Universidade do Minho
Escola de Engenharia
Departamento de Informática

SIMPÓSIO DOUTORAL 2006

Relatório de Actividades do Doutorando

Anália Maria Garcia Lourenço

2006

Índice

A. IDENTIFICAÇÃO

A.1 Doutorando

Nome:	Anália Maria Garcia Lourenço
Data e Local de Nascimento:	11 de Fevereiro de1975, Melgaço, Portugal
Nacionalidade:	Portuguesa
Bilhete de Identidade:	10623763 de 2002/09/09 (Viana do Castelo)
Morada Institucional:	Departamento de Informática Escola de Engenharia Universidade do Minho Campus de Gualtar 4710-057 Braga – Portugal
Telefone:	+351 253 604470
Fax:	+351 253 604471
Residência:	Rua da Calçada, nº10, Vila 4960-529 - Melgaço
Telemóvel:	+351 961 283 994
Correio electrónico:	analia@di.uminho.pt

A.2 Título da Tese

O título provisório da tese é "Web Crawlers: Detecção e Previsão Comportamental baseadas na análise de *clickstreams*".

A.3 Orientador

Nome:	Orlando Manuel de Oliveira Belo
Morada Institucional:	Departamento de Informática Escola de Engenharia Universidade do Minho Campus de Gualtar 4710-057 Braga – Portugal
Telefone:	+351 253 604470
Fax:	+351 253 604471
Correio electrónico:	obelo@di.uminho.pt

A.4 Data Início

Data oficial – Outubro de 2002.

Data efectiva - Outubro de 2002.

A.5 Data Término

Data oficial – Setembro de 2006.

Data efectiva - Setembro de 2006.

B. RESUMO

B.1 Área de Investigação e Desenvolvimento (I&D)

As principais áreas de I&D são: os Sistemas de *Data Webhousing*, a Mineração de Dados Web e os Web *Crawlers*. Contudo, áreas mais abrangentes como sejam os Sistemas de *Data Warehousing*, a Extracção de Conhecimento de Bases de Dados e a Qualidade de Dados também se incluem entre as suas áreas de I&D.

B.2 Resumo

Actualmente, os Web *crawlers* constituem uma das comunidades mais activas da rede, encontrando-se envolvidos em praticamente todas as actividades de recolha automática de conteúdos. A simplicidade do conceito e a facilidade de implementação deste tipo de programas tornou-os extremamente populares e rapidamente os seus objectivos ultrapassaram os propósitos definidos pelos motores de pesquisa. De acordo com a literatura da área, os Web *crawlers* são responsáveis pela recuperação de inúmeros recursos da rede, desde dados com vista à manutenção de índices de pesquisa ao *browsing offline*, passando pela validação de hiper-ligações e esquemas DTD, a monitorização de alterações em páginas Web ou a duplicação de sítios (*mirroring*). Mais, novos tipos de programas têm sido desenvolvidos com vista a actividades ditas focalizadas, destinadas às mais variadas áreas de negócio e investigação.

A relevância destes programas no cenário actual da rede é inquestionável, mas, mais uma vez, a sua evolução não foi suficientemente pensada: não só não se regulamentou as suas acções como também nunca se procedeu à sua devida catalogação. Sempre se pressupôs que os programas criados seriam éticos, ou seja, respeitariam as regras de acesso aos sítios, preservariam os recursos da rede e dos sítios e se auto-identificariam. Contudo, a realidade dista muito deste propósito, existindo já relatos de incidentes relacionados com a sobrecarga da rede e de servidores Web e a violação de direitos de autor e privacidade, entre outros. De facto, embora a maioria dos Web *crawlers* actuais tenha propósitos lícitos e benéficos para os sítios, apenas uma minoria se auto-identifica, convertendo a detecção das suas visitas, particularmente enquanto estas ainda estão a decorrer, em tudo menos numa tarefa trivial.

Este trabalho de dissertação teve por objectivo estudar este problema afim de se conceber uma abordagem apropriada quer à detecção e à contenção de visitas quer à análise de padrões de navegação e propósitos dos Web *crawlers*. Eles não devem ser encarados do mesmo modo que os utilizadores regulares (humanos), pois as suas tendências e os seus propósitos escapam muitas vezes à lógica habitual de interesses, preferências e associações entre páginas. Eles não têm preferência por um determinado tipo de *layout* nem são sensíveis a uma dada promoção ou campanha de *marketing*, eles não têm sítios favoritos nem mantêm *bookmarks*, eles simplesmente pretendem varrer o espaço de pesquisa que lhes foi imposto. Assim sendo, faz todo o sentido separar as águas, isto é, efectivar a diferenciação do fluxo de *clicks* (dados relativos às travessias dos utilizadores na Web) em duas vertentes: o fluxo de *clicks* dos Web *crawlers* e o fluxo de *click*s dos utilizadores Web ditos convencionais ou regulares. Esta diferenciação permitirá uma análise conscienciosa dos perfis de utilização, caracterizando cada grupo de utilizadores convenientemente, indagando acerca dos seus propósitos e tendências e detectando eventuais situações anómalas ou imprevistas. Se por um lado isto proporcionará um retrato muito mais fiel dos utilizadores reais de um sítio, algo extremamente rentável em termos de *marketing* e personalização e renovação de sítios, por outro lado, representará também um passo em frente na segurança e preservação dos recursos dos sítios, monitorizando-se e detendo-se actividades que ponham em perigo o serviço a utilizadores regulares e a privacidade de conteúdos.

B.3 Objectivos Estratégicos

Esta dissertação pretende estabelecer uma abordagem capaz de identificar, interpretar e controlar as actividades de Web *crawling* ao nível de um sítio. Nesta medida, os objectivos que foram estabelecidos ao longo deste trabalho passaram necessariamente por distintos estágios de aquisição conhecimentos e desenvolvimento: o estudo do estado da arte dos Web *crawlers*, identificando os principais esquemas de actuação, a generalidade de propósitos actuais e os estratagemas usados para passar despercebidos; a definição de uma nova abordagem capaz de promover a detecção das suas actividades com precisão e em tempo útil; e, a manutenção de um repositório retrospectivo de todos os eventos de *crawling* detectados no sítio, o qual permite a análise comportamental deste tipo de agentes de software.

C. CONTRIBUIÇÕES

C.1 Principais contribuições técnico-científicas

Acima de tudo, a principal contribuição deste trabalho é a concepção de uma abordagem genérica de detecção, análise e contenção de actividades de *crawling*. De acordo com o tipo de sítio em questão e, consequentemente, com o tipo de *crawling* a que é submetido diariamente, uma abordagem desta natureza não só se agradece como é extremamente necessária. Dentre os pouquíssimos trabalhos de investigação publicados nesta área, é a primeira vez que se propõem uma abordagem com uma plataforma de implementação subjacente, em que há uma preocupação clara e efectiva de suportar todas as etapas do processo de um modo genérico, mas aberto a extensões que abarquem especificidades pontuais dos sítios. Paralelamente, asseguram-se repositórios de informação quer para actividades associadas às actividades de detecção quer para actividades de natureza analítica, os quais reportam informação válida quer para o estudo comportamental dos *crawlers* quer para estudos ditos convencionais.

C.2 Publicações

- Anália Lourenço e Orlando Belo. "Detection and Containment of Web Crawler Activities based on Clickstream Navigation Pattern Mining". Nas actas da Data Mining and Information Engineering 2006 - Seventh International Conference on Data, Text and Web Mining and their Business Applications and Management Information Engineering. Praga, República Checa. 11-13 Julho, 2006.

The Web is a very appealing arena for information gathering. Web crawlers are only some of the agents that visit it periodically for collect data according their sponsors' requirements. Usually, these activities impel the generation of additional clickstream and pattern data that will raise the necessity for extra processing and filtering. As we know, Web crawlers are not conventional Web users. However, some of them intentionally pretend to be so. Their requests flood Web server logs, preventing the discovery of real patterns and trends and forcing the application of additional filtering clickstreams procedures. In order to detect the influence of crawlers confusing usage pattern analysis, we designed and developed a clickstream crawler identification and filtering system: the ClickTips. It is an analysis platform integrating clickstream navigation pattern mining supported by data webhousing systems. This paper describes the system, as well as illustrating its applicability to major Web analysis problems: Web crawler detection and containment. Our main concern was to develop a platform that was generic enough to embrace the analysis of any given Web site, but also capable of integrating site-specific directives whenever seemed fittest. ClickTips supports major standard Web server log formats, provides basic processing in terms of host, user agent, referrer and request identification and several Web session reconstruction schemas. It models clickstream data into three distinct data webhouse grains, namely hit, session path and session grains, which provide a broad scope of site activities and allow the deployment of most frequent studies. In particular, it is possible to evaluate site usage as a whole as well as to perform usage profiling, analysing user general patterns or focusing specific user groups.

- Anália Lourenço e Orlando Belo. "Applying Clickstream Data Mining to Real-Time Web Crawler Detection and Containment". Nas actas da 30th Annual Conference of the German Classification Society (GfKl 2006) - Special Track on Web and Text Mining. Berlim, Alemanha. 8-10 Março, 2006.

Web crawler identification and analysis is quite relevant for preserving Web server performance and Web site privacy and copyrights. This is a widely recognised problem, although a very challenging one. There are few published papers in this area and standard detection heuristics face major limitations trying to cope with Web crawler constant evolving. In this paper, we propose a novel approach to the problem that tackles Web crawler real-time detection and containment based on clickstream data mining. This approach allows us to detect not only well-known Web crawlers, but also camouflaging and previously unknown Web crawlers, minimising potential server and site hazard. The work has two main goals: the enforcement of Web crawler detection and containment in real-time with high accuracy and the capture of Web crawler trends and purposes, gathering relevant knowledge about crawler evolution. In order to achieve such goals we designed and developed a specific object-oriented platform named ClickTips. This platform embraces three main components: the detection and containment component, the data webhousing component and the clickstream data mining component. The first component is based on a decision model that tags Web sessions on a hit-based inspection and on permanent Web server monitoring. Even when a camouflaged crawler manages to pass by undetected, Web server metrics will trigger an alert whenever server looses performance issuing for containment. Web visits are deviated to trap doors whenever they consume too many server resources, snoop around private contents or seem to be attacking the server. Web server shut down is issued only as a last option. On the other hand, the data webhousing component is in charge of clickstream processing, sustaining procedures such as: user, host, referrer and user agent identification, request parsing and session delimitation. As the idea is to distinguish Web crawler and human sessions, we choose a delimitation schema based on IP address and user agent information, outputting single-agent sessions. Hit and Web session data are arranged into a three fact table schema with hit, session sequence and session grains, respectively, and stored in a specific data webhouse. The clickstream data mining component acts over these contents, generating decision models based on crawler behaviour. These models sustain the convenient update of detection heuristics and capture deeper insights about Web crawler purposes and trends. The focus is set on classification techniques, although sequence mining and clustering techniques are also issued as potentially useful.
In order to sustain our claims and methodology, we have been performing several experiments with distinct real-world Web sites. Our experiments showed that it is possible to obtain highly accurate detection models with our platform and that these models are able to detect crawler activities at web site visiting earliest stage, minimizing potential server overload or site privacy or copyrights violation.

- Rita Serra, Anália Lourenço, Orlando Belo e Armando Venâncio. "Portuguese regional differences in the mycoflora of Portuguese wine grapes – a 3 year-study". International Workshop Ochratoxin A in Grapes and Wine: Prevention and Control. Marsala, Itália. 21-22 Outubro, 2005.

The mycoflora of healthy ripe berries from Portuguese winemaking regions was studied between the years 2001 and 2003. Four regions were selected: Alentejo, Douro, Ribatejo and Vinhos Verdes, located in the southeast, northeast, southwest and northwest of the Portuguese mainland, respectively. All the regions have Mediterranean climates apart from Vinhos Verdes, which is Submediterranean, more humid than the other regions due to Atlantic influences. Eleven vineyards were analysed (2 to 3 vineyards per region), and a total of 32 grape samples were taken, of 50 berries each. The mycoflora of grapes was evaluated by plating methods. During this study a total of 3623 strains were isolated and identified to genus level. The Aspergillus and Penicillium strains were identified to species level. The differences in the mycoflora of grapes between the 4 regions were analyzed using the non-parametric test Kruskal-Wallis H. Ostensibly, the classification of the grapes into their geographical origin based on its mycoflora was attempted using a decision tree algorithm (C4.5) based on the Shannon Information Theory. The success of the models to classify and predict the region of origin of the samples was compared. Furthermore, due to the increasing interest on the presence of A. carbonarius in grapes due to ochratoxin A (OTA) production, the relationship between the presence of this species in grapes and the remaining mycoflora was studied trough the Spearman correlation coefficient (rs).Twenty-seven (27) genera of fungi were identified. The most frequent genera in grapes by descending order were Cladosporium, Botrytis, Alternaria, Aspergillus, Penicillium, Aureobasidium, Rhizopus, Epicoccum and Trichothecium, with a mean frequency in the samples of 54%, 36%, 36%, 34%, 32%, 10%, 8%, 6% and 2%, respectively. The mean frequency of the remaining 18 genera in the samples was below 2%. Three genera varied its incidence significantly according to the region of origin of the samples: Aspergillus, Botrytis and Ulocladium. The mean incidence of Aspergillus and Botrytis in Vinhos Verdes samples was significantly lowest and highest than in the other regions, respectively. Ulocladium was significantly higher in Alentejo than in the other regions. The 524 Aspergillus strains identified belonged to 14 species. The most frequent were by far black aspergilli, namely A. niger aggregate (79% of the isolated strains) and A. carbonarius (13%). The only species that varied significantly its frequency between regions was A. niger aggregate, in the same way as the genus Aspergillus. The 446 Penicillium strains identified belonged to 25 species. The most frequent were P. brevicompactum (32%), P. thomii (29%) and P. glabrum/spinulosum (14%). Six Penicillium species differed significantly between regions. Using decision trees it was possible to classify successfully 91% of the samples according to 3 sample classes: Vinhos Verdes, Douro and South samples (Alentejo and Ribatejo). In the model it was used as decision criteria the low incidence (≤ 8%) of A. niger aggregate in grapes to classify the Vinhos Verdes samples and the highest incidence of P. thomii (> 4%) in Douro grape samples to separate this region from the South, but it could not discriminate accurately between the two south regions, Alentejo and Ribatejo. The estimated predictive ability of the model in the 3 classes was 82%.Regarding the OTA producing species A. carbonarius, its presence varied between regions, but not significantly with the statistics tests used. The strongest significant correlation found between A. carbonarius and the remaining species was with the A. niger aggregate ($r_s = 0.539$, $P<0.001$). The data presented here indicate that grapes are consistently exposed to a particular mycoflora that varies according its geographical origin. The positive significant correlation of A. carbonarius with the low OTA producing A. niger aggregate species suggests that the A. niger aggregate presence in grapes may indicate the presence of the rarer species A. carbonarius.

- Rita Serra, Anália Lourenço, Orlando Belo e Armando Venâncio. "Portuguese regional differences in the wine grapes mycoflora". Abstract Book of One Day Meeting "Trends in Mycology", Micoteca da Universidade do Minho. 23 Setembro de 2005.

The recent discovery of mycotoxins in wine, in particular ochratoxin A, caused concern and motivated an extensive survey to the mycoflora of Portuguese grapes. It is known that the mycoflora of agricultural commodities can vary according to the geographical origin, and therefore, regional differences in the mycoflora of Portuguese were investigated. Four regions were selected for a 3-year study: Alentejo, Douro, Ribatejo and Vinhos Verdes. The mycoflora of grapes was evaluated by plating methods. A total of 32 grape samples were taken, of 50 berries each. The differences in the mycoflora of grapes between regions were analyzed using the non-parametric test Kruskal-Wallis H. Ostensibly, the classification of the grapes into their geographical origin based on its mycoflora was attempted using a decision tree algorithm (C4.5) based on the Shannon Information Theory. Of the 27 fungal genera identified, 3 varied its incidence significantly according to the region of origin of the samples: Aspergillus, Botrytis and Ulocladium. The only species that varied significantly its frequency between regions was A. niger aggregate. Six Penicillium species differed significantly between regions: P. brevicompactum, P. citrinum, P. glabrum/spinulosum, P. expansum, P. implicatum and P. thomii. Using decision trees it was possible to classify successfully 91% of the samples according to 3 sample classes: Vinhos Verdes, Douro and South samples (Alentejo and Ribatejo).
The classification was based on the incidence of A. niger and P. thomii in the grape samples. The estimated predictive ability of the model in the 3 classes was 82%.
The data presented here indicate that grapes are consistently exposed to a particular mycoflora that varies according its geographical origin, which may be of importance to establishing risk areas for mycotoxin contamination of grapes and wine.

- Anália Lourenço e Orlando Belo. "When the Hunter Becomes the Prey – Tracking down Web Crawlers in Clickstreams". Nas actas 1ª Data Gadgets Workshop – Bringing Up Emerging Solutions for Data Warehousing Systems –, realizada em conjunto com a IX Conference on Software Engineering and Database (JISBD'04). Malaga, Spain. 9 de Novembro, 2004.

Clickstreams are the latest acquisition of decision support systems. They are an amazing opportunity in terms of analysis, opening up the area of usage profiling. However, there are Web-related specificities and issues that have to be taken care of meanwhile. In particular, in order to perform proper usage profiling it is necessary to differentiated conventional users from non-conventional users – specifically, Web crawlers. By definition, they are information hunters that traverse the Web trying to perform some task, like gathering and indexing information about some topic. Usually, their visits to a Web site are brief, but intensive, becoming a routine over time. Clickstreams mixture the activities concerning real users with the ones related to Web crawlers. Therefore, any usage study requires some pre-processing to filter, or at least tag, Web crawlers' activities, preventing that these might mislead any further analysis. After delimiting their activities, it is possible to deploy regular Web usage analysis as well as Web crawler analysis. With the latest, one can conduct deeper studies about Web crawlers' behaviours and purposes, learning to distinguish between harmless crawlers and pervasive ones, and making purpose-based clustering in order to establish communities of Web crawlers. In this paper, we review Web crawlers' primary characteristics, research in the area of crawler detection and pattern analysis and present one case study to highlight the relevance of the task and the use of certain heuristics.

- Anália Lourenço e Orlando Belo. "Attenuating the effect of data abnormalities on Data Warehouses". Nas actas da 6ª International Conference on Enterprise Information Systems (ICEIS 2004), pp. 411-415. Porto, Portugal. 14-17 Abril, 2004.

Today's informational entanglement makes it crucial to enforce adequate management systems. Data warehousing systems appeared with the specific mission of providing adequate contents for data analysis, ensuring gathering, processing and maintenance of all data elements thought valuable. Data analysis in general, data mining and on-line analytical processing facilities, in particular, can achieve better, sharper results, because data quality is finally taken into account. The available elements must be submitted to an intensive processing before being able to integrate them into the data warehouse. Each data warehousing system embraces extraction, transformation and loading processes which are in charge of all the processing concerning the data preparation towards its integration into the data warehouse. Usually, data is scoped at several stages, inspecting data and schema issues and filtering all those elements that do not comply with the established rules. This paper proposes an agent-based platform, which not only ensures the traditional data flow, but also tries to recover the filtered data when an data error occurs. It is intended to perform the process of error monitoring and control automatically. Bad data is processed and eventually repaired by the agents, integrating it again into the data warehouse's regular flow. All data processing efforts are registered and afterwards mined in order to establish data error patterns. The obtained results will enrich the wrappers knowledge about abnormal situations' resolution. Eventually, this evolving will enhance the data warehouse population process, enlarging the integrated volume of data and enriching its actual quality and consistency.

- Anália Lourenço e Orlando Belo. "Abnormal data formats identification and resolution on Data Warehousing populating process". Nas actas do 4º ICSC Symposium on Engineering of Intelligent Systems (EIS 2004). Funchal, Madeira, Portugal. 29 Fevereiro-2 Março, 2004.

Data Warehousing systems are perhaps one of the most valuable assets that organisations possess today. They manage and sustain crucial, strategic information, granting the urging decision support. However, before taking advantage of this magnificent resource, there has to be set a plan to ensure its population. The process of extracting, transforming and loading data into the data warehouse is anything less straightforward. These scenarios are inherently heterogeneous. The idea of gathering every piece of information that is available and thought useful brings along different data models and data schemas to conciliate. Besides, within each single source, it is likely that several kinds of conflicts, inconsistencies and errors pump up. Therefore, tools capable of identifying and resolving these situations are in order.
This paper aims to bring some light into the subject, covering basic issues related with data cleaning, as well as, proposing a new computational platform - an agent-based abnormal data formats identification and resolution platform. The aim was set on assisting the process, learning from past experiences and thus, evolving wrappers knowledge about abnormal situations' resolution. Eventually, this evolving will enhance the data warehouse population process, enlarging the integrated volume of data and enriching its actual quality and consistency.

- Anália Lourenço e Orlando Belo. "Qualifying Web data to meet expectations". Nas actas da 4ª International Conference on Data Mining including Building Applications for CRM & Competitive Intelligence (Data Mining 2003), pp. 401-410. Rio de Janeiro, Brasil. Dezembro, 2003.

The World Wide Web is today's richest medium of information and services. Interchanging data freely "without" any time constraints is much more than appealing. However, the Web has to be restructured in order to address today's needs and expectations, favouring users' searches and navigation and, to do so, Web analysis tools are in the order of the day. There is a similarity between the analytical Web data processing and the analytical processing of "conventional" data. The truth is that the enormous volume of data along with the high transactional activity strongly prohibits the use of raw data. In such scenario, Web server logs emerge as main sources of information and there are many Web analysis tools today that work over these raw elements. However, the obtained results are far from being satisfactory, since they do not help to understand users' navigational patterns. In a near future, data webhouses will be in charge of any analytical work, requiring well-defined extraction, transforming and loading processes, capable of reuniting all relevant data and ensuring its consistency and quality. Only when these processes are implemented there is room for thinking about Web Mining. This paper aims to target the problem of establishing data webhouses as primary Web Mining sources, highlighting the excellence of the Web domain towards mining and defining the guidelines that populating processes must comply to. In this sense, there will be presented a case study, identifying its available data sources, establishing a proper data webhouse and specifying all the processing required to populate it.

- Orlando Belo, Anália Lourenço e J. Santos. "Profiling in an augmented-commerce environment". Nas actas da 4ª International Conference on Data Mining including Building Applications for CRM & Competitive Intelligence (Data Mining 2003), pp. 377-385. Rio de Janeiro, Brasil. Dezembro, 2003.

Retail systems are changing. Day after day customers require new forms of services, demanding better shopping attendance and support. The Grocer project brings to real world retail scenarios an easy-to-use augmented-commerce platform, combining mobility, efficiency, shopping commodity, availability of services "in-hand" with a 24 hours a day service assistance in shopping platforms. One of the most promising areas inside Grocer's efforts is profiling. Grocer already provides services and means to help store managers in daily management and customer satisfaction services, developed over advanced mobile communication infrastructures and sophisticated promotion and selling mechanisms based on agent technology. However, we need to provide Grocer with new agents having abilities to extract knowledge about customers, establishing their sales' patterns, which makes possible prediction of future trends in customer's sales and shopping activities. In this paper, we will focus on aspects concerned with customer profiling generation inside Grocer system, giving particular attention to the problem of data gathering and preparation inside the system, and presenting a knowledge discovery classification model that will be implemented in a specific class of profiling agents.

- Anália Lourenço e Orlando Belo. "Assessing web usage profiles". Nas actas da IADIS International Conference WWW/Internet 2003. Algarve, Portugal. Novembro, 2003.

Today the World Wide Web is seen as an unique medium for interaction. Every organisation has or is about to have its own Web site, aiming to attract as much visitors as each one cans. So, it is no surprise that Web analysis is in the order of the day. Competitiveness rises up the need for deep customers' understanding, otherwise, sites' impact will be low and non profitable. This kind of knowledge requires for specialised analysis skills facilitated by Web Usage Mining. This area, a part of the Web Mining research area, is set to provide insights on personalisation, business intelligence, usage characterisation and system and site improvement. This paper aims to clear out some of the concerns and availabilities that the Web presents towards this analysis area, highlighting its main excellences and pitfalls. A case study concerning an educational support site is studied. Intentionally, there were chosen a common Web analysis tool and a Web Usage Mining tool in order to assess what kind of knowledge each one of them is able to produce. Finally, some comments are made both upon tools achievements and the potential of the available informational resources.

- Anália Lourenço e Orlando Belo. "Promoting Agent-Based Knowledge Discovery in Medical Intensive Care Units". Nas actas da 3ª WSEAS International Conference on Information, Science and applications. Malta. Setembro, 2003.

Decision support has become vital in major scientific and organisational areas. Data analysis and knowledge acquisition are no longer means to an end, they have become the end itself. Knowledge discovery is a priority, constantly demanding for new, better suited efforts. Systems or tools capable of dealing with the steadily growing amount of data presented by information systems and digging through their structure, relations, contexts and ultimately, their contents, are in order. In this paper, an agent-based decision support system based in data warehousing systems contents is described, analysing its actual potential and skills in knowledge discovery as a whole and, in particular, in data preparation for data mining. Different requirements, algorithm dependent and non-dependent, are pointed, analysing valid approaches. An algebraic model is used to specify the core of some of the embraced steps, as well as, to demarcate the data flow throughout the knowledge acquisition process. A case study is presented in order to demonstrate the application of the system to real world scenarios. The application area concerns neonatal intensive care units, choosing a neonatal mortality data sample concerning very low birthweight newborns. A large number of variables influencing the survival of newborns is taken under consideration, and decision-making has to be quick and sharp. Thus, an automatic or semi-automatic system capable of applying different mining approaches, combining efforts and absorbing the most relevant knowledge presents a valuable advantage in this area.

- Anália Lourenço, Eurico Borges e Orlando Belo. "Evaluating the Impact of Information Sources Heterogeneity inside Data Webhousing Systems". Nas actas da 7ª WSEAS International Conference on Circuits, Systems, Communications and Computers. Corfu, Grécia. Julho, 2003.

People use to access electronic commerce sites, searching for goods and services and making their regular shops. It is a very common situation today. The commodity and advantages that electronic commerce sites brought transformed them into an essential and permanent services provider to common people. However, day after day, people become more demanding, requesting more quality, user-friendly functionalities, and efficiency in all kinds of services provided by such sites. In order to answer appropriately, organizations need to have new and effective means to gather information about their sites' users, giving the chance to follow their preferences and tendencies. Data webhousing systems help organizations in the hard task of daily decision support activities, providing information about user's preferences and navigation paths that will be very useful in the improvement and effectiveness of their sites. This paper presents a general overview of the different types of data that we can gather in electronic commerce sites and integrate into a data webhouse. Additionally, it is given an extensive description of the extraction, transformation and loading process of a data webhouse, identifying its main tasks and problems, as well as highlighting the most used approaches to work around them. The overall idea is evaluating the impact of information sources heterogeneity inside Data Webhousing Systems, assessing what can be expected and what can be accomplished.

- Anália Lourenço, Ana Cristina Braga e Orlando Belo. "Mortality Risk Classification for Decision Support in Neonatal Critical Care". Nas actas da IEEE International Conference on Systems, Man and Cybernetics (SMC 2004). Hammamet, Tunísia. Outubro, 2002.

Feature selection is perhaps one of the most challenging topics embraced by the Knowledge Discovery pre-processing phase, but it is also one of the most complex ones. After acknowledging that noisy, contaminated, irrelevant and redundant data jeopardize the outcome of the process, issues rose by large sets of features and data overload have gained increasing relevance. In this paper, it is given a description of the problem of data preparation as a whole and it is presented a study of some of the available feature selection techniques (both wrapper and filter approaches), applying them to a knowledge discovery classification task. The case of study concerns the evaluation of several factors that might influence or even cause strokes. Specifically, there were considered four different types of strokes, and analysed the predominant factors for each case. Patients clinical history concerning diabetes, cardiac diseases, previous occurrence of strokes, high cholesterol levels, high blood pressure and kidney insufficiencies data, among others, were taken into account. Additionally, other features like gender and age were also included. Basically, the aim was to build an accurate classification model, in order to gather some kind of advice on clinical scenarios. Physicians require both precision and comprehensibility. So, the model has to expose the facts with rigour and conviction, but it has also to be simple enough in order to be understood, studied and, ultimately, applied. In this sense, feature selection becomes quite important, as it allows and urges the detection of irrelevant or dependent features, forcing a dimensionality reduction. Normally, a classification problem involving fewer features implies a simpler model, which can be ideal (an optimal commitment between precision and comprehensibility) if the most relevant features are selected. In this particular case, the model must indicate which factors tend to cause or are associated to the occurrence of each type of stroke. It is interesting to detect not only irrelevant features, but also relations or dependences among features, trying this way to assess their weight in the final outcome - their relevance in terms of the prediction of a certain stroke type.

- Anália Lourenço, Ana Cristina Braga e Orlando Belo. "Strokes Risk Factors Classification Modelling". Nas actas da 3ª International Conference on Data Mining Methods and Databases for Engineering, Finances and Other Fields (Data Mining 2002), pp. 187-196. Bolonha, Itália. Setembro, 2002.

Feature selection is perhaps one of the most challenging topics embraced by the Knowledge Discovery pre-processing phase, but it is also one of the most complex ones. After acknowledging that noisy, contaminated, irrelevant and redundant data jeopardize the outcome of the process, issues rose by large sets of features and data overload have gained increasing relevance. In this paper, it is given a description of the problem of data preparation as a whole and it is presented a study of some of the available feature selection techniques (both wrapper and filter approaches), applying them to a knowledge discovery classification task. The case of study concerns the evaluation of several factors that might influence or even cause strokes. Specifically, there were considered four different types of strokes, and analysed the predominant factors for each case. Patients clinical history concerning diabetes, cardiac diseases, previous occurrence of strokes, high cholesterol levels, high blood pressure and kidney insufficiencies data, among others, were taken into account. Additionally, other features like gender and age were also included. Basically, the aim was to build an accurate classification model, in order to gather some kind of advice on clinical scenarios. Physicians require both precision and comprehensibility. So, the model has to expose the facts with rigour and conviction, but it has also to be simple enough in order to be understood, studied and, ultimately, applied. In this sense, feature selection becomes quite important, as it allows and urges the detection of irrelevant or dependent features, forcing a dimensionality reduction. Normally, a classification problem involving fewer features implies a simpler model, which can be ideal (an optimal commitment between precision and comprehensibility) if the most relevant features are selected. In this particular case, the model must indicate which factors tend to cause or are associated to the occurrence of each type of stroke. It is interesting to detect not only irrelevant features, but also relations or dependences among features, trying this way to assess their weight in the final outcome - their relevance in terms of the prediction of a certain stroke type.

D. ENQUADRAMENTO

D.1 Enquadramento Científico

Dadas as dimensões actuais da rede, é inviável deixar ao acaso a promoção e a divulgação dos sítios. Não é suficiente deter informação, é imprescindível saber localizar essa informação sempre que algum utilizador a procure, o que implica, inevitavelmente, uma aposta na recolha e estruturação de índices relativos aos conteúdos da rede. Neste sentido, os motores de pesquisa e os directórios fazem já parte do quotidiano dos utilizadores, assistindo-os na descoberta dos sítios mais adequados aos seus objectivos. A implantação deste tipo de sistemas, com *interfaces* declaradamente gráficas e *user-friendly*, tem contribuído significativamente a melhorar a situação. Porém, o problema está longe de estar resolvido, pois nenhum motor de pesquisa ou directório é capaz de cobrir todo o espectro da rede, indexando os conteúdos somente de uma determinada porção, mais ou menos vasta consoante os seus propósitos e as suas possibilidades. A maioria dos motores de pesquisa recorre a algoritmos e técnicas de Recuperação de Informação (RI) e diversos estudos estimam que a abrangência destes sistemas varia entre os 5% e os 30% sensivelmente, sendo que o conjunto dos 11 maiores motores de pesquisa actuais cobre menos de metade do espaço de pesquisa da rede. Neste sentido, o Web *crawling* é apontado como a grande aposta da RI em termos do cenário Web, constituindo uma abordagem inovadora de perscruta e recuperação automática de informação.

Presentemente, os Web *crawlers* constituem uma das comunidades mais activas da rede, encontrando-se envolvidos em praticamente todas as actividades de recolha automática de conteúdos. A simplicidade do conceito e a facilidade de implementação deste tipo de programas tornou-os extremamente populares e rapidamente os seus objectivos ultrapassaram os propósitos definidos pelos motores de pesquisa. De acordo com a literatura da área, os Web *crawlers* são responsáveis pela recuperação de inúmeros recursos da rede, desde dados com vista à manutenção de índices de pesquisa ao *browsing offline*, passando pela validação de hiper-ligações e esquemas DTD, a monitorização de alterações em páginas Web ou a duplicação de sítios (*mirroring*). Novos tipos de programas têm sido desenvolvidos com vista a actividades ditas focalizadas, destinadas às mais variadas áreas de negócio e investigação.

A relevância destes programas no cenário actual da rede é inquestionável, mas, mais uma vez, a sua evolução não foi suficientemente pensada: não só não se regulamentou as suas acções como

também nunca se procedeu à sua devida catalogação. Sempre se pressupôs que os programas criados seriam éticos, ou seja, respeitariam as regras de acesso aos sítios, preservariam os recursos da rede e dos sítios e se auto-identificariam. Contudo, a realidade dista muito deste propósito, existindo já relatos de incidentes relacionados com a sobrecarga da rede e de servidores Web e a violação de direitos de autor e privacidade, entre outros. De facto, embora a maioria dos Web *crawlers* actuais tenha propósitos lícitos e benéficos para os sítios, apenas uma minoria se auto-identifica, convertendo a detecção das suas visitas, particularmente enquanto estas ainda estão a decorrer, em tudo menos numa tarefa trivial.

O ritmo acelerado de início de actividade. e adaptação de crawlers sugere a aplicação de técnicas de análise avançadas a este problema, designadamente técnicas de Mineração de Perfis de Utilização Web. Dado não ser praticável manter um catálogo actualizado dos programas em activo, resta-nos aprender a detectá-los com base na informação fornecida pelas sessões Web. Deste modo, poder-se-á identificar crawlers já conhecidos, bem como, aprender a reconhecer crawlers até então desconhecidos e crawlers que procuram passar despercebidos. Os esforços de controlo têm sido débeis e ineficientes, sendo muito difícil etiquetar os *user agents* (isto é, as *strings* de identificação de programas usados no acesso ao sítios) como sendo exclusivos de Web *crawlers* ou Web *browsers* ou, pior ainda, distinguir quando um mesmo user agent é usado por cada um dos grupos de programas. Na realidade, mesmo se fosse praticável a análise manual dos *user agents* (algo impensável dada a diversidade existente), muitas vezes a dúvida continuaria a persistir, pois muitos dos crawlers actuais não só não se auto-identificam como procuram fazer-se passar por utilizadores convencionais simulando os seus padrões de navegação.

D.2 Motivação

O estudo das actividades de *crawling* é relativamente recente e existem muito poucos investigadores a trabalhar na área. Quando se efectua uma pesquisa bibliográfica acerca dos Web *crawlers* constata-se que as linhas de investigação estão normalmente ligadas ao desenvolvimento de novos programas e à adaptação de programas a tarefas específicas. Há muito poucos trabalhos publicados que realmente versem o estudo do comportamento dos *crawlers* ao nível dos sítios. Os estudos existentes relatam o desempenho dos *crawlers* em distintas tarefas, comparam diferentes técnicas de *crawling* e ilustram possíveis implementações, ou seja, instruem acerca de como construir e afinar programas deste tipo. Poucos destes estudos reflectem o impacto que estas actividades têm sobre um sítio e muito menos analisa a ética dos *crawlers* actuais. Assim, muito embora se constate que a maioria destes programas não representa uma ameaça para os sítios e a rede, a verdade é que

não há garantias neste sentido nem planos de contenção. Quando ocorre um ataque, seja ele fortuito ou intencionado, o sítio não se encontra preparado para o anular, sendo que a maior parte das vezes o ataque só é detectado quando o servidor ou a rede já sofreram danos consideráveis.

Assim sendo, a relevância da análise dos perfis de navegação dos Web *crawlers* assenta em dois motivos primordiais: a regulamentação das suas actividades nos sítios, prevenindo potenciais ataques; e, a diferenciação das suas visitas das visitas dos utilizadores regulares, tornando a análise comportamental a nível de ambos grupos de utilizadores muito mais focada e próxima da realidade. A detecção e a contenção de actividades de *crawling* assegurará o direito dos sítios a determinar quais os *crawlers* que aceitam receber e quais as áreas dos sítios passíveis de perscruta. Paralelamente à política de conduta, permitirá também uma monitorização constante que garanta a efectividade das restrições impostas e a identificação de programas que forçam tais restrições usando identificações falsas. Genericamente, preservar-se-á os recursos dos servidores Web, a sua operacionalidade e a da própria rede, dando garantias aos sítios que aceitam este tipo de programas e dando credibilidade o trabalho dos próprios *crawlers*. Implicitamente, estar-se-á a aumentar a satisfação e a segurança dos utilizadores regulares, visto que será possível restringir o consumo abusivo de recursos dos servidores, privilegiando as visitas destes utilizadores em detrimento das visitas dos Web *crawlers*. Por outro lado, a análise comportamental diferenciada permite capturar de uma forma muito mais precisa as tendências e propósitos dos diferentes tipos de utilizadores. Assim, poder-se-á desenvolver campanhas de *marketing*, personalizar os sítios ou até mesmo planear a sua reestruturação tendo em mente as necessidades dos utilizadores convencionais. Do mesmo modo, estudando as suas visitas isoladamente, aprender-se-á muito mais acerca dos Web *crawlers*, proporcionando um retrato fiel do cenário actual e contribuindo para a sua detecção e contenção, sempre que necessário.

D.3 Objectivos Detalhados

A lista de objectivos estipulados por ordem cronológica de definição e persecução é a seguinte:

- O estudo dos principais esquemas de *crawling* actuais, analisando os padrões de navegação resultantes e o impacto causado sobre os sítios;
- A manutenção de um catálogo dos *crawlers* actuais - as suas identificações (*user agents*) e, sempre que possível, os seus propósitos - procurando tais informações ao nível de entidades oficiais e fóruns de debate da especialidade;

- A definição de um esquema primário de detecção automática deste tipo de programas, capaz de providenciar um conjunto inicial de dados que permita desenvolver o processo de análise;
- A concepção e a implementação de um sistema de *data webhousing*, assegurando as etapas básicas de processamento de fluxos de *clicks*, visando a construção de um *data webhouse* específico às tarefas de Classificação associadas à actualização do esquema de detecção e *data webhouses* diferenciados de Web *crawlers* e utilizadores convencionais sobre os quais é possível realizar análises comportamentais;
- A criação de uma abordagem semi-automática de construção de conjuntos de dados de treino para Classificação baseada na informação obtida através do esquema primário de detecção e na inspecção dos dados por parte do analista;
- A aplicação de técnicas de Classificação, designadamente a construção de árvores de decisão, ao problema da detecção em tempo útil das visitas dos Web crawlers;
- A actualização do esquema de detecção com base as informações obtidas da mineração da informação relativa às visitas do sítio;
- A implementação de esquemas de desvio de visitas como medida de contenção de visitas consideradas ilícitas ou que de algum modo põem em perigo o bom funcionamento do sítio e do servidor Web;
- E, o estudo comportamental dos Web *crawlers*, procurando capturar os principais padrões de navegação e distinguir os diferentes propósitos das suas acções.

D.4 Trabalhos Alternativos

Tal como já foi referido anteriormente, existem pouquíssimos trabalhos de investigação publicados na área da detecção e interpretação das actividades de *crawling*. Existem uma infinidade de trabalhos associados às diversas técnicas de *crawling* e sua aplicação à distintos domínios, mas quase não existem referências na área de detecção e análise de actividades de *crawling*. Neste sentido, e de acordo com a orientação que a investigação que suporta esta dissertação tomou, o único trabalho alternativo que faz sentido referir foi realizado por Pang-Ning Tan e Vipin Kumar do Departamento de *Computer Science* da Universidade de Minnesota. Em linhas gerais, a abordagem destes investigadores serviu de inspiração ao doutorando, em particular, no que diz respeito à tarefa de mineração. Contudo, o seu trabalho não refere a criação ou sequer pretensão de criação de uma plataforma de implementação nem a tentativa de generalizar a abordagem descrita para aplicação em

distintos sítios. Mais ainda, as preocupações em termos de processamento de fluxos de *clicks* são consideravelmente distintas, por vezes, talvez contrárias.

D.5 Bibliografia Principal

- G. Chang, M. J. Healy, J. A. M. McHugh e J. T. L. Wang. "Mining the World Wide Web: an Information Search Approach". Kluwer Academic Publishers. 2001.
- Oren Etzioni. "Moving up the information food chain: Deploying softbots on the world-wide web". In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI96), pp. 1322-1326. Portland, OR, Estados Unidos da América. 1996.
- Oren Etzioni. "The World-Wide Web: Quagmire or Gold Mine?". In Communications of the ACM 39(11), pp. 65-68. 1996.
- Lan Huang. "A survey on web information retrieval technologies". Technical report, ECSL. 2000.
- G. W. Flake, S., Lawrence, and C. L. Giles. "Efficient identification of web communities". In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000), pp. 150-160. Boston, MA, USA. Agosto, 2000.
- R. Ford e H. Ray. "Googling for gold: Web crawlers, hacking and defense explained". Network Security, Volume 2004, Issue 1, pp. 10-13. Janeiro, 2004.
- R. Kimball e R. Merz. "The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse". Wiley Press. 2000.
- Mei Kobayashi e Koichi Takeda. "Information Retrieval on the Web". ACM Computing Surveys, vol. 32, Num. 2, pp. 144-173. 2000.
- Ron Kohavi. "Mining E-commerce Data: the Good, the Bad, and the Ugly". Invited talk at PAKDD 2001. Hong Kong, China. Abril, 2001.
- Ron Kohavi e Foster Provost. "Applications of Data Mining to Electronic Commerce". Data Mining and Knowledge Discovery journal, 5(1/2). 2001.
- Martijn Koster. "Evaluation of the Standard for Robots Exclusion". 1996. http://www.robotstxt.org/wc/eval.html, Novembro de 2005.
- Martijn Koster. "Robots in the Web: threat or treat?". ConneXions, Volume 9, No. 4. Abril, 1995. http://www.robotstxt.org/wc/threat-or-treat.html, Novembro de 2005.
- Martijn Koster. "Guidelines for robots writers". 1993. http://www.robotstxt.org/wc/guidelines.html, Novembro de 1995.

- G. Pant, P. Srinivasan e F. Menczer. "Crawling the Web". In M. Levene and A. Poulovassilis (Eds.), Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer-Verlag. 2004.
- G. Pant, S. Bradshaw e F. Menczer. "Search engine-crawler symbiosis: Adapting to Community Interests". In Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries. Trondheim, Noruega. Agosto, 2003.
- Pang-Ning Tan e Vipin Kumar. "Discovery of Web Robot Sessions based on their Navigational Patterns". Data Mining and Knowledge Discovery, 6(1):9-35. 2002.
- Pang-Nin Tan e Vipin Kumar. "Modeling of Web Robot Navigational Patterns". Nas actas da ACM WebKDD Workshop 2000. 2000.
- Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". Em ACM SIGKDD Explorations Newsletter, (1) 2, pp. 12-23. Janeiro, 2000.
- Mark Sweiger, Mark R. Madsen, Jimmy Langston e Howard Lombard. "Clickstream Data Warehousing". Wiley Computer Publishing, John Wiley & Sons, Inc. 2002.

E. DESENVOLVIMENTO

E.1 Macro-planeamento das Actividades

No sentido de realizar com sucesso, e de forma efectiva, o projecto de doutoramento apresentado no ponto anterior, o plano de trabalhos foi organizado em cinco fases distintas:

- 1ª Fase.
  Estudo da área dos sistemas de data webhousing, análise do "estado-da-arte" deste tipo de sistemas e aquisição de conhecimento sobre a implementação e manutenção de *clickstreams* de sítios Web de comércio electrónico.
  Duração: 6 meses

- 2ª Fase.
  Análise, estudo e definição de modelos de extracção de conhecimento para sistemas de dados heterogéneos de um sistema de *data webhousing*. Estudo da adequação dos modelos desenvolvidos sobre cenários de aplicação real.
  Duração: 12 meses

- 3ª Fase.
  Desenvolvimento das bases de conhecimento e agentes de processamento relacionados com os modelos de extracção definidos durante a etapa anterior, implementação do sistema de extracção de conhecimento e sua integração num sistema de data *webhousing* real.
  Duração: 12 meses

- 4ª Fase.
  Implementação de uma plataforma de monitorização da actividade dos servidores Web que serve de suporte à detecção e à contenção de actividades de *crawling* com base no impacto causado.
  Duração: 6 meses

- 5ª Fase.

  Análise e preparação de amostras de dados real para teste e validação da plataforma em termos globais, bem como, das sub-plataformas independentemente. Análise dos resultados obtidos.

  Duração: 6 meses

- 6ª Fase.

  Escrita da tese de doutoramento.

  Duração: 6 meses

E.2 Recursos Necessários

No decurso da investigação realizada houve a preocupação de recorrer, sempre que possível, a ferramentas *free* e open-*source*. Dada a necessidade óbvio de sistemas de bases de dados e de ferramentas estatísticas, de mineração de dados e que possibilitassem a construção de gráficos, a ideia era manter a plataforma o mais independente e portável que fosse possível. Era relevante para os investigadores assegurar a implantação da plataforma em meios variados, sem condicionantes de maior.

Obviamente, caso se pretendesse a avaliação da plataforma tendo por base Sistemas de Gestão de Bases de Dados comerciais, seria necessária a licença de utilização de plataformas tais como Microsoft SQL Server e Oracle. De igual forma, seria interessante a disponibilização de licenças de utilização de ferramentas de *Data Mining* comerciais tais como SPSS Clementine, IBM Intelligent Miner ou SAS Enterprise Miner.

E.3 Recursos Disponibilizados

No decurso das actividades de doutoramento, foi fornecido pelo Departamento de Informática, uma máquina que serve de bancada de trabalho.

F. AVALIAÇÃO

F.1 Avaliação Comparativa

Comparativamente, o trabalho desenvolvido abre um novo horizonte de investigação no que diz respeito à detecção e à análise de perfis de programas de *crawling*. Em termos das áreas científicas que suportam o trabalho, a sua aplicação a este domínio permitiu reportar mais uma área de aplicação, bem como, identificar as especificidades envolvidas e equacionar a respectiva resolução. O trabalho é independente de casos de estudo ou *softwares*, garantindo uma aplicabilidade, uma flexibilidade e uma inter-operabilidade até ao momento nunca alcançadas, ou pelo menos, nunca referidas.

F.2 Auto-Avaliação do Trabalho Realizado

Há três vertentes principais a salientar no trabalho realizado, as quais se reflectem nas correspondentes plataformas: a plataforma de monitorização e contenção de actividades de *crawling*, a plataforma de processamento diferenciado de fluxos de *clicks* e a plataforma de análise de dados Web. Os objectivos básicos impostos a cada um destes níveis foram cumpridos, mas há a consciência que quer a monitorização de servidores Web quer a análise de dados Web podiam e deviam ser estendidas, recorrendo a investigação de vanguarda e, eventualmente, ao enriquecimento dos pacotes de software implementados.

F.3 Auto-Avaliação da Documentação Produzida

A documentação produzida procurou relatar e, em certa medida cunhar, o trabalho desenvolvido. Dada a ausência de investigação corrente nesta área, por vezes, é necessário um balanceamento entre a contextualização do trabalho e o trabalho propriamente dito o que retira extensão aos relatos pretendidos. De igual modo, a escolha dos fóruns para onde encaminhar tais trabalhos requereu algum cuidado extra, pois constatou-se que em fóruns exteriores ao domínio Web a sensibilidade à problemática é bastante pequena e manifestamente insuficiente.

G. DIFICULDADES

G.1 Dificuldades Técnico-Científicas

A maioria das dificuldades que se fizeram sentir prendeu-se com o processamento de fluxos de *clicks*. Este facto não é inusual, pois o processamento de dados é sempre uma das áreas em que a aquisição de sensibilidade passa por uma experiência prática considerável. Aprender a reconhecer potenciais pontos de conflito ou obstáculos ao processamento dos fluxos de *clicks* levou tempo e implicou o manuseamento de distintos *logs*. No decurso deste processo de aprendizagem adquiriu-se também conhecimentos suficientes para a asserção das técnicas mais apropriadas para os distintos casos, antevendo riscos e resultados e assumindo as limitações inerentes.

G.2 Outras Dificuldades

Uma das maiores dificuldades encontradas foi a inexistência de recursos económicos que suportem a assistência e a apresentação de trabalhos em fóruns de discussão relevantes, algo que restringe a amplitude de escolha ao nível de eventuais publicações e deslocações.