Personalization of information delivery based on web mining and DNA classification

M. Santos[1] & J. Amado[2]
[1]*Department of Information Systems, University of Minho, Portugal*
[2]*Department of Information Systems, Court of Auditors, Portugal*

Abstract

It is hard to find any kind of media with a growth-rate as high as the World Wide Web. At the same time, it is hard to find one that stores within itself such an amount of metadata, useful for an in-depth study. It is wrong to look at the WWW simply as a kind of information store. Although all its contents are information one way or the other, the truth is there are quite a few ways of letting the users interact with that information, either to manipulate it (via ajax-based applications), to alter it (through the use of wikis), to add to it (via blogs and web sites themselves) or to transform and amplify its meanings. These are only a few examples of what can be done today. Web site access logs are the main information source on how the WWW is used. Rather than asking the users if they viewed the pages (such as a TV station might do), any web site has the means to keep a permanent record about its visitors. By analyzing these logs, we are able to get a better understanding of the roles played by the web site. In this paper we borrow a few concepts from biology, in order to establish a kind of 'DNA' for each document on the web site of the Portuguese Tribunal de Contas (Court of Auditors). We do this by looking at the WWW as an information source and by processing what we find. At the same time, we try to extend the same approach to the users who looked for those documents, by processing the web access logs. The results of such an approach might enable future uses of automatic document classification, as well as an effective personalization of information delivery.

Keywords: web mining, DNA, access logs.

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090041

1 Introduction

Given the abundance of information available in the World Wide Web, it is surprising that some of the underlying structures are so fragile. At the same time, there are so many ways of using that content, and so many people using it that it is virtually impossible to find a "one size fits all" kind of solution.

The perfect web site should be able to provide the adequate information to all its users every time. Unfortunately, such a thing seldom occurs. Providing the correct answers and the appropriate information is a task that demands permanent care and attention from those who manage the web site. Apart from content itself, web usage analysis is one of the most used technical resources to ensure some degree of knowledge on usage habits and needs.

Web mining techniques provide a large amount of (often) useful information on the way people use a web site. Exploring such information may give insights on subjects such as:

- Web site structure optimization, by studying web page navigation patterns (either manually or automatically [8]);
- Database optimization on dynamic content web sites, by studying information needs;
- Web site usability evaluations [17];
- Establishing the adequacy of content vs. objectives (something rather useful for e-commerce web sites who gain a lot from knowing their users very well [10]).

In this paper we show one way of using web mining techniques to discover web site access patterns, and to classify its users, based on the content of the documents they access. We do so over the information gathered from and provided by the Portuguese Court of Auditors web site (http://www.tcontas.pt) for the six-year period between March 14, 2001 and March 14, 2007. To present our findings we borrow a concept from biology and classify both users and documents according to an artificially constructed "DNA". Given the fact that this web site does not have any kind of user tracking or identification in place (only basic web usage processing), this work presents an opportunity of getting to know better a whole class of users and their needs.

2 The Portuguese Court of Auditors web site

2.1 Characterization of the Portuguese Court of Auditors

The Portuguese Court of Auditors is a sovereign body of the Portuguese Republic, defined as "the supreme body which examines the legality of public expenditure and rules on the accounts which the law has ordered to be submitted to the Court" [14]. Its attributions correspond to the need for auditing public funds, public revenue and expenditure and public assets. The Court of Auditors exercises the function of financial and jurisdictional control in relation to those entities which are part of the Public Administrative Sector and of the Public Business Sector and, in general, to all entities administrating or using public

money. Documents created by the Court of Auditors are called Acts and are addressed to several entities, such as the Portuguese Parliament, the Regional Legislative Assemblies of the islands of Madeira and Azores, as well as all audited entities.

2.2 Act processing

All acts created by the Court of Auditors are made available on the court intranet and are managed by an internally developed application, called **TCJure**. This application manages both the document contents and additional metadata, allowing for some flexible information retrieval operations. Record management is performed manually – there is a department responsible for the insertion of new acts in the underlying database, as well as updating content as necessary. Document classification is supported by a dedicated thesaurus, also managed within the same application.

A subset of all the acts is made available on the Court of Auditors web site (http://www.tcontas.pt). Unlike the intranet, there are no advanced document retrieval options, only basic browsing. Some types of acts (decisions and sentences) have their own search interface on the web site, and are also subject to a specific set of processing rules before being made available to the public. This also means that there are differences in the way the same content is presented and accessed inside and outside the institution. Although the **TCJure** application is the cornerstone of act processing for the whole Court, it is not used on the web site. All document classification available on the intranet is lost when the same documents are published on the web site.

2.3 The thesaurus

The thesaurus used by the **TCJure** application contains a total of 8.556 terms, headed by 32 Top Terms. When following thesauri-specific rules (which contemplate a number of relations between terms, such as Top Term, Equivalent Term, Related Term and Use), there are only 6.349 terms available. All the others are “loose” within the thesaurus (in the sense that they do not relate to any of the 32 Top Terms). Document classification within the **TCJure** application is done manually, a method which has several known limitations [15, 18, 21].

2.4 The web site

The Court of Auditors web site is active since May 1998. Until March 2001 it was used mainly to show what the institution was and how it worked, while providing a small amount of documentation (although, for a brief period until 1999, the **TCJure** application was available to the general public). From March 14, 2001 until the present, the web site publicizes a large number of acts.

As it was mentioned before, there is no data collection on the site regarding specific user identification. Web usage data is permanently collected and processed on a daily basis (albeit in a rather basic fashion, presented in a report available at http://www.tcontas.pt/diario.html). However, the amount of data

collected this way was vital for our work by showing the most active users on the web site as well as the most visited documents. It made possible further processing in order to reach our goal – to classify both documents and users according to the same classification framework.

3 Web mining

The field of web mining is rather recent, due to the age of the underlying technologies (the WWW itself has not turned 20, yet). The explosive growth of the WWW guarantees an almost endless amount of data sources which, in turn, resort to web mining techniques in order to answer its users' needs. Information is abundant in the WWW and that abundance demands new approaches.
According to [4] web mining comprises four different steps:

- Resource identification, in which the resources needed for information extraction are identified (either through search engines or through dedicated spiders and scrapers [20]). Web sites of a dynamic nature are quite often ignored in this process [7];
- Pre-processing, in which relevant information is selected from found information sources. This step is directly related to information extraction techniques ([2, 4]);
- Generalization, in which automatic pattern discovery is made on several web documents. This step uses data mining techniques as well as clustering and classification trees;
- Analysis, in which pattern discovery is validated and interpreted.

These four steps are put together and applied in different ways, according to the type of information source upon which they are made to act.

3.1 Web usage mining

Web usage data is easily collected. It only takes a web server with logging functions enabled to gather a large amount of useful information concerning the way users interact with content. Such data not only shows what was seen on the web site, but also when it was seen, from which place, using which operating platform (amongst other details). When further processed, it may also shed some light on web server behaviour (by showing web page loading answer times), on web server security conditions (by showing possible attack fingerprints disguised as malformed web page requests), on user needs and motivations (by analyzing information requests and the paths taken through the web site).

Web usage mining is not a magic solution for all information needs concerning the way users access a web site. In fact, there are quite a few obstacles creeping along the. What is being measured – page views, hits, user sessions? What exactly do each of these terms mean? How about search engine spiders, which leave trails through web access logs that can be used to study their evolution and behaviour [6] but must be cleaned before analyzing human access to the same web pages [19]. Proxy servers may remove traffic from a web site, by providing users with web pages more quickly, avoiding them the need to go

live on the WWW, but also making possible to get the wrong idea from the web usage logs ("my web site is known and famous but my usage logs are empty..."). Corporate local networks hiding behind firewalls, with hundreds of unique users, may appear in web usage logs as a single user with a single IP address.

Web usage mining techniques provided us with the information necessary to identify both the top users of the web site, and the most popular documents.

3.2 Web content mining

Given the size of all the resources available on the WWW, human users are hardly expected to be able to apprehend everything. It is difficult to control such an amount of data, either due to time constraints, or to our natural inability in finding useful patterns among huge volumes of seemingly unrelated pieces of information.

Web content mining techniques allow for the automated discovery of knowledge. Most of these techniques are highly complex and demand abundant technical resources, two characteristics that put them away from most users. However the same techniques have been put to god use and some of their results are used daily by millions of users worldwide. Every major search engine (Google, Yahoo and Live Search) as well as most of the smaller ones, gathers information through web content mining, processing hundreds of millions of web pages as well as all sorts of documents in many different formats.

Extracting information from web documents this way opened a world of opportunities in areas such as

- Geo-location, in which web page content is used to identify its geographic location [9];
- Content extraction for automated structuring of naturally unstructured documents (such as the majority of existing web pages [16]);
- User privacy [13];
- Keyword extraction, not only for automatic document classification, but also as helpers for thesauri and ontology creation.

We used web content mining techniques for trying to find a way to automatically classify documents, through text mining and keyword selection.

3.3 Web structure mining

Web structure mining techniques try to discover models underlying the hyperlink structure of the WWW [4]. Such models can be used to categorize web pages as well as to provide information on the degree of similarity between web pages and the way pages are related to each other. The results obtained are extremely important and can be found in every major algorithm in use by the most important WWW search engines, such as the HyperSearch [5], PageRank [11], HITS [12] and TrustRank [3] algorithms.

4 Establishing a DNA for both documents and users

Our first approach was based on the information provided by the TCJure application and its associated thesaurus. Since it is the only document classification method actually used in the Court of Auditors, it provides a stable framework on which to base a new way to look at things.

4.1 Document DNA

From a grand total of a 1003 documents available on the Court of Auditors web site, only 742 are registered in the TCJure application. Steps taken to establish the document DNA included

- Identifying every term of the thesaurus used to classify each of the 742 documents. 6018 terms were found, from which only 3266 were considered (the remaining 2752 did not relate to any Top Term);
- Grouping terms according to their Top Term within the thesaurus. Each document was then "given" a Top Term, according to the terms used in its classification. The same document could belong to several Top Terms at the same time;
- Weighting each Top Term according to the ratio between the amounts of terms related to it inside the document, and the full amount of terms used in its classification. The weight stands between 0 and 9.

The result of these calculations is shown partially in Figure 1.

ADN baseado nas classes do TCJure
(representação numérica)

	Actividade política	Direito	Economia	Educação	Empresa e Concorrência	Entidade reguladora	Defesa	Finanças (Direito Financeiro)	Indústria e Energia	Organizações Internacionais	Política Internacional	Administração Interna	Questões Sociais	Trabalho. Emprego	Transportes e Comunicações	Tribunal (Organização Judiciária	Outros	Administração Pública	Agricultura, Pecuária e Pesca	Agro-Alimentar	Ambiente	Associação e Fundação Privada	Ciência e Tecnologia	Cargos	Empresas	Geográficos	Organismos	Intervenções Operacionais	Programas	Regulamentos	Sistemas
Instruções nº1/95 - 2ª Secção	0	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Instruções nº1/97 - 2ª Secção	0	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Instruções nº2/97 - 2ª Secção	0	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Instruções nº1/00 - 2ª Secção	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Instruções nº2/99 - 2ª Secção	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Instruções nº2/00 - 2ª Secção	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 21/1999 - 2ª Secção	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 1/2004 - 1ª Secção	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Relatório de Auditoria nº 38/1999 - 2ª Secção	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 2/2000 - 2ª Secção	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 29/2000 - 2ª Secção	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 31/2000 - 2ª Secção	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 33/2000 - 2ª Secção	0	2	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Relatório de Auditoria nº 30/2000 - 2ª Secção	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1: Document DNA.

Given the fact that the document classification provided by the TCJure application is not used on the web site, next we tried to establish another kind of document DNA, based on automatic keyword discovery, using text mining techniques. Software used was SAS 9.1. Keyword discovery and extraction was an iterative process. From a total of 63.022 keywords initially found through SAS, several rounds of manual optimization reduced this total to 2.208 keywords. Further processing steps included:

- Weighting each keyword according to the amount of occurrences inside the 742 documents being processed. The weight stands between 0 and 9;
- Randomly selecting 20 documents classified under each of the five most used Top Terms of the TCJure thesaurus. As the same document may belong to more than one Top Term, we used only 90 documents;
- Counting the occurrences of each keyword inside these 90 documents. The total number of occurrences for each keyword was then multiplied by the weight found on the previous step. By doing this we tried to value multiple occurrences of keywords with low weights.

Due to time constraints, our analysis was further confined to documents belonging to the most representative Top Term inside the thesaurus, the one called *Direito* (Law). From the original 90 documents, 15 related to Law, having between them 40 keywords. For comparison, the same calculations were made using all of the 493 documents belonging to the Law Top Term. Table 1 shows the results for both sets of documents, with keywords ordered by their relative importance in the 15 document set.

The top 40 keywords from the 493 document set only include 19 from the 15 document set. Table 1 shows the relation between them, according to their relative importance. For instance, the most important keyword in the 15 document set, *autarquias locais*, ranks only in the 8th position in the 493 document set. We found it difficult to use these values as a basis for document classification.

Table 1: Comparing documents.

Keyword	# in the 15 document set	# in the 493 documents
autarquias locais	1	8
auditorias orientadas	2	3
adicional	3	20
…	…	…
acto	21	24
capacidade instalada	23	18
conclusão da análise	24	10
avaliação da qualidade	25	5
comparticipação comunitária	30	11
contabilidade analítica	31	7

4.2 User DNA

Considering the TCJure thesaurus as a viable source for classifying documents, we extended its usage in order to classify users. Web usage mining results were

used to discover the top five users of the web site (we restricted ourselves to five users in order to keep processing operations down to a manageable level). These five users are institutional users, the first four being entities belonging to the central government, and the fifth belonging to the local government. User DNA is determined according to the following rules:

- After viewing the first document, the user is assigned the document's DNA;
- For every next document viewed by the same user, the DNA from the new document is added to his own DNA, the result being a rounded average of both values.

The result of these operations when applied to access data from the top user of the web site is shown in Figure 2.

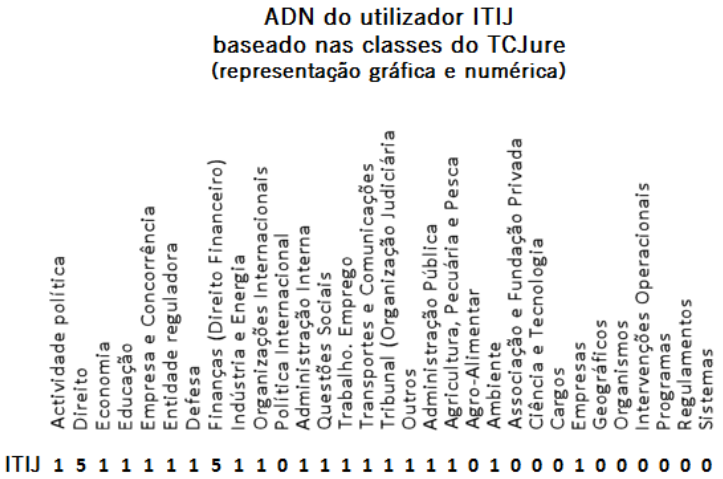

Figure 2: DNA for the top user.

User classification becomes possible using both web usage data and the TCJure document classification.

5 Conclusion

By studying the Court of Auditors web site we were able to gather information on the way documents are classified and used. There are document processing rules active only inside the institution, without benefits towards the web site. Although this situation is expected to change in the near future, there is quite a sizeable amount of access data which cannot be easily interpreted, unless several pre-processing steps, such as the ones we described in this paper, are taken.

The TCJure thesaurus is, at present, the main tool being used for document classification. Such classification is performed manually, by human operators.

We tried to establish the grounds for an automatic classification procedure, based on keyword discovery. However, in order for such an approach to succeed, it would be necessary to establish an adequate keyword list, a task for which some intensive human input would be necessary. Time constraints prevented us from doing that. As such, the thesaurus stood as the only viable option to classify documents.

We used the terms of the thesaurus in a new way, at least as far as document classification and grouping is concerned. Some of its limitations soon became apparent – for instance, including quite a large number of terms not related to any other, something that is quite contrary to the objectives of any thesaurus. Other limitations have to do with the way the thesaurus itself is used. If not enough terms are used to classify a document, then a whole document collection may seem limited in scope or subject. By using a biological metaphor such as DNA, we tried to convey the notion of personalization. Although this specific DNA (both for documents and users) has little to do with the biological version, it keeps its symbolic meaning. By assigning a DNA to a document, we are classifying it in a manageable way. By using the same information as the basis for user DNA, we are establishing a direct relationship between document content and user interest. Given the way user needs evolve through time, user DNA also evolves, keeping in line with those interests. Future work may include thesaurus optimization, as well as „better resolution‟ in document classification. Input from people who use the thesaurus should prove vital in helping to define a keyword list for using in automatic document classification procedures. Real world usage of such concepts as „document DNA‟ and „user DNA‟ will depend on including new interfaces on the web site, both for user registration, and for defining searches based on these classification schemes.

References

[1] Constitution of the Portuguese Republic, article 214.

[2] CHANG, C.H.; KAYED, M.; GIRGIS, M.R.; SHAALAN, K.F., “A survey of web information extraction systems”, IEEE Transactions on Knowledge and Data Engineering 18 (10): 1411-1428, Oct. 2006

[3] GYÖNGYI, Zoltán; GARCIA_MOLINA, Hector, PEDERSEN, Jan, “Combating Web Spam with TrustRank”, Proceedings of the International Conference on Very Large Data Bases, 30:576, 2004.

[4] KOSALA, Raymond, BLOCKEEL, Hendrik, “Web mining research: a survey”, SIG KDD Explorations, Vol. 2, pp. 1-15, July 2000.

[5] MARCHIORI, Massimo, “The Quest for Correct Information on the Web: Hyper Search Engines”, Proceedings of the Sixth International World Wide Web Conference (WWW6), 1997.

[6] MAYR, Philipp: “Website entries from a web log file perspective – a new log file measure” in Proceedings of the AoIR -ASIST 2004 Workshop on Web Science Research Methods, Brighton, 2004.

[7] MÉNDEZ-TORREBLANCA, A., MONTES-Y-GÓMEZ, M., LÓPEZ-LÓPEZ, A., “A Trend Discovery System for Dynamic Web Content

Mining", XI International Conference on Computing CIC-2002, Mexico City, Mexico, November 2002.
[8] PERKOWITZ, Mike; ETZIONI, Oren: Adaptive Web Sites: Concept and Case Study, Department of Computer Science and Engineering, University of Washington, Seattle, 1999.
[9] SILVA, M.J.; MARTINS, B.; CHAVES M.; AFONSO, A.P.; CARDOSO, N., "Adding geographic scopes to web resources", Computers, Environment and Urban Systems 30 (4): 378-399, Jul. 2006.
[10] SPILIOPOULOU, Myra; POHLE, Carsten: "Data Mining for Measuring and Improving the Success of Web Sites", Data Mining and Knowledge Discovery, 5, 85–114, 2001.
[11] United States Patent 6.285.999 – Method for node ranking in a linked database.
[12] United States Patent 6.112.202 – Method and system for identifying authoritative information resources in an environment with content-based links between information resources.
[13] VAN WEL, Lita; ROYAKKERS, Lambèr, Ethical issues in web data mining, 2004.
[14] Constitution of the Portuguese Republic, article 214.
[15] BASTIAN, Alistair; BRIFFETT, Adam; COOK, Daniel; YEOMAN, Eric, SOMbrero - Document Classification with Large Document Sets: The SOM utilising Layout Information, CO600 Project Report, University of Kent at Canterbury.
[16] CHANG, C.H.; KAYED, M.; GIRGIS, M.R.; SHAALAN, K.F., "A survey of web information extraction systems", IEEE Transactions on Knowledge and Data Engineering 18 (10): 1411-1428, Oct. 2006.
[17] GAYO-AVELLO, Daniel; ÁLVAREZ-GUTIERREZ, Dário; GAYO-AVELLO, José; Naïve Algorithms for Keyphrase Extraction and Text Summarization from a Single Document Inspired by the Protein Biosynthesis Process, Bio-ADIT 2004 - The First International Workshop on Biologically Inspired Approaches to Advanced Information Technology. A.J. Ijspeert et al. (Eds.): BioADIT 2004, LNCS 3141, pp. 440-455.
[18] GOLLER, Christoph; LÖNING, Joachim; WILL, Thilo; WOLFF, Werner, "Automatic Document Classification: A thorough Evaluation of various Methods", Zweiter Workshop des MK2 - Proceedings, 25.7.2000.
[19] HAIGH, Susan; MEGARITY, Janette: "Measuring Web Site Usage: Log File Analysis", Network Notes #57, August 4, 1998, Information Technology Services, National Library of Canada.
[20] HEMENWAY, Kevin; CALISHAIN, Tara: Spidering Hacks, O'Reilly, 2003.
[21] ROY, Devshri; SARKAR, Sudeshna; GHOSE, Sujoy, "Document Type Identification for use with intelligent tutoring system", ICDE International Conference, November 19-23, 2005, New Delhi.