Improved Weight based Web Page Ranking Algorithm

Megha Bhawsar
Department of Computer Science
Sushila Devi Bansal College of Technology, Indore (M.P), India

Shraddha Kumar
Assistant Professor, Department of Computer Science,
Sushila Devi Bansal College of Technology, Indore (M.P), India

ABSTRACT
Web page ranking is a technique to optimize the search engines for finding the more relevant content according to the user search query. In this context the web pages are evaluated in such manner by which the appropriate position of a web page is decided in a World Wide Web graph. In literature several web page ranking techniques are available but most of them requires significant amount of time and memory resources for evaluation of web page rank. Therefore, the proposed work is motivated for designing and development of the efficient technique of web page rank. The proposed web page rank evaluation technique is a weight-based page rank technique. The weights are basically the page rank value based on which the web pages are organized on web graph. To compute the web page rank in the proposed technique the web page TF (Term Frequency), IDF (Inverse Document Frequency), Inbound and Outbound links are considered. Therefore, the proposed technique utilizes the techniques of web structure mining and web content mining for developing web page rank of a given web page. After computation of the considered factors the combined weight for all web pages are computed and most higher weight-based page is ranked first for any given query. The implementation of the proposed web page rank computation technique is performed on visual studio technology. After implementation of the proposed technique the performance of system is measured in terms of time and space complexity. In addition of that the experimentation is extended for finding the optimal weighted factor therefore it is concluded that the weighting factors 0.25, 0.25, 0.25, 0.25 is the most suitable weighting factor for web page rank calculation.

Keywords
web page ranking, web graph, weighted page rank, optimal weighting factor selection, implementation, and performance evaluation.

1. INTRODUCTION
The information technology is rapidly increased in recent years. The web technology is also affected with these changes. Several new internet users are appeared in recent years. For every kind of data and information search the users are completely dependent on different web search engines. The web search engines are basically an information retrieval tool that process entire web data and result most appropriate results for the submitted query to the search engines [1]. To find the user query relevant data in entire web the search engine makes efforts and produces results in very fewer time. In less amount of time search engine generate more relevant results because the search engines usages a special kind of data structure. This web data structure is known as the web graph. In this web graph the nodes are appeared as the web pages in the entire World Wide Web and the edges of this graph are decided according to the available content and Meta description of the web pages. Basically, search engines are pre-process the new web pages first and then place it into the web graph additionally during the data search the search engines directly approach those pages that ranked in the web graph. This work is about web page ranking and proposing a new technique to evaluate the web pages for ranking in web graphs. The proposed technique is based on the concept of web mining domain i.e. structure mining and the content mining. Both the concepts are aggregated in the proposed work to rank the data in more relevant manner for finding better outcomes and minimizing the resource consumption during the pre-processing of web documents.

2. PROPOSED WORK
This chapter provides the detailed explanation about the proposed technique of web page ranking for finding appropriate content during the search. In this context a new data model is proposed, and their functional overview is provided in this chapter.

2.1 System Overview
Web mining is a technique by which the web-based data analyzed, and the valuable patterns are recovered. In this context the data mining techniques on web data is employed. The web data can be obtainable from web access logs, web data structures such links and their organization, and the web pages. In this presented work the web structure mining and web content mining is key area of study and investigation. Basically, the web structure mining is technique where the web page structure and their composition with the other pages are analyzed. On the other hand, the data available in web pages such as text, images and other is part of web content mining. In this presented work the web page text and their linked organization is used for designing the proposed technique.

The web pages are organized in web in form a web graph. This web graph is collection of the web pages and the edges which link these pages according to their importance in web graph. The unique and innovative text is place in higher position in this web graph. This technique of placement of web data in this web graph is known as the web page ranking. That is useful for the search engines to find the relevant content from the web during the user query execution. Additionally, based on the query the most relevant data is ranked first. There are several web page ranking techniques and algorithms are available in literature recently. These techniques are not much efficient and a significant resource consuming. Therefore, the proposed work is focused on minimizing the resource consumption and time required to find the suitable rank for a given web page. The proposed work includes the different techniques and methodologies for optimizing the computation of the existing web page ranking technique. Additionally, a weight-based rank calculation approach is proposed for design and development. This

section provides an overview of the proposed web page ranking technique. In next section the proposed methodology is explained in detail.

2.2 Methodology

The basic overview of the proposed methodology is demonstrated in Fig. 1. The different components of this model are given as follows:

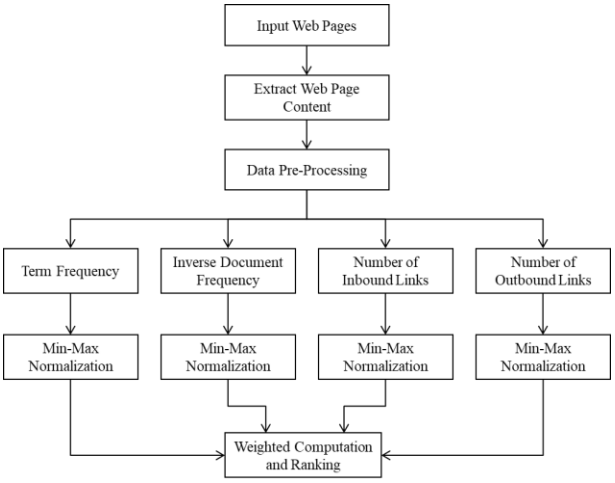

Fig. 1 Proposed System Architecture

2.2.1 Input web pages

The web page ranking techniques are works on web pages. Therefore, need to provide the web pages as input. Basically, in web page ranking techniques web search engines first crawl the web pages or download web pages into their repository and then the analysis is performed for rank these web pages into the web graph. In the similar manner here, a web page directory is considered as complete web sites additionally the pages of this website are placed in a directory.

2.2.2 Content extraction

After downloading of the web pages the content of web pages is extracted. Therefore, a HTML parser is implemented with the proposed system which removes the HTML tags nd extracts the text content in the given web pages. In this presented work the text content is considered for analysis images and other kinds of contents are not considered.

2.2.3 Data preprocessing

The main aim of data preprocessing is to optimize the content or data to become enable the algorithm to work effectively. Therefore, the text data is preprocessed in this phase for reducing the unwanted content and filter out the valuable content which can be utilized with algorithm for application point of view. Therefore, in this phase two key techniques are implemented. In first the stop words from the text data is removed. Additionally, in second method the special characters are removed. After filtering the web content, the remaining data is utilized in further phases for developing web page rank

2.2.4 TF (Term Frequency)

It is basically a probability of document words by which based on this probability the effective words can be selected by the algorithm. That can be computed using the total amount of words available in document and the occurrence of a target word the document. The following formula can be used for measuring the term frequency.

$$TF = \frac{total\ times\ a\ word\ found\ in\ a\ document}{total\ words\ avilable\ in\ document}$$

2.2.5 IDF (Inverse Document Frequency

It is a measurement of word, to find how much information a word provides in a document. In other words, the importance of a word in a given document is measured using the inverse document frequency. For example, the word "THE" frequently found in document as compared to all other words but is the word "THE" provides much valuable information in the document. Therefore, the inverse document frequency is computed to distinguish between informative and non-informative words. That can be computed using the following formula:

$$IDF = -log\frac{n_t}{N}$$

2.2.6 Inbound links

The web pages in a web graph are represented as a node. Additionally, the edges coming to the node is considered as the inbound link:

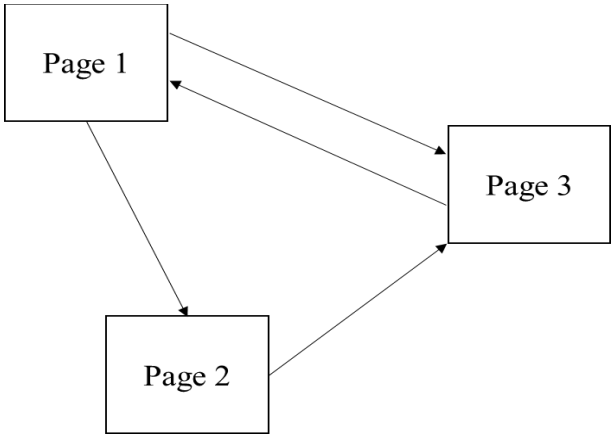

Fig. 2 Example: Web Pages Organization

The above given diagram contains the three web pages web page 1 contains direct link to visit page 2 and 3. Similarly the web page 2 contains a link directly for page 3 and finally page contains a single link to visit page 1 therefore the page 1 having only one inbound link.

2.2.7 Outbound links:

Links in a web graph going away from the given node is termed as the outbound link. In the given Fig. 2 the web page 1 contains 2 outbound links and 1 inbound link.

2.2.8 Min-max normalization

The normalization technique is used when the data available to combine is measured in different scales. Therefore, to scale all the data measured is scaled between defined scales using the min-max technique. To normalize all the data in a common scale the following formula is used.

$$normalized\ value = \frac{current\ value - minimum}{maximum - minimum}$$

2.2.9 Weight computation

To combine all the measured factors of the web pages for ranking them a weight is computed. The weight computation is performed using the following formula:

$$W = TF * w_1 + IDF * w_2 + IN * w_3 + OT * w_4$$

Where, w_1, w_2, w_3 and w_4 is the scaling factors. These intermediate weights are user defined weights and can be selected randomly between 0-1. But the only consideration is the sum of all the weights is equal to 1 i.e. $w_1 + w_2 + w_3 + w_4 = 1$.

To implement the above given weight computation function can be extended as:

$$W = \frac{1}{N}\sum_{i=1}^{N} TF_i * w_1 + \frac{1}{N}\sum_{i=1}^{N} IDF_i * w_2 + \frac{1}{N}\sum_{i=1}^{N} IN_i * w_3 + \frac{1}{N}\sum_{i=1}^{N} OT_i * w_4$$

The higher weighted page is ranked first for any given user query

2.3 Proposed Algorithm

The above given web page ranking methodology is described in this section as the algorithm steps. The Table 1 shows the algorithm steps.

Table 1 Proposed Algorithm

Input: web pages P_n Output: web page rank W
Process: 1. $for(i = 1; i \leq n; i++)$ a. $R = readWebpage(P_i)$ b. $PP = preProcessData(R)$ c. $T[M] = tokenizeData(PP)$ d. $for(j = 1; j \leq M; j++)$ i. $TF = countTF(T[j])$ ii. $IDF = CountIDF(T[j])$ e. $TF = MinMaxNorm(\frac{1}{N}\sum_{k=1}^{N} TF_k)$ f. $IDF = MinMaxNorm(\frac{1}{N}\sum_{k=1}^{N} IDF_k)$ g. $IN = MinMaxNorm(countinBound(P_i))$ h. $OT = MinMaxNorm(countOutBound(P_i))$ 2. $end\ for$ 3. $W = \frac{1}{N}\sum_{i=1}^{N} TF_i * w_1 + \frac{1}{N}\sum_{i=1}^{N} IDF_i * w_2 + \frac{1}{N}\sum_{i=1}^{N} IN_i * w_3 + \frac{1}{N}\sum_{i=1}^{N} OT_i * w_4$ 4. $return\ W$

3. RESULT ANALYSIS

This chapter provides the analysis of the performance of the proposed approach of page ranking technique. Therefore, the different parameters and experimental details are reported in this chapter

3.1 Time Consumption

The consumption of the implemented system is termed as the amount of time required to process the entire web data for computing the page rank of the page. In other terms the time consumption is the time complexity of the implemented algorithm. That can be calculated using the following formula:

$$time\ required = algorithm\ end\ time - start\ time$$

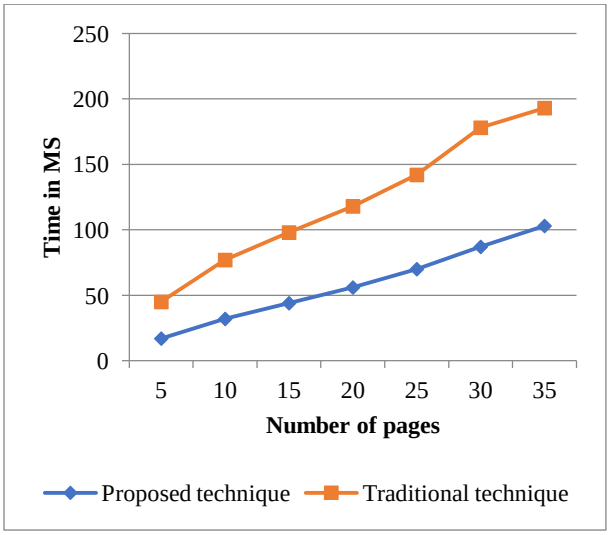

Fig. 3 Time Consumption

Table 2 Time Consumption

Number of Pages	Proposed Technique	Traditional Technique
5	17	45
10	32	77
15	44	98
20	56	118
25	70	142
30	87	178
35	103	193

The time consumption of both the page rank computation techniques is reported in Table 2 and Fig. 3. The table contains the amount of time for both the approaches and similarly the Fig. 3 contains the graphical representation of both the techniques performance. The X axis of the Fig. 3 shows the number of pages in database for computing the page rank and the Y axis shows the corresponding time consumed for processing the given data. According to the given performance the proposed approach requires less amount of time as compared to the traditional technique.

3.2 Memory Usages

The memory usages of a process are sometimes also termed as the space complexity of the algorithms. The amount of main memory required for processing the data is termed as the memory consumption of the given process. In the dot net-based approaches the memory usages can be computed using the following formula:

$$memory\ usages = total\ allocated\ memory - free\ memory$$

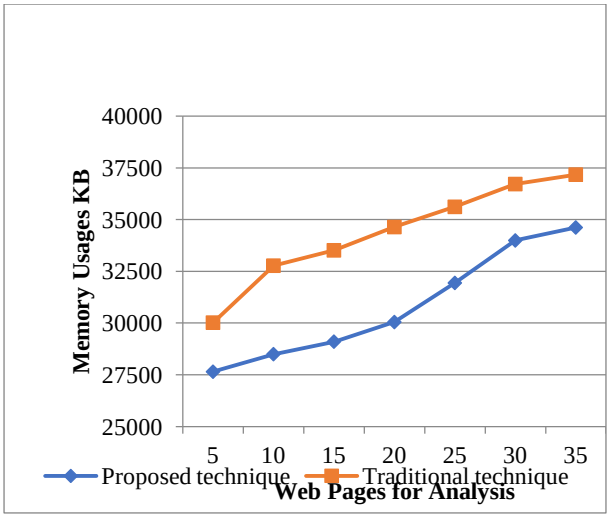

Fig. 4 Memory Usages

Table 3 Memory Usages

Number of Pages	Proposed Technique	Traditional Technique
5	27648	30019
10	28497	32771
15	29101	33514
20	30048	34641
25	31939	35615
30	33991	36716
35	34618	37173

The Table 3 and Fig. 4 contains the memory consumption of both the approaches for page rank computation. In the given experiments the memory usages are computed in terms of kilobytes (KB). The Fig. 4 contains the performance of the algorithms in terms of memory usages. Therefore, the X axis of the diagram represents the amount of data produced for algorithm processes and the Y axis of the diagram includes the corresponding amount of main memory consumed during processing of request. According to the obtained performance the proposed approach is lightweight technique for page rank calculation.

3.3 Effect of Weights

In this section we are investigating about the most appropriate weights combination by which most optimal page rank can be obtainable. Therefore, different combinations of weighting factors are applied with the implemented technique and final rank of the web pages are computed. The Table 4 and Fig. 5 shows the combinations of different web page weighting factors and the relevant obtained rank of web page.

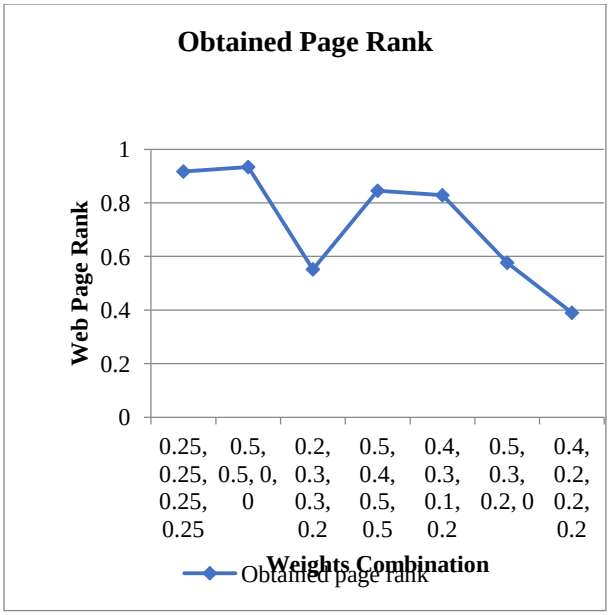

Fig. 5 Effect of Weights

Table 4 Effect of Weights

Weighting Factor Combination	Obtained Page Rank
0.25, 0.25, 0.25, 0.25	0.9168
0.5, 0.5, 0, 0	0.9336
0.2, 0.3, 0.3, 0.2	0.5519
0.5, 0.4, 0.5, 0.5	0.8451
0.4, 0.3, 0.1, 0.2	0.8283
0.5, 0.3, 0.2, 0	0.5765
0.4, 0.2, 0.2, 0.2	0.3898

4. CONCLUSION

This chapter provides the summary of the entire effort performed for designing and developing the proposed page rank technique. Therefore, this chapter includes the observations and the experiments-based conclusion, and the future extension of the work is also reported in this chapter

4.1 Conclusion

The web page raking is a concept of search engine for finding more relevant information in less amount of time. In this context entire world web pages are considered nodes for a large web graph. Additionally, a newly appeared page is evaluated to place that web page into this huge web graph. This technique is termed as the web page ranking. When the user produces query to the search engine the search engine traverses this web graph and the relevancy is measured with the help of user query and the available contents in the web pages. In literature several different techniques are available that claim to improve the web page ranking but either these methods are not much efficient or complex in computation.

Therefore, in this presented work for improving the performance of existing web page ranking technique a new model with the help of web page content mining and structure mining technique is proposed. The proposed data model is promising to minimize the effort and time for computing the

web page rank for a given web page. The proposed data model first evaluates the web page content therefore the TF (Term Frequency) and IDF (Inverse Document Frequency) is computed and meaningful words are selected which are representing the importance of a web page. On the other hand, the inbound and outbound links are computed for finding the trajectory of the web pages and for finding the frequency of user visits for a web page. Finally, all the computed factors are combined using the weight computation technique for ranking of web pages.

The implementation of the proposed page rank approach is performed in .NET technology. During this the performance analysis of the proposed approach is also conducted and compared with the traditional page rank technique. The obtained performance based on different experiments is summarized in Table 5.

Table 5 Performance Summary

S. No.	Parameter	Proposed Technique	Traditional Technique
1	Time Consumption	Low	High
2	Memory Usages	Low	High
3	Weight Factors	In this experiment different combinations of web page weighting factors are applied and then it is concluded that the 0.25, 0.25, 0.25 and 0.25 is the most optimal combination for weighting factor selection	

According to the Table 5 the performance of the proposed approach enhances the relevancy of the user query during the data search. Therefore, the proposed work is acceptable for future extension and real-world application usages

4.2 Future Work

The main aim of the proposed work is to enhance the web page rank computation technique for optimizing the performance of information retrieval form web pages is accomplished successfully. In near future the following work is proposed for more improvements.

The proposed work is considered only the tokens available in the web pages and the structure links it currently not works with the semantics of the tokens. In near future the semantics is also considered for rank computation

The proposed work currently not considers the click streams in near future the click stream is also considered for positioning of a page in web graph

5. REFERENCES

[1] Claudia Elena Dinuca, "Web Structure Mining", Annals of the University of Petroşani, Economics, 11(4), 2011, 73-84

[2] Claudia Elena Dinuca, Dumitru Ciobanu, "Web Content Mining", Annals of the University of Petroşani, Economics, 12(1), 2012, 85-92

[3] Rini John, Sharvari S. Govilkar, "Survey of Information Retrieval Techniques for Web using NLP", International Journal of Computer Applications (0975 – 8887) Volume 135 – No.8, February 2016

[4] Renu Gupta, Ankita Shah, Amit Thakkar, Kamlesh Makvana, "A Survey on Various Web Page Ranking Algorithms", COMPUSOFT, An international journal of advanced computer technology, 5 (1), January - 2016 (Volume-V, Issue-I)

[5] Wanying Chiu, Kun Lu, "Random Walk on Co-word Network: Ranking Terms Using Structural Features", ASIST 2015, November 6-10, 2015, St. Louis, MO, USA.

[6] Chengxiang Zhai, John Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval", SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA, Copyright 2001 ACM 1-58113-331-6/01/0009

[7] R. Campos, G. Dias, A. M. Jorge, A. Jatowt, "Survey of Temporal Information Retrieval and Related Applications", ACM Computing Surveys, Vol. 6, No. 3, Article 9, Pub. date: April 2014.

[8] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, Qiaozhu Mei, "Beyond Ranking: Optimizing Whole-Page Presentation", WSDM'16, February 22–25, 2016, San Francisco, CA, USA. c 2016 ACM. ISBN 978-1-4503-3716-8/16/02

[9] Maria Pershina, Yifan He, Ralph Grishman, "Personalized Page Rank for Named Entity Disambiguation", The 2015 Annual Conference of the North American Chapter of the ACL, pages 238–243, Denver, Colorado, May 31 – June 5, 2015. c 2015 Association for Computational Linguistics

[10] Turgay Celik, "Spatial Mutual Information and PageRank-Based Contrast Enhancement and Quality-Aware Relative Contrast Measure", IEEE Transactions on Image Processing, Vol 25, No. 10, October 2016

[11] Peter Lofgren, Siddhartha Banerjee, Ashish Goel, "Personalized PageRank Estimation and Search: A Bidirectional Approach", WSDM'16, February 22–25, 2016, San Francisco, CA, USA. c 2015 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-3716-8/16/02.c.

[12]