Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Scaled Approach of Web Usage Mining and Clustering to Perk up Website Design

Dr. Abha Choubey*, Reena Dewangan

CSE & CSVTU, Bhilai,
Chhattisgarh, India

Abstract—Web usage mining is the giant field that helps to comprehend range of concepts of diverse fields. It consists of three main phases, namely Data Pre-processing, Pattern Discovering and Pattern Analysis. Server log files become a set of raw data where it's must go through with all the Web Usage Mining phases to producing the final results. The growing reputation of e-commerce websites makes data mining requisite technology for several applications, especially online business competitiveness. In this paper web usage mining combined with association rule mining and clustering algorithm to optimize the content of the web log data to optimize the content of the web log data. Finally, this paper will provide association Rule from web log for user defined cluster which are useful for improvement of website design

Keywords— Business Intelligence, CRM, e-Commerce, e-publishing, Web mining, Web usage mining.

I. INTRODUCTION

As in classical data mining, the aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. There has been huge interest towards web mining. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web documents data, web structure data, web log data, and user profiles data. Two different approaches are proposed on the definition of web mining. One approach is process-based and the other is data-based. Data-based definition is more widely accepted today. In this perspective, web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. There are no differences between web mining and data mining compared in general. All of web data can be mined mainly in three different dimensions, which are web content mining, web structure mining, and web usage mining.

There are several reasons for the emergence of web mining [10]. First of all the World Wide Web is huge and effective source for data mining and data ware housing. The size of the web is very large on the orders of terabytes and it still growing rapidly. Many organizations, individuals or societies provide their public information through web. Also, the content of the web pages are much more complex than any other traditional text documents. Today, web pages lack standard structure; they contain more complex style than standardized formats. Web Mining can be broadly divided into three categories as shown in fig 1 according to the kinds of data to be mined:-

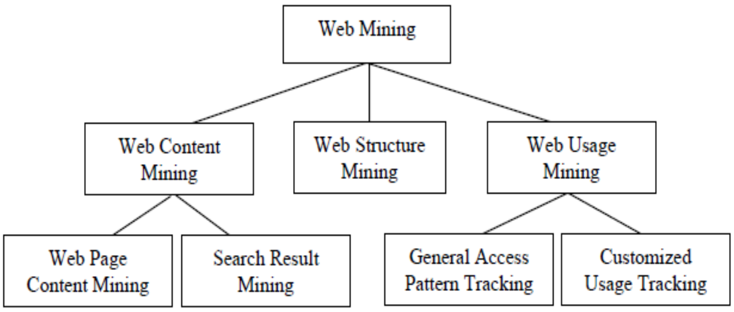


FIGURE 1. TAXONOMY OF WEB MINING

WWW can be accepted as a huge digital library. By the same analogy, Web Mining can be viewed as a digital library's librarian. There are several application areas of web mining; the important ones are listed in Figure-2

The most popular application area of web mining is e-commerce (business-to-customer) and web based customer relationship management. Web usage mining is most dominant application in this context. With the web mining, it is possible to record customer behaviour for web-based business. It is also feasible to adapt web sites based on interesting patterns as a result of analysis on user navigation patterns [12]. Web site topology can be customized to provide better facilities for the site user [10].

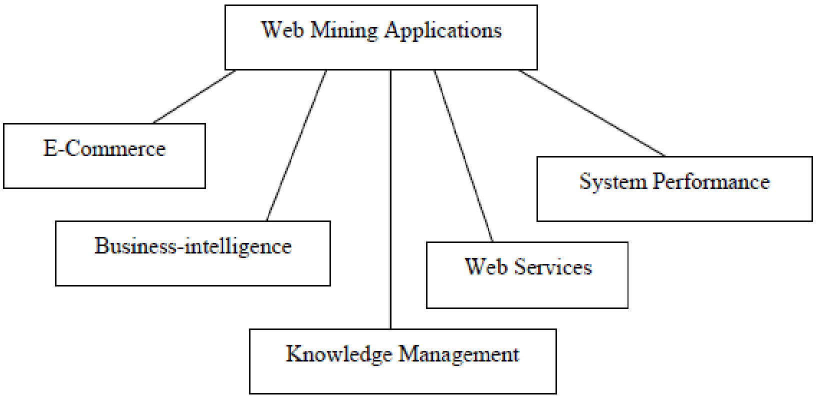


FIGURE 2. APPLICATION AREAS OF WEB MINING

A lot of previous work has focused on Web data clustering (e.g. [14,15]).Web data clustering is the process of grouping Web data into "clusters" so that similar objects are in the same class and dissimilar objects are in different classes. Its goal is to organize data circulated over the Web into groups / collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding users' navigation behavior, improving information retrieval and content delivery on the Web. The aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. A way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. Proposed algorithm is used to generate association rules that associate the usage pattern of the clients for an e-commerce website. In the proposed work we have combined the association mining with the clustering instead of mining association rules from the web log data directly we have mined the clusters. The goal of clustering is to organize data circulated over the Web into groups / collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding users' navigation behaviour, improving information retrieval and content delivery on the Web.

The remainder of this paper is organized as follows: Section 2 provides a brief review of the related work. In Section 3, we explain problems in existing systems. In Section 4, we introduce our proposed algorithm and an illustration of the algorithm. Finally, we concluded our work

II. RELATED WORK

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between group of users with specific interest. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together.

Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For example [5] has introduced a web usage mining intelligent system to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service which enables direct access of website functionalities to the third party.

Santosh kumar et. al. [6] concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of memory usage and time usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm.

Patel et al [7] discusses the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented Web Usage Mining applications and some Web Mining software

In paper [8] provides an introduction to the field of Web mining and examines existing as well as potential Web mining applications applicable for different business function, like marketing, human resources, and fiscal administration. Suggestions for improving information technology infrastructure are made, which can help businesses interested in Web mining hit the ground running.

In [9] an overview of the web mining concept has been presented and how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. The contribution of this paper is towards the various areas containing web sites on internet, which can make best use of different web mining techniques to improve their business decisions based on the user behavior analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth.

Kharwar et al [11] implements the high level process of Web Usage Mining using basic Association Rules algorithm call Apriori Algorithm. Web Usage Mining consists of three main phases, namely Data Pre-processing, Pattern Discovering and Pattern Analysis. Server log files become a set of raw data where it's must go through with all the Web Usage Mining phases to producing the final results. Here, Web Usage Mining, approach has been combining with the basic Association Rules, Apriori Algorithm to optimize the content of the serve log data. Finally, this paper will present a finding association Rule from server log which are useful in many application like cache for web page, Marketing, Targeted Advertising etc.

III. PROBLEM IDENTIFICATION

The explosive growth of the World Wide Web (WWW) in recent years has turned the web into the largest source of available online data.

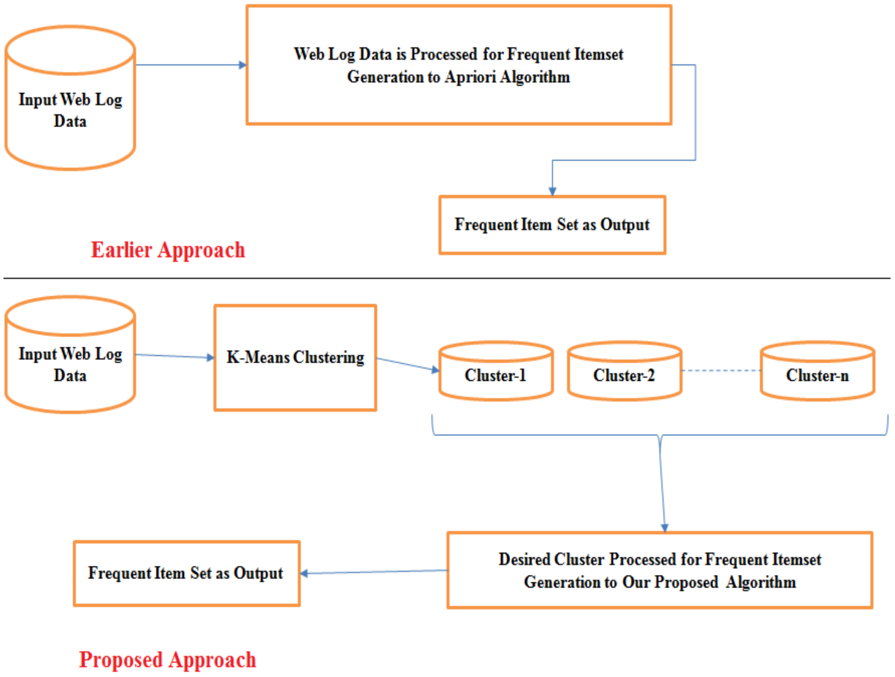
- Situations like several unrelated topics in a single web page may lead to confusion and make harder to reach the information that the visitors are looking for.
- The design of the whole site (interface, content, structure, usability, etc.) is one of the most important aspects for any institution that wants to survive in the cyberspace.
- Understand the way user browses the site and find out which is the most frequent used link and pattern of using the features available in the site.

All these information is available online but are hidden for the users. Presently, there is no powerful that can analyze this hidden information and this Research work uses web usage mining (WUM) Apriori based approach for analyzing the visitor browsing behavior.

Apriori algorithm, in spite of being simple, has some limitation. They are,

1. It is costly to handle a huge number of candidate sets. For example, if there are 104 frequent 1-item sets, the Apriori algorithm will need to generate more than 107 length-2 candidates and accumulate and test their occurrence frequencies. Moreover, to discover a frequent pattern of size 100, such as {a1, . . . , a100}, it must generate 2100 - 2 ˜ 1030 candidates in total. This is the inherent cost of candidate generation, no matter what implementation technique is applied.
2. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns.

IV. COMPARISION


Input Web Log Data
Web Log Data is Processed for Frequent Itemset Generation to Apriori Algorithm
Frequent Item Set as Output
Earlier Approach
Input Web Log Data
K-Means Clustering
Cluster-1
Cluster-2
Cluster-n
Desired Cluster Processed for Frequent Itemset Generation to Our Proposed Algorithm
Frequent Item Set as Output
Proposed Approach

V. PROPOSED SYSTEM

The main goal of the proposed system is to identify usage pattern from web log files of a website. collections of items bought by customers, or details of a website frequentation).In this paper we proposed a new algorithm which combines the concept of association mining and Clustering instead of mining association rules from the web log data directly we have mined the clusters selected by user. Figure 3 represents proposed approach.

Algorithm Description

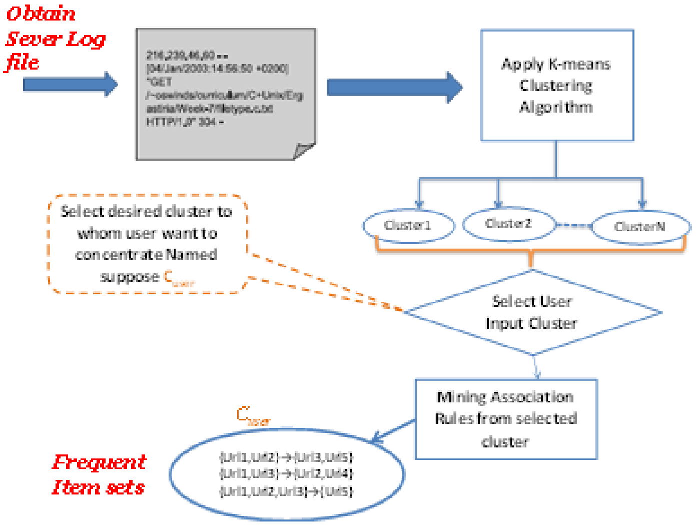
Input: A web log database

The Minimum-Support threshold

Output: Frequent item sets

Method:

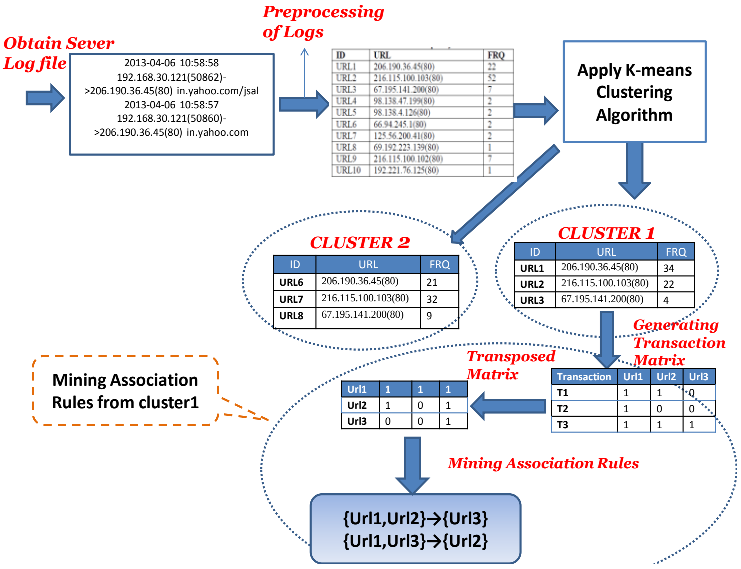
1) Scan the database D and partition the transaction table into clusters using K-means algorithm. Apply the method from step 2 to 6 on user selected cluster.
2) The set of frequent 1 item sets say L1, can then be determined. It consists of candidate 1 item Sets which satisfy minimum support
3) To discover the set of frequent 2- item sets
4) The algorithm iterates to find upto n- frequent item sets
5) From user selected cluster find out the n-frequent item sets.


Obtain Sever Log file
Apply K-means Clustering Algorithm
Cluster1
Cluster2
ClusterN
Select desired cluster to whom user want to concentrate Named suppose C_{user}
Select User Input Cluster
Mining Association Rules from selected cluster
C_{user}
Frequent Item sets

```
216.239.46.60 - -
[04/Jan/2003:14:56:50 +0200]
"GET
/~oswinds/curriculum/C+Unix/Eng
astria/Week-7/filetype.c.txt
HTTP/1.0" 304 -
```

```
{Url1,Url2}→{Url3,Url5}
{Url1,Url3}→{Url2,Url4}
{Url1,Url2,Url3}→{Url5}
```

Figure 3: Proposed Approach


Obtain Sever Log file
Preprocessing of Logs
Apply K-means Clustering Algorithm
CLUSTER 2
CLUSTER 1
Generating Transaction Matrix
Transposed Matrix
Mining Association Rules from cluster1
Mining Association Rules

```
2013-04-06 10:58:58
192.168.30.121(50862)-
>206.190.36.45(80) in.yahoo.com/jsal
2013-04-06 10:58:57
192.168.30.121(50860)-
>206.190.36.45(80) in.yahoo.com
```

ID	URL	FRQ
URL1	206.190.36.45(80)	22
URL2	216.115.100.103(80)	52
URL3	67.195.141.200(80)	7
URL4	98.138.47.199(80)	2
URL5	98.138.4.126(80)	2
URL6	66.94.245.1(80)	2
URL7	125.56.200.41(80)	2
URL8	69.192.223.139(80)	1
URL9	216.115.100.102(80)	7
URL10	192.221.76.125(80)	1

ID	URL	FRQ
URL6	206.190.36.45(80)	21
URL7	216.115.100.103(80)	32
URL8	67.195.141.200(80)	9

ID	URL	FRQ
URL1	206.190.36.45(80)	34
URL2	216.115.100.103(80)	22
URL3	67.195.141.200(80)	4

Transaction	Url1	Url2	Url3
T1	1	1	0
T2	1	0	0
T3	1	1	1

Url1	1	1	1
Url2	1	0	1
Url3	0	0	1

{Url1,Url2}→{Url3}
{Url1,Url3}→{Url2}

Figure 4: Example illustration

A. 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 partitions the input dataset into k clusters. First select k initial centers based on desired number of clusters. The user can specify k parameter value. Each data point is assigned to nearest centroid and the set of points assigned to the centroid is called a cluster. Each cluster centroid is updated based on the points assigned to the cluster. The process will be repeated until the centroids remain the same or no point changes clusters. In this algorithm mostly Euclidean distance is used to find distance between data points and centroid.

Algorithm: The k -means clustering algorithm

```
Input:   D:{d1,d2....dn}\\set of n items
         K //Number of desired clusters
Output: A set of k-clusters.
Steps:
1. Arbitrarily choose k-data items from
   D as initial centroids;
2. Repeat assigns each item di to the cluster which has the closest centroid,
         Calculate new mean for each cluster;
   until convergence criteria are met.
```

Association Rule Mining

Given a server log files that represent user activities, the main purpose of Association Rules is to generate all Association Rules that have support and confidence greater than the user specified minimum support (called min_sup) and minimum confidence (called min_conf) respectively. An algorithm for finding all Association Rules, henceforth, referred to as the Apriori algorithm[2].

In Apriori algorithm, discovery of association rules require repeated passes over the entire database to determine the commonly occurring set of data items. Therefore, if the size of disk and database is large, then the rate of input/output (I/O) overhead to scan the entire database may be very high. We have proposed a new Algorithm, which improves the Apriori algorithm for repeated scanning of large databases for frequent itemsets generation. In our algorithm, transaction dataset will be used in the transposed form and the description of proposed algorithm is discussed in the following sub-sections

Procedure Gen_candidate_itemsets (Lk-1)

```
    Ck = $\Phi$
     for all itemsets $I_1 \in L_{k-1}$ do
     for all itemsets $l_2 \in L_{k-1}$ do
     if $I_1[1] = I_2[1] \wedge I_1[2] = I_2[2] \wedge \ldots \wedge I_1[k-1] < I_2[k-1]$ then
               c = $I_1[1], I_1[2] \ldots I_1[k-1], I_2[k-1]$
               Ck = $C_k \cup \{c\}$
End Procedure
```

Procedure Prune (Ck)

```
for all c $\in$ Ck
        for all (k-1)-subsets d of c do
           if d $\notin$ Lk-1
     then Ck = Ck – {c}
End Procedure
```

Algorithm: Association Rule Mining for each cluster

```
1. Read the database to count the support of C1 to determine L1 using sum of    rows.
2. $L_1$= Frequent 1- itemsets and k: = 2
3. While (k-1 $\neq$ NULL set) do
      Begin
      $C_k$: = Call Gen_candidate_itemsets ($L_k$-1)
      Call Prune ($C_k$)
      for all itemsets i $\in$ I do
            Calculate the support values using dot-multiplication of array;
            $L_k$ : = All candidates in Ck with a minimum support;
            k:=k+1
      End
4. End of step-3
End Procedure
```

VI. CONCLUSION

Massification of the use the internet has made automatic knowledge extraction from Web log files a necessity. Information provided are interested in techniques that could learn Web users' information needs and preferences. This can improve the effectiveness of their Web sites by adapting the information structure of the sites to the users' behavior.

The aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. A way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. Proposed algorithm is used to generate association rules that associate the usage pattern of the clients for an e-commerce website. In the proposed work we have combined the association mining with the clustering instead of mining association rules from the web log data directly we have mined the clusters. The goal of clustering is to organize data circulated over the Web into groups / collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding users navigation behavior, improving information retrieval and content delivery on the Web.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. N. Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD, International Conference on Management of Data, pp.207- 216, 1993.

[2] Agrawal. R., and Srikant. R., Fast Algorithms for Mining Association Rules, Proceedings of 20th International Conference of Very Large Data Bases. pp.487-499,1994.

[3] Kosala and Blockeel, "Web mining research: A survey," SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000

[4] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in PADKK '00: Proceedings of the 4th PacificAsia Conference on Knowledge Discovery and Data Mining,

[5] B.Naveena Devia, Y.Rama Devib, B.Padmaja Ranic, R.Rajeshwar Raod, Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce International Conference on Communication Technology and System Design 2011

[6] B.Santhosh Kumar,K.V.Rukmani, Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms Int. J. of Advanced Networking and Applications Volume:01, Issue:06, Pages: 400-404 (2010)

[7] Ketul B. Patel,Dr. A.R. Patel, Process of Web Usage Mining to find Interesting Patterns from Web Usage Data International Journal of Computers & Technology www.ijctonline.com ISSN: 2277-3061 Volume 3, No. 1, AUG, 2012

[8] Rahi, Priyanka. "Business Intelligence: A Rapidly Growing Option through Web Mining." arXiv preprint arXiv:1208.5875 (2012).

[9] Pradnya Purandare, WEB MINING: A KEY TO IMPROVE BUSINESS ON WEB IADIS European Conference Data Mining 2008.

[10] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques

[11] Patel, Premal. "Implementing APRIORI Algorithm on Web serve log."