ON TWO ALGORITHMS USED IN WEB STRUCTURE MINING

Claudia Elena Dinucă Ph. D Student
University of Craiova
Faculty of Economics and Business Administration
Craiova, Romania
Dumitru Ciobanu Ph. D Student
University of Craiova
Faculty of Economics and Business Administration
Craiova, Romania

Abstract: Due to the continuous growth and spread of the internet using Web Mining to improve the quality of different services has become a necessity. Web Mining is nothing else than applying data mining techniques and algorithms on web data. In this work we present two algorithms used in Web Structure Mining namely Page Rank and HITS. Both algorithms draw their origin from social networks analysis and they are modeled based on the Theory of Markov Chains. Page Rank is used by the search engine GOOGLE and HITS by the search engine CLEVER. We present their strengths, weakness and other areas of applicability.

JEL classification: M15, M21

Key words: critical; Web Mining, Web Structure Mining, Algorithms, Page Rank, HITS.

1. INTRODUCTION

The Web is a critical channel of communication and promoting a company image. E-commerce sites are important sales channels.

The Web has become very popular in the last decade, bringing a strong platform for dissemination of information and knowledge extraction and analysis of information.

Today the Web is known as a repository of data containing a wide variety of data and knowledge base in witch are hidden Web information (Guandong et all. 2011).

Features of Web applications are hypertext links and 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 "lead" 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 the system documents through Web browsers is called browsing. Note that currently most web applications are due to electronic publications and the possibilities the Web offers: an information fast 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 billions of Web pages created by millions of authors, organizations, World Wide Web is a great source of knowledge. The knowledge comes

not only from the content itself, but the unique features of the Web, hyperlinks and the diversity of content and languages.

Web size and dynamic unstructured content makes extracting useful knowledge a challenge for research. Web site generates a large amount of data in various formats that contain valuable information. For example, Web server logs information about user access patterns 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 volume of data in the Web is huge it has become a necessity to apply new techniques to extract information and knowledge much needed for future developments.

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.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.

It is important to use data mining methods to analyze data from the activities performed by visitors on websites (Dinucă, 2011).

Web mining methods are divided into three categories (Cai et al., 2004; Chakrabarti et al., 1999):

• 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 Structure Mining (Web Mining Linkage) 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.

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.

Traditional information retrieval systems and search engines first extract relevant documents for users based on content similarity query entered and indexed pages. In the late 1990s, it was concluded that the methods used that 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.

2. PAGE RANK ALGORITHM

These Page Rank algorithm was created in 1998 by Sergey Brin and Larry Page. Based on this algorithm works 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 N_u 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 $1/N_u$. 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 $1/N_u$ from the prestige of *u*. This is also the idea behind the web page ranking algorithm Page Rank (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 elements

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

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

$$P_1 = A^T P_0, \qquad (2)$$

$$P_k = A^T P_{k-1}, \qquad (3)$$

$$\lim_{k\to\infty} P_k = P. \qquad (4)$$

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 formula:

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

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.

Pseudo code algorithm for calculating the rank of web pages is presented below

$Page\,Rank$

$P_0 \leftarrow \frac{e}{n};$

$k \leftarrow 1;$

repeat

$P_k \leftarrow (1-d)e + dA^T P_{k-1};$

$k \leftarrow k+1;$

until $\|P_k - P_{k-1}\|_1 < \varepsilon;$

display P_k.

where e is the vector with all elements 1, ε is the accuracy threshold and $\|\cdot\|_1$ is the norm of the vector calculating by summing up its elements.

3. HITS ALGORITHM

The 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 1 shows the hubs and authorities in web (Kosala & Blockeel, 2000).

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:

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

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

where *(x,y)* denotes that there is a hyperlink from page *x* to page *y*.

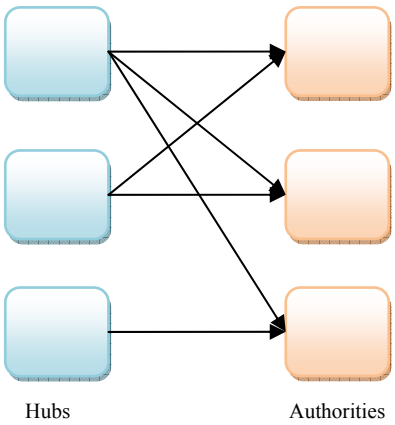


Fig. 1. Hubs and Authorities

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 $A_0(x) = H_0(x) = 1$ and normalized so that they sum up to one before each iterations. 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 "Related Pages" functionality (Dean & Henzinger, 1999).

The HITS approaches combine 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:

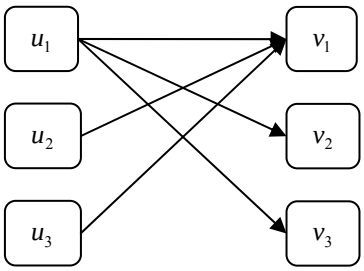
1. Sampling Step - in this step a set of relevant pages for the given query are collected;
2. Iterative Step - in this step Hubs and Authorities are found using the output of sampling step.

An important difference between Page Rank 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. 2. 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.

$$H(u_1) = A(v_1) + A(v_2) + A(v_3)$$



$$A(v_1) = H(u_1) + H(u_2) + H(u_3)$$

Fig. 2. Computing hub *(H)* and authority *(A)* scores (Markov & Larose, 2007).

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

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

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

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.

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

4. CONCLUSIONS

Various algorithms are used in Web Structure Mining to rank the relevant pages. Page Rank, and HITS treat all links equally when distributing the rank score for web pages. These algorithms were originally designed to help web information retrieval methods based on indexes as old ones could not be used in conditions of exponential growth in the size of the Internet in recent years. Page Rank was superior in this area because of the possibility of using in real time.

Page Rank calculates ranks web pages and save them in tables that are permanently updated. When it is received a request it reads only ranks of pages from the table. Hits calculates page ranks after receiving the request leading to a response time of the order of minutes, which makes it unusable for the current search. We note that in the search engines implementations are used also other methods for filtration and removal of spam and for grouping pages, etc, but these are trade secrets. The two algorithms have found other important applications. Page Rank was heavily used in an attempt to predict the next page that will be visited by a web user. We mention here two works of the authors of this article. They have proposed a method of predicting the next pages that will be visited using the algorithm Page Rank (Dinuca & Ciobanu, 2011a) and a further improved work by applying the algorithm only on subset of sessions which contains current page. (Dinuca & Ciobanu, 2011b). Predictions obtained can be used for creating systems of recommendation, preloading pages to increase speed of navigation and to optimize the structure of Web sites.

HITS was used to determine the impact factor of scientific journals, in this case magazines being associated to web sites, web pages being replaced by articles and citations in journals representing connections between pages.

Another area where HITS algorithm may find commercial application is the telephone network. In the last two examples, HITS is better than Page Rank because it takes into account both the inputs and outputs in the calculation of the ranking. For example, in the telephone network by using Page Rank algorithm for providing users ranks it would mean that users are important only if they are contacted by other people. Thus, a user who initiates multiple calls but receives little can be classified incorrectly as unimportant by using Page Rank. By applying HITS algorithm the user would be considered a good hub.

We propose the use of HITS algorithm for trade between countries. In this case they would replace web sites, companies will be web pages and trade between firms will be in place of hyperlinks. For this model, ranks must be given to share links that represent the (value) of transactions. So, the authority score will designate imports and hubs score exports.

In this paper we had 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. 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, 1998.
2. Cai, D. Block-Level Link Analysis., Proceedings of the 27th Annual

	He, X. Wen, J. R. Ma, W. Y.	International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR04), pp. 440–447. ACM Press, 2004.
3.	Chakrabarti, S. Dom, B. Gibson, D. Kleinberg, J. Kumar, R. Raghavan, P. Rajagopalan, S. Tomkins, A.	Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.
4.	Chakrabarti, S. Dom, B. Raghavan, P. Rajagopalan, S. Gibson, D. Kleinberg, J.	Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks, 30(1–7):65–74, Proceedings of the 7th InternationalWorldWide Web Conference (WWW-7), Brisbane, Australia, 1998.
5.	Dean, J. Henzinger,M.R.	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, 1999.
6.	Ding, C. He, X. Husbands, P. Zha, H. Simon, H.	Link analysis: Hubs and authorities on the world, 2001.
7.	Dinucă, C. E.	USING WEB MINING IN E-COMMERCE APPLICATIONS, Annals of the University "Constantin Brâncuşi" from Târgu Jiu, Economic Sciences Series, 2011.
8.	Dinucă, C. E. Ciobanu, D.	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, 2011a.
9.	Dinucă, C. E. Ciobanu, D.	Improving the prediction of next page request by a web user using Page Rank algorithm, 1th WSEAS International Conference "TOURISM AND ECONOMY DEVELOPMENT" (TED '11), Drobeta Turnu Severin, Romania, October 27-29, 2011b.
10.	Kleinberg, J.M.	Authoritative sources in a hyperlinked environment. In Proc. Of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '98), pp 668-677, 1998.
11.	Klienberg, J.M.	Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5):604-632, 1999.
12.	Kosala, R. Blockeel, H.	Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1, 2000.
13.	Markov, Z.; Larose, D. T.	DATA MINING THE WEB, Uncovering Patterns in Web Content, Structure and Usage, USA: John Wiley & Sons, 2007.
14.	Palau, J. Montaner, M. Lopez, B. de la Rosa, J.L.	Collaboration analysis in the recommender system using social networks, In CIA, pages 137-151, 2004.
15.	Guandong, X. Yanchun, Z. Lin, L.	Web Mining and Social Networking, Techniques and Applications, Australia, Springer, 2011.