Analysis of WebLog Usage Mining For Intelligent Recommendation

Priyanka D.Doltade
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Sarika K. Bodage
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Ajinkya V.Bochare
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Rahul A. Patil
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

***Abstract*— Nowadays weblog usage is tremendously increased. There is need of storing the log for further requirements and analysis. This huge amount of weblog is stored into Web Server Log. Web server log keeps the records of web usage of web users. All this information which is saved in web log will be used for different purposes. For example, different preprocessing techniques can be applied on recorded weblog; from this we can get useful patterns and analysis of web usage in a particular area or location. In this paper we are doing the analysis of web usage of users based on the location, age along with other parameters which helps in improving the quality of web services in terms of providing faster web browsing to clients and based on their frequent usage of a particular website, ISP will provide buffering to those particular websites only.**

***Keywords*— Weblog, Web Mining, K-Means Algorithm, Apriori Algorithm, Clustering, Association Rules, ISP (Internet Service Provider).**

I. Introduction

Internet usage is increased day by day and every individual needs a faster response. It is the responsibility of ISP to provide buffering to frequent websites which are used in particular location or area, from that clients can get faster response for their browsing. In actual the ISP doesn’t know which website should be given high buffering. So, this paper is concerned with usage of websites using web mining process. The dataset gathering being used is based on location and age wise collection of user interaction record.

Weblog is a type of repository (dataset) which stores a large amount of user’s data with respect to web usage like URL, request time, response time, finish time, IP address, age, location etc. Basic introduction of web mining and web usage mining is given by [2, 3 and 4].

By collecting these datasets, we are applying clustering and association mining techniques for the improving quality of web services. On this dataset we will cluster the related data in particular location and on that clusters we are applying association rules in order to obtain graphical analysis. Depending on the confidence of the result set obtained via association rules, we can determine which website is more frequent in a particular location or area. We can represent these generated rules graphically and these results will be provided to the ISP for buffering purpose.

II. Overview of web usage pattern analysis

A. Need of web usage mining:

In field of information technology we have huge amount of data available that need to be turned into useful information. This information can be used for various applications such as market analysis, fraud detection, improving quality of service etc.

Web usage mining mainly consist following phases:

1) Web usage data collection:
Web usage log store the activities of user from web sites. These activities can collect from web server or web browser

2) Data preparation:
The data collection from the logs may be partial, deafening and conflicting so, the objective of preprocessing is to transfer raw log files in particular format which data mining algorithm can handle easily. The main tasks of preprocessing is data cleaning, user identification, session identification, path completion, data integration and formatting.

3) Pattern discovery:
Statistical method as well as data mining methods are applied inorder to detect interesting results.

4) Pattern analysis:
Extracted patterns are analyzed through OLAP tools, knowledge management query techniques and intelligent agents to sort out the patterns or rules.

III. Proposed Methodology

In this section we present an overall description of how we can develop the solution. Our solution consists of following main modules at client and server side.

A. Chrome extension:

For storing browsing history to local machine.

B. Servlets(local):

Called by the chrome extension for storing and accessing browsing data in database.

C. Client application:

Where in the client can perform the following tasks:

1) View their own browsing data.
2) Filter the data and upload to server
3) View server log
4) Fetch calculated average time for accessing a particular *website*

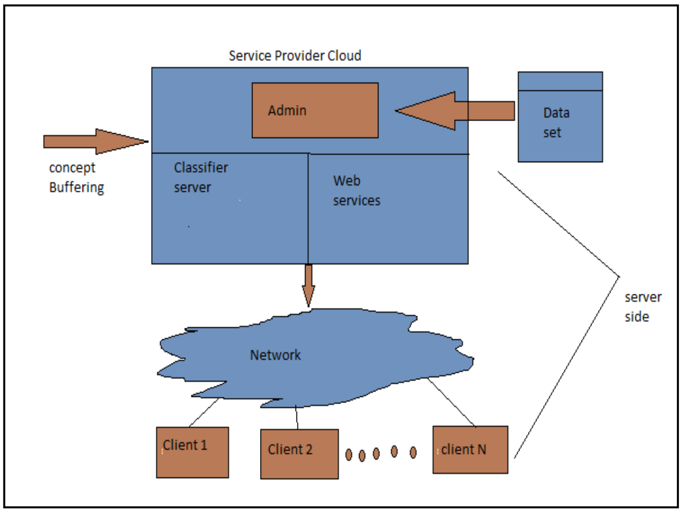

Fig. 1. Server side architecture

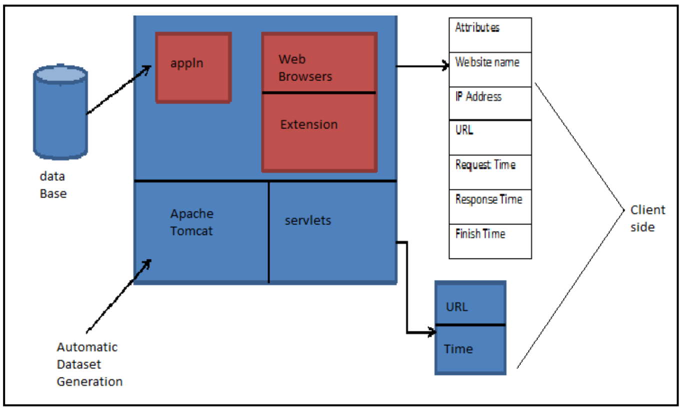

Fig. 2. Client side architecture

D. Server database:

Where the browsing data collected from the client is to be stored.

E. Admin application:

Where the admin/ISP can apply data mining and view usage patterns graphically.

IV. ALGORITHMIC STRATEGY

The algorithmic strategies used in the paper as follows:

A. *Why* K-Means?

- K-means is one of the simplest unsupervised learning algorithms that partition feature vectors into k clusters so that the within group sum of squares is minimized.
- Mean Shift clustering is able to produce clusters with shapes that depend upon the topology of the data and does not need an a priori estimate of the number of clusters to find.
- K-means on the other-hand assumes the isotropy of the clusters and need to be taught the number of clusters to extract in advance.

B. Why Apriori?

- Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules.
- Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending then to larger and larger item sets as long as those items sets appear sufficiently often in the database. The frequent item set determine by Apriori can be used to determine association rules which highlight general trends in the database.[2]

V. IMPLEMENTATION PLANNING

Figure 1,2 shows the backend architecture or rough idea about the implementation of the web log usage mining for intelligent recommendation. In which we are going to used k-means and the Apriori algorithm. Web mining is an appliance of data mining techniques to large web log data repositories [3].

Working of algorithm is described as below:

A. K-Means:-

1. K= centroid
2. Make initial guesses for the means m1, m2,..., mk Until there are no changes in any mean .
3. Use the estimated means to classify the samples into clusters for I from 1 to k.
4. Replace mi with the mean of all of the samples for cluster i end for or end until.

B. Example of k-means in our implementation

- Set of age is, Age=1, 2, 6, 4, 2, 3 ………n.
- K is the centroid value which need in the cluster formation which is first predefined in the algorithm. Take k=3 (small, middle, upper) ages. set is applied 7, 1 2, 3, 9, 10, 5, 12, 13, 16, 5, 4, 11.
- Apply k-means.
- The k values will be taken randomly. In the k-means depend on the k values the sorting is done and on that basis the values gets sorted.

The following figure shows the sorting and the cluster.

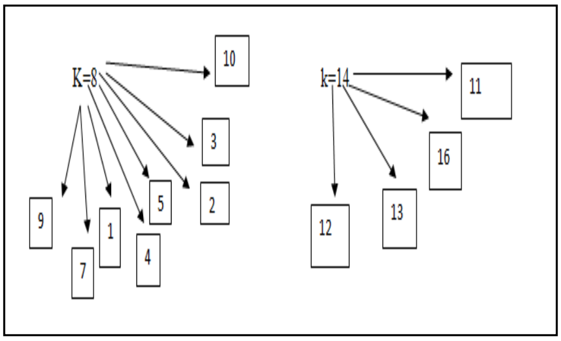

Fig. 3. Clustering with k values

- From this the avg. is calculated that is as below
  For k=8 Avg. =9+7+1+4+5+2+3+10/8
  For k=14 Avg. =12+13+16+11/4.
- In the next step k values gets changed till the same k values are not get in algorithm.
- At last the step comes in which the k values are remain same and the final we get sorted clusters.
- Then stops the iteration.

C. Association:

1.The output obtained by clustering will be provided as an input for the association.
2. For association in apriori, calculatete support using following formula:
Supp(A ->B)= (A U B) / Total no. of tuple in A, B.
3. For association in apriori calculate confidence using following formula:
Conf(A->B)= (A U B) / Total no. of tuple in A.

D. Example of Apriori in our implementation:

- The input for the association is based on the output of the clustering.
- In the Apriori algorithm we calculate the support and confidence as mentioned above formula.

TABLE I. SAMPLE DATASET FOR ASSOCIATION

	Age wise partition		
Transaction Id	***small***	***middle***	***upper***
1	1	1	0
2	0	1	1
3	0	0	0
4	1	1	1
5	0	1	1
6	1	1	1
7	1	1	1

- From the above table calculations, Support of age that is small, middle and upper will be,
- Let x=small, middle ,upper
- Supp(x)= no.of transaction which contains the itemset x / Total no. of transaction.

Supp(x)=3/7 =0.7

- Conf(x->y) = supp(x U y) / Supp(x)

Let x= m ,s ,u
Where,
m=middle age
s=small age
u=upper age
Y= m ,s
Supp(x)= 3/7=0.42
Supp(y)= 4/7=0.57
Conf(x->y)=0.42/0.57=0.73
That means for 73% of the transaction containing small, middle and upper ages who search for the same websites. This rule is accepted in association

E. Output:

From the above rules which are valid from the association and are finally accepted to provide a graphical representation of the analysis of web usage.

VI. Conclusion

We have examined that each and every type of data is not useful, thus examining the most relevant and useful information in the weblog data may provide more specific information about the patterns for visitors of the web site. This is the identifying information that the client browser reports about itself. The extracted knowledge of user's navigational behavior from web log file, may be used to answer different queries like efficiency of web site in delivering information, users view point about website structure, prediction of users next visit, fulfillment of needs of different users, user satisfaction and many more such type of information to facilitate a web administrator in taking a decision.

References

[1] Dilip Singh Sisodia, Shrish Verma," Web Usage Pattern Analysis Through Web Logs" ,IEEE,2012.

[2] Cooley, R., Mobasher, B., and Srivastava, J, "Web mining: information and pattern discovery on the World Wide Web", International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997, pp. 558-567.

[3] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava," Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information System,1999,pp. 1-27.

[4] Robert Cooley, Bam shad Mobasher, and Jaideep Srivastava." Grouping Web page references into transactions for mining World Wide Web browsing patterns", Knowledge and Data Engineering Workshop, New port Beach, CA.IEEE, 1997, pp.2-9.