WEB STRUCTURE MINING

CLAUDIA ELENA DINUCĂ *

ABSTRACT: *The World Wide Web became one of the most valuable resources for information retrievals and knowledge discoveries due to the permanent increasing of the amount of data available online. Taking into consideration the web dimension, the users get easily lost in the web's rich hyper structure. Application of data mining methods is the right solution for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering and Web based data warehousing. In this paper, I provide an introduction of Web mining categories and I focus on one of these categories: the Web structure mining. Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. It offers information about how different pages are linked together to form this huge web. Web Structure Mining finds hidden basic structures and uses hyperlinks for more web applications such as web search.*

KEY WORDS: *web mining; internet; web structure mining; link mining.*

JEL CLASSIFICATION: *L86.*

1. INTRODUCTION

The caracteristics of Web applications are the existance of hypertext links and of some procedures that allow real-time dialogue between client and server. Hypertext links are indicated by different marking 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 due to electronic publications and 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.

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

With several billion Web pages created by millions of authors, World Wide Web is a great source of knowledge. We can find information about almost anything.

The information provided can be divided into two broad categories: documents and services. Most important characteristiques of information contained on the web are:

- Most data are semi-structured form because the structure of HTML code.
- There are links between information in the pages of a site and between pages from different sites.
- More information is 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 there are queried through different interfaces).
- The dynamic nature of information is obvious. Constantly monitoring 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.

Web size and dynamic unstructured content, makes the extracting of useful knowledge a challenge for researchers. Web site generates a large amount of data in various formats that contain valuable information. For example, Web server logs contain information about user access patterns and 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 is huge, the application of new techniques for extracting information and knowledge from the web is much needed.

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.

2. WEB MINING OVERVIEW

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of the following tasks (Kosala & Blockeel, 2000):

1. Resource finding: It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.

2. Information selection and pre-processing: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be

renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in training corpus.

3. Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization.

4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

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) using techniques borrowed from: data mining, text mining, databases, statistics, machine learning, multimedia etc.

Web mining, when looked upon in data mining terms, can be said to have 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.

3. WEB MINING METHODS

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 (Raymond & Henrik, 2000; Srivastava et al., 2000):

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

Web mining tasks can be classified into three categories (Kosala & Blockeel, 2000; Srivastava, et al., 2000): Web content mining, Web structure mining and Web usage mining. All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web - see Figure 1. In the following section we provide a brief introduction about each of these categories.

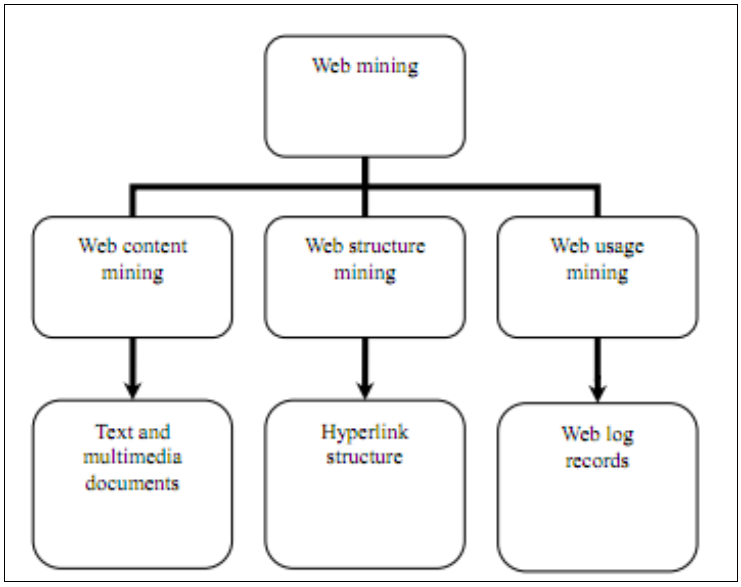

Figure 1. Web Mining Structure

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 data mining techniques are applied on the Web, where most data are in text form.

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;
- Integration of information;
- Extracting opinions online sources (forums, chat, surveys, etc.);
- Knowledge synthesis;
- Web page segmentation and detection of redundant information

Web usage mining is the most relevant in terms of marketing, because it explores ways of navigation and behaviour during a visit to the website of a company. With the continued growth of e-commerce, Web services and Web-based information systems, volume and clickstream data collected by Web-based organization in its daily operations have reached astronomical proportions (Mobasher et al., 2006).

The benefits obtained by analysis of log files are related to classification of users, improved site design, prediction and detection of fraud actions among users. Benefits of clickstream can be seen in the way the content is viewed by website users. The clickstream analysis provides information on: the number of site visitors, the site showing the most interest, the region from where the visitors came, pages or parts of pages that are more or less populated, sites or web pages that offer the highest current advertising for the current website.

Data mining applications that are best suited for log files analysis 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 a page, the countries from which the site' users came from, together with the percentage of users visiting the country properly, and the most used search engines.

4. WEB STRUCTURE MINING

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also used to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links. This enables clustering of connected Web pages to establish the relationship of these pages.

In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites. The determined connection brings forth a useful tool for mapping competing companies through third party links such as resellers and customers. This cluster map allows for the content of the business pages placing upon the search engine results through connection of keywords and co-links throughout the relationship of the Web pages. This determined information will provide the proper path through structure mining to improve navigation of these pages through their relationships and link hierarchy of the Web sites. With improved navigation of Web pages on business Web sites, connecting the requested information to a search engine becomes more effective.

Therefore, Web mining and the use of structure mining can provide strategic results for marketing of a Web site for production of sale. The more traffic directed to the Web pages of a particular site increases the level of return visitation to the site and recall by search engines relating to the information or product provided by the company. This also enables marketing strategies to provide results that are more productive through navigation of the pages linking to the homepage of the site itself. In order to utilize the website as a business tool, web structure mining is a must.

The role of web page links is similar to the role of citations in scientific literature, for example. Popular articles are often cited. Many hyperlinks pointing to a page draw the attention of web users just as citations to an article do for academics.

In fact, the Web is an example of a *social network*, a network of entities such as individuals or organizations that connect (or interact with) each other in various ways. The notions of *popularity*, *authority*, and *prestige* are central to social networks. There is an approach called *bibliometrics*, which is used in library and information science to analyze the merit of scientific publications. An example of bibliometrics approach is the citation indices. For example, the *impact factor* (number of citations in the preceding two years) uses the in-degree of the nodes in the network of scientific journals to evaluate the merit of a publication. The measures of popularity, authority,

and prestige can be used for ranking web pages retrieved by a search engine. The idea is to assign to each page on the Web a rank based on the hyperlink structure. This ranking can be done off-line (e.g., when pages are indexed) because it is independent of any query and web page textual content and is then used to rank the pages returned by the keyword search query.

We present two important web page ranking algorithms, PageRank and HITS, which further develop that notion and combine it with a content-based web search.

Web Structure Mining (Web Mining Linkage) offer information about how different pages are linked together to form this huge web. Web Structure Mining finds hidden basic structures and uses hyperlinks for more web applications such as web search.

The Links pointing to a document indicates the popularity of the document, while links coming from a document indicate the richness and variety of topics contained in that document. Web Structure Mining describes the organization of web content, web structure is defined as "hypertext links between pages and HTML formatting commands within a page" (Cohen, 2003). Understanding the relationship between content and structure of the site is useful to keep an overview on Web sites. Thus, it can be indicate whether a page fix within the structure of links and content and help identifying topics that extend over several connected web pages, thereby helping web designers by comparing their intentions with real content and structure of the site. Other studies deal with the web page as a collection of blocks or segments. The authors in (Cai, et al., 2004) uses an algorithm to partition the web page into blocks, by extracting the page-to-bloc relations, block-to-page link structure and page layout analysis, a semantic graph can be built on the Web, which represents each node by exactly one semantic topic, this graph can better describe the structure of semantic Web.

5. WEB SEARCH AND HYPERLINKS

Traditional information retrieval systems and search engines first extract relevant documents to users based on content similarity query entered and indexed pages. In the late 1990s, it was concluded that the methods used are based on content alone are not sufficient due to the large volume of information available on the Internet. When applying a query using a search engine the page numbers results relevant to this query is very high. Thus, to meet the satisfaction of users, search engines must choose the first 30-40 pages results of relevant query. Thus, there are used hyperlinks that connect pages together. In 1998, there were created two very important algorithms based on hyperlinks, PageRank and HITS. Both algorithms, PageRank (Palau, et al., 2004) and HITS (Kleinberg, 1998), draw their origin from social network analysis.

They use the hyperlinks structure of the web pages to give ranks according to the degree of prestige or authority. Page Rank algorithm was created in 1998 by Sergey Brin and Larry Page. Based on this algorithm it works the most successful Internet search engine, Google. Page Rank is rooted in social network analysis, it basically

provide a ranking of each web page depending on how many links from other sites leading to that page.

The key idea is to use the probability that a page is visited by a random surfer on the Web as an important factor for ranking search results. This probability is approximated by the so-called *page rank*, which is again computed iteratively. The popularity (or prestige) of a web page can be measured in terms of how often an average web user visits it. To estimate this, we may use the metaphor of the "random web surfer," who clicks on hyperlinks at random with uniform probability and thus implements the *random walk* on the web graph. Assume that page *u* links to *Nu* web pages and page *v* is one of them. Then once the web surfer is at page *u*, the probability of visiting page *v* will be 1*/Nu*. This intuition suggests a more sophisticated scheme of propagation of prestige through the web links also involving the out-degree of the nodes. The idea is that the amount of prestige that page *v* receives from page *u* is 1*/Nu* from the prestige of *u*. This is also the idea behind the web page ranking algorithm PageRank (Markov & Larose, 2007).

It is assumed that we have *n* pages, each page containing a number O_i of links to other websites. Let *A* be the adjacency matrix associated to the web regarded as a directed graph *G = (V, E)* where pages are vertices and links between pages are arcs of the graph.

Associated graph adjacency matrix will have the elements:

$$A_{ij} = \begin{cases} \dfrac{1}{O_i}, \; if \; (i,j) \in E \\ 0, \quad if \; (i,j) \notin E \end{cases}. \tag{1}$$

Starting with an initial probability vector and using an irreducible stochastic and aperiodic matrix, according to Ergodic Theorem of Markov chains it is obtained a convergent series of vectors of probabilities to a unique equilibrium state:

$$P_1 = A^T P_0, \tag{2}$$

$$P_k = A^T P_{k-1}, \tag{3}$$

$$\lim_{k \to \infty} P_k = P. \tag{4}$$

The probability vector obtained will give us rank web pages. To apply the Ergodic Theorem of Markov chains the adjacency matrix is transformed to meet conditions for irreducibility and aperiodicity.

The formulate:

$$P(i) = (1-d) + d \sum_{(j,i) \in E} \frac{P(j)}{O_j}, \tag{5}$$

gives us the rank of page *i*, where *P(i)* is the rank of page *i* and *d* is a damping factor which takes values between 0 and 1.
The pseudocode algorithm for calculating the rank of web pages is presented below.

$$
\begin{aligned}
&PageRank\\
&\quad P_0 \leftarrow \frac{e}{n}\\
&\quad k \leftarrow 1\\
&\quad repeat\\
&\qquad P_k \leftarrow (1-d)e + dA^T P_{k-1};\\
&\qquad k \leftarrow k+1;\\
&\quad until \left\| P_k - P_{k-1} \right\|_1 < \varepsilon\\
&display\ P_k
\end{aligned}
$$

where *e* is the vector with all elements 1, ε is the accuracy threshold and 1 is the norm of the vector calculating by summing up its elements.

HITS (Hyper-link Induced Topic Search) Algorithm

Kleinberg (1999) suggests that there are two types of pages that could be relevant for a query: *authorities* are pages that contain useful information about the query topic, while *hubs* contain pointers to good information sources. Obviously, both types of pages are typically connected: good hubs contain pointers to many good authorities, and good authorities are pointed to by many good hubs.

A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time (Ding, et al., 2001; Klienberg, 1999).

The HITS algorithm treats WWW as directed graph *G(V,E),* where *V* is a set of vertices representing pages and *E* is set of edges corresponds to link. Figure 2 shows the hubs and authorities in web (Kosala & Blockeel, 2000).

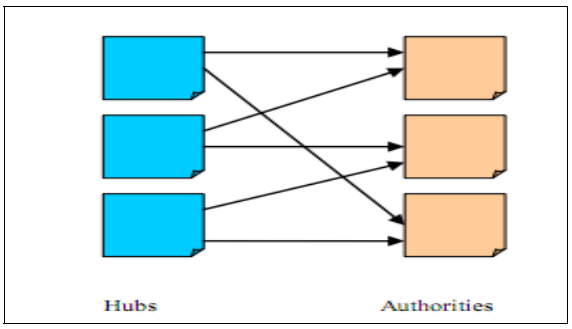

Figure 2. Hubs and Authorities

Kleinberg (1999) suggests to make practical use of this relationship by associating each page x with a hub score $H(x)$ and an authority score $A(x)$, which are computed iteratively:

$$H_{i+1}(x) = \sum_{(x,s)} A_i(s) \tag{6}$$

$$A_{i+1}(x) = \sum_{(p,x)} H_i(s) \tag{7}$$

where (x,y) denotes that there is a hyperlink from page x to page y. This computation is conducted on a so-called *focused subgraph* of the Web, which is obtained by enhancing the search result of a conventional query (or a bounded subset of the result) with all predecessor and successor pages (or, again, a bounded subset of them). The hub and authority scores are initialized uniformly with $A_0(x) = H_0(x) = 1.0$ and normalized so that they sum up to one before each iteration. It can be proved that this algorithm (called HITS) will always converge (Kleinberg, 1999), and practical experience shows that it will typically do so within a few (about 5) iterations (Chakrabarti et al., 1998). Variants of the HITS algorithm have been used for identifying relevant documents for topics in web catalogues (Chakrabarti et al., 1998, Bharat & Henzinger, 1998) and for implementing a "Related Pages" functionality (Dean & Henzinger, 1999).

The HITS approaches combines content-based search with link-based ranking. It makes the basic assumption that if the pages from the root set are closed to the query topic, the pages belonging to the base set (one link farther) are, by their content, similar to the query (Markov & Larose, 2007).

The Hits algorithm has two steps:

1. Sampling Step - in this step a set of relevant pages for the given query are collected.
2. Iterative Step - in this step Hubs and Authorities are found using the output of sampling step.

An important difference between PageRank and HITS is the way that page scores are propagated in the web graph. In HITS the hub collects its score from pages to which it points (Markov & Larose, 2007). The graph shown in Fig. 3. illustrates this. At each step, page u_1 collects its hub score $h(u_1)$ as a sum of the authority scores of the pages to which it points (v_1, v_2, and v_3). At the next step, page v_1 collects its authority score $a(v_1)$ as a sum of the hub scores of the pages that point to it. This process continues until all scores reach some fix point.

Following expressions (8) and (9) are used to calculate the weight of Hub (H_p) and the weight of Authority (A_p).

$$H_p = \sum_{q \in I(p)} A_q \tag{8}$$

$$A_p = \sum_{q \in B(p)} H_q \qquad (9)$$

where H_q is Hub Score of a page, A_q is authority score of a page, *I(p)* is set of reference pages of page *p* and *B(p)* is set of referrer pages of page *p*, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

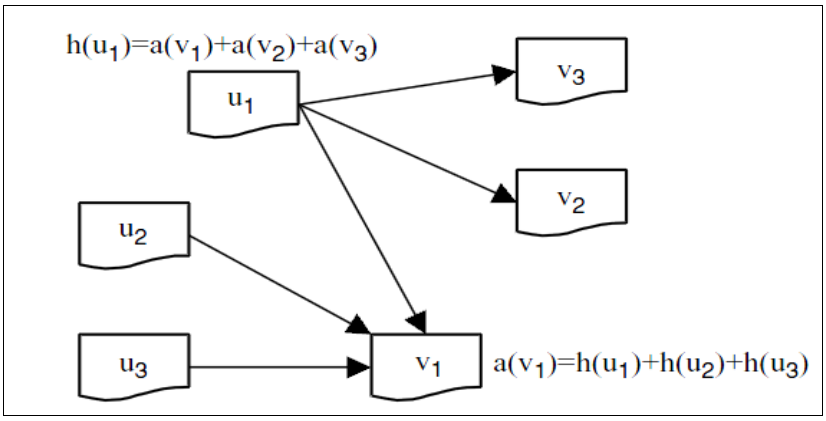

Figure 3. Computing hub *(h)* and authority *(a)* scores (Markov & Larose, 2007)

The following are some constraints of HITS algorithm (Chakrabarti, et al., 1999): it is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities, sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights, HITS gives equal importance for automatically generated links which may not have relevant topics for the user query, HITS algorithm is not efficient in real time.

HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints, HITS could not be implemented in a real time search engine.

The main drawback of this algorithm is that the hubs and authority score must be computed iteratively from the query result, which does not meet the real-time constraints of an on-line search engine.

Complexity of PageRank algorithm is *O(log N)* whereas complexity of HITS algorithms are less than *O(log N)*.

6. CONCLUSION

Web Mining is a powerful technique used to extract the information from past behaviour of users. Web Structure Mining plays an important role in this approach. Various algorithms are used in Web Structure Mining to rank the relevant pages. PageRank, and HITS treat all links equally when distributing the rank score. In this

paper, we approach the research area of Web mining, focusing on the category of Web structure mining. We had introduced Web mining. Later in the paper when we had discussed Web structure mining, and introduced Link mining, as well as block-level link mining issues. We had also reviewed two popular algorithms to have an idea about their application and effectiveness. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES:

[1]. Bharat, K.; Henzinger, M.R. (1998) *Improved algorithms for topic distillation in a hyperlinked environment*, in Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98), pp.104-111

[2]. Cai, D.; He, X.; Wen, J. R.; Ma, W.Y. (2004) *Block-Level Link Analysis*, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR04), pp.440-447. ACM Press

[3]. Chakrabarti, S.; Dom, B.; Gibson, D.; Kleinberg, J.; Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, A. (1999) *Mining the Link Structure of the World Wide Web*, IEEE Computer, Vol. 32, pp.60-67

[4]. Chakrabarti, S.; Dom, B.; Raghavan, P ; Rajagopalan, S.; Gibson, D.; Kleinberg. J. (1998) *Automatic resource compilation by analyzing hyperlink structure and associated text*, Computer Networks, 30(1-7):65-74, Proceedings of the 7th International World Wide Web Conference (WWW-7), Brisbane, Australia

[5]. Cohen, W.W. (2003) *Learning and Discovering Structure in Web Pages*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 26(1):3-10

[6]. Dean, J.; Henzinger, M.R. (1999) *Finding related pages in the World Wide Web*, in A. Mendelzon, editor, Proceedings of the 8th International World Wide Web Conference (WWW-8), pp 389-401, Toronto, Canada

[7]. Ding, C.; He, X.; Husbands, P.; Zha, H.; Simon, H. (2001) *Link analysis: Hubs and authorities on the world*

[8]. 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

[9]. 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

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

[11]. Dinucă, C.E. (2011d) *Using web mining in e-commerce applications*, Analele Universităţii "Constantin Brâncuşi" din Târgu Jiu, Seria Economie

[12]. 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

[13]. Ganciov (Miculescu), A.; Pribac, L.I. (2010) *Knowledge and Information - Factors of Economical and Social Development*, Annals of the University of Petroşani, Economics, 10(1), pp.91-102

[14]. Kleinberg, J.M. (1998) *Authoritative sources in a hyperlinked environment*, in Proc. Of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '98), pp.668-677

[15]. Klienberg, J.M. (1999) *Authoritative sources in a hyperlinked environment*, Journal of the ACM, 46(5), pp.604-632

[16]. Kosala, R.; Blockeel, H. (2000) *Web Mining Research: A Survey*, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1

[17]. Markov, Z.; Larose, D.T. (2007) *Data mining the web*, Uncovering Patterns in Web Content, Structure and Usage, USA: John Wiley & Sons

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

[19]. Palau, J.; Montaner, M.; Lopez, B.; de la Rosa, J.L. (2004) *Collaboration analysis in the recommender system using social networks*, in CIA, pp.137-151

[20]. Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P.N. (2000) *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, ACM SIGKDD Explorations Newsletter, Volume 1 Issue 2

[21]. Stuparu, D.; Vasile, T.; Dăniasă, C.I. (2010) *Bayesian Approach of Decision Problems*, Annals of the University of Petroşani, Economics, 10(3), pp.321-332

[22]. Țarcă, N.; Vătuiu, T.; Popa, A. (2009) *The Importance of the Web Technologies During the Communication Process between a Company and its Clients*, Annals of the University of Petroşani, Economics, 9(2), pp.307-313