International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Implementation Web Usage Mining Using D-Apriori

Mrs.S.Geethamani[1] Ms.S.Pradeepa[2]

[1]*Assistant Professor, Department of Computer Science,*
Sri Ramakrishna college of Arts & Science for Women, Coimbatore (India)
[2]*Assistant Professor, Department of Computer Science,*
Nirmala College for Women, Coimbatore (India)

ABSTRACT

Web usage mining is the branch of web mining. The data is assembled has result in awfully large information in web. The data is grouped according to the number of visitors in the search engine. The data is grouped using the Divisive clustering method. The Divisive Analysis is one of the types of Hierarchical method. It is used to separate each datasets from the clustered data sets. The Divisive clustering method is used in Apriori and FP Growth algorithm to proposed new algorithm called D-Apriori and DFP.

Keywords: Apriori, FP Growth, Clustering, D-Apriori, DFP Algorithm and Web Usage Mining.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Web mining is the process of pattern extraction from the web. The web mining is divided into three categories which are web structure mining, web usage mining and web contents mining. Web usage mining integrates the techniques of two popular research fields.

Web usage mining is the discovery and analysis of user access patterns from log files and associated data from a particular website. The web usage mining has three steps to process. There are Data Preprocessing, Pattern Discovery and Pattern Analysis. In Data Preprocessing the unwanted data is removed. In Pattern Discovery the frequent occurring itemsets are mined. In Pattern analysis the most occurring frequent itemsets are mined.

This work proposed an iterative method which uses Divisive based clustering is applied in Apriori that mine frequent clustered web usage data from weblog files. The system aims at developing an effective technique for finding frequent data using iterative association rule mining algorithms with divisive clustering. In recent years there has been an increasing interest of work in Web usage mining. So Improving and utilizing web usage mining techniques such as clustering which rely on pattern discovery from user transactions to be aimed to improve the scalability of collaborative rule verification.

II. PROPOSED METHODOLOGY

1) ***PREPROCESSING***

1) ***Hierarchical clustering:***

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Cluster Analysis, also called data segmentation, has a variety of goals. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. There are two basic approaches to generating a hierarchical clustering:

1) Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

2)Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

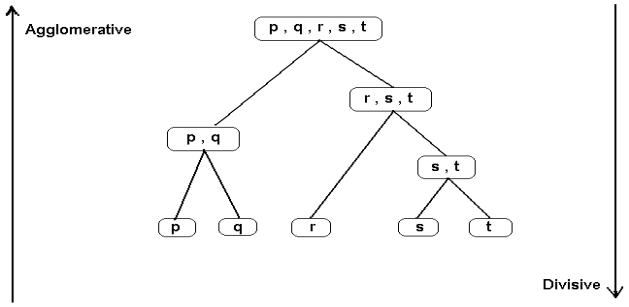


Fig 1: Hierarchical Clustering

The following diagram shows the structure of divisive technique. This applies the top down approach which iteratively performs until every object in single.

Steps included in divisive clustering

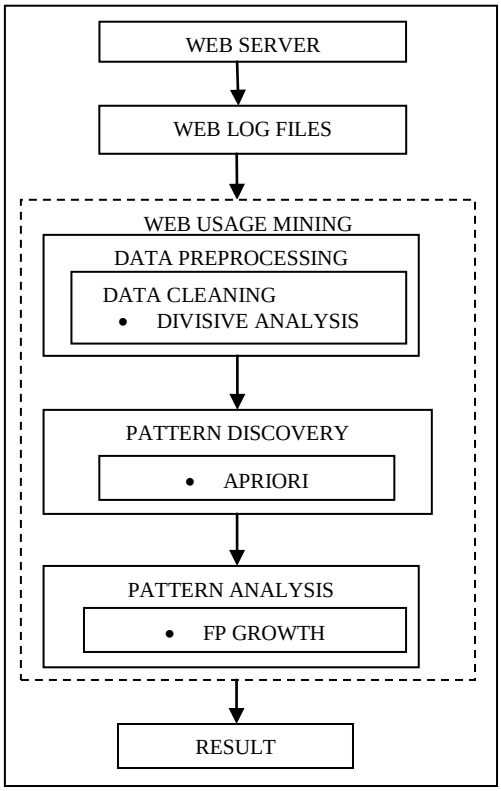
1. Put all objects in one cluster
2. Repeat until all clusters are singletons
   a) Choose a cluster to split with a criterion
   b) Replace the chosen cluster with the sub-clusters and decide
   • selects number of clusters
   • Criterion to split
   • "reversing" agglomerative => split in two
   • cutting operation: cut-based measures seem to be a natural choice.
   – Focus on similarity across cut
3. End

The process of preprocessing in the proposed approach also having the divisive technique, this starts at the top with all documents (weblog) in one cluster. The cluster is split using a divisive clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster. Top-down clustering is conceptually more complex than bottom-up clustering since this need a second, divisive clustering algorithm as an iterative. There is evidence that divisive algorithms produce more accurate hierarchies than bottom-up algorithms in some circumstances.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

2) PATTERN DISCOVERY

The dataset is get from the web server through web log files. First the Data Preprocessing, in this process the itemset is cleaned and clustered frequent itemsets and cluster using Divisive Analysis. In Patter Discovery, the frequent itemsets are mined using D-Apriori algorithm.



OVERLALL ALGORITHM STEP
Step 1: Read all web log files (Wf) from web server Ws.
Step2: Combine all the weblogs and perform step 3
Step3: Preprocess by cleansing unwanted transactions
Step4: Apply divisive clustering technique D (Wf)=∑(I to n) {Trans $_i$ (split(divisive pattern p)}
Step5: Perform pattern discovery p
Step 6: Apply apriori algorithm for every discovery of pattern p

D-Apriori algorithm

D-Apriori Algorithm: The algorithm can be used to generate all frequent item sets. J- Frequent itemsets.

Pass 1

1. Find the object, which has the highest average dissimilarity to all other object. This object initiates a new cluster – a sort of a splinter group.
2. For each object i outside the splinter group compute.
3. D_i = [average d(i , j) j $R_{splinter\ group}$] – [average d(i , j)j $R_{splinter\ group}$]
4. Find an object h for which the difference D_h is the largest. If D_h is positive, then h is on the average close to the splinter group.
5. Repeat step 2 and 3 until all difference D_h are negative. The data set is then split into clusters.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

6. Select the cluster with the largest diameter. The diameter of a cluster is the largest dissimilarity between any two of its objects. Then divide this cluster, following steps 1 to 4.
7. Repeat step 5 until all clusters contain only a single object.
8. Generate the candidate itemsets in K_1
9. Save the frequent itemsets in S_1

Pass j

1. Generate the candidate itemsets in Kj from the frequent itemsets in S_{j-1}
   i) Join S_{j-1} p with S_{j-1q} as follows **insert** into K_j **select** $p.item_1, q.item_1, \ldots p.item_{k-1}, q.item_{k-1}$ **from** S_{j-1} p,S_{j-1q} where $p.item_1 = q.item_1, \ldots p.item_{j-2} = q.item_{j-2}$, $p.itemj_{-1} < q.itemj_{-1}$
   ii) Generate all (k-1)-subsets from the candidate itemsets in K_j
   iii) Prune all candidate itemsets from K_k where some (k-1)-subset of the candidate item set is not in the frequent itemset S_{j-1}
2. Scan the transaction database to determine the support for each candidate itemset in K_j
3. Save the frequent itemsets in S_j.

3) PATTERN ANALYSIS

In Pattern analysis the uninteresting patterns are removed from the patterns identified during pattern discovery phase. In this phase, the FP is applied on the cluster datasets to extract the most frequent accesses web page of the user.

Algorithm: FP–Growth Algorithm
Input: filtered transaction from Apriori
Output: Frequent item set
Description: FP-Growth: Allows frequent itemset discovery without candidate itemset generation. Two step approach:
Steps:
(i) Build a compact data structure called the FP-tree built using 2 passes over the data-set.
(ii) Extracts frequent itemsets directly from the FP-tree traversal through FP-Tree.

II. RESULT

The sample dataset are taken from URI dataset which the proposed implementation has been used. The system used 25668 entries. The proposed system applies the D-Apriori algorithm for effective frequent item set mining.

T1	{ABDE}
T2	{ABECD}
T3	{ABEC}
T4	{BEBAC}
T5	{DABEC}

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table 1: Sample Web transactions involving page views A, B, C, D, and E

Size 1	Size 2	Size 3	Size 4
{A}(5)	{A, B}(5)	{A, B, C}(4)	{A, B, C, E}(4)
{B}(6)	{A, C}(4)	{A, B, E}(5)	
{C}(4)	{A, E}(5)	{A, C, E}(4)	
{E}(5)	{B,C}(4)	{B, C, E}(4)	
{B,E}(5)			
{C, E}(4)			

Table 2: Frequent Itemsets generated by the Apriori algorithm

The processing time reported includes the CPU time consumed in the preprocessing steps (after frequent items have been selected), the template generation, and the complete process for calculating frequent items. The I/O time spent on the index construction and the database modification is excluded in order to highlight the impact of the database scale on this indexing mechanisms and the proposed method for frequent pattern mining. The proposed experiments uses 25688 entries.

Documents(thousands)	Existing	DApriori	DFP
1	**1.30**	0.5	**0.3**
3	**2.03**	1.02	**1.00**
5	**6.0**	4.5	**3.9**
10	**9.0**	6.5	**5.4**

Table 3:Scalability on the database size

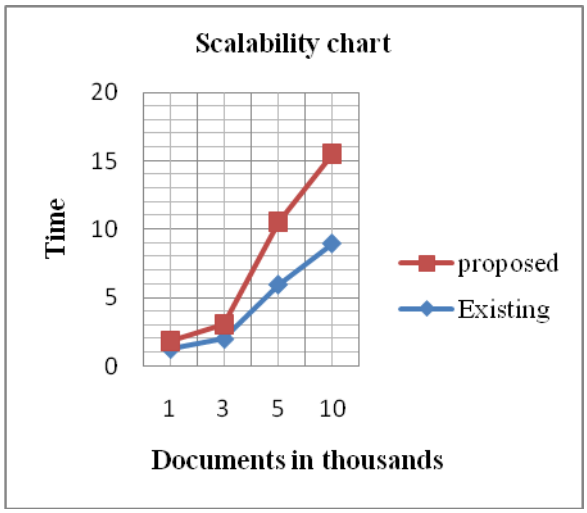


Fig 3. Scalability on the database size

The results of the CPU time under varied database sizes are plotted in Fig.3. Each of the variations is scalable in terms of the database size. In this implementation, the table keeping the template information can be fully loaded into the main memory. The indexing mechanisms and the prime-number representation are the major reasons for the good scalability of the approach. The former supports fast data access with the available association techniques. Moreover, the frequent items are mapped to the prime numbers in a reverse order of their frequencies. In this way, the product of prime numbers for representing a frequent itemset will not be too large. In addition, the results plotted in Fig.3 indicate that all the variations are scalable in terms of the number of frequent items.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

III. CONCLUSION

This research has attempted for the purpose of web usage mining. The proposed methods were successfully tested on the web log files. In this research, the problem is solved easily in server log files. The simulation result shows that the D-Apriori and DFP-Growth algorithm is used for finding the most frequently access pattern generated from the web log data, by using the concept of web usage mining and the problem can easily find out that the user's interest. So that our web site can be improve and more easily accessible for the users. By using clustering method in these algorithms time will be reduced. The main goal of the proposed system is to identify usage pattern from web log files.

REFERENCE

[1.] R. Cooley, B. Mobasher, and J. Srivastava "Web mining information and pattern discovery on the World Wide Web", 8 Nov 1997.

[2.] Jiawei Han and Michelin Kamber, "Data mining Concepts and Techniques", Elsevier publication, Edition 2006.

[3.] Rajan Chattamvelli, "Data Mining Methods", Narosa publications, Edition 2009.

[4.] Santhosh Kumar and Rukumani,"web usage mining", ijana publication, vol.1, pages 400-404, Edition 2010.

[5.] Rahul Mishra and Abha Choubey, "FP from web log data using FP Growth for web usage mining", ijarcsse publications, vol.2, Edition 2012.

[6.] Divya and Vinod Kumar,"AIS,Apriori and FPTree Algorithm", ijcsmr publication, vol.2, paper 30.

[7.] G.Sudamathy and C.Jothi venkateshwaran, "An efficient hierarchical frequent pattern analysis approach for web usage", ijca publication, vol.43, Edition 2012.

[8.] Jianhan Zhu,Jun Hong and John G. Hughes, "Page clustering: Mining conceptual link hierarchical from web log files for adaptive websites navigation", ACM publication, vol.4, Edition 2004.

[9.] Harish Kumar and Anil Kumar," Clustering Algorithm Employ in Web Usage Mining: An Overview", INDIA Com publication, Edition 2011.

[10.] Idams, "Divisive analysis (DAINA) Algorithm", Uneseo publication, chapter 7.1.5, Edition 2005.

[11.] Hussain T, "A hierarchical cluster based preprocessing methodology for web usage mining", IEEE publication, Edition 2010.

[12.] Ashok Kumar D,Loraine Charlet Annie M.C, "Web log mining using K-Apriori algorithm", ijca publication, vol.41 Edition march 2012.

[13.] Shyam Sundar Meena, "Efficient discovery of frequent pattern using KFP-Tree from web logs", ijca publication, vol.49, Edition July 2012.

[14.] J Han, J Pei, Y Yin, R Mao, "Mining frequent patterns without candidate generation: A Frequent pattern tree Approach" Data mining and knowledge discovery, 2004 – Springer.

[15.] J Minghai, Y Ping, J Huiyan - Wuhan, " Research and application on web information retrieval based on improved FP-Growth algorithm" Journal of Natural , 2006 – Springer

[16.] Mr.Ravindra gupta,Prateek gupta,"Fast preprocessing of web usage mining with customized web log pre-processing and modified frequent pattern tree" International Journal of Computer Science- 2012 vol 1.

[17.] Bharat Bhustan Agrwal,Dr.M.H.Khan and Shivangi Dhall, "Web mining information and pattern discovery on the world wide web" Ijstm journal,2010 December edition.

[18.] Jaideep Srivastava,Robert Cooley,Mukund deshpande,Pang-Ning tan,"Webusage mining:Discovery and application of usage pattern from web data",SIGKDD Publication, volume 1, issue 2, Jan 2000.

[19.] Florian verhein, "Frequent pattern growth(FP-Growth) Algorithm ", Edtion 2008.

[20.] BS Kumar, KV Rukmani, "Implementation of web usage mining Apriori and FP growth", 2010 – ijana.