ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

IJETMR
International Journal of Engineering Technologies and Management Research
A Knowledge Repository

AN OPTIMIZED PAGE RANK ALGORITHM WITH WEB MINING, WEB CONTENT MINING AND WEB STRUCTURE MINING

Kwame Boakye Agyapong[1], Dr. J.B.Hayfron-Acquah[2], Dr. M. Asante[3]
[1] PhD Candidate, Computer Science, K.N.U.S.T, Ghana
[2, 3] Senior Lecturer, Computer Science, K.N.U.S.T, Ghana

Abstract:
With the rapid increase in internet technology, users get easily confused in large hypertext structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs. In order to achieve this goal, they use the concept of web mining. Web mining is used to categorize users and pages by analyzing the users" behaviour, the content of the pages, and the order of the URLs that tend to be accessed in order. Most of the search engines are ranking their search results in response to users' queries to make their search navigation easier. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks. It is very difficult for a user to find the high quality information which he wants. Page Ranking algorithm is needed which provide the higher ranking to the important pages. In this paper, we discuss the improvement of Page ranking algorithm to provide the higher ranking to important pages. Most of the search engines are ranking their search results in response to user's queries to make their search navigations easier.

Keywords: *PageRank; Web Content Mining; Web Mining; Web Structure Mining; Web Usage Mining.*

Cite This Article: Kwame Boakye Agyapong, Dr. J.B.Hayfron-Acquah, and Dr. M. Asante. (2017). "AN OPTIMIZED PAGE RANK ALGORITHM WITH WEB MINING, WEB CONTENT MINING AND WEB STRUCTURE MINING." *International Journal of Engineering Technologies and Management Research*, 4(8), 22-27.
DOI: 10.5281/ zenodo.914660.

1. Introduction

Today, the World Wide Web is the popular and most interactive medium to disseminate information. The Web is huge, diverse and dynamic. The Web contains vast amount of information and provides an access to it at any place and at any time. Most of the people use internet for retrieving information. But most of the time, they gets lots of insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

When we search any information on the Google, there are many URL's that opens. The bulk amount of information becomes very difficult for the users to find, extract and filter the relevant

ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

information. So web mining techniques are used to solve these problems. Web mining is the application of Data Mining technique to find useful information from web data. With the help of web, we can access multiple data. In the distributed information environment, document or objects are usually linked together to facilitate interactive access so that we can easily access information.

2. Components of Web Search Engine

The components of a web search engine are the User Interface, Parser, Web Crawler, Database and Ranking Engine.

The User Interface -It is the part of Web Search Engine interacting with the users and allowing them to query and view query results.

The Parser- It is the component providing term (keyword) extraction for both sides. The parsers determine the keywords of the user query and all the terms of the Web documents which have been scanned by the crawler. Term extraction procedure includes the following sub procedures: Tokenization, Normalization, Stemming and Stop word handling.

The Web Crawler - A web crawler is a relatively simple automated program, or script that methodically scans or "crawls" through Internet pages to create an index of the data it is looking for. Alternative names for a web crawler include web spider, web robot and automatic indexer. When a web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich meta tags. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. Lastly, the website is included in the search engine's database and its page ranking process.

The Database - It is the component that all the text and metadata specifying the web documents scanned by the crawler are kept.

The Ranking Engine - This component is mainly the ranking algorithm operating on the current data, which is indexed by the crawler, to be able to provide some order of relevance, for the web documents, with respect to the user query. The Web is perhaps the single largest data source in the world. Due to the heterogeneity and lack of structure, mining and integration are challenging tasks.

3. Web Mining

Web Mining is the use of data mining techniques to automatically discover and extract Information from web documents and services. The World Wide Web, www or web is becoming a complex universe. Web mining methodologies are Web Content Mining, Web Structure Mining and Web Usage Mining

ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

4. Web Content Mining

Web content mining is the improvement of information easy to get to on the Web into more structured forms, resting on its indexing for simple tracing of information site. Web content may be unstructured (plain text), semi structured (HTML documents), or structured (removed from databases into active Web pages) (Khan et al, 2015)

A. Data Preprocessing

Web content mining is effectively connected to the ground of Text Mining, consequently in authority to track and bring as one Web page their content ought to be foremost rightfully administer sequentially to take away goods of special treatment. These designated possessions are consequently used to signify the leaflets and assist the gathering or cataloging procedures (Rajman and Vesely, 2003).

B. Web document representation models

To reduce the complicatedness of the leaflets sequentially and make them easier to grasp, for the duration of the assembly and/or classification measures, one ought to initially choose the kind of facial appearance or character (e.g. arguments, expressions, or relations) of the leaflets that are of standing, and how these ought to be signified. In the interim leaflets are signified in a smooth method, the similarity stuck between two leaflets can then be easily calculated.

5. Web Structure Mining

The process by which we discover the model of link structure of the web pages is termed as Web Structure Mining. We list the links; produce the information such as the resemblance and relatives among them by captivating the benefit of hyperlink topology. PageRank and hyperlink analysis also fall in this category. The aim of Web Structure Mining is to produce prepared synopsis about the web site and web page. It tries to find out the link arrangement of hyperlinks at bury file level. The web documents contain links and they use both the actual or most important data on the web so it can be established that Web Structure Mining has a relation with Web Content Mining. It is quite often to combine these two mining tasks in an application.

6. Web Usage Mining

The procedure used to determine the user's browsing behavior is called Web usage mining. There are three phase process, consisting of the data preparation, pattern discovery and pattern analysis phases (Srivastava et al, 2000). In the first phase, Web data are preprocessed in order to recognize users, sessions, page views, and the like. The hits registered in the Web usage logs of the site are mainly the input data, from time to time joint with additional information such as registered user profiles, referrer's logs, cookies (Eirinaki & Vazirgiannis, 2003).

A. Web Server Data

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

ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

B. Identifying Navigational Patterns

The Web logs sites' registers the users' activity when browsing through Web sites. Taking into consideration the normal figure of visits to a medium-sized Web site per day, one can guess that the quantity of information unseen in the site's Web logs is enormous, yet worthless if they're not properly processed. By giving out these data, either using easy numerical methods, or by using more difficult data mining techniques, one can recognize motivating trends, and patterns concerning the activity in the Web site. Site administrators are able to then use this information to revamp or modify the Web site according to the wellbeing and performance of its guests, or get better the performance of their systems.

C. Web Usage Logs

The access log is used to record each access to a Web page in the Web server that hosts it. The entries of a Web log file comprise of fields that go after a predefined arrangement. The fields of the widespread log arrangement are: remote host 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. (Khan et al, 2015).

D. Data Preprocessing

Data preparation is the first issue in the preprocessing phase. Web log data may require to be cleaned from entries concerning pages that returned an error or graphics sleeve accesses. In addition, crawler action can be cleaned out, for the reason that such entries do not give useful information regarding the site's usability. An additional difficulty to be met has to do with caching. Accesses to cached pages are not recorded in the Web log; as a result such information is lost. Caching is deeply reliant on the client-side technologies applied and as a result cannot be dealt with without difficulty. In such cases, cached pages can more often than not be incidental using the referring information starting the logs.

7. Pagerank

This algorithm was developed by Brin and Page at Stanford University which extends the idea of citation analysis (Kleinberg, 1999). In citation analysis the arriving links are treated as credentials but this method could not give productive outcome because this gives some estimate of significance of page. So PageRank gives an improved move toward that can calculate the significance of web page by merely counting the number of pages that are linking to it. These links are called as backlinks. Page ranking algorithms are used by the search engines to present the search outcome by bearing in mind the significance, meaning and content score and web mining techniques to order them according to the user attention. The link from one page to another is considered as a vote. The significance of pages that casts the vote is also significant but not only is the number of votes that a page receives considered as significant. Page and Brin (ref) proposed a formula to calculate the PageRank of a page A as stated below:

PR(A)= (1-d)+d(PR(T1)/C(T1)+…..+PR(Tn/C(Tn))- c ....(1) where PR(Ti) is the PageRank, Ti are links to page A, C(Ti) is number of out links on page Ti and d is damping factor. It is used to stop other pages having too much influence. The total vote is “damped down” by multiplying it to 0.85.

ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

The PageRank forms a probability distribution over the web pages so the sum of PageRank of all web pages will be one.

The PageRank of a page can be calculated without knowing the final value of PageRank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. PageRank of a page depends on the number of pages pointing to a page.

Limitation

- Result come at the time of indexing but not at the query time
- New pages have less page rank and they take much time to be listed and gain high ranks.
- PageRank scores do not reflect current events

We have therefore developed an algorithm that will solve these problems. A Simplified version of PageRank is defined in the flowchart shown in fig 1.

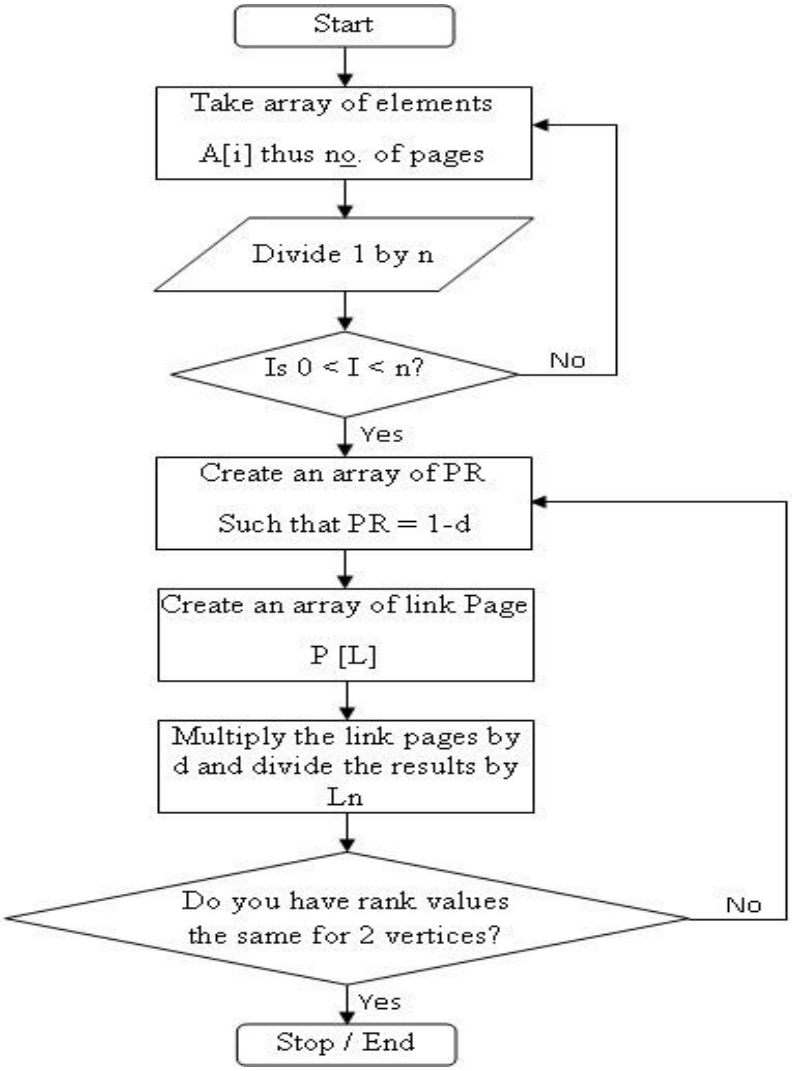

Figure 1: Flowchart of the PageRank

ISSN: 2454-1907
DOI: 10.5281/zenodo.914660

A [i] - Array of elements to be ranked
n - Total number of elements in the array
d - dumping factor 0 < d < 1
1 - d - To avoid some page rank, we make room for pages that do not have on this
PR - Page rank
L - Link pages
Ln - Total number of outgoing edges of L

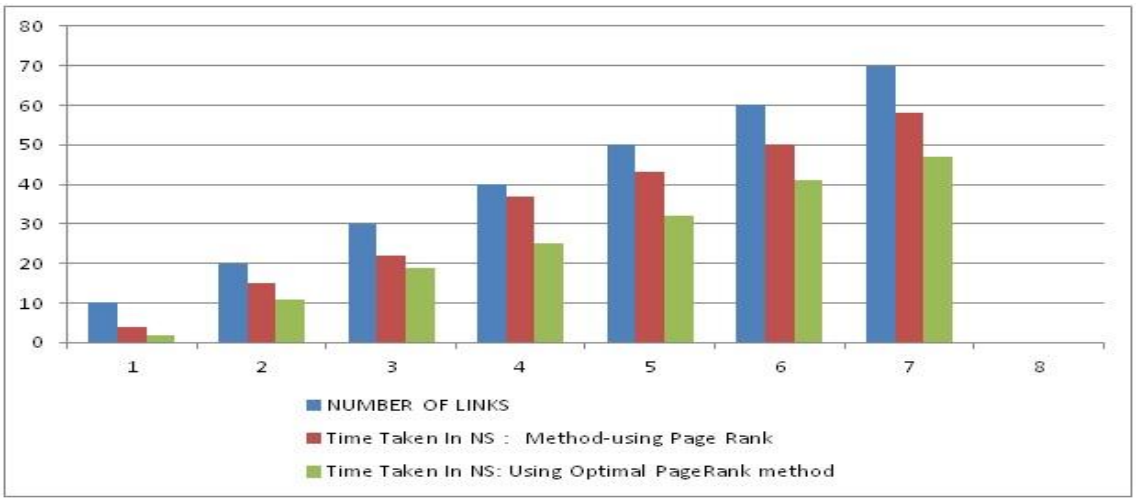

Figure 2: Comparison as a result of time taken to retrieve records/data (Nanoseconds)

8. Conclusion

The Page Rank, one of the algorithms used for link analysis shows that the taken in retrieving a data is much slower in the Page Rank algorithm than the proposed optimal PageRank algorithm. The new approach aims at obtaining information that may assist web site recognition to better suit the user. The logs include information about the referring pages, user identification, time a user spends at a site and the sequence of pages visited.

References

[1] Rajman, M., Vesely, M. (2003). 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.
[2] Srivastava, J, Cooley, R. Deshpande, M. Tan, P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, l (2):12-23.
[3] Eirinaki, M. Vazirgiannis, M. (2003). Web Mining for Web Personalization, in ACM Transactions on Internet Technology (TOIT), 3(1)
[4] Kleinberg, M. Hubs, Authorities, and Communities, ACM Computing Surveys, 31 (4), December (1999)
[5] [5]. Kosala, R. Blockeel, H. (2000). Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2(1):1-15.

*Corresponding author.
E-mail address: opanin007@ yahoo.com