REDUCING EXECUTION TIME IN GOOGLE PAGERANK ALGORITHM WITH PRECALCULATED RANKING

Htet Aung Kyaw [1], Khaing Thanda Swe [2]

Mandalay Technological University (MTU), 05072, Myanmar, Burma

Abstract

Information Retrieval on the web is very important and also very complex operation for web mining. Web mining is the application of data mining techniques to extract the useful information from the web. Web data are web content, web structure, web usage. This paper is basically focused on Google Search Engine. Google's heart is Google Pagerank Algorithm. Because of enormously increased the number of web site on the Internet, the execution of Pagerank Algorithm should be easy and faster in operation.

In this paper, the original Google Pagerank Algorithm was split into two algorithms in distribution operation.To demonstrate our proposed method, we have been created simple website using mostly HTML, CSS and also bootstrap, front-end framework for web design. Then using Python programming, we created link number extractor (simple crawler) program that can extract every anchor tag and also total number of <a> tag in specific webpage. And finally, take the convergence value to specify the iteration count during the operation of proposed Pagerank Algorithm. The strong point of Proposed Pagerank Algorithm is faster in program execution time more than original pagerank algorithm.

Keyword: Google pagerank algorithm, Google's heart, Information Retrieval, Web Mining.

1.INTRODUCTION

Today world is becoming closer with the aid of technologies. These technologies are also improved days after days. Especially most researchers are dependent on the world wide web [1] (www) such as Yahoo, Bing, Ask, Google etc. Among them, Google search engine is plentiful of resources than the other search engines. Therefore, most users, especially researchers, used Google throughout their search.

Actually, web technology is based on the information retrieval [2], which is very important and very complex operation. Because extracting web data is different from traditional information retrieval techniques. Especially in this paper, using about information retrieval in extraction of links which included in every single webpage. For that operation, the program needed webpage URLs in order to crawl throughout those webpages. Link number and links text extraction program will produce those two results for the next operation. The next operation needed the total number of webpages to calculate iteration count value in a specific.

The proposed system is about to enhance Google Pagerank Algorithm's operation [3] especially in iteration step. Also learn about SEO[11] for Google and Web Crawler, Spider, which are crawled websites in order to specify how important it is. Proposed Pagerank Algorithm was implemented by Python programming [4]. By separating original pagerank algorithm, the execution time of program is obviously faster than original. The program flow is not dependent on the total number of web page.

2.EXISTING OPERATIONS

Sergery Brin and Lawerence Page [5] were developed original of Google Pagerank Algorithm at Standford University as PHD thesis that are based on hyperlink structure in 1998. At the beginning of the research, they did not tend to be a commercial product. But finally, it becomes because of outstanding powerful operation in web search results. Their algorithm was better

performance than existing Search Engine's Algorithm at that day. The operation of Google search engine has two features in order to produce the high precision results. The first one is using of the web structure of web graph to calculate the ranking quality of every single web page which is called Pagerank score. The second one is utilizing links to improve the search results.

My proposed algorithm is better than original in the operation of pre-calculated PR value, execution time, because of splitting two operations. Therefore, we can reduce the operation of program execution time, no matter how many web pages.

Dennis Johansson [6] from Uppsala University was published a research paper which studies the relation between keywords and website ranking in Google Search. In this paper, iPropest organization was supported to him for analysis website in order to test his results. The purpose of this paper is placing keywords in webpage. In doing so, how much improve the quality of that webpage in Google's important list for search results. He presented his result by using keyword optimization method, how improve the ranking of webpage. As the future work, he was introduced by testing for difference languages such as English, French, India in keyword value.

In this paper, the author is focused on the query keywords, which can effective on Google's important lists. But for us, we focus on the pre-calculated PR value and does not depend on query keywords.

Ritu Sachdeva (Sharma) [7] was discussed about various ranking algorithm as a survey on Pagerank Algorithm in 2018. In this paper, the author described about web mining technologies and about various algorithms which are adopted from original Google Pagerank Algorithm. The author also described about 10 author proposed methods with their features. Finally, the author tested those various proposed methods and compared with original pagerank algorithm's results by using difference damping factor values.

This paper is survey of various PR Algorithms, which are original and advanced PR Algorithm. The author evaluated the differences between these algorithms using various damping factor values. Our proposed one is not depended on damping factor values. Our algorithm is focus on the pre-calculated page rank value.

Ao-Jan Su [8], who was developed the myths and reality for the improvement of Google Ranking in October 2010. The system of this paper can predict the improvement of Google ranking score based on keywords and also in just content-only. In here, the system can predict 7 out of the top ten pages for 78% in keywords evaluation. For content-only ranking, their system can correctly predict 9 or more page out of the top ten one for 77% of search query results. Actually, they focus on the keyword placement in the website where domain name, title, header, body, image, heading tag etc.

In this paper, the author went to improve in actual Google page ranking result of first tenth pages list. Therefore, this paper is included knowledge of SEO (Search Engine Optimization). But for our proposed one is only the calculation of pre-calculated PR value operation in program execution time.

Aritra Banerjee [9] was published a paper which was advanced to Google Pagerank Algorithm. In this paper, they modified the existing pagerank algorithm's ranking mechanism as an advanced which based on Semantics, In-links, Out-links and Google Analytics. Google Analytics is used to store the hits rate of a website in a particular variable and also for adding the required percentage amount of ranking procedure. For their advanced method, they have been created some webpage to calculate the rank score by using their methods. Firstly, they take the total number of out-links, in-links and Google Analytics hits rate of a specific webpage. Then, they used the following equation in order to calculate the rank score.

Rank score = c * s + d * vl * (inlinks + outlinks) + (ga * No. of hits)/ 1000

Where:

- c is 0 or 1 that depend on Meta data not matched or matched.
- s is used for Semantics.
- d means dmaping factor.
- vl stand for visit of links.
- ga is for Google Analytics amount

This paper is closely related with our proposed algorithm but not at all. In this paper, the author care about Semantic, In-Links, Out-Links and Google Analytics. But our proposed system is needed on In-Links, Out-Links and total number of web pages in a specific website.

3.METHODOLOGY

The main function of Web Mining [10] Techniques is to discover useful information from the Web based on hyperlink structures, web page contents and data usage on the web. Web mining is not an application of data mining technique because of vast data of the web such as structured, semi-structured and even unstructured nature of web pages.

Web mining techniques can be broken into three portions: Web Structure Mining, Web Content Mining and finally Web Usage Mining (sometimes called Web Logs Mining). This paper is focus on Web Structure Mining because PageRank Algorithm is depended on structure of website.

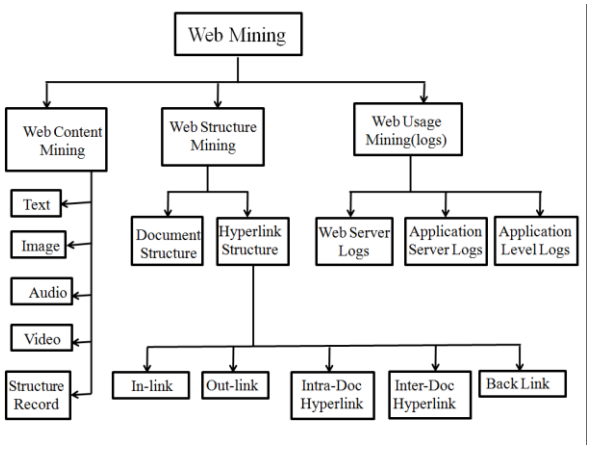

Figure 1Basic Structure of Web Mining Technique

3.1. Web Structure Mining

Mining to the structure of the website is called the "Web Structure Mining". This mining is based on either with or without the description of the links. Background theory is based on the Markov Chain [7] Model, in Google, in order to categorize the web page. Typical web graph structure consists of webpage as nodes and hyperlinks as edges as shown in Figure 2.

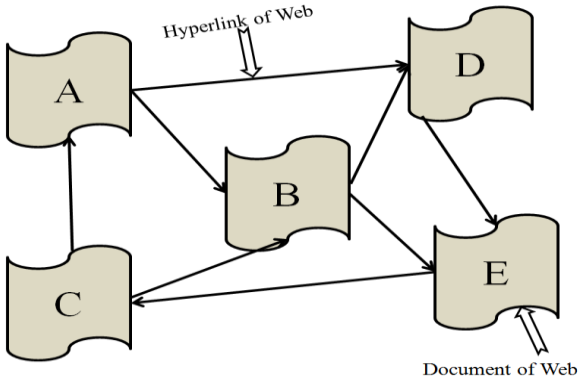

Figure 2Typical Web Graph

3.2. Web Content Mining

Web Content Mining is basically extracting useful information from the content of website, such as text, images, videos, audios and also structured format like lists and tables.

3.3. Web Usage Mining (Web Logs Mining)

Web Usage Mining sometimes called Web Logs Mining technique used to discover the usage patterns from the web.

4.PROPOSED PAGERANK ALGORITHM

In this paper, create a website in order to check and test for proposed pagerank algorithm. That website included five pages which are tightly linked to each other, to calculate the pagerank score using original pagerank calculation equation (pre-calculated PR score). To validate the calculated PR scores which are compared with the result of online web master tool's output. For evaluation, create another website which has eleven pages same linked structure like the previous one.

In this paper, proposed algorithms are implemented in Python [8] programming according the hosted website that used to demonstrate. There are three portions in this propose method. The first one is to get the total number of links on those two websites. The next one is

about to get the specific iteration count for the next step. The final one is the calculation of original Pagerank Formula using total number of In-Links, Out-Links, number of webpage and iteration count result to specific in the iteration.

A. Calculating Specific Iteration Count Result Algorithm: This diagram is about getting specific iteration count value using Original Google PageRank Algorithm.

Step 1: get link numbers of every webpage (inlinks, outlinks)

Step 2: calculate specific iteration count base on original PR Algorithm

Step 3: Output Si (Specific result) value

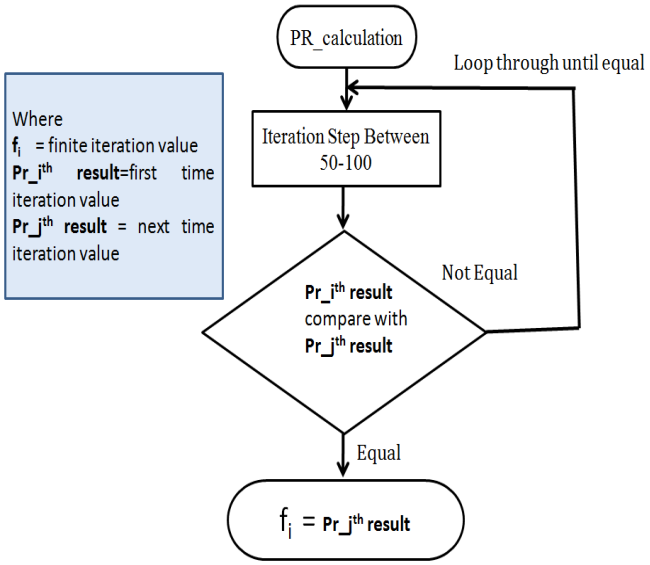

Figure 3 Getting Specific Iteration Count Flow (fi = Si)

B. Calculating Proposed PR Algorithm:

This diagram is represented about calculation of pre-calculated Page Rank value using specific iteration count value.
Step 1: Getting total number of webpages (Web Crawler [12])

Step 2: Getting links number of every webpage
Step 3: Getting specific count result from previous one

Step 4: Calculation of proposed PR Algorithm

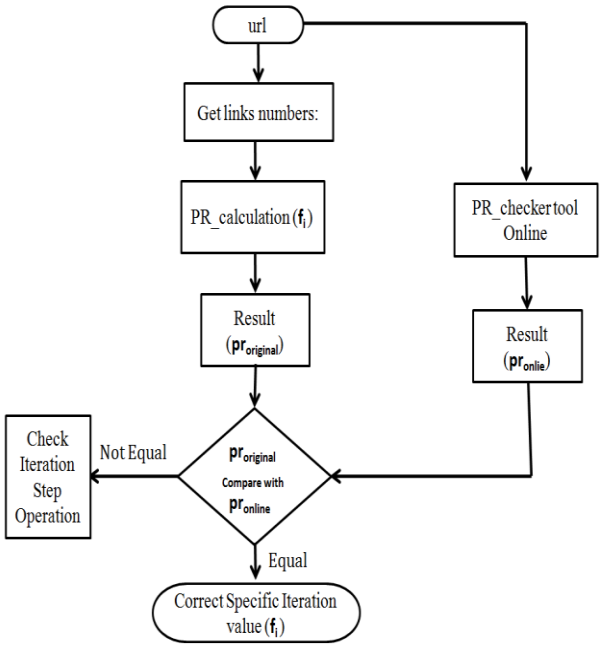

Figure 4 Working Flow of Proposed Algorithm

5.RESULTS AND PERFORMANCE

By separating original Google Pagerank Algorithm, the executing time of proposed Pagerank Algorithm is much faster than original. In here, comparison of program's executing time in seconds of original and propose Pagerank Algorithm as bellow. There are two websites. The first one has five webpage and the next has eleven webpages. We extract every links number for those two websites. Then we calculate the specific iteration count value using Original Google Page Rank Algorithm. By the final result of proposed PR Algorithm execution time is checked as below.

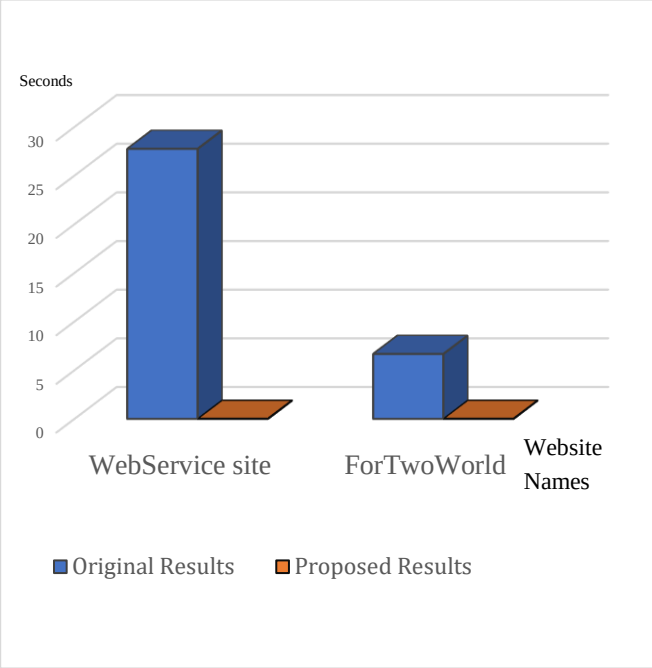

Figure 5.1 Comparison results of Original PR and Proposed PR Algorithm (in seconds)

In figure 5.1, Web-Service (Ori:) means the website name of Web-service calculated with original pagerank algorithm and time taken is 27.906549 seconds long. The next one is clear that for Proposed PR method calculation and which result is 0.0 seconds. For the next website, for-two-world (website name), in original, 6.12506890 seconds and in proposed, it takes 0.0 seconds.

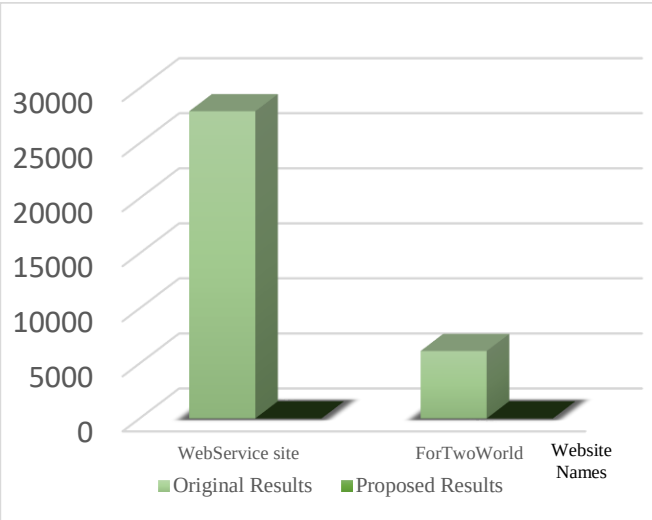

Figure 5.2. Comparison results of Original PR and Proposed PR Algorithm (in Milliseconds)

In figure 5.2, test results for Millisecond are 27906.54969215 milliseconds long in original for "Webservices" website and also 6125.068902969 millisecond in original result for "Fortwoworld" website. As a proposed result for both websites are just 0.0000000 milliseconds times.

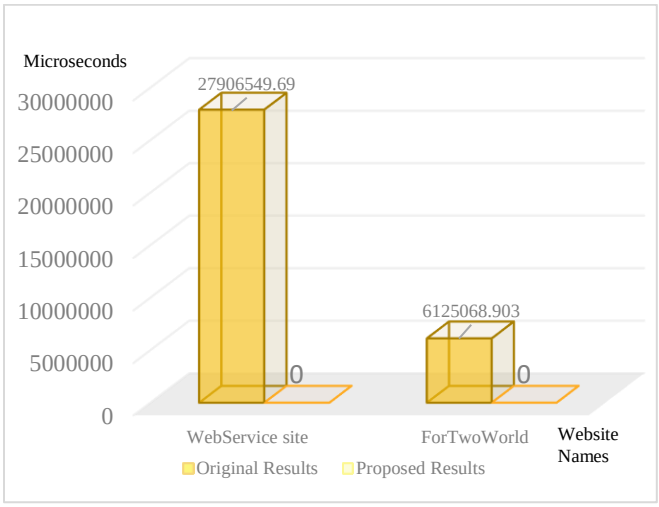

Figure 5.3. Comparison results of Original PR and Proposed PR Algorithm (in seconds)

Figure 5.3, is represent for time taken in microseconds for two websites. Original results are 27906549.6921539 microseconds for "Webservices" website and 6125068.902969360315156 microseconds for "Fortwoworld" website. The proposed results are same as above wo results, 0.0000microseconds.

6.ACKNOWLEDGEMENT

In preparation of my research, I had to take the help and guidance of some respect persons, who deserve my deepest gratitude. Firstly, I would like to thanks to my parents who were support me in here. As the completion of this paper gave me much pleasure, I would like to thanks to the prime minister of Science and Technology, Dr.Myo Thein Gyi and also thanks to rector of Mandalay Technological University, Dr.Sint Soe. In addition, I would like to thank to all of teachers in CEIT department

of MTU who were sitting in every seminar of me and giving me an inspiration to improve the quality of my research.

REFERENCES

[1] https://web.stanford.edu/class/cs344g/www-1992.pdf
The world-wide web by T.J. Berners-Less, R. Cailliau and J.-F. Groff, Computer Networks and ISDN Sysems 25 (1992).

[2] https://webis.de/downloads/publications/papers/stein_2009b.pdf.
Information Retrieval: Concepts and Practical Considerations for Teaching a Rising Topic by Deniel Blank, Norbert Fuhr, Andreas Henric, Thomas Mandl, Thomas Rolleke, Hinrich Schutze, Benno Stein

[3] https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf
Google pagerank Algorithm by Eric Roberts in November 9, 2016.

[4] https://www.python.org/doc/
Python Programming's Documentation

[5] http://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.
The Antomy of Large-Scale Hypertexual Web Search Engine by Sergey Brin and Lawrence Page at Computer Science Department, Stanford University, Stanford, CA 94305, USA

[6] http://www.diva-portal.org/smash/get/diva2:937802/FULLTEXT01.pdf
"Search Engine Optimization and Long Tail of Web Search" by Dennis Johnsson in June 16, 2016.

[7] https://www.ijser.org/researchpaper/A-Literature-Survey-on-Page-Rank-Algorithm.pdf
"A Literature Survey on the PageRank Algorithm" by Ritu Sachedeva in May 2018 and published at ISSN.

[8] https://sci-hub.tw/https://ieeexplore.ieee.org/abstract/document/5616194
"How to Improve Your Google Ranking: Myths and Reality" by Ao-Jan in October 2010.

[9] https://arxiv.org/ftp/arxiv/papers/1709/1709.02858.pdf
"Advanced Page Rank Algorithm with Semantics, In Links,Out Links and Google Analytics' by Aritra Banerjee in August 2017.

[10] https://www.researchgate.net/publication/274708192_CURRENT_LITERATURE_REVIEW_-_WEB_MINING
CURRENT LITERATURE REVIEW - WEB MINING by K.Dharmarajan-Scholar, Dr.M.A.Dorairangaswamy at Bharathiar University, Coimbatore- 641 046, India on 12 October 2017.

[11] http://www.seobook.com/seo-tools.pdf
Search Engine Optimization Book by Aaron Matthew Wall, about how to make informed observations and decision as search engine continue to change.

[12] https://www.ijcttjournal.org/Volume13/number-3/IJCTT-V13P128.pdf

Web Crawler: Extracting the web data by Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnic in India. Published at IJCTT in July 2014.