Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Improved Algorithm for Optimization of Web Content Mining

Arti Sharma
Shobhit University, Meerut,
U.P., India

Dr. Shashank Awasthi
G.L.Bajaj Institute of Technology &
Management, India

Abstract*— By this paper we would like to reduce the crisis which are substance held in the web pages. The elementary point of our paper is to assess, recommend and develop the exercise of web data clustering techniques that is mainly used by special advent of mining huge content based data sets that allows data analysts to conduct more efficient execution of large scale web data search. The data is available in the search space in a random fashion which may cause trafficking when searched multiple times. Consequently, inside this provided paper we make available a better algorithm which may diminish the search space using clustering techniques in search engines.

***Keywords*—Web mining, clustering algorithm, web document, data sets, K-mean clustering and Web usage mining.**

I. INTRODUCTION

This paper is primarily focussed in describing the most important issues related to "Using highly improved clustering techniques improving the efficiency of textual static web content mining techniques".This Clustering has set up many applications in Web search for example if we search a keyword it might return a large number of hits i.e. is pages relevant to the search due to extremely large number of web pages. Therefore we can conclude that clustering techniques can be highly recognised to arrange data in groups and then and then present the result in precise and concise available method However when the content of data is large clustering can be used to cluster data into topics which are commonly used in information reclamation practices. At the same span of time as a data mining function cluster analysis can be used as a data mining tool to gain imminent into splitting up of data, to supervise the characteristic of each data cluster and helps to focus in clusters of different kinds for further analysis.

1.1 WEB MINING CATEGORIES

Web mining is highly divided into following categories upon which our research is focused these are web content mining, web structure mining and web usage mining.

A. Web Content Mining

Web content mining, also known as text mining, is normally the subsequent step in Web data mining. Content may be referred to as scanning and mining of pictures and graphs and text of a Web page to determine the relevance of the substance realted to the search query. This kind of scanning is done after the clustering of web pages using structure mining and it also makes available the consequences depending upon the relevance to the suggested query. The massive quantity of information facilitated on the World Wide Web, the results lists to search engines is available by the content mining in order of highest relevance to the keywords in the query.

B. Web Structure Mining

It is a process by which we discover the model of link structure of web pages. We catalog the links, to generate the web pages produce the information such as the resemblance and relations among them by taking the advantage of hyperlink topology. The main objective of Web Structure Mining is to engender structured summary regarding website and corresponding web page. Page Rank and hyperlink analysis also fall in this category. It tries to find out the link structure of hyper links at inter document. Since it is very pervasive that the web documents and credentials contain links and as well as they use both the real or crucial data on the web so it can be accomplished that Web Structure Mining is related to Web Content Mining.

C. Web Usage Mining

Web Uasge mining is the procedure by which we can recognize the browsing patterns by means of analyzing the navigational actions of user. It focuses on techniques that can be used to predict the user activities while the user is having an interaction by means of the web. It formulates use of the secondary data available on the web. Such kind of activity involves the involuntary discovery of user access patterns from single or more web servers.
Through this mining technique we can ascertain what users are looking for on Internet. It consists of three phases, amely preprocessing, pattern analysis and discovery.

1.2 EXISTING SYSTEM

Text mining is an emerging technology for extracting meaning from the "unclustered and unstructured text "that constitutes a mainstream of endeavor information resources. Clustering can be referred as a technique to group together a set of items having similar characteristics. Clustering can be performed on either the users or the page views. Clustering analysis in web usage mining intends to find the cluster of page, or sessions from log file, where each and every cluster correspond to a group of objects with common interest or features. User clustering is intended to discover user groups that have common interests depending on their behaviors, also it is critical for user community construction. Page clustering can be said as the process of clustering pages according to the users' access above them. Such kind of knowledge is especially useful for inferring user demographics so as to perform market segmentation in e-Commerce applications or provide personalized web content to the various users.Alternatively, the clustering of pages will discern groups of pages having related content. Permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information requirements. The discernment is that if the chance of visiting page, given page has also been visited, and is high, and then maybe they can be grouped into lone cluster.All the sessions are processed to find some interesting session clusters for session clustering. All session cluster possibly will be one interesting topic within the web site. Abraham et al [AR2003] proposed an ant-clustering algorithm to discover web usage patterns and a linear genetic programming approach to analyze the visitor trends. They proposed fusion framework, which makes use of an ant colony optimization algorithm to cluster or group Web usage patterns. Cleaning and preprocessing of raw data is done from log files and the ACLUSTER algorithm is used to identify the convention patterns. The urbanized clusters of data are fed to a linear genetic programming model to analyze the usage trends.

The rationale of knowledge discovery from users profile is to find clusters of similar interests among the users [SZAS1997].

II. PROBLEM PROCLAMATION

It is known that whenever we have the procedure of web content mining we have to access quite huge quantity of textual information from unlike heterogeneous sources and thus this task becomes very unwieldy as data recovery is complicated and time consuming.

Assume that we have a large association and have singular managers who watch over the various operations in that association so our basic task is to assemblage the employees in dissimilar clusters. For such kind of cluster formation we use unique techniques available in the literature as we are concerning on the partitioning techniques of clustering which is basically of two types these are K -means and K-medoids. These two algorithms basically work on repeated number of scans in the database for cluster formation and thus give us approximate results. Some problems with existing techniques are as follows:

- Optimizing the within cluster variation is computationally challenging.
- The k-means method is not guaranteed to converge to the global optimum and often terminates at a local optimum.
- The necessasity for users to identify k, the figure of clusters in advance can be seen as a disadvantage.
- The web pages at different servers are similar thus clustering the data according to the relevance requires a large amount of query system.
- Filters are used widely and outliers may not be detected and if detected may not be recognized.
- Superfluous web pages might be outliers and therefore they have to be out clustered and sorted as per their relevance.

2.1 ALGORITHM REVIEWS

The k-means algorithm is a distance – based clustering algorithm that partitions the data into a predetermined number of clusters.

The K-means method is the clustering algorithm :

- It takes a parameter k as input, which point to the quantity of clusters the user wants to structure.
- Initially, k values (points) are chosen at random from the Set of all data points to represent the centre (mean) value of all cluster. Then all other point on the plane is allocated to the cluster it is closest to. The "closest cluster" is determined by the shortest Distance from a point to the mean value of each cluster, Using formula: d = √ (α(x1 - x2)2 + β (y1 - y2)2) Where α and β are coefficients with a default value of 1.000 The k-means algorithm works only with numeric aspects. The distance base algorithms considers a distance metric to measure the similarity among data points. therefore the distance metric is either Euclidean, Cosine or Fast Cosine distance. According to the distance metric the distance points are assigned to the nearest cluster.

The k-mean algorithm:

Begin with the data sets.

1. k-means algorithm for partitioning , where the centre of each cluster's is represented by the mean value of the objects within the cluster.
2. Input:
   Initialize NS, KN,CC1,CC2….CC_k;
   Where NS is the size of the dataset,
   KN is the number of clusters, $CC_{1,}CC_{2,\dots}CC_k$ are cluster centers.

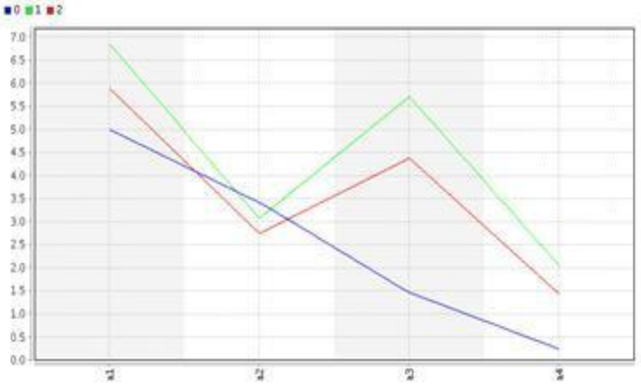
3. Do assign the n data points to the closest CC_i; Recomputed CC1,CC2….C_k using Simple mean Function; Until no change in CC1,CC2….$CC_{k;}$
4. Return CC1,CC2….$CC_{k;}$
5. End

Now when taking sample data of market analysis and check whether the cluster formation we come into the conclusion of getting the following cluster formation with the following graphical representation.The transactional data with numeric attributes forms three clusters and with four frequently occurring elements thus these elements with three mean values are clustered according to the following representation.

Table 1: cluster organization with attribute set.

Attribute	Cluster 0	Cluster 1	Cluster 2
a1	5.005	6.853	5.883
a2	3.418	3.076	2.740
a3	1.464	5.715	4.388
a4	0.243	2.053	1.434

Therefore, when the cluster formation is represented graphically we get the desired curve that is given below where each one attribute is exposed with a particular cluster formation where we met our goal of high intracluster resemblance and low intercluster resemblance.



III. PROPOSED CONCEPT

Clustering is dependent on likeness. In clustering assessment it is required to work out the relationship or remoteness. So when the data is bulky or scattered manner it is quite difficult to as it should be arrange them in a cluster. The mean problem with the above k-mean algorithm is the repeated number of scans and then selecting the minimum mean value which is highly affected by extreme standards. To surmount this predicament a new method is proposed which highly groups the elements and reduces the redundancy which in turn when clustered by the above algorithm will limit the number of scans and produce a highly acceptable minimum clusters with more number of elements in a compressed form thus eliminating the risk of high outliers and low intracluster similarity.

Proposed Algorithm: Input

The set of data of numeric form and we have to apply the following process.

Procedure

1. The data set is arranged as per the probability of occurrence of elements from the rate of recurrence of occurrence within a specified data set
2. This procedure is recur until no new large item sets are identified (this is to eradicate unnecessary data and absent values).
3. faction the rudiments according such as no two elements are repeated and thus multiple the probabilities such that every elements is paired with every other element.
4. Initialize no, L, cc1,cc2….cc_k; Where no is the size of the data set, L is the number of clusters, $cc_{1,}cc_{2,\ldots.}cc_k$ are cluster centers.
6. Do allocate the no data points to the closest cc_i; Recomputed cc1,cc2….cc_k using Simple mean Function; Until no change in cc1,cc2….$cc_{k;}$
7. Return cc1,cc2….$cc_{k;}$
8. End

3.1 APPLICATION OF THE PROPOSED APPROACH

The above algorithm has its own reward from the simple k-means algorithm particularly in cluster formation. Since the proposed algorithm apply a different approach of probability distribution of frequently occurring elements and grouping of elements thus the attribute head count is also condensed. Currently taking again the same sample data used in simple k-means algorithm we get some different results as follows.

The probability of occurrence of elements is calculated as number of events occurred upon number of possible outcomes thus whenever we have large dataset or suppose data of credit card collections where a single person can have different credit cards with different fraudulent information with each and he swipes cards at different counters multiple times and every time he swipes a card a transaction is generated and a query is generated for its collection. So, if we want to limit its customer data access according to credit cards or according to transactions or purpose we need to form clusters representing different information parameters.

Now, suppose if we take the same data set as above with k means it is given as follows:

Table 2:Given below are the Example of data set used for analysis within k-means and proposed algorithm.

TID	ITEM SETS
1	1,2,3
2	1
3	4,5,6

With K-means if we take k=2 and finding the within cluster variation we get the following output with two groups: Cluster 1(1, 2, 3) (1, 4, 5, 6) here 1 is an redundant data set in both the given groups. Cluster 2(1, 2) (3, 1, 4, 5, 6) now here we find the within cluster variations by undergoing number of scans we get 15 for cluster 1 and 34.5 for cluster 2. Thus we come into conclusion that smallest the cluster variation high is the intracluster similarity for the object and cluster centroid. Now, when we form the chart for the distribution according to probability we get the following table:

Table 3: computation of frequency and probability of occurrence of elements to compute the cluster analysis with proposed algorithm.

ITEMSETS	FREQUENCY	PROBABILITY
1	2	0.333
2	1	0.167
3	1	0.167
4	1	0.167
5	1	0.167
6	1	0.167

This is the table generated according to above data set with the proposed algorithm. Now, when we divide them into clusters we get the following results according to the probability distribution. Cluster 1(0.333, 0.167) (0.167, 0.167, 0.167, 0.167) we get within cluster difference as 0.332.Cluster number 2(0.333, 0.167, 0.167) (0.167, 0.167, 0.1670 we get within cluster difference as 0.027. Therefore we can say that 0.027 is the accepted cluster when k=2 in this algorithm which has the minimum cluster variation and compressed clusters with very high intracluster similarity and very low intercluster similarity.

Table 4: Results Comparison

Test data	K-means	Proposed Concept
Group 1	15	0.332.
Group 2	34.5	0.027.

IV. CONCLUSION

This paper presents a new method for an enhanced cluster formation as the clusters are good in number and high in quality of information for investigation. The concert of new algorithm neither does nor depends upon the size, scale and values in dataset. The new algorithm has great advantages in error with real results and selecting initial points in almost every case. Thus the advantages above the k-means algorithm are given as follows:

- This approach saves cost and time of scanning highly large database in ETL application of scanning a large database as the probability function reduces the redundancy of frequently occurring elements.
- Whenever we find probabilities of large datasets it is easier to generate fast moving data streams such as real time traffic and network monitoring thus clustering becomes easier and compact.
- This method is highly scalable as optimization of within cluster variation is drastically improved.
- This method can be used in scan global databases from different heterogeneous sources thus and does not terminate at local optimum as increasing the number of groups compresses the data to further extent.
- Outliers can be easily adjusted and removed as redundant and missing values and placed within a cluster as in k-means outliers are not easily detected.
- Since we group according to the probability distribution thus as number of cluster increases web content becomes move centric and unique.

V. FUTURE SCOPE

In view of the fact that in this we assemblage according to the probability distribution, in future we can work on improving the contained by cluster variation by first organization of the objects according to correlation within each and then finding probability and running the scans with any clustering algorithm . This will improve the excellence of clusters and entropy of information gain from each cluster.

REFERENCES

[1] R.Agrawal and R.Srikant. Fast algorithms for mining association rules.In VLDB'94, pp.487 {499.

[2] Data Mining: Concepts and Techniques: Concepts and Techniques -Jiawei Han, Micheline Kamber, Jian Pei.

[3] Optimizing the Web Mining Techniques using Heuristic Approach –Gunjan atral,Vijay Laxmi2 and M.Afshar Alam3.

[4] A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms- Preeti Chopra, Md. Ataullah- (IJEAT)Feb2013.

[5] A Survey Paper on Hyperlink Induced Topic Search (HITS) Algorithms for Web Mining- Mr.Ramesh Prajapati- (IJERT) April-2012.

[6] Web Content Mining Techniques-A Comprehensive Survey- Darshna Navadiya, Roshni Patel- (IJERT) December -2012.

[7] Text Classification Using Data Mining-S. .Kamruzzaman, Farhana Haider,Ahmed Ryadh Hasan-ICTM2005.

[8] [AR2003]. Ajit Abhraham, Vitorino Ramos, Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming, to appear in CEC´03 - Congress on Evolutionary Computation, IEEE Press, anberra, Australia, 8-12 Dec. 2003.

[9] [SZAS1997]. Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah, Knowledge Discovery from Users Web-page Navigation, IEEE RIDE 1997

[10] [MHD2003]. Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall, 2003.

[11] By Using Modified Clustering Algorithm Optimization of Web Content Mining, Srishti Vashisht, Ms. Anshul Tickoo IJARCSSE 4(8)