Web Mining in Search Engines for Improving Page Rank

Sadik Khan, Yashpal Singh, Ajay Kumar Sachan

Abstract: ***An application of web mining can be seen in the case of search engines. Most of the search engines are ranking their search results in response to users' queries to make their search navigation easier. In this research, a survey of page ranking algorithms and comparison of some important algorithms in context of performance has been carried out.So this kind of problem is actual need of this proposed research work. One of the major problems for automatically constructed portals and information discovery systems is how to assign proper order to unvisited Web pages. Topic-specific crawlers and information seeking agents should try not to traverse the off-topic areas and concentrate on links that lead to documents of interest. In this chapter, we propose an effective approach based on the relevancy context graph to solve this problem.Some commonly used link algorithms are page rank, HITS and Weighted Page Content Rank. Most of the search engines are ranking their search results in response to user's queries to make their search navigations easier. In this paper we give a study of page ranking algorithms and description about Pagerank , HITS, based on web content mining and structure mining that shows the relevancy of the pages to a given query is better determined, as compared to the Page Rank and HITS.***

Keywords: ***Web Mining ,Data mining, HITS, Search Engines, web content, Page rank, Web Logs, web structure mining, web content mining.***

I. INTRODUCTION

The Web grows and evolves faster than we would like and expect, imposing scalability and relevance problems to Web search engines. There are three main data types in the Web: content (text, multimedia), structure (links that form a graph) and Web usage (transactions from Web logs). We emphasize the web mining.In this tutorial we present:

- Introduction of Web mining;
- how mining Web data and usage logs allows to improve search engines in several ways (ranking, indexes,queries, and interfaces); and
- a new subfield called query mining that allows to obtain information scent and new content suggestions to improve a Web site.

Server logs of search engines store traces of queries submitted by users, which include queries themselves along with Web pages selected in their answers. Query mining is based in the fact that user queries in search engines and Websites give valuable information on the interests of people.In addition, clicks after queries relate those interests to actual content

Revised Version Manuscript Received on July 09, 2015.

Sadik Khan, Research Scholar, Bhagwant University, Ajmer, India.
Dr. Yashpal Singh, Assoc. Prof. & HOD Department of CS, Bundelkhand Institute of Engineering & Technology, Jhansi, India.
Dr. Ajay Kumar Sachan, Professor & Director, Radharaman Institute of Technology & Science, Bhopal, India.

II. WEB MINING

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Even though it is strongly related to data mining, it is not equivalent to it. Three main axes of Web mining have been identified, according to the Web data used as input in the data mining process, namely Web structure, Web content and Web usage mining.

III. WEB STRUCTURE MINING

Web is a graph. It is a directed labeled graph whose nodes are the documents and the edges are the hyperlinks between them. Web is a huge structure, growing rapidly. This network of information lacks organization and structure, and is only held together by the hyperlinks.

A. Document Structure

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents [12].

B. Citation Analysis

Link analysis has close ties to social networks and citation analysis, the study of the co-citations occurring between scientific papers. The best-known measure of a publication'S importance is the "impact factor" , developed by Eugene Garfield. This metric takes into account the number of citations received by a publication. The impact factor is proportional to the number of citations a publication has. This measure, counts all references equally. However, it is evident that some "important" citations should be given additional weight. The problem is to define what is "important". Pinski et al. [11] overcame this by developing a model for computing the equilibrium for what they defined as "influence weights". The weight of each publication is equal to the sum of its citations, scaled by the importance of these citations. On the Web, the notion of citations corresponds to the links pointing to a Web page. The most simplified ranking of a Web page could be accomplished by summing up the number of links pointing to it. However, this approach favors the most popular Web sites, such as universally known portals, news pages etc. Moreover, the diversity of the content and its quality in the Web should also be taken into consideration. Usually the co-citations are between closed networks of knowledge. The metric proposed by Pinski et al. influenced Brin and Page to develop PageRank [4], the

algorithm hiding behind the most popular search engine, Google [Google]. On the other hand, the notion of important pages in terms of content (authorities) as well as important pages serving as indices (hubs) was introduced in the HITS algorithm [2], which was developed by Kleinberg. This algorithm supports another prototype search engine, Clever [3].

C. The role of hyperlinks in web searching

In order to make navigation in this chaotic structure easier, people use search engines, trying to focus their search by querying using specific terms/keywords. At the beginning, where the amount of information contained in the Web did not yet have these big proportions, search engines used manually-built lists covering popular topics. They maintained an index, containing a list for every word, of all Web pages containing this word. This index was then used in order to answer to the users' queries. However, after a few years, when the Web evolved including millions of pages, the manual maintenance of such indices was very expensive. The automated search engines relying in keyword matching, give results including hundreds (or more) Web pages, most of them of low quality. The need for ranking somehow the importance and relevance of the results was more than evident. The main break-through in the research area of searching the Web came by the realization that the characterization as well as the assessment of a page, i.e. how important it is considered, is enhanced if we take into account how many people consider it important, and how they characterize it. This is performed by using the link structure of the Web, taking into consideration the incoming links to a page. The most important algorithms based on this idea are PageRank [4] and HITS [2].

D. PageRank Algorithm

This algorithm is influenced by citation analysis, considering theincoming links as citations to the Web pages. However, by simply applying citation analysis techniques to the diverse set of Web documents, would not result in as good outcomes. Therefore PageRank provides a more sophisticated way to compute the importance of a Web page than simply counting the number of pages that have a link pointing to it (named as "backlinks"). If a backlink comes from an "important" Web page, then it weighs more than others that come from minor pages. Intuitively, "a page has high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has few highly ranked backlinks" [4]. In otherwords, we may consider that links from a page to another as a vote. However, not only the number of votes a page receives is considered important, but the "importance" of the ones that cast these votes as well. First thing that should be computed is the number of links pointing to every Web page. This is something not known apriori, therefore a technique based on random walk on graphs is employed. Intuitively, they model the behaviour of a "random surfer". The "random surfer" visits a page and then follows its links to other pages equiprobably. In the steady state each page will have a "visit rate" which will be what defines its importance.

The following equation calculates a page's PageRank:

$$PR(A) = (1- d) + d(PR(tl)/C(tl) + ... + PR(tn)/C(tn)),$$

where tl - tn are pages linking to page A, C is the number of outbound links that a page has and d is a damping factor, usually set to 0.85.

PageRank : Examples

Simple calculations

In the following we will illustrate PageRank calculation. For this, we are using the normalisation (equation) M * PR = (1 - d). Most of the calculations are done analytically. To get numerical results one has to insert numerical values for the different parameters, e.g. taking d = 0.85 for the damping factor.

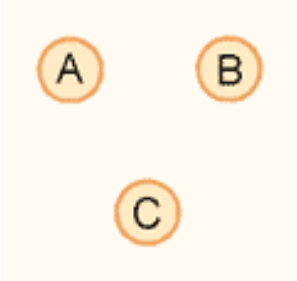

Fig-1.0

Example:

Not connected pages are the simplest case. One gets

$$PR_A = PR_B = PR_C = (1 – d)$$

In fig1.0 all pages have the same PageRank. 1 - d is the minimal PageRank value. The solution is independent from the number of (not connected) web pages.

E. HITS Algorithm

He identifies two different forms of Web pages: hubs and authorities. Authorities are pages bearing important content. Hubs are pages that act as resource lists, directing users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject. We should stress that a page might be a good hub and a good authority in the same time. This circular relationship leads to the definition of an iterative algorithm, HITS. Hyperlink-Induced Topic Search (HITS) (also known as Hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. A page may be a good hub and a good authority at the same time. 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. Attempts to computationally determine hubs and authorities on a particular topic through analysis of a relevant sub

graph of the web.

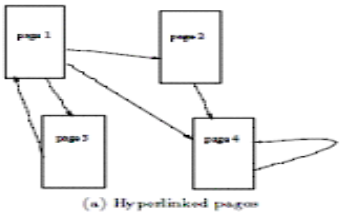

Fig-2.0

The algorithm converges after a few iterations. However, since HITS makes iterative computations from the query result, it makes it difficult to meet the real-time constraints of an online search engine.

F. Search Engines

As already mentioned PageRank is the supporting algorithm of the most popular search engine at this time, Google, and HITS is used from Clever. Google and Clever have two main differences: Google is query-independent, whereas Clever computes the rankings according to the query terms. Secondly, Google looks only to the forward direction, whereas Clever also takes into consideration the backward direction too, leading through its computation to the creation of web communities. Therefore Google works well on answering specific queries, whereas HITS works well on answering broad-topic queries. To be more specific, in Google, when a query is given, all pages meeting the query (i.e. containing the search term) are firstly retrieved, but are presented to the end user ranked according to their PageRank. Therefore, the order is query independent. Of course, PageRank is not the only algorithm hiding behind Google. Even if it's not clearly stated, a number of heuristics are also used to support the ranking of the results presented to the end user.

IV. WEB CONTENT MINING

Web content mining has to do with the retrieval of information (content) available on the Web into more structured forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semistructured (HTML documents), or structured (extracted from databases into dynamic Web pages).

A. Data Preprocessing

Web content mining is strongly related to the domain of Text Mining, since in order to process and organize Web pages their content should be first appropriately processed in order to extract properties of interest. These selected properties are subsequently used to represent the documents and assist the clustering or classification processes. We discriminate four stages of data preprocessing, based on techniques used in text mmmg, namely Data Selection,Filtering, Cleaning, and Representation [9].

B. Web document representation models

In order to reduce the complexity of the documents and make them easier to handle, during the clustering and/or classification processes, one should first choose the type of characteristics or attributes (e.g. words, phrases, or links) of the documents that are of importance, and how these should be represented. Since documents are represented in a uniform way, the similarity between two documents can then be easily calculated.

V. WEB USAGE MINING

The process of analyzing the user's browsing behavior is called Web usage mining. It can be regarded as a three-phase process, consisting of the data preparation, pattern discovery and pattern analysis phases [8]. In the first phase, Web data are preprocessed in order to identifY users, sessions, pageviews, and so on. The input data are mainly the hits registered in the Web usage logs of the site, sometimescombined with other information such as registered user profiles, referrer's logs, cookies, etc [1].

A.Web Server Data

The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

B. Identifying navigational patterns

The users' activity when browsing through Web sites is registered in these sites' Web logs. Considering the average number of visits to a medium-sized Web site per day, we can presume that the amount of information hidden in the site's Web logs is huge, yet meaningless if they're not appropriately processed. By processing these data, either using simple statistical methods, or by using more complicated data mining techniques, we can identify interesting trends, and patterns concerning the activity in the Web site. Site administrators can then use this information to redesign or customize the Web site according to the interests and behavior of its visitors, or improve the performance of their systems.

C. Web usage logs

Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format. The fields of the common log format are: remotehost rfc931 authuser date "request" status bytes Except for Web server logs, which are the main source of information, usage data can also be acquired by proxy server logs, browser logs, user profiles, registration data, cookies, mouse clicks etc.

D. Data Preprocessing

The first issue in the preprocessing phase is data preparation. Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. Furthermore, crawler activity can befiltered out, because such entries do not provide useful information about the site's usability. Another problem to be met has to do with caching. Accesses to cached pages are not recorded in the Web log, therefore such information is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be dealt with easily. In such cases, cached pages can usually be inferred using the referring information from the logs.

VI. CONCLUSIONS

The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key computer science contributions made by the field, a number of prominent applications, and outlined some promising areas of future research.. Web mining is a very broad research area trying to solve issues that arise due to the WWW phenomenon. In this paper, after analyzing the three separate categories of Web mining, we tried to make a prediction concerning its future. The distinctions between the three axes of Web mining (especially of Web content and Web structure mining) are in many cases ambiguous. This information is usually expressed using the expressive power of the Semantic Web backbone, which are the ontologies. This representation , closes the gap between Semantic Web and Web Mining areas, to create a fastemerging research area, that of Semantic Web Mining.Therefore the need for discovering new methods and techniques to handle the amounts of data existing in this universal framework will always exist.

REFERENCES

1. M Eirinaki, M Vazirgiannis, Web Mining for Web Personalization, in ACM Transactions on Internet Technology (TOIT), 3(1), February (2003).
2. 1.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5):604-632, September (1999).809
3. S. Chakrabarti, B. Dom, D. Gibson, 1. Kleinberg, R Kumar, P. Raghavan, S. Rajagopalan, A Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer (1999) Vol.32 No.6.
4. S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks, 30(1 7): 107-117, 1998, Proceedings of the 7th International World Wide Web Conference(WWW7).
5. 1.M. Kleinberg, Hubs, Authorities, and Communities, ACM Computing Surveys, 31 (4), December (1999).
6. D7. Gibson, J. Kleinberg, P. Raghavan, Inferring Web Communities from Link Topology, in the Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, (1998).
7. R Kumar, P. Raghavan, S. Rajagopalan, A Tomkins, Trawling the Web for Emerging Cyber-Communities, in Proceedings of the 8th WWW Conference (WWW8), (1999).
8. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, January 2000Nol. 1, Issue 2, pp. 12-23.
9. M. Rajman, M. Vesely, From Text to Knowledge: Document Processing and Visualization: a Text Mining Approach, in Proceedings of the NEMIS Launch Conference, International Workshop on Text Mining & its Applications, Patras, Greece, April(2003).
10. N. Oikonomakou, MVazirgiannis, A Review of Web Document Clustering approaches, in Proceedings of the NEMIS Launch Conference, International Workshop on Text Mining & its Applications, Patras, Greece, April (2003).
11. G. Pinski, F. Narin, Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, in Information Processing and Management. 12, (1976).
12. S. Shearin and H. Liebermann, Intelligent Profiling by Example, Proc. of Intern. Conf. of Intelligent User Interfaces (IUI2001), p. 145-152, Santa Fe, NM, Jan. 14-17, 2001.
13. WEB MINING: A ROADMAP, Magdalini Eirinaki, Dept. of Informatics Athens University of Economics and Business.
14. Evaluating the datamining techniques and their roles in increasing the search speed data in web, Ayatollah Amoli Branch, Comput. Dept., Islamic Azad Univ., Amol, Iran , DOI: 10.1109/ICCSIT.2010.5563818 Conference: Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, Volume: 9