IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Web Usage Mining using Classification and Clustering technique

Ganga Gudi
Department of Computer Science,
KLE's S.Nijalingappa College, Bangalore, India.

Abstract: *Web usage mining is an application of data mining technique to discover to find the usage patternand serve the requirement of Web-based applications. Analyzing the data through web miningeffectively helps in website mangement, creating website, network traffic flow analysis and so on. The aim of this paper is to find the find user access pattern based on user's session and behaviour.Web usage mining includes three phases pre-processing, pattern discovery and pattern analysis. We propose this technique which cluster the user session based on K-Mediods and DBSCAN technique to find the effective usage pattern.*

Keywords: Web usage mining, clustering, user-session, K-mediods and DBSCAN.

1. 1. INTRODUCTION

Web mining is one of the data mining techniques to automatically discover and extract information from internet documents and services. Web usage mining is helpful in extracting the information from the server logs so as to know the user behavior and fulfill the web user by applying data mining techniques. Web mining is divided into three types, like web usage mining, web content mining and web structure mining.

1.1 Web Content Mining

Web content mining is the process to discover useful information from text, image, audio or video data in the web. It is also called as text mining and the technologies used in this are NLP(Natural Language Processing) and IR(Information Retrieval).

1.2 Web structure mining

Web structure mining operates on the hyperlink structure. It is the process of using the graph theory to analyze the node and connection structure of a web site. This graph structure provides information about rankings enhance the search results of a page.

1.3 Web structure mining

Web usage mining also known as web log mining, aims to discover user access pattern from web browsing data that are stored in web server log, proxy server log or browser log. The usage data records the user behavior when the user browser makes the transaction from the web site.

2. PROPOSED METHODOLOGY

In order to improve the web page prediction accuracy the following system is used.

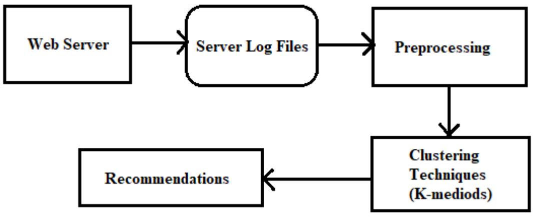

Figure 1 System Architecture

2.1 Web Server Data

When a user agent hits an URL in a domain, the information is recorded in an access log file. The log file on the server side contains log information of user that opened a session. These records contain:

- User's IP address
- Access date and time
- Request method(GET and POST)
- URL of the page accessed
- Transfer Protocol
- Success of return code
- Number of bytes transmitted

2.2 Preprocessing

Web usage data contains information about the internet address of web users.

The cleaning process of web log data is conducted to remove the unwanted data items. This cleaning process is done based on four criteria:

- File extension
- Respond code from web server

- Access methods
- User access frequency

2.3 Clustering

Based on the access of the user the documents are clustered. This kind of clustering makes the search easier. They are clustered by matching the documents with the previous accessed document.

K-mediods

The K-mediods algorithm is a clustering technique related to k-means and mediods shift algorithm. K-means and K-mediods algorithm are partitioned and both attempt to minimize squared error. K-mediods algorithm is a partitioned technique of clustering that clusters the data set of *n* object of *k* clusters. It removes noise and outliers and minimizes a sum of squared Euclidean distances. The common realization of K-mediods clustering is the Partitioning around Mediods (PAM) algorithm and is follows:

a) Initialize: randomly select *k* of *n* data points as the mediods.
b) Assignment: Associate each data point to the closet mediod.
c) Update step: For each mediod *m* and each data point *o*, compute the total cost of the configuration.
d) Select the mediod *o* with the lowest cost of the configuration.
e) Repeat the step b and c until there is no change in the assignment.

Results and Performances

Clustering of web usage data, which is most useful in finding the user access patterns and the order of visits of the hyperlinks of the each user. The suggested approach was used for efficiency contained a hard clustering of the data set and as the analysis indicate each of the individual clusters seems to contain observations with specific common features and improve the algorithm efficiency with help of k-medoids clustering algorithm. Experiments prove that this system has high prediction accuracy by using the appropriate data mining tool orange. As a further improvement, we can still enhance the quality of data by applying two level clustering techniques.

3. Pattern Discovery

Association Rule Mining

Let I=I1, I2,…, Im be a set of m distinct attributes, T be transaction that contains a set of items such that T ⊆ I, D be a database with different transaction records Ts. An association rule is an implication in the form of X -> Y, where X, Y⊆ I, are set items sets. There are two important basic measures for association rules, support(s) and confidence(c).

Support(s) of an association rule is defined as the percentage of records that contains X U Y to the total number of records in the database. The count for each item is increased by one every time the item is encountered in a transaction T in a database D. Support(s) is calculated using the formula:

$$\text{Support}(XY) = \frac{\text{Support count of } XY}{\text{Total number of Transaction in } D}$$

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X U Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence interesting association rules X -> Y generated.

$$\text{Confidence}(XjY) = \frac{\text{Support}(XY)}{\text{Support}(X)}$$

Limitation of Association Rule Mining

One of the major drawbacks of associations rule mining is that too many rules are generated and no guarantee for all generated rules to be relevant. Minimum support and minimum confidence parameters are set in such a way to eliminate false discoveries. When minimum support is too small, every rule will get a chance to be true, leading to wrong recommendation and when minimum support is too large, for small data set, wrong predictions may occur.

Clustering Technique

Cluster is collection of data objects. Objects that are similar are in same cluster. A good clustering method will produce high quality clusters with high inter-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. Similarity is expressed in terms of a distance function, which is typically metric: d(i, j). The commonly used distance measure is the Euclidean distance.

DBSCAN Clustering Algorithm

The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of the database elements. It also determines what information should be classified as noise or outliers. It is density-based clustering algorithm as it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN can categorize these nodes into separate clusters that define the different classes. It also finds the clusters of arbitrary shape. DBSCAN requires two parameters eps(ε) and minimum number of points to form a cluster. It starts with an arbitrary point that has not been visited. This point ε-neighborhood is retrieved and if contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. If a point is found to be dense part of a cluster, its ε-neighborhood is also the part of cluster. Hence all the points that are found are added in ε-neighborhood. This process continues until the density-connected cluster is completely found. Then a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

The proposed approach for web mining usage is a three step process:

a) Collect the web log file and perform preprocessing operation and store it in a database.
b) Here combined approach of clustering and association rule mining is used. DBSCAN algorithm is used to find the user's common behavior and access pattern.

c) In step 3 association rule mining technique will be used to find user's access patterns from this clustered group of data.

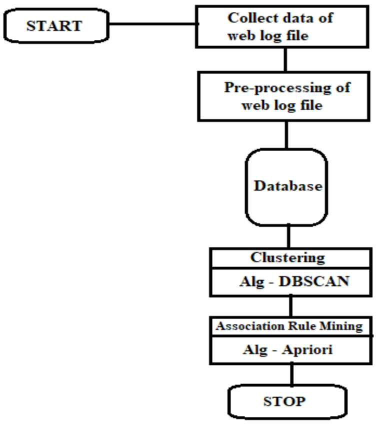

Figure 2 Proposed Approaches for Pattern Discovery in Web Usage Mining

Clustering will be performed and every user will be assigned to a specific cluster according to their behavior and access patterns. Applying association rule mining technique on this clustered data will help us to find results having less computing time and better accuracy.

4. Conclusion

Web usage mining techniques are great area of research. In this association rule mining technique is used on clustered data. Basic association rule mining technique may have drawback of irrelevant rules leading to reduction of accuracy. Minimum support and minimum confidence parameters can be set in such a way that it eliminates false discovery. Clustering frequent access pattern reduce data set for Association Rule Mining and improve result accuracy and produce results of pattern discovery of web usage mining process effectively.

5. References

[1] J. kay, "Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning," IEEE Trans. Learning Technology, vol. 1, no. 4, pp. 215-228, Oct. 2008.

[2] Liu, F., Lu, Z. & Lu, S. (2001), `Mining association rules using clustering', Intelligent Data Analysis (5), 309 - 326.

[3] M. Salehi, M. Pourzaferani, and S.A. Razavi, "Hybrid Attribute-Based Recommender System for Learning Material Using Genetic Algorithm and a Multidimensional Information Model," Egyptian Informatics J., vol. 14, no. 1, pp. 67-78, 2013.

[4] Lai, H. & Yang, T. C. (2000), "A group-based inference approach to customized marketing on the web integrating clustering and association rules techniques" Hawaii International Conference on system sciences pp. 37 – 46.

[5] [Cooley 99-2] R. Cooley, P. Tan and J. Srivastava (1999), Discovery of interesting usage patterns from Web data. Advances in Web Usage Analysis and User Profiling.

[6] Cooley, BamshedMobasher, and JaideepSrivastava, "Web mining: Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence.

[7] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In Proe. ACM SIGCOMM, pages 241-253.

[8] Bernardo Huberman, Peter Pirolli, James Pitkow, and Rajan Kukose. Strong regularities in world wide web surfing.

[9] R.Agrawal and R.Srikant Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, pages 487-499.

[10] B.Mobasher, N.Jaln, E. Hart, and J. Srivastava. Web mining: Pattern discovery from World Wide Web transactions.