International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 01, Issue – 07, October – 2016, PP – 14-23

GENERALIZED WEIGHTED PAGE RANKING ALGORITHM BASED ON CONTENT FOR ENHANCING INFORMATION RETRIEVAL ON WEB

Ms. H. Bhargath Nisha, M.Sc., M.Phil.
School of Computer Science,
CMS College of Science and Commerce

Mrs. R. Sumithra, MCA. M.Phil..,
Assistant Professor,
School of Computer science,
CMS College of Science and Commerce

Abstract: World Wide Web is a distributed diverse information resource which includes hyperlinks and data with exponential growth. It has been developed not to be easy to access preferred information that matches with user needs and interest. Web has become a most popular trend in terms of ease of use rich contents related to almost every field of life. However, it retrieves more number of documents, which are all related to the search topics and to retrieve the most meaningful documents related to search topics, ranking algorithm is used in information retrieval process. The existing work improved the web information retrieval, used to find out the importance of particular web page that is being evaluated by the user click and as well as the content available on the web. The proposed work extends the previous work by generalizing the user click based relevant content information. In weighted page ranking based on the higher value, web page is ranked to enhance information retrieval on web. The result derived using the approach demonstrates the effective and efficiency of page ranking algorithm.

Keywords: Web mining, web page, page rank, weighted page rank.

1. INTRODUCTION

Google's Page Rank algorithm is one of the best-known algorithms in web search. With the increasing number of Web pages and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and its output [1]. For efficient search results as according to user's query, many ranking algorithms are used which calculate Page Rank of web pages and the goal of the Page Ranking is to make the user get the desired result at the top of the list[2]. A new approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and page ranking provided by search engines. Based on the user query, search engine results are retrieved. Every result is individually analyzed based on keywords and content.[3] .The more popular web pages are, the more linkages that other web pages tend to have to them or are linked by them. The proposed extended Page Rank algorithm–Weighted Page Rank Algorithm assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). [4]. The page ranking is an algorithm developed by Sergey Brin and Lawrence Page in 1998 and used by Google search engine. The algorithm assigns a numerical value to each element in the world wide web for the purpose of measuring the relative importance of each page, the idea being to give a higher page rank value to a page that is frequently visited by users.[5].

1.2 OVERVIEW OF WEB MINING

In 1996 it's Etzioni who first coined the term web mining. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outliers the subtasks of web mining.[6].It refers to overall process of discovering potentially useful and previously unknown information from web document and services web mining could be viewed as an extension of standard data mining to web data.

Web Mining Process: Web Mining can be decomposed into the following subtasks:-

Resource Finding: the function of retrieving relevant web documents.

b) Information Selection and Pre-processing: the automatic selection and pre-processing of specific information from retrieved web resources.
c) Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine Learning are used in generalization.
d) Analysis: The validation and interpretation of the mined patterns. It plays an important role in pattern mining.

1.2.1 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in web content mining. It is also related to text mining because much of web contents are text based. However, it is also different from these because web data is semi structured in nature and text mining focuses on unstructured text.

1.2.2 Web Usage Mining

Web usage mining is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage. Since Web is a reaction media between Web users and Web pages, user navigational behavior needs to be fully concerned during Web mining.

1.2.3 Web Structure Mining

The goal of web structure mining is to generate structural summary about the website and web page. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure.

1.3. LITERATURE REVIEW

This paper proposed a technique on the basis of analyzing previous work for mining web dataset. In the existing technique the weighted page rank, hyper induced topic search, page rank algorithm are not able to perform efficient solution for the page rank. Thus Enhanced page rank algorithm technique is proposed in paper [7]. This paper proposed a new algorithm, the Simplified Weighted Page Content Rank (SWPCR) for page rank, based on combination of web structure mining and web content mining [8]. In this survey paper they analyzed various improvements of Page Rank which uses web content mining for efficient ranking. Relative strengths and limitations of some algorithms are explored to find out further scope of research [9]. In this paper, a new algorithm is proposed, WPUCR (Weighted Page User Content Rank) algorithm which is a combination of web usage Mining, web content mining and Web structure mining. It is the extension of weighted page content rank algorithm and shows the relevancy of the pages to a given query in a much more refined manner and works on the user behavior [10]. The proposed work investigates web page ranking methods and recently-developed improvements in web page ranking Further, a new content-based web page rank technique is also proposed for implementation. The proposed technique finds out how important a particular web page is by evaluating the data a user has clicked on, as well as the contents available on these web pages. The results demonstrate the effectiveness of the proposed page ranking technique and its efficiency [11].

1.4. MOTIVATION

Based on the literature survey conducted, the research is motivated in the direction of web content mining, further it aims to retrieve meaningful information from the web based on the user specific search need. To retrieve effective search results, page ranking algorithms are applied, so that the research concentrated on proposed algorithm. Instead of using conventional algorithms, the research finds out the fact that it is better to have weighted approach in the search process. Hence the research investigates weighted page ranking algorithm and moves on the way of finding out generalized approach for the web.

1.5. ORGANIZATION OF THE PAPER

The rest of the paper is organized as follows. Section 2 describes different page ranking Algorithms. Section 3 discusses the proposed technique. In Section 4, the results are analyzed and discussed. Section 5 focuses on the conclusion of the research.

2. DIFFERENT PAGE RANKING ALGORITHMS

2.1 Page rank algorithm

Page Rank algorithm is used by the famous search engine that is Google. This algorithm is the most commonly used algorithm for ranking various pages. Working of the Page Rank algorithm depends upon link structure of web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. It is assumed that page A has been $T_1...T_n$ which point to it are links. The variable d is a damping factor, whose value can be set between 0 and 1. Usually it has been set the value of d to 0.85. PR (T1) is the incoming link to page A and C (T1) is the outgoing link from page T1 (such as PR (T1)). The Page Rank of page A is given by the following equation(1)

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(Tn/C(Tn))$$

Equation 1

2.2 Weighted page rank

Weighted Page Rank algorithm (WPR) is the modification of original Page Rank algorithm. WPR decides the rank score based on the popularity of pages by taking into consideration the importance of both the in links and out links of the pages. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among its out link pages. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}(m, n)$ and $^{W\ out}(m, n)$ respectively. $^{W\ in}(m, n)$ as shown in equation(2) the weight of link(m, n) is computed depending on the number of incoming links of page n and the number of incoming links of all reference pages of page m.

$$W^{in}_{(m,n)} = I_n / \sum_{P \in Re(m)} I_p$$

$$W^{out}_{(m,n)} = O_n / \sum_{P \in Re(m)} O_p$$

Equation 2

The formula as proposed for the WPR is as shown in equation (3) which is a modification of the Page Rank formula.

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m)\, W^{in}_{(m,n)}\, W^{out}_{(m,n)}$$

Equation 3

2.3 Weighted page user content rank

Weighted Page User Content Rank algorithm (WPUCR) is the modification of the original Page Rank (PR) algorithm followed by Weighted Page Content Rank (WPCR). WPUCR decides the rank score based on the popularity of pages by taking into consideration the importance of both in links and out links of the pages along with the user behavior. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among it's out link pages. Every out link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links along with user behavior. is given by following equation(4)

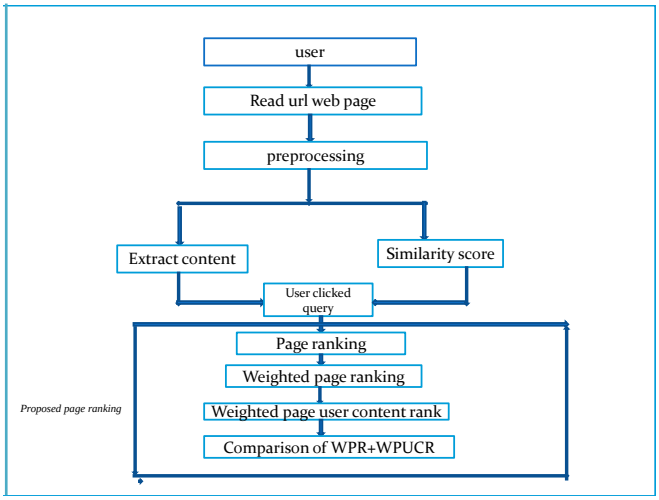
$$WPUCR(n) = F*\{(1-d) + d \sum_{m \in b(u)} PR(m)^{Win}(m, n)\ ^{Wout}(m, n)*(CW+PW)\}$$

Equation4

3. PROPOSED PAGE RANKING ALGORITHM

Figure 1 represents the overall working of the proposed page ranking algorithm using a generalized weighted page ranking algorithm based on content and user click based method. To evaluate a different page ranking for the web, a user click and content-based page rank algorithm is proposed. Content analysis is evaluated based on the available content on a webpage. The entire page ranking model can be simulated using three key steps: first, a user query interface by which a user sends a request for a query. The extracted results are pre-processed using Porter's stemming algorithm, and user-clicked contents are traced using user clicks. The contents of user-needed data are extracted, analyzed and similarities measured using the similarities score for accuracy. The comparison of weighted page rank and weighted page user content ranks are formulated in the new algorithm to get efficient results for proposed page ranking algorithm and finally performance is evaluated using precision, recall, fallout and f-measure.

Figure1.Block diagram for the Proposed Page Ranking Algorithm



3.1 Pre-processing

Data in the real world is dirty and incomplete, lacking both attribute values as well as certain attributes of interest, or containing only noisy and inconsistent aggregate data. There is no quality in data-extracted results and no quality mining. Further, the queries asked by users are particularly short. After retrieving the results for a user's query, the snippets are mixed with unwanted content and are pre-processed using information-retrieval techniques. The aim is to generate highly relevant results for a given search query, which can be achieved by stemming and stop word removal.

3.1.1 Stemming algorithm

Stemming refers to the process of removing affixes (prefixes and suffixes) from words. In the information retrieval context, stemming is used to conflate word forms to avoid mismatches that may undermine recall. As a simple example, consider searching for a document entitled "How to write". If the user issues the query "writing" there will be no match with the title. However, if the query is stemmed, so that "writing" becomes "write", then retrieval will be successful. In many languages stemming is imperative for retrieval performance is satisfied then suffixes S1 is replaced by suffix S2.

3.1.2 Stop word elimination

This method is used to find out the root/stem of a word. For example, the words "user, users, used, using all can be stemmed to the word "USE". The purpose of this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time. The stemming process is done using various algorithms. Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions.

3.1.3 Extract content from web page

Content from the refined results is extracted by finding frequent item sets in data mining. When a user types a query, a set of relevant web snippets are returned. If a keyword or phrase exists frequently in web snippets relating to a particular query, it represents important content related to the query because it exists with the query in the top documents. To measure interest in a particular keyword or phrase k_i extracted from web snippets its followed by equation(5)

$$\text{support}(k_i) = \frac{sf(k_i)}{n} . |k_i|$$

Equation 5

Where $sf(k_i)$ is the snippet frequency of the keyword or phrase k_i , n is the number of web snippets returned, and $|k_i|$ is the number of terms in the keyword or phrase k_i If the support of a keyword or phrase k_i is greater than the threshold s, then k_i acts as a concept for the query q.

3.1.4 Find unique words

User's identification is, to categorize who access web site and which pages are accessed. Different users may have same URL in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows: Each URL address represents one user. For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, An URL represents a different user. Using the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same URL.

3.2 Similarity score

In extracting content-based similarity, the two keywords from a query q are similar if they co- exist frequently in web snippets arising from the query q its following by equation (6)

Equation 6

Similarity=unique words of the page ∩ user query % total number of unique words

3.3 Generalized page rank

In this paper, it is proposed to derive a new kind of algorithm through comparison of WPR and WPUCR. In WPR algorithm the higher value page is ranked. WPUCR algorithm takes user content alone to calculate content in the weighted page .Then compare the algorithm weighted page and user content to get the new algorithm GPR to give efficient result which is better than existing algorithm. The calculation of Generalised Page Ranking is given in the following equation(7)

$$GPR= [(1-d) + d (w^{in}* w^{out})] + [(1-d) + d*(w^{in}*w^{out})*(CW+PW)] \quad \text{equation 7}$$

GPR=generalized page ranking

D=damping factor which can be set between 0 and 1

W^{in}=in weight of link

W^{out}=out weight of link

CW=content weight (user query)

PW=page weight

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Performance metrics

4.1.1 Precision:

Precision is the fraction of documents retrieved that are relevant to a user's information needs. It takes all retrieved documents into account its following by equation(8)

Precision=relevant document ∩ retrieved documents/ retrieved documents equation 8

4.1.2 Recall:

Recall is the fraction of documents successfully retrieved and relevant to a query. Also called sensitivity, it can be looked at as the probability that a relevant document is retrieved by the query its following by equation(9)

Recall = relevant document ∩ retrieved documents/ relevant document equation 9

4.1.3 Fallout:

Fallout is the proportion of non-relevant documents retrieved from all the non-relevant documents available. It can be looked at as the probability that a non-relevant document is retrieved by a query its following by equation(10)

Fallout= non-relevant document ∩ retrieved documents/non-relevant document equation 10

4.1.4 F-measure:

F-measure is the harmonic mean of precision and recall, and provides good results when precision and recall provide good results its following by equation(11)

F-measure=2*(precision*recall)/(precision +recall) equation 11

4.2 EXPERIMENTAL RESULTS

In this result, Generalizing different URL web pages has been ranked in different page ranking algorithm. Proposed system uses page ranking, weighted page ranking, weighted page user content and the comparison of WPR and WPUCR are given to get better result.

Precision values for different page ranking

1. URL:'https://www.google.co.in/search?newwindow=1&q=normal+ecg+database&sa=X&ved=0ahUKEwjP8ordk_rOAhWJLo8KHdx5A0UQ7xYIGigA&biw=1242&bih=606';

Queries	WPR	WPUCR	GPR
Information	0.9196	0.9406	0.9944
Applications	0.9264	0.9529	0.9943
Butterworth	0.9162	0.9769	0.9942
Collections	0.9157	0.9423	0.9944
Conferences	0.8843	0.9472	0.9945

Table 2 precision values for different page ranking algorithms

Recall values for different page ranking

Queries	WPR	WPUCR	GPR
Information	0.8507	0.8927	0.9029
Applications	0.8775	0.9028	0.9297
Butterworth	0.8661	0.9027	0.9941
Collections	0.8421	0.8944	0.9027
Conferences	0.7775	0.8995	0.9030

Table 3 recall values for different page ranking algorithms

Fallout for different page ranking

Queries	WPR	WPUCR	GPR
Information	0.0409	0.0391	0.0251
Applications	0.0424	0.0376	0.0252
Butterworth	0.0434	0.0366	0.0253
Collections	0.0407	0.0392	0.0251
Conferences	0.0397	0.0393	0.0254

Table 4 fallout values for different page ranking algorithms

F-measure values for different page ranking

Queries	WPR	WPUCR	GPR
Information	0.8838	0.9160	0.9465
Applications	0.9013	0.9412	0.9463
Butterworth	0.8904	0.9462	0.9854
Collections	0.8774	0.9177	0.9465
Conferences	0.8275	0.9227	0.9465

Table 5 F-measure values for different page ranking algorithms

2.URL:https://www.google.co.in/search?newwindow=1&q=advantages+of+weighted++page+ranking+algorithm&oq=advantages+of+weighted++page+ranking+algorithm&gs_l=serp.3...989427.1001520.0.1002657.13.13.0.0.0.0.238.1801.3j7j3.13.0....0...1c.1.64.serp..0.12.1700...0i13i30k1j30i10k1.bjQhOwQWoz8

Precision values for different page ranking

Queries	WPR	WPUCR	GPR
Country	0.9195	0.9407	0.9945
Articles	0.9254	0.9539	0.9963
Advantages	0.9182	0.9789	0.9952
Weighted	0.9177	0.9433	0.9954
Page	0.8883	0.9482	0.9965

Table 6 precision values for different page ranking algorithms

Recall values for different page ranking

Queries	WPR	WPUCR	GPR
Country	0.8527	0.8967	0.9129
Articles	0.8875	0.9128	0.9277
Advantages	0.8671	0.9227	0.9931
Weighted	0.8321	0.8964	0.9227
Page	0.7975	0.8985	0.9130

Table 7 recall values for different page ranking algorithms

Fallout for different page ranking

Queries	WPR	WPUCR	GPR
Country	0.0402	0.0374	0.0271
Articles	0.0414	0.0396	0.0222
Advantages	0.0448	0.0421	0.0395
Weighted	0.0399	0.0296	0.0236
Page	0.0412	0.0403	0.0341

Table 8 fallout values for different page ranking algorithms

F-measure values for different page ranking

Queries	WPR	WPUCR	GPR
Country	0.8725	0.9148	0.9459
Articles	0.9103	0.9445	0.9487
Advantages	0.8924	0.9491	0.9732
Weighted	0.8871	0.9273	0.9515
Page	0.8475	0.9327	0.9565

Table 9 F-measure values for different page ranking algorithms

According to the results obtained in table 2, the performance of the proposed technique is optimal, when compared to other traditional page ranking approaches. It is easy to achieve a recall of 100 percent by retrieving all documents in response to a query. Hence recall alone is not enough, and the number of non-relevant documents is required to be measured as well. Hence fallout is considered and from table 4, that the proposed page ranking has fewer non -relevant documents retrieved than other page ranking algorithms. Table 5 makes it clear that the proposed method provides good results compared to other existing methods because precision and recall for the existing methods are fewer, compared to the proposed method. Hence the f-measure value for content and user click-based page ranking gives good results.

In this paper, four types of performances graph are discussed and it is proved that compared to the existing graph the proposed graph gives good result. In the graph measure the weighted page rank, weighted page user content rank, and proposed GPR generalized page rank it's given the efficient result.

1. Precision graph

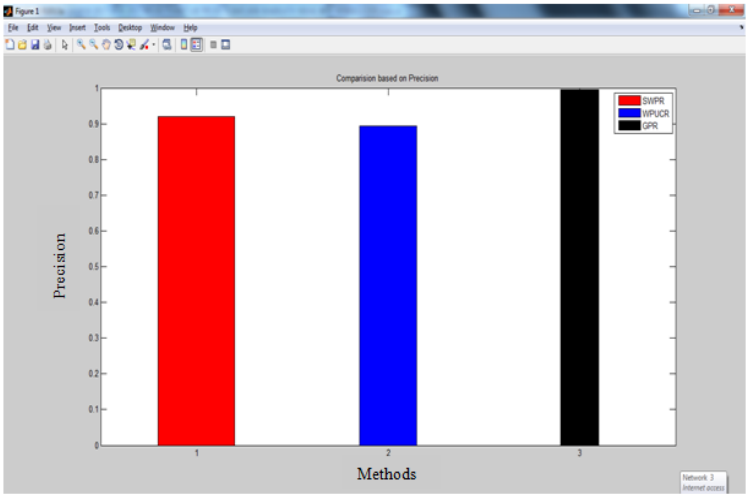


Figure10. Comparison based on Precision in Different Page Ranking

2. Recall graph

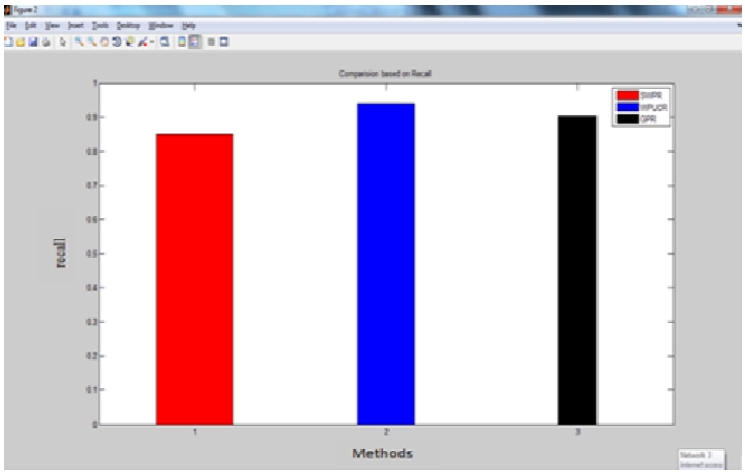


Figure11. Comparison based on Recall in Different Page Ranking

3.fallout graph

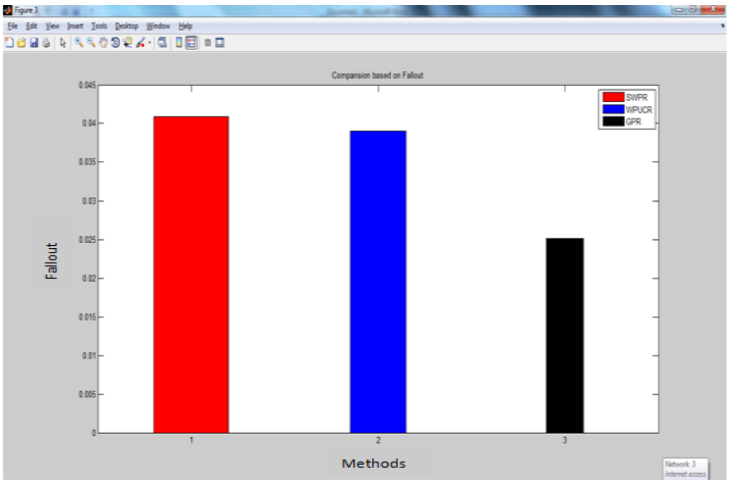


Figure12. Comparison based on Fallout in Different Page Ranking

4. f-measure graph

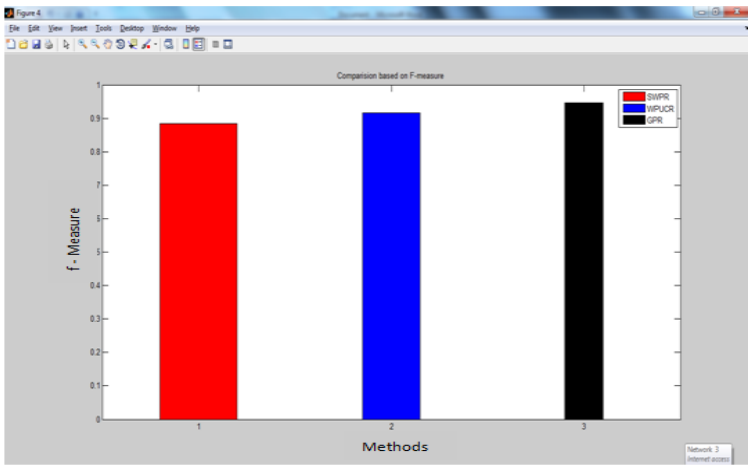


Figure13. Comparison based on F-measure in Different Page Ranking

5. CONCLUSION

This research work studies, generalized web pages ranks through different page ranking techniques and proposes a novel page ranking algorithm through a comparison of WPR and WPUCR. This page rank algorithm provides a rank according to the relevance of a user's query and the contents available in web pages. In addition, the relevancy of search results is measured in terms of precision, recall and F-measure. These results demonstrate the efficient of relevant ranks for the search results available. The proposed work is intended to provide an efficient page rank technique using an analysis of web page contents. The page rank technique presented ranks results according to the importance of generalized web page, user search query and the content available in the web page. The proposed technique is efficient to generalize the web page, but is not providing efficient, scenario-specific page ranking. Therefore, in the near future, the proposed technique is to be extended to derive a Scenario-specific framework for page rank estimation.

6. REFERENCES

[1]. s. Prabha, k. Duraiswamy, j. Indhumathi world academy of science, engineering and technology international journal of computer, electrical, automation, control and information engineering vol:8, no:8, 2014 'COMPARATIVE ANALYSIS OF DIFFERENT PAGE RANKING ALGORITHMS'

[2]. Gupta1, Ankita shah2, amitthakkar3, kamlesh makvana4 comp soft, an international journal of advanced computer technology, 5 (1), January - 2016 (volume-v, issue-i) 2046 'A SURVEY ON VARIOUS WEB PAGE RANKING ALGORITHMS'

[3]. P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, Member IAEN' Content Based Ranking for Search Engines'

[4]. Wenpu Xing and Ali Ghorbani 'Weighted Page Rank Algorithm'

[5]. Dhiliphan rajkumar.t1,Suruliandi.a2 and selvaperumal.p3international journal on computational science & applications (ijcsa) vol.5,no.6, december 2015doi:10.5121/ijcsa.2015.5608 111'CONTENT AND USER CLICK BASED PAGE RANKING FOR IMPROVED WEB INFORMATION RETRIEVAL'

[6]. Ekta bhardwaj1, shiv kumar2, kuldeep tomar3international journal on recent and innovation trends in computing and communication issn: 2321-8169volume: 3 issue: 5 3381 – 3385 3381 ijritcc | may 2015, available @ http://www.ijritcc.orgenhancing 'PAGE RANK ALGORITHM' Ekta bhardwaj1, shiv kumar2, kuldeep tomar3

[7]. Seifedine kadry and Ali kalakech 'ON THE IMPROVEMENT OF WEIGHTED PAGE CONTENT RANK' seifedine kadry and ali kalakech journal of advances in computer networks, vol. 1, no. 2, june 2013

[8]. Jaroslav pokorny page content rank: 'AN APPROACH TO THE WEBCONTENT MINING'

[9]. Taruna kumari1 , Ashlesha gupta2, Ashutosh dixit3international journal of innovative research in computer and communication engineering(an iso 3297: 2007 certified organization)vol. 2, issue 2, february 2014copyright to ijircce www.ijircce.com 2929'COMPARATIVE STUDY OFPAGE RANK AND WEIGHTEDPAGE RANK ALGORITHM'

[10]. Preetibala deshmukh, vikram garg international journal of computer applications (0975 – 8887) volume 139 – no.1, april 2016 39 'AN ENHANCED PAGE RANK ALGORITHM OVER DOMAIN '

[11]. pooja sharmaa ,Deepak Tyagpawan ,Bhadana pooja sharma et al. / international journal of engineering science and technology vol. 2 (12), 2010, 7301-7310'WEIGHTED PAGE CONTENT RANK FOR ORDERINGWEB SEARCH RESULT'

