Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach for Ranking and Clustering the Research Papers

Manpreet Kaur Gill*
M.Tech, Research scholar
CSE, RIMT-IET, Mandigobindgarh,
India

Nidhi Bhatla
Assistant Professor
CSE,RIMT-IET, Mandigobindgarh,
India

Abstract— In today's world, the huge information is stored on World Wide Web. Every user wants to get relevant information according to its query. There are number of techniques are used to provide the best results to the user. Various page ranking algorithms are used by the search engines to provide relevant results to the user by ranking their results. This paper proposed a new method for providing the relevant results to the user using ranking and clustering algorithm. The main aim of this paper is to rank the results, so that the users retrieve the relevant results on the top according to the user query.

Keywords— Clustering, K-means Clustering, Ranking, Page rank, WPCR.

I. INTRODUCTION

WWW is a vast resource of hyperlinked and heterogeneous information including text, image, audio, video and metadata. From early 1990's, in WWW there is an explosive growth. With huge increase in availability of information through World Wide Web, it has become difficult to access the useful information on Internet; therefore many users use Information retrieval tools like Search Engines to search desired information on the Internet. 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. With a search engine, users have to type in "keywords" relating to the information that they need. The search engine would then return a set of results that match best with the keywords entered. A Web Search Engine which takes a software program as input from the user, searches its database and then returns the results. It is important that the search engine does not search the internet; that it searches its database, which is populated with data from the internet. The information about many web pages is stored by Web search engines, which then they retrieve from the World Wide Web. These pages are retrieved by a Web crawler which is followed by every link. To determine how it should be indexed, the contents of each page are analyzed [1].

II. CLUSTERING

Clustering is an automated process for grouping the related records together. Records having similar values for the attributes are grouped together. The objective of clustering analysis is to find segments or clusters and to examine their attributes and values. There are number of algorithms that are used for clustering. The clustering technique defines the classes and puts objects in each class to which it is associated. The objective of clustering analysis is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest and the groups themselves stand apart from one another.

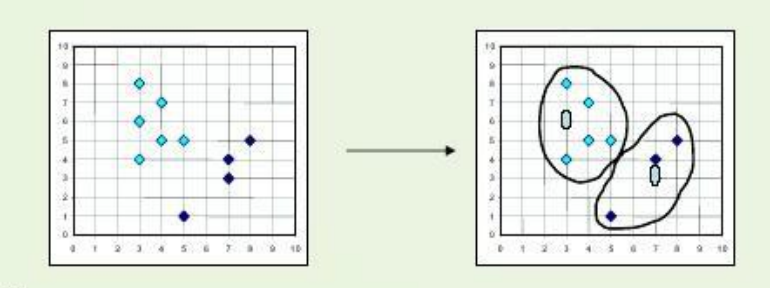

Fig. 1 Formation of Clusters

Clustering is the method by which similar records are collected. This is done to give the end user a high level view of what is going on in the database. Clustering is occasionally used to do segmentation [2]. Different types of Clustering Algorithms are under below:

A. Hierarchical Clustering Algorithm

Hierarchical Clustering constructing a hierarchy of clusters by dividing the data set. Construction of clusters is step by step. This technique takes series of partition not a single step partition. This hierarchy graphically represented by a diagram called a Dendrogram or by binary tree. Root node represents whole data set and leaf represents a data object.

Height of dendrogram represents distance between an object and a cluster, between each pair of an object or between each pair of clusters [7].

B. K-means Clustering Algorithm

K-means clustering is a popular method for cluster analysis in data mining. Here, n observations are partitioned into k clusters, where k is the number of clusters defined by the users where the value of k is fixed. In clustering process, firstly the 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 [8].

C. Density Based Clustering Algorithm

Density-based partitioning techniques are one-scan technique. It finds clusters according to the regions which grow with high density. Clusters are high density area than remaining data set. Density is the number of objects in a cluster. It finds arbitrary shaped clusters. It is applicable to spatial data [7] [9].

D. Self-organisation maps(SOM)

The SOM net is same as two layers neural network model. Each neuron is represented by n-dimensional weight vector where n is equal to the dimension of the input vectors. The neurons of the SOM are themselves cluster centers map units are used to form bigger clusters iteratively. It is robust and deal with missing data values as well as detect easily outlier from the map, but the distance between the input spaces from other units is large [10].

E. EM Clustering Algorithm

The expectation-maximization (EM) algorithm is used for finding maximum a posteriori or maximum like estimates of parameters by number of iteration in statistical model. The EM performing an expectation (E) step, by using the current estimate for the parameter in which the expectation of the log-likelihood is calculated. In maximization step, the parameters for the expected log-likelihood found on the E step are evaluated. The result of the parameters which are estimated is then used for the latent variables in the next E step [11].

- *Expectation:* The missing labels are estimated by fix the model.
- *Maximization:* It finds the model of expected log-likelihood of the data by fix the missing labels.

III. RANKING

Nowadays 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 searching effectively. Because of the size of web and requirements of users creates the challenge for search engine page ranking. Ranking is the main part of any information retrieval system Today's search engines may return millions of pages for a certain query It is not possible for a user to see all the got results. So, ranking of pages is helpful in web searching. Rankers are divided into two groups: Content-based rankers and Connectivity-based rankers. Content-based ranker's works on the basis of number of location of terms, frequency of terms, matched terms, etc. Connectivity-based rankers work on the theory of link analysis technique; links are edges that point to different web pages [3]. There are two famous link analysis methods:

A. Hypertext Induced Topic Search (HITS)

It is a link based algorithm which is used to rank pages that are retrieved from the web according to the given user query based on their textual information. When the user retrieved the required pages then HITS algorithm started ignoring textual information and starts focusing only on the web structure. The algorithm is used to rank the relevant pages and treat all the links equally for the distribution of rank scores. In this, HITS rank the pages by analysing their in-links and out-links. The web pages that points to the hyperlinks are known as hubs but the hyperlinks that points to the web pages are known as authorities [4].

B. Page Rank Algorithm

It is the most commonly used algorithm for ranking. The working of this algorithm is depends upon the link structure of the web pages. In this if there are important links towards a page then the links towards the other pages from it is also considered as important. In this back link is used to provide the rank score and if the rank of the back links is large then the rank given is large rank of particular page [5].

C. Weighted Page Rank Algorithm

It is the extension of the original page rank algorithm In this larger rank values are assigned to more important pages instead of dividing the rank value of a page equally among its outgoing linked pages. The value given to the outgoing links is proportional to the importance of that page [4]. WPR performs better than the conventional Page Rank algorithm in terms of returning larger numbers of relevant pages to a given query [6].

D. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) is a page ranking algorithm in which according to a user query a sorted order to the web pages returned by a search engine. WPCR is a numerical value based algorithm on which the web

pages are given an order. This algorithm considers web structure mining as well as web content mining as their main techniques whereas in weighted page ranking only web structure mining technique is used. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much a page is relevant. Importance means the popularity of the page which means how much number of pages is referred by or is pointing to a particular page [12].

IV. RELATED WORK

Supreet Kaur et al [13], the introduction about K-means clustering and its algorithm is given. The experimental results of K-means clustering and its performance are discussed in terms of execution time. The more time taken for execution is a limitation in K-means Clustering. So in order to reduce the execution time we are using the weighted page rank with k means clustering.

Amar Singh et al [5], this represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this paper work represents those using Page Rank algorithms users may not get the required relevant documents easily, but in new algorithm Weighted Page Rank user can get relevant and important pages easily as it employs web structure mining and web content mining.

Sonal Tuteja [14], represents the standard Weighted PageRank algorithm is being modified by incorporating Visits of Links. The proposed method takes into account the importance of both the number of visits of in links and out links of the pages and distributes rank scores based on the popularity of the pages. Various ranking algorithms have been developed such as Weighted PageRank, PageRank, HITS, Page Content Ranking etc.

Seifedine Kadry et al [4], an algorithm which is the Simplified Weighted Page Content Rank for page rank, in which two classes "Web content mining" and "Web structure mining" are considered. In this paper, introduction of the SWPCR algorithm which is an enhancement to the well-known algorithm "Weighted Page Rank" which is used by the most famous search engine Google by adding to this algorithm a content weight factor (CWF) to retrieve more relevant page.

Amandeep Kaur Mann et al [7], a review of clustering and its different techniques in data mining is done. Clustering can be done by the number of algorithms such as partitioning, hierarchical algorithms etc. Clustering is important in data analysis and data mining applications. It means making a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). There are different types of clusters: Well-separated clusters, Shared Property, Center-based clusters, Density-based clusters, Contiguous clusters or Conceptual Clusters.

Neelam Tyagi et al [15], a preface to Web mining then trying to explain detailed Web Structure mining, and supply the link evaluation algorithms brought into play by the Web. This paper also explores different PageRank algorithms and compares those algorithms used for Information Retrieval. The comparison of these algorithms in context of performance has been carried out. Page Ranks are designed for PageRank and Weighted PageRank algorithm for agreed hyperlink composition. It explores different PageRank algorithms and compares those algorithms used for Information Retrieval.

V. PROPOSED WORK

A. Objective

- To Design an efficient Algorithm using Weighted Page Content Rank and K-means clustering Algorithms.
- To assign ranks to research papers on the basis of popularity of papers.
- To cluster the research papers on the basis of user keyword query.
- To retrieve the most relevant research papers on the top for a given user query.
- To analyze the parameters on the basis of accuracy, execution time, precision and recall.

B. Outline of algorithm

The algorithms will be used are as under:

K-Means Algorithm

K-means clustering is a popular method for cluster analysis in data mining. In this method, n observations are partitioned into k clusters, where k is the number of clusters defined by the users 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 [8]:
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 a numerical value based algorithm on which the web pages are given an order. Web structure mining is used to calculate the importance of the page and web content mining tells that how much a page is relevant. Here importance is defined as the popularity of the page which means how much number of pages is referred by or is pointing to a particular page. It can't be calculated with the help of in

links only, out links are also to be considered. The matching of the page with the user query shows the relevancy of the page. The page is more relevant if it maximally matched to the user query [12].

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 the N strings are occurring in P or not?
Z= Sum of frequencies of all N strings.
3) S= Set of the 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 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.

VI. METHODOLOGY

The methodology of this research is quite simple. The database has to be created for all types of research papers. The weighted page content rank algorithm will be used to rank the research papers. For clustering the well-known k-mean clustering algorithm will used. For discovering the useful knowledge from the database the knowledge discovery process will be used. This methodology is useful for the users to get their relevant papers on the top according to their query. In last the validated results will be compared with other ranking algorithm on the basis of various parameters like accuracy, precision, recall, execution time, relevancy etc.

Fig. 2 Flow Diagram

VII. CONCLUSIONS

In this paper, study of different clustering algorithms such as Hierarchical algorithm, K-means clustering algorithm, Density Based clustering algorithm; EM algorithm etc. and also study of different ranking algorithms Hypertext Induced Topic Search, Page Rank algorithm, Weighted Page Content Rank Algorithm etc. is done . On the basis of the study, concluded that both content and link based algorithms are important to calculate a final score or page rank of a web page. So, in order to rank massive web pages accurately and effectively, we propose an approach using ranking algorithm as well as clustering algorithm which computes the score on the basis of content as well as link structure of the web pages and users retrieve the relevant results on the top according to the query.

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] 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.

[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] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Emanuele Coviello, Antoni B. Chan and Gert R.G.Lanckriet, "*The Variational hierarchical EM algorithm for clustering hidden Markov models"*.

[12] 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.

[13] 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.

[14] Sonal Tuteja, *"Enhancement in Weighted PageRank Algorithm Using VOL"*, ISSN: 2278-8727 IOSR Journal of Computer Engineering (IOSR-JCE) Volume 14, Issue 5 (Sep. - Oct. 2013).

[15] 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.

