ISSN:0975-9646

Ranking of Research Papers using WPCR with Clustering Algorithm

Manpreet Kaur Gill[#1], Jasmeet Singh[*2]

[#]*M.Tech, Research Scholar,*
Department of Computer Science and Engineering,
RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India

[*]*Assistant Professor,*
Department of Computer Science and Engineering,
RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India

***Abstract*—There are number of hyperlinks and data in distributed heterogeneous information that is contained by the World Wide Web. It is necessary to provide the relevant results (research paper's) to the research scholars, scientists etc using search engines according to their query. In order to meets the user's requirements, it becomes important to rank the research papers for providing the relevant results. Web page ranking algorithms play vital role for ranking web pages that the user can retrieve the page which are more relevant to the user's query. The main aim of the research paper is to providing rank to research papers using Weighted Page Content Rank with K-Means clustering algorithm. Users can search the papers according to the different areas like data mining, networking etc. and they got the papers according to rank given to papers. The paper focuses on comparison study of proposed work with previous results. Results are shown in the form of tables and graphs according to various parameters such as precision ,recall etc.**

***Keywords*— Data Mining, Ranking, Clustering, WPCR, K-Mean Clustering**

I. INTRODUCTION

WWW is a big resource of heterogeneous and hyperlinked information including audio, video, text, image, and metadata. In WWW there is an explosive growth from the early 1990's. With huge increase in availability of information through WWW, it is difficult to acquire the useful information on Internet; therefore many users use Information retrieval tools like Search Engines to search desired information on the Internet [13]. A Search Engine is an information retrieval system which helps users finds information on WWW by making the web pages related to their query available [1].

Now-a-days research scholars search the papers at very high level for their purposes. All research scholars want latest and relevant research papers in less time. They need the relevant papers for the efficient result of the work. But the search of relevant and latest papers in less time on the top is such a difficult task. So to retrieve the papers on top in less time, we applied different algorithms on them. Here research scholars retrieve the research papers according to the specific area also such as data mining, networking, cloud computing etc.

Text mining is the process of discovery of text from the text documents or interesting knowledge. It is a challenging task to help the users in finding what the user' actually want from the number of text documents. It is quite difficult to deal with the text which is in unstructured form. The purpose of the text mining is to finding ―nuggets‖ of interesting information from the natural language text [4]. To answer the complicated questions and to do the web searches with intelligence is the main aim of the text mining tools. Text mining uses the automated methods for achieving the unusually knowledge which is available in text documents. Techniques of text mining are [16]:

- Natural Language Processing (NLP)
- Information Extraction (IE).

Web mining comes under the application of data mining. It refers to overall process of finding the useful and previously unknown information from the web services and documents.

There are three categories of web mining as shown in Fig. 1 which are: Web Structure Mining, Web Content Mining, and Web Usage Mining [13] [17].

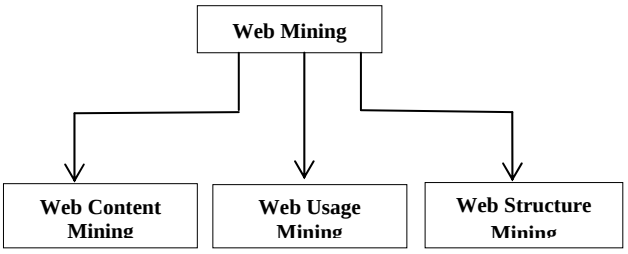

Fig.1 Web Mining Categories

1) Web Usage Mining (WUM): Web usage mining is used to find usage patterns from Web data in order to understand Web based applications. It consists of three phases which are pattern analysis, pattern discovery, and preprocessing. In this, servers, proxies, and client applications can easily capture data about Web usage.

2) Web Content Mining (WCM): Web Content Mining is the process of extracting meaningful information according to the contents of Web documents. Content data consists of facts like images, texts, audio, video, HTML documents, and data in tables that are included in a web page. The individual pages which are mined in web content mining is the primary Web resources can be used to grouping, categorizing, analyzing, and retrieving documents.

3) Web Structure Mining (WSM): The aim of web structure mining is to create a structural summary about the web page and website. In this, web structure mining is used to extract the patterns in the web from hyperlinks. To connect the web page to another location hyperlink is used which is a structural component.

Clustering is the method which includes the grouping of similar type objects into one cluster and a cluster which includes the objects of data set is chosen in order to minimize some measure of dissimilarity [15]. Clustering is not a type of supervised learning but it is supervised learning unlike Classification. In the clustering method, objects of the data are gathered into clusters, in the way that groups are very differ from one other and the objects which are in the same cluster are very similar to each other[2]. In Classification there are predefined set of classes are presented, but in clustering no predefined set of classes which results that resulting clusters are not given before the execution of clustering algorithm. Here the clusters from the dataset are extracted by grouping the objects in the process [10].

In Ranking method searching on the internet is most widely used operation on the World Wide Web. The amount of information is increasing day by day rapidly that creates the challenge for information retrieval. There are so many tools to perform the effective searching here. The challenge for search engine page ranking is because of the size of web and requirements of users [3].

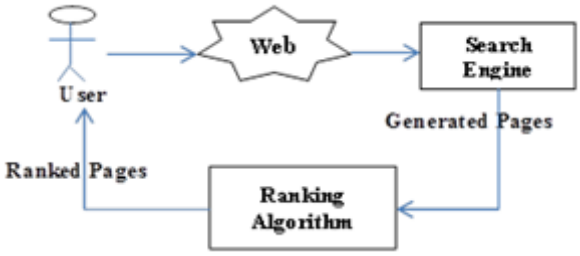

Fig. 2 Working of Search Engine

The main part of any information retrieval system is ranking. Today's search engines may return millions of pages for a certain query but it is not possible for a user to see all the concluded results [14]. So, ranking of pages is helpful in web searching. Rankers split into these two groups: Content-based rankers and Connectivity-based rankers. Content-based ranker's works on the basis of number of matched terms, location of terms, etc. Connectivity-based rankers work on the basis of link analysis technique; links are edges that point to different web pages [11].

II. Overview Of Clustering And Ranking Method

K-Means Algorithm

K-means clustering is a popular method for cluster analysis in data mining. In this method, n observations are partitioned in k clusters, where the users defined, k is the number of clusters but the value of k is fixed. In clustering process, first of all centroid of the each cluster is selected then on the basis of selected centroid, data points having minimum distance from the given cluster are assigned to the particular cluster. Its main steps are [4] [12]:

Let a document set D (d1, d2, d3……dm).

- Firstly choose k-data points as initial centroids.
- Then Find out the distance between each d ε D and the chosen centroid.
- Assign d to the closest cluster.
- Recomputed the centroid until it becomes stable.

Weighted Page Content Rank Algorithm

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm in which according to a user query a sorted order to the web pages returned by a search engine. WPCR is an algorithm based on the numerical value on which the web pages are given in an order. To calculate the importance of the page, web structure mining is used and how much a page is relevant given by web content mining. The popularity of the page defined by the importance which means how much number of pages is pointing to that particular page. Importance cannot be calculated on the basis of in links only, out links are also to be considered here. The matching of the user query with the particular page shows the relevancy of the page. The page is more relevant if it maximally matched to the user query [5].

Algorithm: WPCR calculator
Input: Page P, in link and Out link Weights of
All back links of P, Query Q, d (damping factor).
Output: Rank score
Step 1: Relevance calculation:

1) Find all meaningful word strings of Q (say N)
2) Find whether in P the N strings are occurring or not?

Z= Sum of frequencies of all N strings.

3) S= Set of maximum possible strings occurring in P.
4) X= Sum of frequencies of strings in S.
5) Content Weight (CW) = X/Z
6) C= No. of query terms in P
7) D= No. of all the query terms of Q while ignoring stop words.
8) Probability Weight (PW) = C/D

Step 2: Rank calculation:

1) Find all back links of P (say set B).
2) PR (P) = (1-d) +d
3) Output PR (P) i.e. the Rank score.

III. Proposed Work

Existing work has been implemented for the ranking of web pages, web links etc. but this work is not time efficient. Previous work did not give the relevant web pages according to the user's query. Also Weighted Page Rank Algorithm did not implemented on text search, it only implemented on the search engines. For the ranking of research papers, there is no such efficient algorithm exists which provides better results in terms of relevancy and execution time. So, we ranked the research papers using Weighted Page Content Rank algorithm and K-means Clustering algorithm for clustering of research papers. Also the user can retrieved the most relevant paper according to

user query. Compare the results with other algorithms in terms of relevancy, precision, recall and accuracy.

The methodology of this research is quite simple. The research methodology is divided into steps to achieve our desired goal:

- **Step 1:-** In this step, we upload the research papers to provide rank.
- **Step 2:-** This step include the implementation of Weighted Page Content Rank algorithm which assigns rank to all the research papers and store them in database.

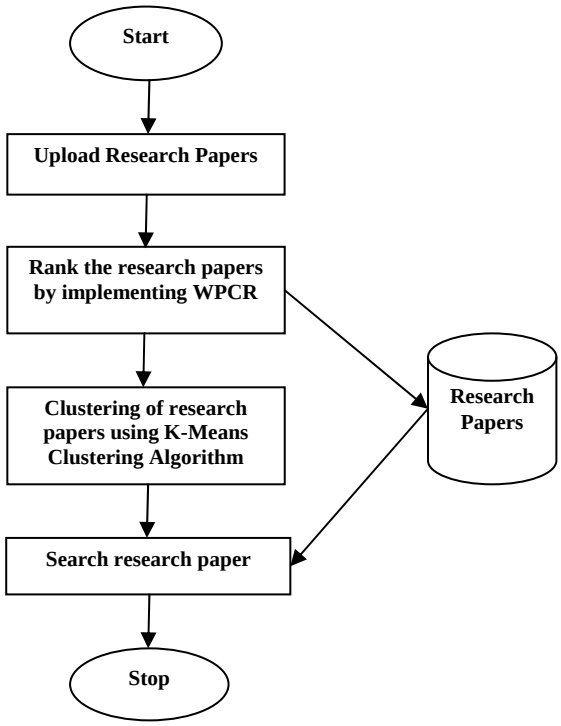

Fig. 3 Flow Diagram

- **Step 3:-** In this step, K-means clustering algorithm is implemented to cluster the research papers on the basis of different research areas.
- **Step 4:-** In this step, relevant research papers from the database are searched by user according to the research area.
- **Step 5:-** Final result compared on the basis of various parameters like accuracy, precision, recall, F-measure, execution time etc.

IV. Results And Comparison

To determine the quality of proposed work, it was necessary to compare it with another algorithm .The performance of the proposed technique is based on the parameters i.e. Recall, Precision, F-Measure, Execution Time.

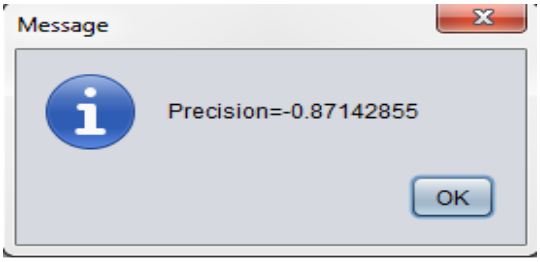

Fig. 4 Precision Value

TABLE I
COMPARISON OF PRECISION VALUES ON THE BASIS OF NO. OF PAPER'S

No. of Papers	Precision of Proposed Work	Precision of Previous Work
5	0.73	0.41
10	0.78	0.52
15	0.68	0.73
20	0.71	0.61
25	0.87	0.56

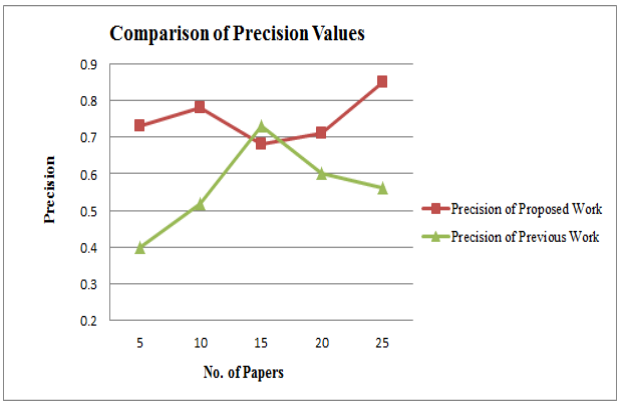

Fig. 5 Comparison of Precision Values

The above graph represents the comparison of precision values of the proposed work with the previous work and concluded that the values of precision of proposed are better than previous work.

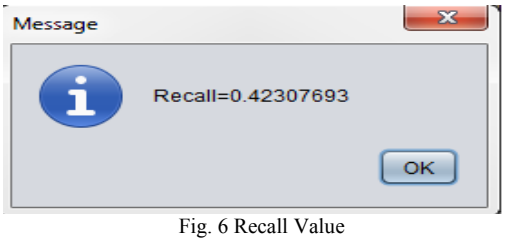

Fig. 6 Recall Value

TABLE II
COMPARISON OF RECALL VALUES ON THE BASIS OF NO. OF PAPER'S

No. of Papers	Recall of Proposed Work	Recall of Previous Work
5	0.19	0.20
10	0.32	0.36
15	0.41	0.68
20	0.35	0.38
25	0.42	0.47

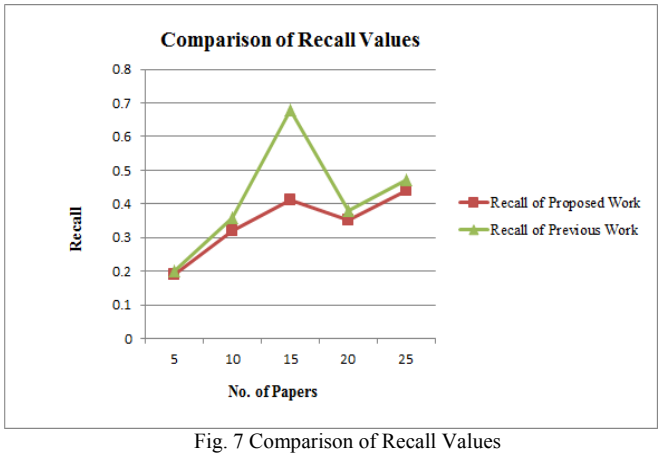

Fig. 7 Comparison of Recall Values

The above graph represents the comparison of recall values of the proposed work with the previous work and concluded that the values of recall of proposed are better than previous work.

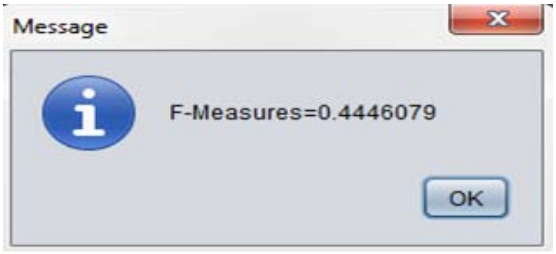

Fig. 8 F-Measure Value

TABLE III
COMPARISON OF F-MEASURE VALUES ON THE BASIS OF NO. OF PAPER'S

No. of Papers	F-Measure of Proposed Work	F-Measure of Previous Work
5	0.30	0.26
10	0.45	0.42
15	0.52	0.70
20	0.35	0.46
25	0.55	0.51

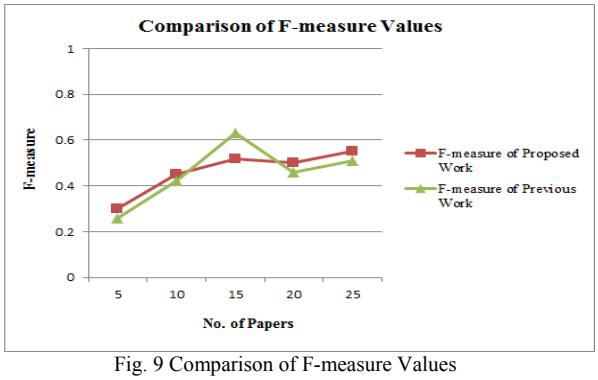

Fig. 9 Comparison of F-measure Values

The above graph represents the comparison of F-Measure values of the proposed work with the previous work and concluded that the values of F-Measure of proposed are better than previous work.

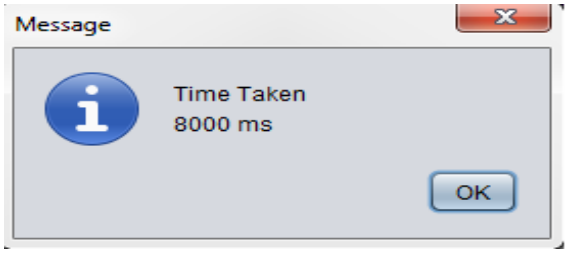

Fig. 10 Comparison of F-measure Values

TABLE IV
COMPARISON OF EXECUTION TIME ON THE BASIS OF NO. OF PAPER'S

No. of Papers	Time Taken by Proposed Work (in sec.)	Time Taken by of Previous Work (in sec.)
5	8	9
10	12	19
15	21	27
20	36	42
25	43	48

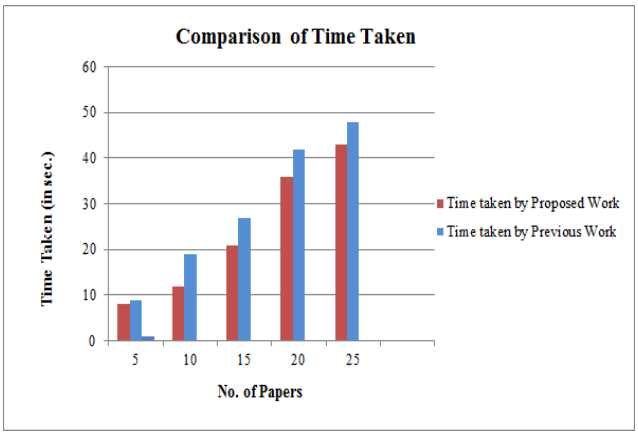

Fig. 11 Comparison of Execution Time

The above graph represents the comparison of execution time of proposed work with the previous work and concluded that the values of execution time of proposed are better than previous work.

V. CONCLUSION AND FUTURE WORK

This paper proposed a technique for ranking the research papers using weighted page content rank and k-means clustering algorithms. Users can search their papers, articles according to rank given to them. Ranking of research paper's is done for providing the relevant paper's to user's according to their queries. The proposed method provides the better and efficient results as compared to the previous results on the basis of various parameters such as precision, recall, execution time etc.

In future work will be extended by providing the results or ranking using the other algorithms and also the rank the papers or articles according to the published date of the research papers. Also we can try to achieve the better results.

ACKNOWLEDGMENT

The author would like to thank the RIMT Institutes, Mandi Gobindgarh-147301, Fatehgarh Sahib, Punjab, India. Author also extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

REFERENCES

[1] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, "*Comparative Study of Page Rank and Weighted Page Rank Algorithm*", International Journal of Innovative Research in Computer and Communication Engineering, IJIRCCE, ISSN(Online): 2320-9801, ISSN (Print): 2320-9798, Vol. 2, Issue 2, February 2014, pp-2929-2937.

[2] Prof. Neha Soni, and Prof. Amit Ganatra, "*Categorization of Several Clustering Algorithms from Different Perspective: A Review*", IJARCSSE: International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), vol. 2, Issue 8, August 2012.

[3] Hema Dubey ,Prof. B. N. Roy, "*An Improved Page Rank Algorithm based on Optimized Normalization Technique"*, International Journal of Computer Science and Information Technologies, IJCSIT, ISSN:0975-9646, Vol. 2 (5) , 2011, pp-2183-2188.

[4] Divya Nasa, "*Text Mining Techniques- A Survey*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 2, Issue 4, April 2012, ISSN: 2277 128X pp. 50-54.

[5] Pooja Sharma, Deepak Tyagi, Pawan Bhadana, "*Weighted Page Content Rank for Ordering Web Search Result*", International Journal of Engineering Science and Technology (IJEST), ISSN: 0975-5462, Vol. 2 (12), 2010, pp. 7301-7310.

[6] Supreet Kaur, Usvir Kaur, "*An Optimizing Technique for Weighted Page Rank with K-Means Clustering*", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSSE, ISSN: 2277 128X, Volume 3, Issue 7, July 2013, pp. 788-792.

[7] Amar Singh, Navjot Kaur, "*To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm*", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSSE, ISSN: 2277 128X, Volume 3, Issue 8, August 2013, pp. 143-148.

[8] Amandeep Kaur Mann, and Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science and Technology (ISSN (Online): 0975-4172), vol. 13, Issue 5, Version 1.0, 2013.

[9] Neelam Tyagi, Simple Sharma, "*Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)*", ISSN: 2278-3075 International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-1, Issue-1, June 2012.

[10] Madhuri V. Joseph, Lipsa Sadath, Vanaja Rajan, "*Data Mining: A Comparative Study on Various Techniques and Methods*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 3, Issue 2, February 2013, pp. 106-113.

[11] Seifedine Kadry and Ali Kalakech, "*On the Improvement of Weighted Page Content Rank*", Journal of Advances in Computer Networks, DOI: 10.7763/JACN.2013.V1.23, Vol. 1, No. 2, June 2013, pp-110-114.

[12] Rama.B, Jayashree.P, Salim Jiwani, "*A Survey on Clustering*", IJCSE: International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 2976-2980.

[13] Shruti Aggarwal, Gurpreet Kaur, ―*Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method*‖ International Journal of Computer Science & Communication Networks (IJCSCN), ISSN:2249-5789, Vol 3(4), pp. 231-239.

[14] Wenpu Xing and Ali Ghorbani, "*Weighted PageRank Algorithm*", Proceedings of the Second Annual Conference on Communication Networks and Services Research, IEEE, 2004.

[15] Amandeep Kaur Mann, and Navneet Kaur, "Survey Paper on Clustering Techniques", IJSETR: International Journal of Science, Engineering and Technology Research (ISSN: 2278-7798), vol. 2, Issue 4, April 2013.

[16] Shaidah Jusoh, Hejab m. Alfawarch, ―T*echniques, Applications and Challenging Issue in Text Mining*‖, International Journal of Computer Science Issues (IJCSI), ISSN (Online): 1694-0184, Vol. 9, Issue 6, No 2, November 2012, pp. 431-436.

[17] T.Munibalaji, C.Balamurugan, ―*Analysis of Link Algorithms for Web Mining*‖, International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume 1, Issue 2, February 2012, pp-81-86.