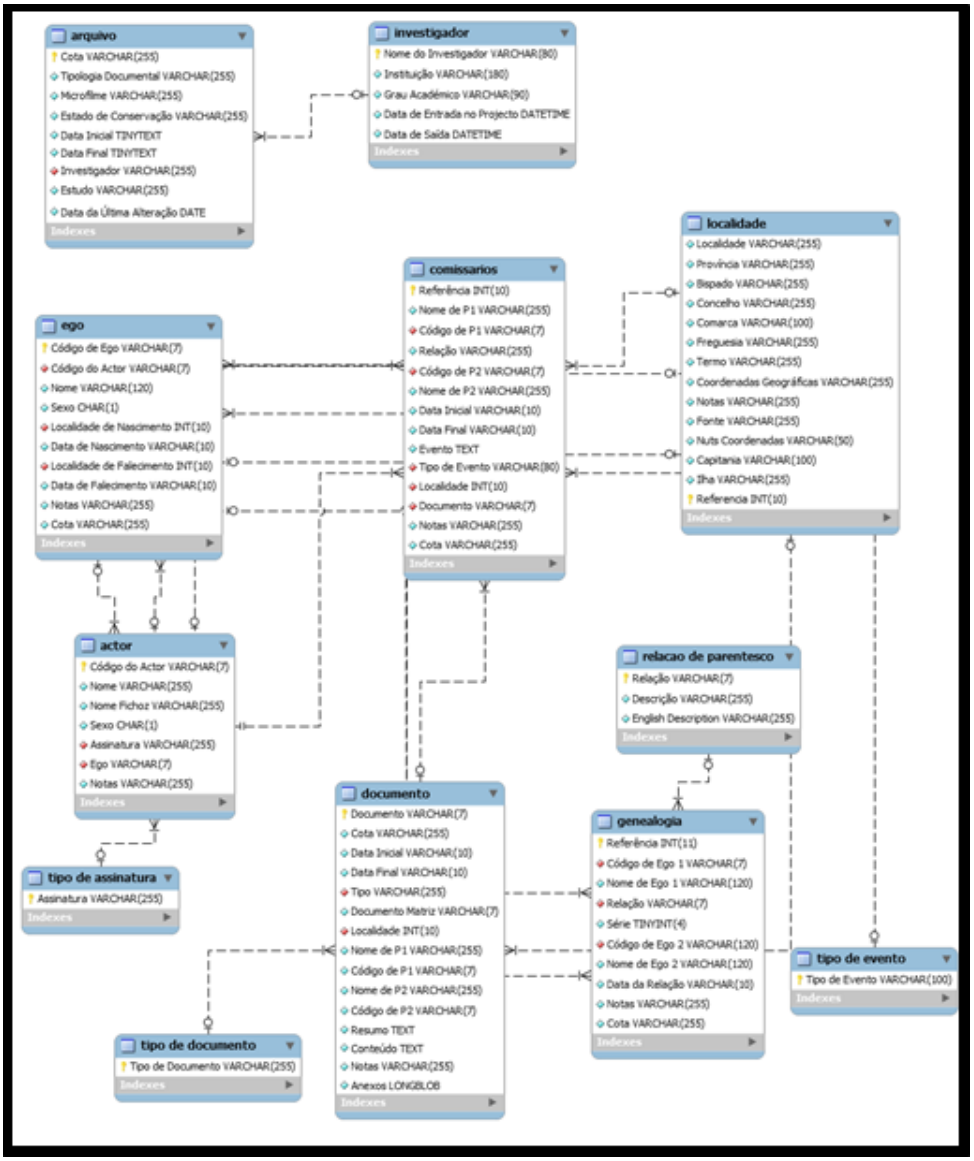


Fig. 1. Modelo de dados do sistema SPARES.

consiga analisar redes e obter dados, tanto analíticos como gráficos, que podem ser explorados por outras aplicações, pesquisámos outras ferramentas Open Source. Como futuramente pretendemos integrar na mesma plataforma o pré-processamento dos dados e a análise de rede, consideramos que o GEPHI (disponível em: `http://gephi.org/`) [3] permitirá uma maior interoperabilidade.

Para preparar os dados de modo a poderem ser utilizados no software de rede foi necessário:

- criar uma tabela com os códigos e nomes dos primeiros intervenientes (P1);
- acrescentar a essa tabela os códigos e nomes dos segundos intervenientes (P2);
- criar uma tabela com os vértices da rede;
- criar uma tabela com as relações da rede;
- gerar as listagens que vão ser exportadas.

Para que o ficheiro obtido pudesse ter o formato que o PAJEK lê foi ainda necessário:

- criar um procedimento e uma pesquisa para atribuir uma numeração sequencial;
- criar um procedimento para eliminar linhas em branco do ficheiro de output.

Para que se possa vir a analisar a rede por intervalo de tempo foi também necessário preparar os ficheiros com a informação da década a que cada uma das relações corresponde. A data está num formato de texto que varia entre o padrão 1709=11=08 (exatamente esta data) ou 1742<06<09 (pensa-se que tenha ocorrido antes desta data) e vagamente nesta data (1709-00-00). Os historiadores precisam de trabalhar deste modo, pois nem sempre têm a certeza da cronologia exata da ocorrência. Foi necessário alocar a cada vértice (indivíduo) as décadas em que ele interveio nos processos de familiaturas do Santo Ofício. O intervalo de tempo resultante para cada vértice deverá ter o formato [década x-década y] ou por exemplo [2-4], que significa que o comissário interveio durante 30 anos, ou seja, 3 décadas (2, 3 e 4).
No decorrer deste trabalho identificaram-se e corrigiram-se dados que foram introduzidos de modo incorreto, nomeadamente:

- datas negativas;
- datas inferiores a 1579 (primeira relação conhecida);
- comissários que mantinham relação com eles próprios;
- o mesmo código (único para cada um dos indivíduos) atribuído a dois indivíduos diferentes;
- o mesmo indivíduo com nomes diferentes.

A identificação destas ocorrências foi feita através de pesquisas quando se identificou que o ficheiro final possuía mais relações do que as originais. Tivemos de rever todo o percurso e identificámos as situações atrás referidas. Relativamente às 4 primeiras situações, foram corrigidas manualmente, pois é necessário conhecer a base de dados, nomeadamente as relações envolvidas. No que diz respeito à última relação, criou-se um procedimento que atribuísse ao nome do indivíduo referenciado em P1 o nome que ele possuía em P2.
Os ficheiros obtidos como inputs vão ser utilizados no software de análise de redes, evidenciando a importância destes no processo de análise.

3 Resultados e Discussão

Apresentamos nas Figuras 2 e 3 exemplos de ficheiros construídos para utilização nos softwares de análise de redes.
Dos cerca de 500 eventos relacionais identificados na base de dados SPARES, vamos analisar os comissários do concelho de Arraiolos, no que diz respeito às relações de: "Ouvida como testemunha na extra-judicial SO pelo comissário ad hoc com rol do pároco"; "Ouvida como testemunha na habilitação SO pelo comissário SO". Escolhemos este concelho por ser de pequenas dimensões e estes eventos por se encontrarem já estudados do ponto de vista estatístico para esta localidade. Assim, havia todo o interesse em confirmar graficamente o que analiticamente já se conhecia. Por

```
*Vertices    12
    1    "António Correia Bethencourt"    ic Red bc Red"    [13-15]
    2    "António de Noronha e Meneses [Dom]"    ic Red bc Red"    [15-18]
    3    "António Mouzinho [Doutor]"    ic Blue bc Blue"    [1]
    4    "Bartolomeu César de Andrade"    ic Red bc Red"    [11-15]
    5    "Bento Pais do Amaral"    ic Green bc Green"    [16-19]
    6    "Cristóvão de Sousa e Lira [Licenciado]"    ic Red bc Red"    [13-14]
    7    "Diogo Fernandes Branco"    ic Red bc Red"    [8-12]
    8    "Jacome Esteves Nogueira"    ic Blue bc Blue"    [14-18]
    9    "José de Sousa Castelo Branco [Dom]"    ic Blue bc Blue"    [12-14]
    10    "Mariana Isabel de Mesquita e Noronha [Dona]"    ic Red bc Red"    [18]
    11    "Martim Filter"    ic Blue bc Blue"    [8]
    12    "Mateus da Silva"    ic Red bc Red"    [1]

*arcs
    9    1    1    [13]
    5    2    1    [18]
    9    4    1    [14]
    8    5    1    [18]
    9    6    1    [13]
    11    7    1    [8]
    5    10    1    [18]
    3    12    1    [1]
```

Fig. 2. Input para PAJEK.

Vértices

```
Id;Label

91;Bento Pais do Amaral
151;Bartolomeu César de Andrade
152;Cristóvão de Sousa e Lira [Licenciado]
153;José de Sousa Castelo Branco [Dom]
154;António Correia Bethencourt
2035;António de Noronha e Meneses [Dom]
2351;Mariana Isabel de Mesquita e Noronha [Dona]
2375;Jacome Esteves Nogueira
2613;Diogo Fernandes Branco
4731;Martim Filter
6076;Mateus da Silva
6094;António Mouzinho [Doutor]
```

Arestas

```
Source;Target;Label;Weight

91;2351;Patrocínio;1.0
91;2035;Patrocínio;1.0
153;154;Patrocínio;1.0
153;152;Patrocínio;1.0
153;151;Patrocínio;1.0
2375;91;Patrocínio;1.0
4731;2613;Patrocínio;1.0
6094;6076;Patrocínio;1.0
```

Fig. 3. Inputs para GEPHI.

outro lado, as relações "Ouvida como testemunha ..." fazem parte daquelas que têm um papel preponderante na cotação dos atores sociais.

Relativamente ao primeiro evento (relações a castanho) (Figura 4), podemos verificar que os dois comissários (vértice azul e vértice verde) que não são de Arraiolos, e que ali atuaram entre 1743 e 1746, solicitaram ao pároco uma lista das testemunhas (vértices vermelhos) que deviam auscultar. Estas listas não se cruzam entre si, até porque se referiam a freguesias diferentes (S. Gregório e Igrejinha). A partir desta análise, podemos observar se as testemunhas fornecidas são das que mais intervêm nos concelhos de Arraiolos. Fez-se assim uma nova análise que veio confirmar que efetivamente estas são das testemunhas mais vezes chamadas, apesar de uma delas (Leonor Marques) ser mulher – eram menos requeridas – e não saber ler nem escrever (Figura 5). Nesta visualização destaca-se a presença de vértices com diferentes tamanhos. Estes vértices representam as testemunhas ouvidas nos processos, quanto maiores os vértices mais vezes a testemunha foi ouvida. Esta visualização salienta ainda algumas relações onde as testemunhas foram ouvidas mais de uma vez pelo mesmo comissário. Pode observar-se a direção do relacionamento pelo sentido que as relações apresentam. Analiticamente consegue-se confirmar o grau dos diversos vértices. Como se trata de uma representação com grafos dirigidos pode-se obter também o grau de entrada e de saída. Apresenta-se, como exemplo, na Figura 6 estas medidas relativamente às 7 testemunhas mais ouvidas no concelho de Arraiolos.

No que diz respeito ao segundo evento, observamos que existe um maior número de intervenientes, dos quais o comissário Gaspar Barreto de Ladim é o que tem um maior número de relações (55). Este resultado veio confirmar a sua importância. Tratou-se de um comissário ativo durante 30 anos nesta rede. Na Figura 7 apresenta-se a rede, realçando os comissários mais ativos nos concelhos de Arraiolos. Atribuímos a cor amarela àqueles que moravam em Arraiolos. O GEPHI permite facilmente efetuar este tipo de manipulações, de modo a realçar pormenores que o investigador pretende destacar. Neste caso, era relevante destrinçar os ali moradores (conheciam toda a gente) dos

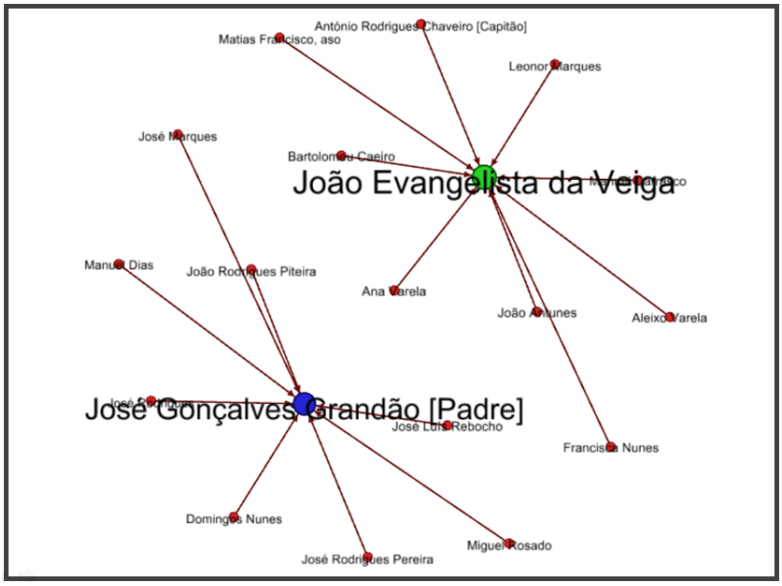


Fig. 4. Evento relacional "Ouvida como testemunha na extra-judicial SO pelo comissário ad hoc com rol do pároco".

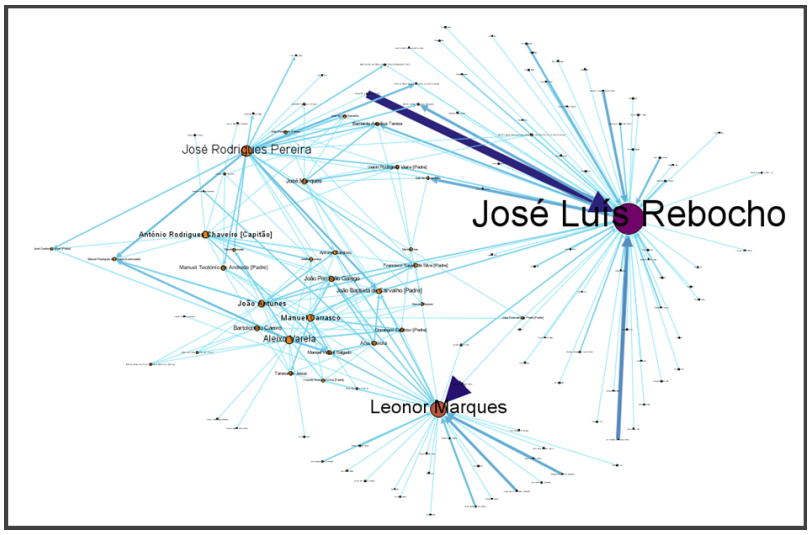


Fig. 5. Testemunhas utilizadas nos concelhos de Arraiolos.

Testemunhas	In-Degree	Out-Degree	Degree
José Luís Rebocho	56	14	70
Leonor Marques	22	13	35
José Rodrigues Pereira	0	23	23
Aleixo Varela	0	16	16
António Rodrigues Chaveiro [Capitão]	0	14	14
Manuel Carrasco	0	14	14
João Antunes	0	13	13

Fig. 6. Medidas de centralidade para as 7 testemunhas mais ouvidas no concelho de Arraiolos.

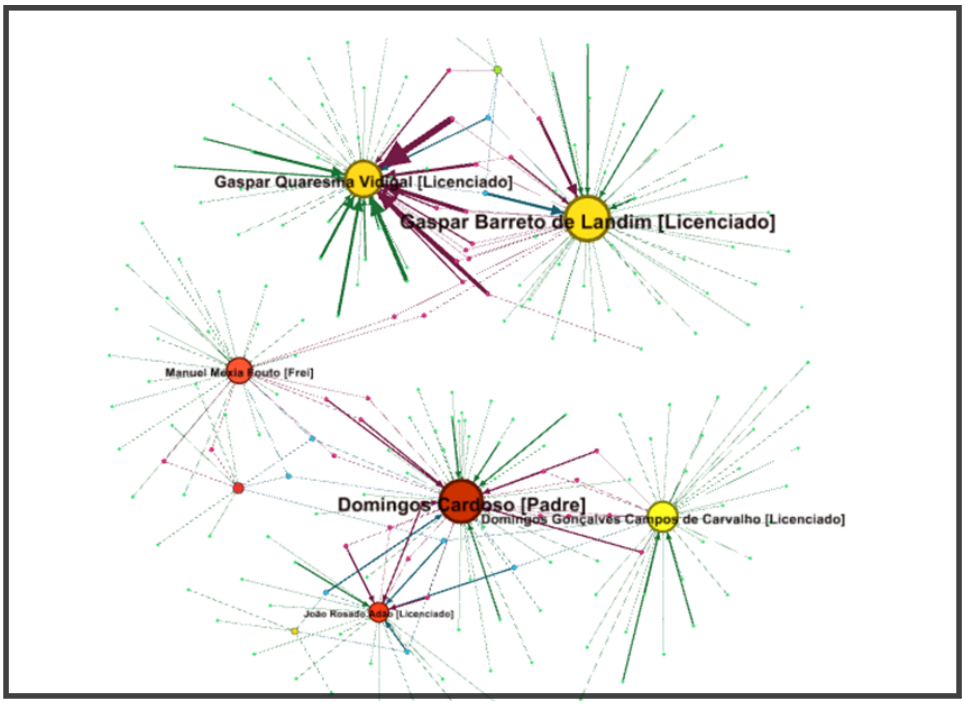


Fig. 7. Comissários mais ativos no concelho de Arraiolos.

que não moravam no concelho, pois tinham menos probabilidades de conhecerem minuciosamente todos. Também nesta figura o tamanho dos vértices é diretamente proporcional ao seu grau, bem como a espessura das relações proporcional ao grau de entrada e de saída (consoante a direção). Na Figura 8 apresenta-se, relativamente aos comissários referidos na Figura 7, o total de comissários por grau. Pode-se contatar que 4,1% dos comissários efetuaram 97,7% das relações.

Total de comissários	172	33	8	1	1	1	1	1	1	1	1	1
Degree	1	2	3	5	6	10	22	30	36	44	52	55

Fig. 8. Total de relações por grau.

4 Conclusões e Trabalho Futuro

Pretendemos com o presente trabalho mostrar o interesse do tratamento e análise dos dados qualitativos, de modo a possibilitar a sua utilização na análise de redes sociais.
A preparação e manipulação dos dados efetuadas até ao momento permitiu-nos confirmar dois eventos relacionais que estatisticamente já se encontravam analisados. Este facto permite, pois, validar a atuação desta rede. No decorrer do nosso trabalho, e em estreita colaboração com os membros do nosso projeto, foi também possível identificar e corrigir algumas situações anómalas. Estas têm de ser feitas automaticamente, pois com grandes números quem introduz perde facilmente o controlo dos dados.

Relativamente à escolha do software, esta recaiu sobre o GEPHI, em detrimento do PAJEK. Com efeito, aquele software revelou-se mais adequado aos nossos objetivos, possibilitando uma integração mais amigável na aplicação que nos encontramos a desenvolver.
Em termos de futura investigação, consideramos pertinente trabalhar na produção de redes dinâmicas do ponto de vista da variação cronológica, dado que na nossa investigação nos deparamos constantemente com esta última, a par do caráter fragmentário dos dados que exploramos. Será também alvo do nosso trabalho a exploração do parentesco horizontal como rede, pois consideramo-lo muitas vezes mais importante que o parentesco vertical.
Uma das nossas metas é ainda construir uma aplicação que permita a adequação entre a base de dados prosopográfica SPARES e o software de redes GEPHI, tornando possível que qualquer utilizador de Ciências Sociais, e como tal menos familiarizado com a Estatística e a Informática, possa efetuar rápida e facilmente uma análise à rede social que estuda.

References

1. Abraham, A., Hassanein, A.-E., Snasal, V.: Computational Social Network Analysis: Trends, Tools, and Research Advances. Springer. London. (2010)
2. Andery, G. F., Lopes, A. A., Minghim, R.: Exploração visual multidimensional de redes sociais. 2nd International Workshop on Web and Text Intelligence. São Carlos. 1–9 (2009)
3. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks. In Proceedings of the Third International ICWSM Conference. California, USA. 361–362 (2009)
4. Batagelj, V., Mrvar, A.: Pajek: Program for Analysis and Visualization of Large Networks. Reference Manual List of commands with short explanation version 2.00. University of Ljubljana. Slovenia. (2010)
5. Borgatti, S. P., Halgin, D.: Analyzing Affiliation Networks. The Sage handbook of social network analysis. Carrington P, Scott J (eds). Sage Publications. (1996)
6. Borgatti, S. P., Mehra, A., Brass, D. J., Labianca, G.: Network Analysis in the Social Sciences. Science. **323** 892–895 (2009)
7. Caldeira, C.: A Arte das Bases de Dados. Edições Sílabo, Lisboa. (2011)
8. Carvalho, J. N.: Time link: a evolução de uma base de dados prosopográfica. Tese de mestrado em Património Europeu, Multimédia e Sociedade de Informação. Universidade de Coimbra. (2010)
9. Hanneman, R. A., Riddle, M.: Introduction to social network methods. University of California, Riverside. (2005) `http://faculty.ucr.edu/~hanneman/`
10. Newman, M. E. J.: The Structure and Function of Complex Networks. SIAM review. **45** 167–256 (2003)
11. Nooy, W, Mrvar, A, Batagelj, V.: Exploratory Network Analysis with Pajek. Cambridge University Press. New York. (2005)
12. Snijders, T.A.B., Steglich, C.E.G., van de Bunt, G.G.: Introduction to Actor-Based Models for Network Dynamics. Social Networks. **32** 44–60 (2010)
13. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge Univ. Press. Cambridge. (1994)

Enterprise Intelligence based on Metadada and Enterprise Architecture Model Integration

Francisco Santana Guimarães

University of Évora

Abstract. Organization management demands from information system a set of functionalities to support the business, but also specific solutions to data reuse and structure, to represent the organization itself in a systemic/Holistic context approach that can be used to improve the organization capabilities to survive in a continuous change environment. With this organization integrated view is possible to improve the opportunities/strength and manage the threats/weaknesses, in the context of strategy process.
Information System domain has presented solutions in these domains, as evolution of operational systems and Business Intelligence systems. One critical component on Business Intelligence systems is the knowledge about the data, known as Metadata. This component can be used to guide implementations and reused by several systems to support the intelligent knowledge to take decisions based on this Organizational Ontology. Research in this domain includes metamodels like OIM (Open Information Model) and CWM (Common Warehouse Model), but none consider the integration with EA (Enterprise Architecture).
This paper presents the state of art and describes the problem and hypothesis about Metadata and Enterprise Architecture integration as organizational ontology.
Keywords: Business Intelligence, DataWareHouse, CPM, BAM, Metadata, Enterprise Architecture, Intelligent Agents, Ontology

1 Introduction

The BI (Business Intelligence) system represents a line of essential evolution within the information system architecture components, being equally reused on a complementary way by management monitoring systems like CPM (Corporate Performance Management), BAM (Business Activity Monitoring), Scorecards/Dashboard with the highlight to BSC (Balance Scorecard) models.

Metadata, one of the BI architectural components, is the responsible to maintain a technical and functional ontology to guide implementation, and to be used as knowledge to decision support. Not only to Business Intelligence but also for all systems and users as part of organizational ontology. However his importance as ontology, has not being used, instead it is a component forgotten in enterprise implementations of Business Intelligence system, despite some line of investigation through metamodels like CWM (Common Warehouse Metamodel) and OIM (Open Information Model).

On the other side, the organization needs to represent knowledge about itself in appropriate structure to allow flexible evaluation, definition and monitorization of strategy and its operationalization, viewing strategic management as a process, with performance indicators to the decision taking. Systems like BPM (Business Process Management) and Workflow systems are used. But, on a more integrated reasoning, EA (Enterprise Architecture) models like TOGAF and Zachman Framework, are used as models capable to represent the holistic vision of components and its relationships, to achieve the best way to understand the interdependencies.

Because the Metadata and EA concepts are conceptually related, some line of investigation exists to explore the way Metadata can be structured based on EA concepts, but not as integrated model of organizational ontology. This concept is critical in the actual organization context where business operational systems and every specific Business Intelligence solution represent a technological amalgama, implemented on different time's period and without integrations. This article considers as hypothesis the integration between Metadata and EA, as Organizational Ontology

concepts, thru the evolution of Metamodels like CWM and TOGAF, creating an integrated business ontology that allows effectively to generate organizational intelligence models, on which every system can be compromised on the domain of the discourse and can support decision taking with learning and knowledge, as true Business Intelligence.

2 State of the art

2.1 Information Systems and Business Intelligence

(Laudon e Laudon) classifies information systems in a hierarchy of Operational Systems, followed by Knowledge Systems, Management Systems and Strategic Systems, fitting the BI systems in the last two groups, with specific components of MIS (Management Information System), DSS (Decision Support Systems) and EIS (Executive Information System).

In specific BI context, (Gangadharan) mentions different definitions of BI, depending on the areas where they are used, like CRM (Customer Relationship Management), DataWareHouse and DataMining. But this author refers that these systems must be looked in more global way as integrative architecture of several components, databases and operational applications, consolidated to give access to organized data with the right sense to decision taking. BI, in this concept, represents a new management discipline where data is finally treated as organizational resources. In order of that,(Kung) refers BI as a possible solution to implement organizational performance systems, like BAM, based on indicators captured on processes and organizational events.

(Berkeley) in roadmap datawarehouse architecture defines BI as a system that includes historical data and a structure based on registered events (facts) and context where they occur (dimensions). This group of facts and dimensions are the basis of Business Intelligence architecture, composed by application architecture, data architecture, security architecture and support architecture (Organic structures and Processes). However, they use Metadata only to DataWareHouse, and not also to be reused by Operational Systems.

Evolution on this domain consider:

- DataWareHouse as part of BI vision is an integration of components like ETL (Extraction, Transform and Load), Metadata, DW (DataWareHouse), ODS (Operational Data Store), DataMart, Data Mining, Scorecard/Dashboard and Reporting. This concept is seen (Gangadharan) as an Enterprise Architecture to integrate operational systems (e.g. business specific systems, Customer Relationship Management, Enterprise Resource Planning) and decision support systems to give access to business information to support decision, with acess by Web and Mobile devices;
- Metadata component of BI architecture is investigated in OIM (Open Information Model) and CWM (Common WareHouse Model) initiatives, based on XML/XMI Technology to support data interchange on import/export integration model. But this metamodels are investigated on context of ETL interoperability tools and specific for DataWareHouse technical Metadata (and no Business Metadata) and "they not provide support to categorize information according to different conceptual levels of metatada as e.g. suggested by Zachman" (Vetterli et al). Other lines of investigation exists to build Enterprise Metadata Repositoy (EMR) to capture business content and Enterprise Conceptual Data Model (ECDM) to capture IT specific components normally in CMDB (Configuration Management DataBase) (Turco);
- The operational systems where management information is captured, is evolved in a classification scale using modern technical architectures in terms of application and data models, not only in order to maintain temporal data but also to be organized to allow specific BI solutions. This cause at enterprise level new challenges in terms of systems heterogeneity at BI level, without the possibility of crossing data between solutions, if we consider the BI component of each solution;
- Complementary solutions to BI have been appearing with proper capture forms and operational data treatment, or using DW/ODS (Kung) data, as the case of CPM (Corporate

Performance Management) and management control online alarms, as the case of BAM (Business Activity Monitoring), that uses online data, instead of depending on periodic loading to DW/ODS.

In any of this evolution lines (Gangadharan), the understanding of organization, structured as data, is the strategic key to create competitive advantages. This data is growing in volume, heterogeneity, syntax and semantic. Giving this complexity, the critical sucess factor is the data integration from several sources and formats, applying the adequate tools and models to its analysis, organization and to guide the decision support based on that information. Not only to business information, but also about organization itself.

2.2 Organizational Architecture and Ontologies

To (Caldeira), the information systems exist in an organizational context, and its implementation based on databases can't be considered just as technology acquisition, without the focus on information architecture. The organization context as systemic/holistic vision can be represented as knowledge using Enterprise Architecture (EA), with several alternative models like Zachman Architecture Framework (Zachman) and TOGAF (The Open Group Architecture Framework) applicable to organizations like enterprises, countries or specific units.

According to the IEEE P1471 (IEEEtd1471-2000) an architecture is "The fundamental organization of a system, embodided in its components, their relationships to each other and the environment, and the principles governing its design and evolution.". According to TOGAF, its utility is related to the fact that organizations have objectives to achieve and the architecture represents the integrated vision of components or the organization resources to achieve these objectives. TOGAF model considers a Framework to design and manage an enterprise architecture structured in components like an applicational architecture (applications and information) and technological architecture. TOGAF considers also a repository of metamodels applicable to several situations, to guarantee the learning and the reutilization, with controlled evolution.

This architecture with organization knowledge representation can be seen as organizational memory to be used on knowledge management and eLearning system, as referred by several authors (Palmer et al), (Maier). As the potential of these reutilizations demonstrates, its utilization is critical to constitute an organizational ontology instanced on technical and functional Metadata that turn a BI system in an Intelligence Enterprise system. With this vision the architecture is seen as Ontology and it is a form of knowledge representation, in a way that can be seen as a logical application of ontologies to build computational models in specific domains that permit its usage in rational processes (Sowa). In this way the knowledge can be represented by ontologies while descriptions of concepts and relations that exist between them, like a vocabulary or universe represented with some declarative formalism where some agents must compromise with, to be able to communicate in this domain, without having to handle with a general knowledge theory (Gruber).

3 The Problem

Therefore, we have a model of Business Intelligence that aggregates architectural components like ETL, DW, ODS, DataMart, Data Mining, Business Visualization, and on a crucial form, the Metadata. We have also other specific systems to support enterprise monitoring like CPM and BAM that need online data, instead of data captured periodically on the DataWareHouse. Finally, any organization need to represents the knowledge about the organization it self, in models of enterprise architecture. But, normally, this models are not related as refered by (Vertelli) when mention the lack of Business Concepts in models like CWM or OIM in Metadata investigations.

To (Moss et al), many organizations have several BI implementations. However there exists a recurrent implementation problem related to the lack of understanding of the complexity and the Transversatility of these systems, particularly the inexistence or the lack of understanding about

the importance of Metadata. At data capturing level, (Kung) refers that many organizations with some dimensions face the complexity of heterogeneity of its IT architecture (applications, technologies); including the using of business support workflows that capture only part of the business processes and its associated events. Particularly in what concern to BAM, (Kung) refers that can be implemented based on DataWareHouses, despite that traditionally they need real time data.

Thus, the problem can be presented to the level of the lack of Transversatility of Metadata, on a context of technological heterogeneity and business support application, without any orientation guide while enterprise ontology that can integrate Metadada and Enterprise Architecture repository. This two concepts, EA and Metadata, must be related as common organizational memory, ontology, enterprise architecture or Metadata, otherwise is not possible to be used as basis for inference and knowledge to improve the system implementations, its management and information consistency.

4 Conclusion/Discussion: The Hypothesis

The problem is part of EA domain, seen as ontologies to knowledge management, and also part of BI domain, on Metadata component.The focus is on data structure as basis of any solution seen as representation model of reality with certain relations, syntax and semantic(Gangadharan). The data can be structured or semi-structured. Can be associated to business processes or strategic internal or external variables, and structured as knowledge to support efficient decisiond (Gangadharan).

But how can we represent these data while enterprise model to guarantee the consistency and as independent definition of any system that will use it? And how can this model represent enterprise ontology to be reused by several systems? Our understanding is that it must be considered as organizational systemic/holistic vision like an enterprise architecture or organizational memory, and be used as Organizational Metadata, that allows the existence of a true Enterprise Intelligence system, if we consider this architecture as organizational ontology.

In this sense, considering the resemblance of concepts, properties and relations, inherent to the two models, the EA and Metadata, the purposed hypothesis is the evolution of CWM or OIM metamodels as part of TOGAF enterprise architectures repository, that can be used not only to BI in specific ETL, but also in business metadata, BI visualization tools and at operational system levels to maintain an integrated vision of dimensions (how we see the business) and facts (what events the business uses and generates).

With this approach as knowledge organization representation ontology, became possible to relate the organization structure, the key functions, the objectives dependency, the application patterns and data, among other organization artefacts. With this model became possible to create a vision of enterprise intelligence. This is the line of investigation that we follow with specific possible usage:

- Common framework and tools to design, implement and manage information architectures with support to impact change analysis
- Renew of BI data architecture model with DataMart over ODS (Operational Data Store) as part of operational system databases with metadata as guide. This can guarantee the online data for business monitoring (like CPM, BAM) and also, more flexible BI architecture.

5 References

(Berkeley) UC-Berkeley DataWareHouse Roadmap. Data WareHouse Architecture. Available at http://edw.berkeley.edu, Retrieved February, 1, 2013.

(Caldeira) Caldeira, Pampulim Carlos. Information Ecology and Domain Definition. Comunicação apresentada ao 6º CONTECSI - International Conference on Information Systems and Technology Management TECSI/EAC/FEA/USP. São Paulo. 03 – 05 June 2009.

(Gruber) Gruber, T.R., (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing, In:Guarino, N., Poli, R.(Eds.):Formal Ontology In Con-ceptual Analysis and Knowledge Representation, Klwer.

(Gangadharan) Gangadharan, .G.R, Sundaravalli, Swami. Business Intelligence Systems: Design and Implementation Strategies. IMIT Class of 2004, Scuola Superiore Sant'Anna, Pisa, Italy.

(IEEE Std1471-2000) Systems and Software engineering - Recommended practice for architectural description of software-intensive systems: ISO/IEC 42010:2007(E).

(Kung et al) Kung, Peter, Hagen, Claus, Rodel, Marisa, Seifert, Sandra. Business Process Monitoring and Measurement in a Large Bank: Challenges and Selected Approaches.

(Laudon e Laudon) Laudon, K. C, Laudon, J. P. Management Information Systems - Organization and Technology in the Networked Enterprise (sixth edition ed.). New Jersey: Prentice - Hall, Inc.

(Maier) Maier, R. Knowledge Management Systems: Information and Conmunication Technologies for Knowledge Management.

(Moss et al) Moss, L.T. , Atre, S. (2003). Business Intelligence Roadmap: The Complete Project Life Cyle for Decision-Support Applications. Addison-Wesley Longman Publishing Co, Inc.

(Sowa) Sowa, John. Knowledge Representation: Logical, Philosophical and computational foundation. Available at www.jfsowa.com, Retrieved February, 1, 2013.

(Togaf) TOGAF v9, The Open Group Architecture Framework. Available at www.Togaf.org, Retrieved February, 1, 2013.

(Turco) Turco, Carl. Enterprise Architecture and Metadata Modelling: A Guide to Conceptual Data Model, Metadata Repositoty, Business and System Re-Enginerring. ISBN-0-7414-5304-5.

(Vetterli et al) Vetterli, Thomas, Vaduva, Anca, Staudt, Martin. Metadata Standards for DataWareHousing: Open Information Model vs Common WareHouse Metamodel.

(Zachman) Zachman, J. P.. The Zachman Framework Evolution. Available at http://www.zachmaninternational.com/index.php/ea-articles, Retrieved February, 1, 2013.

Índice de Autores