International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Survey on K-means Clustering and Web-Text Mining

Aayushi Bindal[1], Analp Pathak[2]

[1]M.Tech Student, Department of computer Science, SRM University, Chennai, India

[2]Assistant Professor, Department of computer Science, SRM University, Chennai, India

Abstract: ***We are presenting a survey on the k-means clustering and web-text mining. Text mining refers to the extracting useful concepts from the text and Web mining refers to finding the useful and previously unknown information from the web. As we required more time to search the research papers. It consumes more time to read a single paper. So it is necessary to move forward new search engine based on fastest reading model .The main problem is that Ranking of research papers does not allotted. K-Means clustering refers in which the given data set is divided into K number of cluster. So in order to reduce the execution time we are using the weighted page rank with k means clustering .We will present the research related to assign the ranks of research papers on the basis of popularity of papers.***

Keywords: Text Mining, Web Mining, K-means Clustering Algorithm, Weighted Page Rank

1. Introduction

Along with the development of new smart technologies, the world is going to be digital. In present, the web is growing day by day and has become a vast source of the information. Looking up for the precise and relevant information, extracting it from the web has now become a time-consuming task. There are different techniques used for the information extraction from the web and text mining is one of them. In today's highly competitive business environment Clustering plays an important role. As K- means Clustering refers to a method for making groups of the data set or the objects that are having similar properties. Data mining means assembling of data which extract the pattern and make the relationship between data and its multiple attributes.

WWW or World Wide Web is a big resource of heterogeneous and hyperlinked information including audio, video, text, images , graphics and metadata. From early 1990's WWW has seen an tremendous growth. It is estimated that Web has expanded by about 2000% since its evolution and is doubling in size every six to ten months [1]. With large increase in availability of information through WWW, it is difficult to acquire the useful or important information on Internet. So many users use Information retrieval tools like Search Engines to search specific 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 . 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 best result of the work. But the search of relevant and latest papers in less time on the top is such a difficult task. For retrieving 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. All the extracted information is linked together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. The classic approach of information retrieval based on keyword search from WWW makes it cumbersome for the users to look up for the exact and precise information from the search results. It is up to the user to go through each document to extract the relevant and necessary information from those search results. This is an impractical and tedious task .Text mining can be a better solution for this as it links all the extracted information together and it also pushes all the irrelevant or not relevant information aside, and keeps the relevant ones based on the question of user interest.

This paper focus on processing of structured and unstructured data mining .With the tremendous growth in website, web portal to provide downloaded data to the user .The semantic web is about machine-understandable web pages to make the web more intelligent and able to provide useful services to the users. The data structure definition and Pattern recognition is to estimate the accurate page ranking and to produce better result while searching operation with web data.

2. Background Theory

2.1 Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a direction that objects in the same group are called a cluster. It is a primary task of explanatory data mining a common technique for statistical data analysis used in various fields including machine learning, pattern, picture analysis, data retrieval & Bioinformatics. In clustering method, targets of the dataset are grouped into clusters, in such a way that groups are almost different from each other and the objects in the same group or cluster are very alike to each other. Unlike Classification, in which previously defined set of categories are faced, but in Clustering there are no predefined set of classes which means that resulting clusters are not recognized before the implementation of

clustering algorithm. These clusters are extracted from the dataset by grouping the objects in it.

2.2 Web Mining

Due to rapid growth of web data, information, files on the internet throughout the world, web mining came into picture. The web data on the internet are heterogeneous, diverse and large. Therefore, the arrangement of different data's must be compulsory to provide these data's to different group of users efficiently. so there is need of data mining on these data by the help of it, user gets relevant information from the web. In web mining many techniques of data mining are applied on the web data's. But it is not the only body that actually fit for this purpose, besides data mining, artificial intelligence, information retrieval, natural language processing technique can be used efficiently[web mining today and tomorrow]. Web mining gives sophisticated result while accessing the web by the users. In research paper the exact definition of web mining given as follows: "Web mining is the application of data mining techniques to find interesting and potentially useful information from the web. It is normally expected that either the hyperlink structure of the web or web log data or both have been used in mining process ."Web mining is based on knowledge discovery from web, extract the knowledge framework represents in proper way. Web mining is like a graph and all pages are node and each connects with hyperlinks. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. " Web mining can be categorized into three :

2.2.1 Web Content Mining

Web Mining is basically extract the useful and important information on the web .in which the process is happen to access the information on the web. So that's why it is called web content mining. There are many pages which are open to access the information on the web. These pages are the contents of web. Searching the information and open search pages is also content of web .So there are various contents present in web in the form of text, image, video, sound etc. But the main resources that are mined in web content mining are individual pages. Web content mining is differentiated from two different points of view that is Information Retrieval View and Database View which summarized the research works done for unstructured data and semi-structured data from information retrieval view.

2.2.2 Web Structure Mining

We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs. It is shown the link one web page to another web page. Web structure mining actually focuses on link information

2.2.3 Web Usage Mining

Web mining usage aims at utilize data mining techniques to discover the usage patterns from web based application. It is technique to predict user behaviour when it is interact with the web. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

2.3 Text Mining

Text mining refers to the process of deriving high quality of information from the text documents.It is a challenging task to help the users in finding what the user's actually want from the number of text documents. It is quite tough to deal with the text which is in unstructured form. The main purpose of the text mining is to finding the interesting information from the natural language text . 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 automation methods for achieving the common knowledge which is available in text documents.

3. Ranking

A ranking is a relationship between a set of items such that, for any two points, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In mathematics it is not necessarily a total order of objects because two different objects can have the same ranking. The rankings themselves are totally merged [2].By reducing detailed measures to a sequence of ordinal numbers, rankings make it possible to evaluate complex information according to certain criteria. Thus, for example, an Internet search engine may rank the pages it finds according to an estimation of their relevance, making it possible for the user quickly to select the pages they are likely to want to see.

With regards to Clustering, ranking operations to estimate the likelihood of the occurrence of data items or the targets. Thus paper proposed to evaluate ranking of overall design of database. Then the ranking function introduces new opportunities to optimize the effects of K-means clustering algorithm.

3.1 Need of Ranking

Search of relevant records or like data search is a most popular function of database to get knowledge .That is why, we need to rank the research papers. The, related answers will be delivered for a given keyword query by the created index and better ranking strategy. And then we applied this Ranking method with K- means clustering method because this method is likewise causing the property to obtain relevant records. . So it is also helpful for creating clusters that are having similar properties between all data points within that bunch.

In recent search engines there are many return millions of pages for a certain query but it is not possible for a user to see all the concluded results. Therefore, 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.

3.2 Methodology

Existing work has been implemented for the ranking of web pages, web links etc. but this work is not time efficient and the 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.

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.

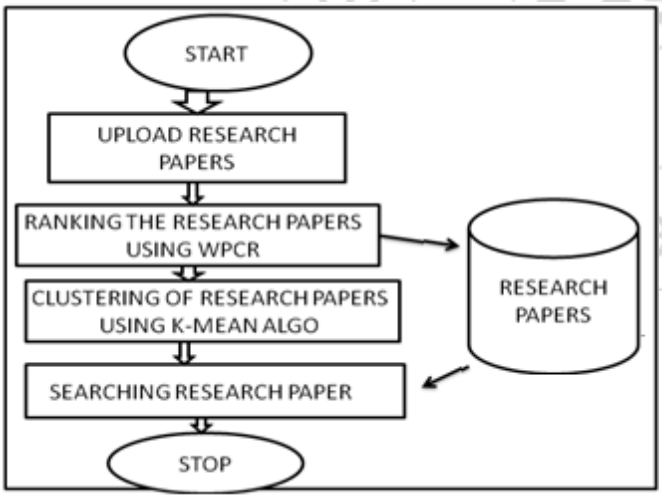

Figure 1: Flow Diagram

The research methodology is divided into steps to achieve our desired goal:

1) Upload the research papers to provide rank.
2) This step include the implementation of Page Content Rank algorithm which assigns rank to all the research papers and store them in database.
3) K-means clustering algorithm is implemented to cluster the research papers on the basis of different research areas.
4) Relevant research papers from the database are searched by user according to the research area.
5) If research paper find then result compared on the basis of accuracy , execution time etc.

4. Literature Review

Syed Thousif Hussain [5]-2012 have proposed the approach which is used to generate a high number object class. This sort of querying the object investigate all type of object and data associated with it. It gives the output based on the re-rank of image and its object class. First it download all the relevant images and the web pages. Then on extracting features it investigate about the downloaded page. and place it in the database and then ranking is done based on text surrounding and metadata features.

The approach is to employ text, metadata and visual features and to use to gather many high quality images from the web. Candidates images are obtained by text based web search. Downloaded page and then keep in track and start classification of extracted feature data. SVM (Support Vector Machine) and Naive Bayes classifier algorithm are compared for ranking. The top rated images are utilized as training data and an SVM visual classifier is learned to improve re-ranking. The main idea of the overall method is in combining text or metadata or visual characteristics in order to reach a completely automatic ranking of images.

Wenpu Xing [7]-2004 discussed a new approach known as weighted page rank algorithm (WPR). This algorithm is an extension of the Page Rank algorithm. WPR performs much better than the conventional Page Rank algorithm in terms of making the larger piece of relevant pages to a passed query.Basically, Page Rank is a way of measuring the importance of website pages.

Neelam Tyagi [6] - 2012 have analyzed that the World Wide Web consists millions of web pages and there is large amount of data available within the web pages. In this report, a page ranking mechanism called Weighted Page Rank Algorithm based on Visits of Links (VOL) is being devised for search engines, which functions along the footing of the weighted Page rank algorithm and calls for a number of visits of inbound links of web pages into account. The original WPR is an extension to the standard Page Rank algorithm. The suggested algorithm is used to obtain more relevant data according to a user's inquiry. Hence, this concept is actually useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour, which shorten the search space to a large plate. The story also presents the comparison between original and VOL method.

Kavita Sharma [4] – 2011 have hit the books about how to extract the important information on the web and also pass the superficial knowledge and comparison about data mining. This paper describes the current, past & future of web mining. This introduces online resources for retrieval Information on the web, i.e. web content mining, & the discovery of user access patterns from web servers, i.e. web usage mining that enhance the data mining drawback and

web structure mining i.e. for analysis the hyperlink structure and document construction. Furthermore, this paper also described web mining through cloud computing i.e. cloud mining.

Summary: in this dissertation various techniques related to the literature been discussed and below table represent various advantages and disadvantage comparison in between discussed technique.

Table 1: Comparison Analysis Table

Methodology	*Advantage*	*Disadvantage*
SVM (Support Vector Machine) And Naive Byes classifier algorithm are compared for ranking	This method works on combining text or metadata or visual features in order to achieve a completely automatic ranking of images.	High computation time
Web page rank algorithm (WPR)	Web page rank algorithm WPR performs better than the conventional Page Rank algorithm in terms of returning larger number of relevant pages to a given query.	Specific to given number of links no route discovery
Weighted Page Rank Algorithm based on Visits of Links (VOL)	It is very useful to find more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour, which reduce the search space to a large scale	High computation and more query execution in process
Hyperlink-Induced Topic Search (HITS).	The scheme therefore assigns two scores for each page its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages and provide better searching results.	Relevant queries are not optimized.

5. Conclusion

K means with page rank algorithm gave results with better result set of various numbers of data-sets. In our case we have worked on k means clustering of database with weighted page rank algorithm. The future work may benefit with less computational time as compared to previous work as database with records are increasing day by day and there is a need of data clustering on large databases.

References

[1] N.Duhan, A. K. Sharma and Bhatia K. K.,'' Page Ranking Algorithms: A Survey'', Proceedings of the IEEE International Conference on Advance Computing, 2009, 978-1-4244-1888-6.
[2] Vikram Garg , Preetibala Deshmukh ,''A Survey Paper of Structure Mining Technique using Clustering and Ranking Algorithm'', proceedings of the International Journal of Computer Applications (0975 – 8887) Volume 119 – No.13, June 2015.
[3] Tamanna Bhatia, "Link Analysis Algorithms For Web Mining" ISSN :2229 - 423 (Print) |ISSN : 0976 - 8491 (Online) IJCST Vol. 2, Issue 2, June 2011.
[4] G. Shrivastava, K. Sharma, V. Kumar " Web Mining Today and Tomorrow" International Conference on Electronics Computer Technology (ICECT) April 2011..
[5] Syed thousifhussain B.N.Kanya "Extracting Images From The Web Using Data Mining Technique",International Journal of Advanced Technology &Engineering Research , March 2012,
[6] NeelamTyagi and Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Computing and Engineering (IJSCE) July 2012
[7] Wenpu Xing and Ghorbani Ali, "Weighted PageRank " IEEE,2004.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] M. Mete, X. Xu, Chun-Yang F., Gal Shafirstein, Head and Neck Cancer Detection in Histopathological Slide, International Workshop on Data Mining in Bioinformatics, Sixth IEEE International Conference of Data Mining (ICDM 2006), December 18-22, 2006, Hong Kong.