Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.111 – 115

SURVEY ARTICLE

A Survey on k-Means Clustering Algorithm Using Different Ranking Methods in Data Mining

Amar Singh[1], Navjot Kaur[2]

[1]Student of Computer and Science Engineering, Shri Guru Granth Sahib World University, Fatehgarh Sahib, India

[2]Assistant Professor, Computer and Science Engineering, Shri Guru Granth Sahib World University, Fatehgarh Sahib, India

[1] er.amar25@gmail.com; [2] navjot_anttal@yahoo.co.in

Abstract— As the volume of information on the internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet. So it is the duty of the service provider to provide the relevant and good information to the internet user when they submit query to the search engine. To support the user to move in the result list various ranking methods are used. In this paper, we choose Page Ranking Method called Weighted Page Rank Algorithm which is based on the popularity of the page by taking the importance of both the inlinks and outlinks of the pages. After that we use time rank algorithm for improving the weighted page rank score by using the visit time of the web page. So this concept is very useful to display most valuable pages on the top of the result list.

Key Terms: - Web mining; Clustering; K-means, Page rank; Weighted Page rank; Time Rank Algorithm

I. INTRODUCTION

The World Wide Web is a rich source of information and continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc. The following challenges [1] in Web Mining are:

1) Web is huge.
2) Web pages are semi structured.
3) Web information stands to be diversity in meaning.
4) Degree of quality of the information extracted.
5) Conclusion of knowledge from information extracted.

II. WEB MINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web.

A. Web Mining Process

The complete process of extracting knowledge from Web data

Fig. 1: Web Mining Process [3]

The various steps are explained as follows.

1. Resource finding: It is the task of retrieving intended web documents.
2. Information selection and pre-processing: Automatically selecting and pre- processing specific from information retrieved Web resources.
3. Generalization: Automatically discovers general patterns at individual Web site as well as multiple sites.
4. Analysis: Validation and interpretation of the mined patterns.

III. WEB MINING CATEGORIES

Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and Web retrieval. The classification is based on two aspects: the purpose and the data sources. Retrieval research focuses on retrieving relevant, existing data or documents from a large database or document repository, while mining research focuses on discovering new information or knowledge in the data. On the basis of this, Web mining can be classified into web structure mining, web content mining, and web usage mining as shown in Fig. 2.

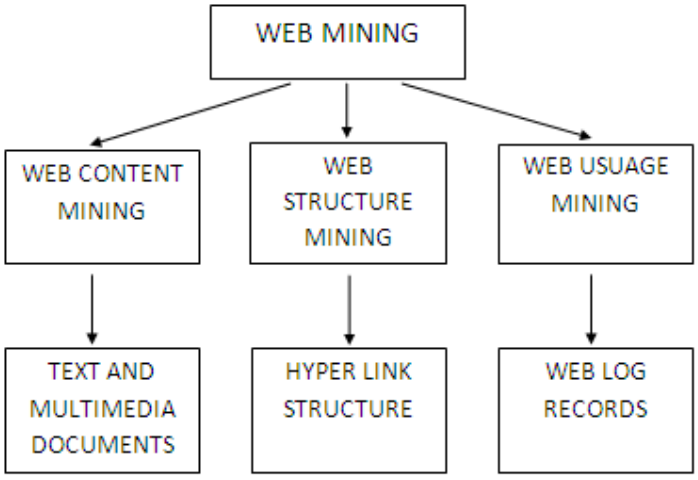

Figure 2: Classification of Web Mining [3]

IV. CLUSTERING

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses is to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

Example to Elaborate the Idea of Clustering

In order to elaborate the concept a little bit, let us take the example of the library system. In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster. For example, books on the database are kept in one shelf and books on operating systems are kept in another cupboard, and so on. To further reduce the complexity, the books that cover same kind of topics are placed in same shelf. And then the shelf and the cupboards are labeled with the relative name. Now when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking in the entire library.

V. K-MEANS [4]

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means Algorithm

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

Choosing the number of clustering

One of the main disadvantages to k-means is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if you had a group of people that were easily clustered based upon gender, calling the k-means algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20, the results might be too generalized to be effective.

VI. DIFFERENT RANKING METHODS

A. Page Ranking

In simple terms, page rank is a measure of how 'important' a web page is. It works on the basis that when another website links to your web page, it's like a recommendation or vote for that web page. Each link increases the web page's page rank. The amount it increases depends on various factors, including how important the voting page is and how relevant it is.

Page Rank is a numeric value that represents how important a page is on the web. Page Rank is the Google's method of measuring a page's "importance." When all other factors such as Title tag and keywords are taken into account, search engine uses Page Rank to adjust results so that more "important" pages move up in the results page of a user's search result display. Search Engine Figs that when a page links to another page, it is effectively casting a vote for the other page. It calculates a page's importance from the votes cast for it. How important each vote is taken into account when a page's Page Rank is calculated. It matters because it is one of the factors that determine a page's ranking in the search results. It isn't the only factor that Google uses to rank pages, but it is an important one.

The order of ranking in Search Engines works like this:

- Find all pages matching the keywords of the search.
- Adjust the results by Page Rank scores.

The algorithm of Page Rank as follows:

Page Rank takes the back links into account and propagates the ranking through links [5]. A page has a higher rank, if the sum of the ranks of its backlinks is high. The original Page Rank algorithm is given in following equation

$$PR(P)=(1-d)+d(PR(T1)/C(T1)+.....PR(Tn) / C(Tn))$$

Where,

PR (P) = Page Rank of page P

PR (Ti) = Page Rank of page Ti which link to page

C (Ti) = Number of outbound links on page T

D = Damping factor which can be set between 0 and 1.

B. Weighted Page Rank [6]

Weighted Page Rank algorithm (WPR): This algorithm is an extension of Page Rank algorithm. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional Page Rank algorithm in terms of returning larger numbers of relevant pages to a given query. According to author the more popular web pages are the more linkages that other WebPages tend to have to them or are linked to by them. The proposed extended Page Rank algorithm–a 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 outlink pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as Win(v,u) and Wout (v,u), respectively. WPR supplies the most important web pages or information in front of users.

C. Time Rank Algorithm [7]

An algorithm named as Time Rank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al. [8] Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

VII. CONCLUSION

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. In this paper we focused that by using Page Rank and Weighted Page Rank algorithms users may not get the required relevant documents easily, but in new algorithm we use two algorithms i.e. Weighted Page Rank and Time Rank in which user can get more relevant and important pages easily on the top list of the results as it employs web structure mining. In this firstly we run Weighted Page Rank algorithm on k means Cluster then after that we apply time Rank on it. By doing this we get more relevant data which satisfied the user query more accurately and give relevant data to the user.

REFERENCES

[1] Rekha Jain, Dr G.N.Purohit, “Page Ranking Algorithms for Web Mining,” International Journal of Computer application,Vol 13, Jan 2011.

[2] Cooley, R, Mobasher, B., Srivastava, J."Web Mining: Information and pattern discovery on the World Wide Web”.In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI’ 97).Newposrt Beach,CA 1997.

[3] Tamanna Bhatia,” Link Analysis Algorithms For Web Mining”, IJCST Vol. 2, Issue 2, June 2011.

[4] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur,” EFFICIENT K-MEANS LUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING”, International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[5] Taher H. Haveliwala, "Topic-Sensitive Page Rank: Context-Sensitive Ranking Algorithms for Web Search", IEEE transactions on Knowledge and Data Engineering Vol.15, No 4 July/August 2003.

[6] Wenpu Xing and Ali Ghorbani, “Weighted Page Rank Algorithm”, In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[7] Dilip Kumar Sharma et al.,” A Comparative Analysis of Web Page Ranking Algorithms”(IJCSE) International

Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676.

[8] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

