International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 4, June - 2012

An Enhanced Clustering Technique for Web Usage Mining

V.CHITRAA
Assistant Professor ,
CMS College of Science and Commerce
Coimbatore, Tamilnadu, India

Dr.ANTONY SELVADOSS THANAMANI
Reader in Computer Science, NGM College
Pollachi, Coimbatore,Tamilnadu, India

Abstract

The World Wide Web has become the default knowledge resource for many years of endeavor, and organizations need to understand their customers behavior, preferences and future needs. Web personalization is an active research topic in which user session clustering is done to understand users activities. Cluster analysis is a widely used data mining algorithmand is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. In this paper an enhanced method to partition into accurate clusters is discussed. The algorithm is carried out in two steps and clusters are with high quality. The experimental results show the performance of the proposed algorithm and comparatively it gives the good results.

Keywords: Clustering, K-Means, Preprocessing, Similarity, Web Usage Mining

I. Introduction

Due to increase in the amount of data in World Wide Web and usage, a lot of research is done in extracting and discovering hidden information related to it. According to Internet World Stats, from 2000 to 2011 there has been a 528.1% increase of web users. However, this plethora often creates problems to the users being unable to retrieve useful and relevant information. Analysing navigational browsing patterns of users is a potential approach to solve this problem. It also help organizations to provide personalized recommendations of web pages according to the current interest of the user.

Web mining is the application of data mining techniques to automatically retrieve , extract and evaluate information for knowledge discovery from web documents and services. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web Content Mining deals with the discovery of useful information from the web contents or data or documents or services. Web Structure Mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites. Web Usage Mining mines the log data stored in the web server. Of late, Web usage mining has gained much attention which analyses the browsing patterns. There are four stages in web usage mining. They are Data Collection, Preprocessing, Pattern Discovery, Pattern Analysis.

Among the four stages Preprocessing is an important step because of its complex nature of web architecture and it takes 80% of mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes tasks such as data cleaning, user identification, session identification, transactions construction. Data Cleaning is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis. User Identification is the process of identifying users by using IPaddress and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site. In Session Identification various methods are used to find set of pages visited by a user within the duration of a particular visit. At last transactions are constructed which are defined as a subset of user session having homogenous pages.

The pattern discovery stage is applying data mining techniques like association rule mining, clustering etc., on preprocessed log data. There are two types of clusters to be discovered: usage clusters and page clusters. Clustering usage data is to find groups of users having similar browsing pattern. The aim of clustering web pages is to divide the dataset into groups of pages which have similar content.This study deals with clustering log data.

The pattern analysis stage is to analyze the patterns found during the pattern discovery step. In this validation of clusters is done and saved for

further recommendations when a new user visits the web site. Uninteresting rules are ruled out and analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

This paper is divided into 5 sections and structured as follows. Section 2 analyzes works related to this paper. Section 3 describes the proposed method. Section 4 presents experimental results. Section 5 reports the conclusions of the authors.

II. LITERATURE REVIEW

Web usage mining is an effective approach in which mining techniques are applied to large web repositories to discover user access patterns automatically. Some of the algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering. Cluster analysis is classified as follows[4]:

Distance Based Clustering : Method according to the distance between data. This algorithm is sensitive to noise data and isolated.

Density Based Clustering: Thisclustering deals with connecting region with the same density. Therefore the density clustering needs the scanning of the entire data set and spatial data will be divided into different small squares, which is approximately express as clusters. It is also used for spatial index structure, clustering by calculating the density in a ball region.

Link Based clustering : This analysis puts the clustering object into a graph model or hypergraph model mapping and then finds out the node set with high connectivity based on the edges or super edges.

Clustering web usage data is different from the traditional clustering due to the data. Therefore, there is a need to develop specialized techniques for clustering analysis based on Web usage data. Some approaches to clustering analysis have been developed for mining the Web access logs. Perkowitz and Etzioni [10] discuss adaptive Web sites that learn from user access patterns. The Page Gather algorithm uses the page co-occurrence frequencies to find clusters of related but unlinked pages. A technique for capturing common user profiles based on association-rule discovery and usage-based clustering is discussed in [8]. Users are classified using a hypergraph partitioning technique by Cooley[3]. Cooley's method is used to identify particularly interesting and similar path histories, but it cannot be used to gain an overall picture of all usage of a Web site. Unsupervised robust multi-resolution clustering techniques to discover Web user groups[9].

Xie and Phoha [11] use belief functions to cluster Web site users. They separate users into different groups and find a common access pattern for each group of users. Unfortunately the approach still needs to identify sessions. An agglomerative hierarchical clustering algorithm is defined by Mayil and Duraiswamy [7]. Banerjee and Ghosh [1] calculated the similarity by using longest common subsequence (LCS) applied the clustering algorithm to cluster the sessions.

It is concluded from the literature review that for session clustering, we so far lack of a complete preprocessing methodology. Hence our proposed methodology will not only improve the quality and efficiency of data for later steps but also to enhance the log file visibility and to structure the information in hierarchical clustering.

III. PROPOSED METHOD

Whenever a user hits a page the log data is collected automatically in Web servers. It represents the accurate navigational behavior of visitors. It is the primary source of data in Web usage mining. Each hit against the server, corresponding to an HTTP request, generates a single entry in the server access logs. There are different forms of log files like Apache, IIS etc., Each log entry may contain fields such as date time s-ipcs-method cs-uri-stem cs-uri-query s-port cs-username c-ipcs(User-Agent) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes. A sample log is given below

```
2007-12-06 05:22:16 ::1 GET /iisstart.htm - 80 - ::1
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.50727;+Media+Center+PC+5.0;+InfoPath.1;+.NET+CLR+1.1.4322;+.NET+CLR+3.5.21022;+.NET+CLR+3.0.04506)
200 0 0 296 336
```

3.1 Data Cleaning

The task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data to be removed [6]. They are.

a. Additional Requests: A user's request to view a particular page often results in several log entries. Graphics and scripts are downloaded in addition to the HTML file, because of the connectionless nature of the HTTP protocol. Since the main intention of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Suffix part of an URL is checked and eliminates suffixes like gif, jpg, GIF, JPEG, css, map etc.

b. Robot's requests: Web robots are software tools that can scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. To remove robot's request, we can look for all hosts that have requested the page "robots.txt", which is checked by robot while browsing.
c. Entries with error: Status code shows the success or failure of a request. Entries with status code less then 200 and greater than 299 are failure entries which are to be removed.

Only necessary fields like date, time, IPaddress, User Agent, URL requested, URL referred, time taken are considered for further experiments to reduce the processing time So attribute subset selection is done.

3.2 User Identification

The strategy of User Identification based on the log entries without considering the topology structure of site. User's IP addresses of two consecutive entries are compared. If the IP address is the same, user's browser and operating system is verified and if both are same, both the records are considered from the same user.

3.3 Session Identification

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a website. A user may have a single or multiple sessions during a period. Once a user has been identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. The simplest methods are time oriented in which one method based on total session time and the other based on single page stay time. The set of pages visited by a specific user at a specific time is called session time. It varies from 25.5 minutes to 24 hours while 30 minutes is the default timeout by Cooley [2]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. In the proposed method the total session time is set as 60 minutes. From the users set entries are divided as sessions. Transactions are derived for each session by using reference length method.

3.4 Clustering Algorithm

Clustering is a technique to search hidden patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. A popular clustering method that minimizes the clustering error is the k-means algorithm. It is attractive in practice, because it is simple and it is generally very fast. It partitions the input dataset into k clusters. Each cluster is represented by an adaptively changing centroid(also called cluster centre), starting from some initial values named seed-points. k-Means computes the distances between the inputs (also called input data points) and centroids, and assigns inputs to the nearest centroid. However, the k-means algorithm is a local search procedure and it is well known that it suffers from two serious drawback, first one is that the number of the clusters is unknown, and the second is initial seed problem [5].

Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair wise similarity or distance. A variety of similarity or distance measures have been proposed and are widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance, Pearson Correlation Coefficient and City Block etc., So an algorithm is devised to enhance k-means in which there are two steps. In the first step dataset is divided into subsets and initial cluster points are calculated. In the second step k-means algorithm is applied to find clusters. Similarity is calculated by using City Block Measures since it is appropriate for log data.

Step I

Let α denote the threshold of similarity between two transactions. The value of α is greater than its actual value, using α to select the initial points and guarantee each cluster at least has one entry to be selected. Choose one transaction from the dataset and select one entry to consider it as centroid. Choose another and compute distance between two. If this distance is smaller than threshold let second session be another cluster. Again choose another entry and compute distance and repeat process.

Input : Dataset D of N log entries and threshold α

Output : Initial points

Algorithm

Select one transaction from dataset
Fix C1 is the center of first cluster
K = 1, $C_{k=} X_1$
For i = 2 to N
$C_m : d(X_i , C_m) = MAX_{i<j<=k} d(X_i,C_j)$
If $d(X_i , C_m) < \alpha$ then
k = k+1 $C_k = X_i$
Else

i= i+1
End if
End for

The above algorithm automatically select initial points based on given threshold α.

Step II:

Assign the initial points as centroids and apply k-means algorithm in which intra cluster comparisons are done. Suppose cj and ck are two centers the similarity between cj and ck is calculated using City Block measure as it is one of the appropriate similarity measure for numerical data. An accurate threshold value is assigned and compare with the resultant similarity value.

$$Sim(c_j,c_k) = \sum_{i=1}^{d} [|cj - ck|]$$

If the value is greater than threshold, then both clusters are same and can be merged as one since while finding initial points by random selection, a number of small clusters are formed. Centers are updated repeatedly until an optimum number of clusters are formed.

IV. Experimental Results

The web log data considered for evaluation is collected from reputed college web server during the period of May to August, 2011. Initially the log file consists of 9464 raw log entries with noisy entries like gif, jpeg etc which are not necessary for web log mining. So data cleaning is performed to remove the unnecessary log which will reduce the processing in determining the web usage pattern. This cleaning phase involves the removal of records with graphics and videos format such as gif, JPEG, etc., and records with robots traversal is also removed. The number of records resulted after cleaning phase is 1476 as shown in the figure.

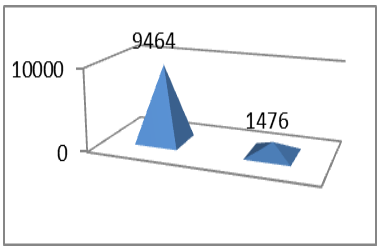

Figure 1:Comparison of initial and Cleaned log

After the data cleaning process is performed, users are identified by using IPaddress and UserAgent fields. There are 124 unique users identified after applying the algorithm and 365 sessions are found whose paths are completed to form transactions. Transactions are given in a user transactions-urls matrix format. A sample isas shown below

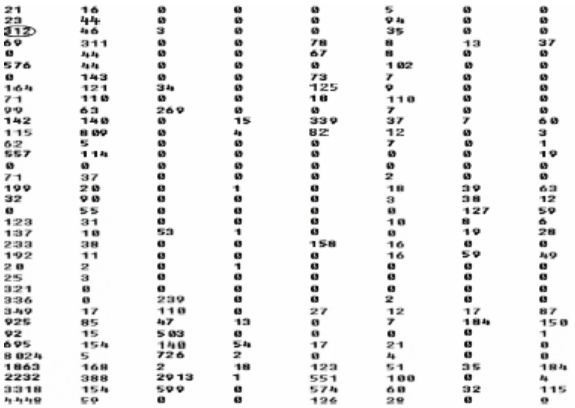

Figure 2 : User Transactions-urls matrix

Clustering technique is applied in the above matrix in 2 steps. Threshold value is taken as initial cluster centroid and the first step is done by comparing eacht row and so on. Initial cluster centroids are found which are used as centroids for second step. In the second step initial points are optimized by using k-means algorithm. Finally well defined clusters with similar intra objects and dissimilar inter objects are obtained. A sample of cluster is shown below.

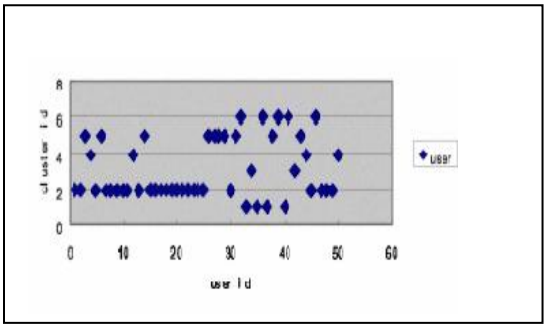

Figure 3 : Clustering result sample

V. Conclusion

The novel method presented, analyzed, and evaluated is to automatically give the actual value of k and select the right initial points based on the datasets objects .The algorithm enhances the k-means clustering algorithm by finding initial points and optimize for accurate results. Our algorithm selecting initial points is more complex than the

random methods, but our algorithm is stable, running it in different times, the clustering results obtained are the same, the random algorithms cannot ensure this, and different initial points lead to different running time on random algorithms, compared with our algorithm, its running time is uncertain and more long.

References

[1]. Banerjee, A. and J. Ghosh (2001). "Clickstream Clustering using Weighted Longest Common Subsequences", in Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, Chicago (2001).

[2]. Cooley, BamshedMobasher, and JaideepSrivastava, " Web mining: Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[3]. Cooley .R, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", Ph.D. Thesis, University of Minnesota, May 2000.

[4]. Houqun Yang, Jingsheng Lei, Fa Fu, "An approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge discovery, IEEE,2007.

[5].Ji He, ManLan, Chew-Lim Tan,Sam-Yuan Sung, Hwee-Boon Low, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on Neural Networks, Budapest,2004

[6].Li Chaofeng, " Research and Development of Data Preprocessing in Web Usage Mining", International Conference on Management Science and Engineering , 2006.

[7]. Mayil, V. V. and Dr.K.Duraiswamy (2008). "Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic Programming." Computer and Information Science, Vol. 1, No. 3, August 2008,

[8]. Mobasher, Cooley R., and Srivastava, J. "Creating Adaptive Web Sites Through Usage-based Clustering of URLs", Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999.

[9]. Nasraoui .O, R. Krishnapuram, "A new evolutionary approach to Web Usage and Context Sensitive Associations Mining", International Journal on Computational Intelligence and Applications-Special Issue on Internet Intelligent Systems, September 2002.

[10].Perkowitz, M., Etzioni, O. "Adaptive Web sites: automatically synthesizing Web pages", Proceedings of Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[11]. Xie.Y, Phoha. V, "Web user clustering from access log using belief function", Proceedings of the ACM K-CAP'OI, First International Conference on Knowledge Capture, Victoria, British Columbia, Canada, (2001).

AUTHORS PROFILE

Mrs. V. Chitraa is a doctoral student in ManonmaniamSundaranar University, Tirunelveli, Tamilnadu. She is working as an Assistant Professor in CMS college of Science and Commerce, Coimbatore. Her research interest lies in Database, Knowledge mining. She has presented 3 papers and published 3 papers in reputed international journals.

Dr. Antony SelvadossThanamani is working as Reader in NGM college, Pollachi with a teaching experience of about 23 years. His research interests includes knowledge management, web mining, networks, mobile computing, telecommunication. He has guided 41 M.Phil scholars, attended 15 conferences, presented 30 papers, published about 8 books and 16 papers.