WEB CONTENT MINING

CLAUDIA ELENA DINUCĂ, DUMITRU CIOBANU [*]

ABSTRACT: *The World Wide Web, or simply the web, is the most dynamic environment.The web has grown steadly in recent years and his content is changing every day. Today, they are several billions of HTML documents, pictures and another multimedia files available on the Internet. There is a need of methods to help us extract information from the content of web pages. One answer to this problem is using the data mining techniques that is known as web content mining, which is defined as "the process of extracting useful information from the text, images and other forms of content that make up the pages".*

KEY WORDS: *internet; web mining; web content mining; data mining.*

JEL CLASSIFICATION: *L86*

1. INTRODUCTION

The advent of the World Wide Web (WWW) has overwhelmed home computer users with an enormous flood of information. To almost any topic one can think of, one can find pieces of information that are made available by other internet citizens, ranging from individual users that post an inventory of their record collection, to major companies that do business over the Web.

To be able to cope with the abundance of available information, users of the Web need assistance of intelligent software agents (often called *softbots*) for finding, sorting, and filtering the available information (Etzioni, 1996). Beyond search engines, which are already commonly used, research concentrates on the development of agents that are general, high-level interfaces to the Web (Furnkranz et al., 2002), programs for filtering and sorting e-mail messages (Payne & Edwards, 1997) or Usenet net news articles (Mock, 1996), recommender systems for suggesting Web sites (Pazzani et al., 1996) or products (Doorenbos et al., 1997), automated answering systems (Burke et al., 1997, Scheffer, 2004) and many more.

[*] *Ph.D. Student, University of Craiova, Romania, clauely4u@yahoo.com.*
Ph.D. Student, University of Craiova, Romania, ciobanubebedumitru@yahoo.com.

Many of these systems are based on machine learning and Data Mining techniques. Just as Data Mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of *web mining* aims at finding and extracting relevant information that is hidden in Web-related data, in particular in (hyper-)text documents published on the Web. Like Data Mining, web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others. Web mining is commonly divided into the following three sub-areas:

- Web Content Mining: application of Data Mining techniques to unstructured or semi-structured text, typically HTML-documents;
- Web Structure Mining: use of the hyperlink structure of the Web as an (additional) information source;
- Web Usage Mining: analysis of user interactions with a Web server.

An excellent textbook for the field is (Chakrabarti, 2002) and an earlier effort was done by (Chang et al., 2001). Brief surveys can be found in (Chakrabarti, 2000, Kosala & Blockeel, 2000). For surveys of content mining, we refer to (Sebastiani, 2002), while a survey of usage mining can be found in (Srivastava et al., 2000).

2. WEB OVERVIEW

The characteristics of Web applications are the existence of hypertext links and of some procedures that allow real-time dialogue between client and server. Hypertext links are indicated by marking different from the rest of the document of words, images or icons that, when selected, cause browser to bring the respective document, regardless of where it is located on the Internet. Assembly of electronic documents that refer to each other led to the name Web.

The process of bringing documents on the system using browsers is named browsing or surfing the web. Note that currently most web applications are electronic publications due to the possibilities the Web offers: a fast information and at a reduced price (actually, only the cost of subscription to the Internet connection), the information is structured, interactive, quickly updated and made available to users.

With several billion Web pages created by millions of authors and organizations, World Wide Web is a great source of knowledge. Knowledge comes not only from the content itself, but also from the unique features of the Web, hyperlinks and the diversity of content, language.

Web size and dynamic unstructured content, makes extracting useful knowledge a challenge for research. Web sites generates a large amount of data in various formats that contain valuable information. For example, Web server logs contain information about user access patterns that can be used to customize information to improve website design.

World Wide Web is certainly the largest data resource in the world. Using global Web network, increasing the role and implications in the daily life of society, has led to a rapid and unprecedented development of many fields such as finance and banking, commercial, educational, social, etc. Because the existing data volume on

Web is huge the application of new techniques for extracting information and knowledge has become a necessity for future evolution.

Web mining is the area that has gained much interest lately. This is due to the exponential growth of World Wide Web and anarchic architecture and the growing importance of Internet in people's lives.

3. WEB MINING

Web mining is the use of data mining techniques for automatic discovery and knowledge extraction from documents and Web services. This new area of research was defined as an interdisciplinary field (or multidisciplinary) that uses techniques borrowed from: data mining, text mining, databases, statistics, machine learning, multimedia, etc.

Web mining has three operations of interests - clustering (finding natural groupings of users, pages etc.), associations (which URLs tend to be requested together), and sequential analysis (the order in which URLs tend to be accessed). As in most real-world problems, the clusters and associations in Web mining do not have crisp boundaries and often overlap considerably. In addition, bad exemplars (outliers) and incomplete data can easily occur in the data set, due to a wide variety of reasons inherent to web browsing and logging. Thus, Web Mining and Personalization requires modelling of an unknown number of overlapping sets in the presence of significant noise and outliers, (i.e. bad exemplars). Moreover, the data sets in Web Mining are extremely large.

The Internet offers unprecedented opportunities and challenges of data mining field due to the diversity and lack of data structures. Thus:

- There are almost all type of data on the web - text, tables, multimedia data, etc. Typically, a Web page contains a mixture of information, such as navigation panels, main content, advertising or copyright notes.
- The amount of existing information is enormous and easily accessible.
- Coverage of the information is very high.

We can find information about almost anything. The information provided can be divided into two broad categories: documents and services.

- Most data are in semi-structured form due to the structure of HTML code.
- There are links between information in the pages of a site and between pages from different sites.
- More information are redundant - meaning that the same information or similar versions of it can appear in multiple pages.
- Information can be found on the surface (in pages that are accessed via browsers) or deep (in the database that are queried through different interfaces).
- The dynamic nature of information is obvious. Monitoring the constant changes in the information is an important issue.
- Above all, the Internet has become a virtual company. In addition to information and services contained, the Internet offers the possibility of

interaction between people, thus contributing to the creation and development of new communities.

The Web is a critical channel of communication and promoting a company image. E-commerce sites are important sales channels. It is important to use data mining methods to analyze data from the activities performed by visitors on websites. Web mining methods are divided into three categories:

- **Web content mining** - extraction of predictive models and knowledge from the contents of Web pages;
- **Web structure mining** - discovering useful knowledge from the structure of links between Web pages;
- **Web usage mining** - analyze the use of Web resources using log files.

Web content mining is the process of extracting useful information from Web documents content. Web content consists of several types of data such as text data, images, audio or video data, records such as lists or tables and structured hyperlinks. Web content mining is closely related to data mining and text mining because many of the techniques are applied for mining the Web, where most data are in text form. Differences resulting from data structure that are analyzed. Thus, if data mining techniques can be applied to structured data sets, text mining focuses on unstructured texts, web data mining operating on semi-structured. Web content mining requires creative applications so both data mining and / or text mining have own unique approach.

In recent years there has been an expansion of mining activities in the field of Web content, this is a natural result of the great benefits arising from such mining activities. However there are still many issues that require further research, such as:

- Extracting data and information;
- Information integration;
- Extracting opinions from online sources (forums, chat, surveys, etc.)
- Knowledge synthesis;
- Web pages segmentation and detection of redundant information.

Web structure mining involves analyzing the links between Web pages and determine the most accessed pages. Such pages can be classified into:

- „authorities" - pages that are accessed many times by links from other websites,
- „hubs" - pages containing a large number of links that access other pages.

Such an analysis in conjunction with the search for certain keywords can be greatly improved results to a search that takes into account only the desired content.

Web usage mining is the most relevant part in terms of marketing, because it explores ways of navigation and behavior during a visit to the website of a company. With the continued growth of e-commerce, Web services and Web-based information systems, the volume of clickstream data collected by Web-based organization in its daily operations has reached astronomical proportions (Mobasher et al., 2006). Methods for extracting association rules are useful for obtaining correlations between different pages visited during a session. Association rules or sequential time series models can be used to analyze data from a website taking into account the temporal dynamics of the site usage. Web usage mining is mainly based on sequence analysis of

pages visited in a session data, analyzing web clicks. Information on buying behavior of visitors can be obtained in the e-commerce web site by analyzing web clicks. Web usage mining aims at extracting the knowledge from user sessions that can be restored using log files.

Log files of Web servers can be in the CLF form (Common Log Format) or ELF form (Extended Log Format). If for the ELF format can be configured the log file, in the case of CLF format, the file will contain information about:
- Remotehost: browser hostname or IP #;
- rfc931: log the user name (always „-" means „unknown");
- AUTHUSER: name of the authenticated user HTTP;
- Date: date and time of application;
- „request": the exact line accessed by the user;
- Status: returned HTTP code (200 is OK, 2xx - successful response, etc.);
- Bytes: size of response content.

Benefits offered by the analysis of log files are related to classification of users, improving site design, prediction and detection of fraud actions among users. Benefits of clickstream can be seen in the way content is viewed by site users. Clickstream provides information about: the number of site visitors, the site showing the most interest, the region from where the visitors come, pages or parts of pages that are more or less populated, sites that offer the highest advertising to their current site.

One problem is to identify users taking into account that they can use different addresses when access the web from different places. Also, log files do not contain actual information accessed by users, and effective reconstruction of a session is often impossible due to the dynamic structure of the sites

Preprocessing data from log files to apply data mining techniques requires very different methods of reconstruction of the sessions and user identification.

Data mining applications that best fit to log files are the association rules, clustering and classification algorithms, and a number of other statistical analysis. Thus, it can be determined by statistical analysis the number of visits in a given period, the average visit of a page, the countries from which are the users of site, together with the percentage of users for every country, the most used search engines, most frequently used browsers etc.

4. WEB CONTENT MINING

Web-mining is an umbrella term used to describe three quite different types of data mining, namely content mining, usage mining and structure mining (Chakrabarti, 2003). In this section, we are concerned with Web content mining, which Linoff and Berry (2001) define as *"the process of extracting useful information from the text, images and other forms of content that make up the pages".*

In the last years the growing of the WWW has overlap any expectations. Today they are several billions of HTML documents, pictures and other multimedia files available on the Internet, and their number is continuous increasing. Taking into consideration the huge variety of the web, extracting interesting contents has become a necessity.

There is a continously expanding amount of information "out there". Moreover, the evolution of the Internet into the Global Information Infrastructure, coupled with the immense popularity of the Web, has also enabled the ordinary citizen to become not just a consumer of information, but also its disseminator. The Web, then, is becoming the apocryphal "*Vox Populi"*. Given that there is this vast and ever growing amount of information, how does the average user quickly find what s/he is looking for - a task in which the present day search engines don't seem to help much!

One possible approach is to personalize the web space - create a system which responds to user queries by potentially aggregating information from several sources in a manner which is dependent on who the user is. As a trivial example - a European querying on casinos is probably better served by URLs pointing to Monaco, whereas someone in North America should get URLs pointing to Las Vegas. A biologist querying on cricket in all likelihood wants something other than a sports enthusiast would.

Thus, Web Content Mining is mining data from the content of web pages (Xu et al., 2011). Web pages consist of text, graphics, tables, data blocks and data records. Web Content Mining uses the ideas and principles of data mining and knowledge discovery process. Using the Web for providing information is more complex than when working with static databases, due to Web dynamics and the large number of documents.

Many researches have been made to cover web content mining problems to improve the way that pages are presented to end users, improving the quality of search results and extract interesting content pages. Thus, in (Lin & Ho, 2002) a system InfoDiscoverer is proposed which extract content information from a set of web pages of a site according to type HTML tags <table> from Web page.

This system partition the web page blocks in redundant and informative. Informative content blocks are distinct parts of pages, whereas redundant content blocks are common parts. This approach helps to improve the accuracy of information retrieval and extraction and reduce the size and complexity index extraction.

In (Morinaga et al., 2002) is presented a system for finding the reputation of products from the Internet. It automatically collects persons' opinions about certain target product of web pages and uses four different techniques for text mining to get the reputation of those products. Research work presented in (Davison, 2001) examine the accuracy of predicting a user's next action based on the content of the pages recently accessed by the user.

Predictions are made using a similarity model of user interests in the text from the content of hypertext anchors and around them of recently requested Web pages. In (Liu, et al., 2003) the authors proposed an algorithm MDR (Mining Data Records) to extract records of contiguous data or not. So, it find all records consist of tags for tables or forms, such as <table>, <form>, <td>, <tr>, these records are important because they contain essential information of host page.

Web Content Mining is related but different from data mining and text mining. Is related to data mining because data mining techniques can be applied in Web Content Mining, but is different from data mining because Web data are semi-structured or unstructured, while data mining deals with the structured data. Web

Content Mining is different from text mining through the structure of semi-structured web, while text mining is focused on unstructured text. Web content can be unstructured (eg text), semi-structured (HTML documents) or structured (data extracted from databases in dynamic Web pages). The dynamic data cannot be classified, forming the so-called „hidden web".

5. CONCLUSIONS

Web content mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data. Another important aspect of Web content mining is the usage of the Web as a data source for knowledge discovery. This offers interesting new opportunities since more and more information regarding various topics is available on the Web. But the use of the Web as a provider of information is unfortunately more complex than working with static databases. Because of its very dynamic nature and its vast number of documents, there is a need for new solutions that do not depend on accessing complete data on the outset.

Research in web mining tries to address this problem by applying techniques from data mining and machine learning to Web data and documents.

REFERENCES:

[1]. Burke, R.D.; Hammond, K.J.; Kulyukin, V.; Lytinen, S.L.; Tomuro, N.; Scott Schoenberg, S. (1997) *Frequently-asked question files: Experiences with the FAQ finder system,* AI Magazine, 18(2), pp. 57–66

[2]. Chakrabarti, S. (2000) *Data Mining for hypertext: A tutorial survey,* SIGKDD explorations, 1(2), pp. 1–11

[3]. Chakrabarti, S. (2002) *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kaufmann

[4]. Chakrabarti, S. (2003) *Mining the Web (Discovering knowledge from hypertext data),* Morgan Kaufmann

[5]. Chang, G.; Healy, M.J.; McHugh, J.A.M.; Wang, J.T.L. (2001) *Mining the World Wide Web: An Information Search Approach*, Kluwer Academic Publishers

[6]. Davison, B.D. (2002) *Predicting Web Actions from HTML Content*, in Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia, College Park, MD, pp. 159–168,

[7]. Dinucă, C.E. (2011a) *The process of data preprocessing for Web Usage Data Mining through a complete example*, Annals of the "Ovidius" University, Economic Sciences Series Volume XI, Issue 1

[8]. Dinucă, C.E. (2011b) *E-Business, a new way of trading in virtual environment based on information technology*, Annals of the "Ovidius" University, Economic Sciences Series Volume XI, Issue 1

[9]. Dinucă, C.E. (2011c) *The need to use data mining techniques in e-business,* Analele Universităţii "Constantin Brâncuşi" din Târgu Jiu, Seria Economie

[10]. Dinucă, C.E. (2011d) *Using Web Mining in E-Commerce Applications,* Analele Universităţii "Constantin Brâncuşi" din Târgu Jiu, Seria Economie

[11]. Dinucă, C.E.; Ciobanu, D. (2011) *Prezicerea următoarei pagini ce va fi vizitată de un utilizator al unui site web utilizând modelul navigării aleatoare,* Cercetarea doctorală în economie: prezent şi perspective, Editura Economică Bucureşti

[12]. Doorenbos, R.B.; Etzioni, O.; Weld, D.S. (1997) *A scalable comparison-shopping agent for the World-Wide Web,* in Proceedings of the 1st International Conference on Autonomous Agents, Marina del Rey, CA, pp. 39–48

[13]. Etzioni, O. (1996) *Moving up the information food chain: Deploying softbots on the world wide web*, in Proceedings of the 13th National Conference on Artificial Intelligence (AAAI- 96), AAAI Press, pp. 1322–1326

[14]. Furnkranz, J.; Holzbaur, C.; Temel, R. (2002) *User profiling for the Melvil knowledge retrieval system*, Applied Artificial Intelligence, 16(4), pp. 243–281

[15]. Kosala, R.; Blockeel, H. (2000) *Web mining research: A survey*, SIGKDD explorations, 2(1), pp. 1–15

[16]. Linoff, G.S.; Berry, M.J.A. (2001) *Mining the Web (Transforming customer data into customer value,* Wiley

[17]. Liu, B.; Grossman, R.; Zhai, Y. (2003) *Mining Data Records in Web Pages*, in Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), ACM Press, pp. 601–606

[18]. Lin, S.H.; Ho, J.M. (2002) *Discovering Informative Content Block from Web Documents*, in Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), ACM Press, pp. 588-593

[19]. Lupu Dima, L.; Edelhauser, E.; Ionică, A. (2010) *E-Learning Platforms in Romanian Higher Education*, Annals of the University of Petroşani. Economics, 10(1), Universitas Publishing House, Petroşani, pp.137-148

[20]. Mobasher, B.; Nasraoui, O.; Liu, B.; Masand B. (2006) *Advances in Web Mining and Web Usage Analysis*, Berlin, Springer Berlin-Heidelberg

[21]. Mock, K.J. (1996) *Hybrid hill-climbing and knowledge-based methods for intelligent news filtering*, in Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), AAAI Press, pp. 48–53

[22]. Morinaga, S.; Yamanishi, K.; Tateishi, K.; Fukushima, T. (2002) *Mining Product Reputations on the Web*, in KDD 2002: Proceedings of the eight ACM SIGKDD international conference on Knowledge discovery and data mining, New York, USA, ACM Press, pp. 341-349

[23]. Payne, T.R.; Edwards, P. (1997) *Interface agents that learn: An investigation of learning issues in a mail agent interface*, Applied Artificial Intelligence, 11(1), pp. 1–32

[24]. Pazzani, M.; Muramatsu, J.; Billsus, D. (1996) *Syskill & Webert: Identifying interesting web sites*, in Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), AAAI Press, pp. 54–61

[25]. Scheffer, T. (2004) *Email answering assistance by semi-supervised text classification*, Intelligent Data Analysis, 8(5)

[26]. Sebastiani, F. (2002) *Machine learning in automated text categorization*, ACM Computing Surveys, 34(1), pp. 1–47

[27]. Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P.N. (2000) *Web usage mining: Discovery and applications of usage patterns from web data*, SIGKDD explorations, 1(2), pp. 12–23

[28]. Stuparu, D.; Vasile, T. (2009) *The Electronic Commerce in the Globalisation Era*, Annals of the University of Petroşani. Economics, 9(2), Universitas Publishing House, Petroşani, pp. 301-306

[29]. Xu, G.; Zhang, Y.; Li, L. (2011) *Web Mining and Social Networking*, Techniques and Applications, Australia: Springer