An Efficient Algorithm for Improved Web Usage Mining

Pooja Sharma
Department Of Computer Science and Engineering
Shree Vaishnav Institute of Technology & Science, Indore
pooja.sharma0804@gmail.com

Asst. Prof. Rupali Bhartiya
Department Of Computer Science and Engineering
Shree Vaishnav Institute of Technology & Science, Indore
rupalibhartiya@rediffmail.com

Abstract:
Clustering is a web mining technique, which is a demanding field of research in which its latent applications create their own special requirements. Clustering is a method of grouping similar data into data sets, called clusters. Cluster analysis is a primary technique in conventional data analysis and many clustering methods have been recognized which requires number of clusters to be precise in advance and is dependent on initial starting points. In this paper, we present a new algorithm to discover data clusters for numerical and nominal data. Apriori algorithm generates large number of candidate set that is not efficient for both data.

Keywords- Web Mining, Clustering, Decision Tree, Apriori algorithm

1. Introduction:
In today's world internet has become very popular and its growth is very fast. The ocean of information is available on internet for people whom they are using for their different propose. The sources for using internet are growing rapidly, it is necessary for users to use automated tools for finding desired information. For finding out the desired information people require systems at client side and server side also. That system used to mine data and extract information from that data. So for a specific user only small information of web is useful and rest of the information not important. Some users interested in content of web that they can browse by using search engines. Information related to web access pattern can be found out by using web log files. These web log files give the information about the behavior of user. In business area, extracting information about user's behavior plays an important role. Web log files are reference data on web pages.

In this paper, we proposed an efficient algorithm for data clustering. This paper aimed to improve the method which is used for accessing web log files. This paper also presents the introduction about web log mining and their classification for mining different data. There is discussion about techniques used in web log mining for extracting information.

Web mining:
We all know that, at the time of accessing internet we deal with lots of data which are not fully useful for us. Mining is the process of excavation for finding out knowledge. Data mining is the process of extracting knowledge from data. Web mining is the process of extracting information and patterns from web. There are three methods [figure1] which are applicable for web mining-

(1) Web content mining
(2) Web structure mining
(3) Web usage mining

Web content mining is the process of extracting knowledge from the content of documents [3]. Web structure mining is the process of reaching at conclusion from referents in the web and web usage mining, which is also known as web log mining, is the process of extracting patterns in web access logs.

Web content mining:
Web content mining is also known as text mining, it is the process that goes beyond keyword extraction. We called it text mining because much of the content is text [4]. On web lots of data is gathered in unstructured format but data mining deals with structured data. Web content mining is also different from text mining in the context of data stored on web is in semi-structured format. So web content mining is related to data mining and text mining and uses its own techniques.
There was rapid expansion in approaches of web content mining in the past few years. This is required because of extraordinary growth of the web contents. Extracting target knowledge through web mining is not fully automated because of lack of structure of web data. In web content mining process, clustering of web pages is made according to structure of data. We can divide strategies of web content mining in two parts-the one that directly mine the content of web and the second that improves the tools which are used for mine content of documents.(search engine)

**Web structure mining:**

Web structure mining is the process of getting knowledge about hyperlinks on the web pages that are link to other web pages. That means web structure mining is useful for knowing the relationship between web pages. This allows users to access the desired information through keyboard association and web content mining. This provides clustering of web pages to establish relationship of these pages. Clustering play a major role to subgroup the web pages according to their structure. The structure of web pages can be seen as a graph which shows pages as nodes and hyperlinks as edges connecting two pages. Web structure mining can be divided into two kinds:

**Hyperlink:**

Hyperlink is the connection between two pages, either within the same page or to a different web page. There are two types of connection possible in hyperlink-one that connects different parts of same page is called an Intra-Document Hyperlink and the other that connects two different pages is called an Inter-Document Hyperlink.

**Document Structure:**

The content of web page can also be organized in a tree structured format; it depends on the HTML and XML within the page.

**Web usage mining:**

Web usage mining is the application of data mining that apply data mining techniques to discover the behavior patterns using web data. Web usage mining process is generally divided into two tasks: Data preparation and pattern discovery. In the task of data preparation server session file build where each session is a sequence of requests of different types made by single user during a single visit to a site. In the pattern discovery task association rules, sequential patterns, usage clusters, page clusters, user classification involve. For finding out the information that is hidden in web logs, several data mining techniques are applied on web server logs. We classify log files in reference of decision tree that is made of root, internal nodes and leaf nodes. There are many decision tree techniques including IDE3, C4.5, CART, SLIQ and SPRINT etc. In decision tree techniques, records are partition using depth first greedy or breadth first approach.

**Problems Description:**

Present research work deals with log files. Web services are available to predict the behavior of users and personalize information to give the access pattern of user. Prediction about access pattern serves for marketing, business and many other different areas.
Web mining provides many algorithms for prediction on the basis of data sets. These algorithms use clustering technique for creating data sets. Apriori algorithm is based on breadth first search to count the candidate data set. From the data sets of length generates (l-1) data sets. After that it removes the unwanted candidates from data sets. Apriori algorithm generates very large candidate data set that affects the execution speed to process data sets [5]. Data coding affects the number and the size of counter vector but it also affects structure of pruned data sets.

There is another method to represent transaction is bit matrix given by Eclat algorithm [6]. In this algorithm each row corresponds to data and each transaction corresponds to column. It works in order of depth first that reaches in boundary for data set. To showing bit matrix there is two ways-one that take bit for each data and transaction and other using for each row a list of those columns in which bit is set. This algorithm generates candidate data set by using only two of its subset. The number of generated data set is much larger than data set generated in Apriori algorithm.

**Proposed algorithm:**

In the previous algorithms data sets generated are very large that affects the time complexity in order to reach at results. To over-come the shortcomings of these algorithms, we propose an algorithm that give efficient result of processing data sets. In network, large number of users connected to the server. Server maintains the log data in sequence order according to time. To identify user in reference of their access pattern of web pages is very difficult, that is why in our propose algorithm we first extract the log data of a specific user and maintain it in sequential order [9]. By doing these we can easily predict personalization individual user.

**Algorithm:**

1. select data set D
2. find list of all attributes in data set
3. check attributes data types
4. if all attributes = numerical data type
   a. get average of each attributes mark as threshold value
   b. compare with all selected attributes
      i. if attribute value <= threshold then
      ii. mark as 0
      iii. else
      iv. mark as 1
5. if all attributes = nominal data type then
   a. find all unique attributes to attributes list
   b. get threshold for each attributes using the given formula
   c. threshold = (total unique values/ total count )logn (total unique values/ total count )
   d. calculate the index of each unique value using the given formula
   e. index=(no of values in list/ total values) * log2(no of values in list/ total values)
   f. Assign label index to the values and compare with threshold
   g. Find distance for all instance
6. end if
7. return classes

**Conclusion:**

In this paper we proposed a clustering algorithm to find out data clusters for both numerical and nominal data by calculating the average and log values of data set. Unlike previous algorithm like apriori algorithm deal with large number of candidate sets, our algorithm deals with small number of candidate sets [10]. Calculating average values is much easier and less time consuming than calculating log values in traditional algorithm. It improves the techniques of Web Usage Mining by first discover the log files of individual users at one place.

**References:**


[7] Qingtian Han, Xiaoyan Gao and Wengue Wu, “Study on Web Mining Algorithm Based on Usage Mining”, Institute of Electrical & Electronics Engineers (IEEE), 2008.

