To Enhance Frequent Closed Pattern Tree using Fuzzy Clustering of Personalized Web-Log in Big Data

Sapana Kumari
Computer Science & Engg.
Lakshmi Narain College of
Technology Excellence
Bhopal, India.

Vikram Garg
Computer Science & Engg.
Lakshmi Narain College of
Technology Excellence
Bhopal, India.

ABSTRACT

In recent time data mining on big data is very tedious task in current scenario because huge data cannot be handling in the memory with different format type of data. In distributed environment web pages access by the user having some patterns, these patterns are merging and finding closed frequent set of web pages. Now do the Fuzzy C-Means clustering of web page access pattern tree. If user need next request page in advance then it search only partial web data not in whole web data. So that this research utilized a personalized weighted recommendation system based on user's interest with less execution time.

Keywords

Web Usage Mining; Closed Sequential Patterns; Sequence Tree; Web Log Data; Fuzzy Clustering.

1. INTRODUCTION

Web Mining has been studied to explore the patterns from networks or databases. In recent years, many applications have applied this method, to discover infrastructure patterns in web database. Web mining methodologies can be classified into one of three distinct categories: web usage mining, web structure mining, and web content mining. Web page usage patterns in order to learn about web users or the relationships between the documents.

Web usage mining is defined as the process of applying data mining techniques to the discovery of usage patterns from web logs data and to identify web users' behavior. In Web usage mining, data can be collected at the server-side, client-side and proxy servers.

Clustering have been useful and active areas of machine learning research that promise to help us cope with the problem of information overload on the Internet. With clustering the goal is to separate a given group of data items (the data set) into groups called clusters such that items in the same cluster are similar to each other and dissimilar to the items in other clusters.

2. RELATED WORK

Yi Pan, HongYan Du [4] presents a novel prefix graph based algorithm for mining closed frequent itemsets. The new approach has constructed an efficient prefix graph structure and use variable length bit vectors to present the relationship between the database and its items.

Ms. Anjali B. Raut et al. [5] proposed fuzzy hierarchical clustering for creating the clusters of web documents based on fuzzy equivalence relation. This technique is used to construct clusters with uncertain boundaries.

Mozhgan Azimpour Kivi et al. [6] proposed a web session clustering based on similar trends. This algorithm clustered the webpages based on similarity measure for web sessions clustering using an agglomerative clustering technique with sequence alignment.

Omar Zaarour et al. [7] proposed an improvement the web log mining procedure for the prediction of online navigational pattern.

Nayana Mariya Varghese et al. [9] are proposed cluster optimization technique using fuzzy logic. Web page access pattern is collected from web log file as input and then eliminate irrelevant data items.

3. PROBLEM DEFINITION

In recent time data mining on big data is very tedious task in current as well as future scenario because huge data cannot be handling in memory with different type. Big data is nothing but just huge data in different format. After few years server having lots of data so that not possible to handle into memory. So the problem in the current scenario is every time the whole database scanning for searching the frequent pattern not partial database.

Web database partition is needed some conditional parameters which is related registered user profile. There is a need to search the data based on similarity or expert condition of Profile User (Income, Age, and Experience etc.) which support as conditional parameter for partition of web database.

4. DESCRIPTION OF PROPOSED WORK

Each web page having some importance is called weight. It is also used personalized system where we having user profile information. As per user profile similarity we merging the frequent pattern not all frequent pattern so that less database scan perform fast response and accurate result in distributed environment.

It uses Fuzzy C-Means Clustering of large web database with some conditional parameter (eg. Age, Experience, Research Area, Sex etc.) of registered user profile. For this it uses combination of two functions Similarity based and Expert of Profile User (Income, Age, and Experience etc.). So only user related information is gathered in web.

In the proposed system it is having three phases –

a) Partition Phase (Step 1-4),
b) Mining Phase (in Step 5),
c) Merging Phase (in Step 6).

Step-1: Collect the web logs of websites.

Step-2: Apply preprocessing to get useful sessional web data.

Step-3: Supply input as number of support by the user and checks support is greater or equal to weight of page and generate frequent sequential pattern.

Step-4: Supply input as number of cluster so web database is clustered based on selected attribute using fuzzy clustering technique. So Partition Phase is completed and each partial frequent databases based on selected attributed is ready for Mining Phase.

Step-5: In Mining Phase it convert frequent sequential pattern into Closed Frequent Sequential Pattern and generate the Pattern-Tree for next item prediction.

Step-6: In Merging Phase it merges all partial frequent databases and apply filter based on user profile. Finally establish good cluster items for prediction into the caching to improve the quality of the results and response time.

The Construction of Closed Sequential Pattern Tree

The constructed pattern tree is used in making the recommendation for a user's web access sequence. The constructed pattern tree is based on the Patricia trie (Radix tree or crit bit tree) data structure. For extraction of sequential web access patterns a dataset (WASDB) of web access pages has been considered, which is shown in Table - 1. Let assume that if min_sup is 2 then complete set Cs is {a:6, b:6, c:6, aa:4, ab:3, ac:3, bb:2, ba:3, bc:3, ca:4, cb:2, cc:2, abb:1, abc:1, aba:2, acc:1, aca:2, bca:1, caa:1, cac:2, cba:1, cbc:1}. Now prune by min_sup and get complete frequent set CFS is { a:6, b:6, c:6, aa:4, ab:3, ac:3, bb:2, ba:3, bc:3, ca:4, cb:2, cc:2, aba:2, aca:2, cac:2}. Here prune item is {abb, abc, acc, bca, caa, cba, cbc} because it having support is only 1. The set CFS consists of 15 frequent sequences and finally generate the closed sequential patterns CCS. For min_sup=2 the various length sequential web access pattern shown in Table 3.8 and their pattern tree is shown in Figure - 1. Here node {cc:2} is merged into {cac:2} node.

Table - 1: Sequential Web Access Patterns with min_sup=2 from the Sample Database

Length of Patterns	Sequential Web Access Patterns with Support
1	a:6, b:6, c:6
2	aa:4, ab:3, ac:3, bb:2, ba:3, bc:3, ca:4, cb:2, cc:2
3	aba:2, aca:2, cac:2

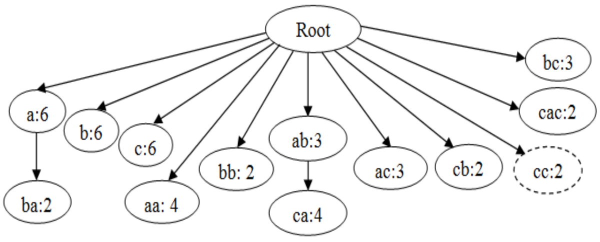


Figure - 1: The pattern-tree constructed from the closed sequential web access patterns

5. EXPERIMENT RESULTS

All the experiments are performed on Intel(R) Dual Core 2.4 GHz Pentium PC machine with 4 GB main memory, running on Microsoft Windows XP. In addition, all the programs are written in Dot Net 2010, 4.0 Framework. So pursue the experiments on real datasets to evaluate the performance of proposed algorithm. The datasets, which contain several months' worth of click sequence data from two e-commerce web sites.

Table - 2: The Example of Page with Weight Range

PageID	Page Name	Support
P1	Page1	9
P2	Page2	7
P3	Page3	6
P4	Page4	5
P5	Page5	3
P6	Page6	2

Table - 3: A Sample Database [WASDB] of Web Access Pattern

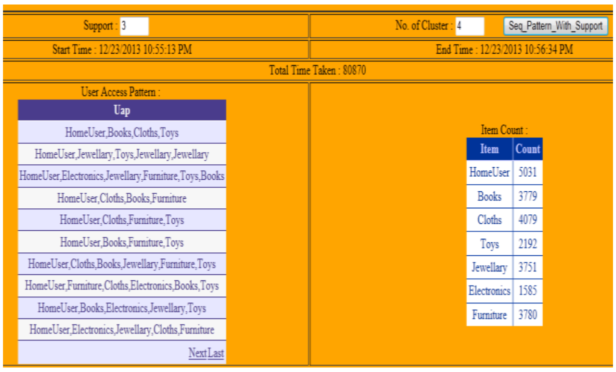
SID	Web Access Pattern
S1	b b c
S2	c a a c
S3	c b c
S4	a c a c
S5	a b a
S6	c b a
S7	a b b
S8	a b c a

Table - 4: Clustered Web Access Pages

Web Access Pages	Cluster-1	Cluster-2
A	Unit-1	
B		Unit-2
C		Unit-2

← 'Winning' output unit →

Now the final conclusion is page b and page c belong the cluster-2 and page a belongs to cluster-1.



Support : 3
No. of Cluster : 4
Start Time : 12/23/2013 10:55:13 PM
End Time : 12/23/2013 10:56:34 PM
Total Time Taken : 80870

User Access Pattern :

Uap
HomeUser,Books,Cloths,Toys
HomeUser,Jewellary,Toys,Jewellary,Jewellary
HomeUser,Electronics,Jewellary,Furniture,Toys,Books
HomeUser,Cloths,Books,Furniture
HomeUser,Cloths,Furniture,Toys
HomeUser,Books,Furniture,Toys
HomeUser,Cloths,Books,Jewellary,Furniture,Toys
HomeUser,Furniture,Cloths,Electronics,Books,Toys
HomeUser,Books,Electronics,Jewellary,Toys
HomeUser,Electronics,Jewellary,Cloths,Furniture

Item Count :

Item	Count
HomeUser	5031
Books	3779
Cloths	4079
Toys	2192
Jewellary	3751
Electronics	1585
Furniture	3780

Figure -2: Web Access Pattern with Minimum Support and No. of Clusters

Table-5: Running Time (in ms) with different size and different min_sup

No. of Records (Sessions)	min_sup (5%)	min_sup (10%)	min_sup (20%)	min_sup (40%)	min_sup (80%)
1000	6223	6810	6623	6803	6304
2000	16950	17989	17986	18452	17972
3000	15610	17953	18564	17963	18384
4000	24503	23769	23925	24802	23946
5000	79618	37578	85431	53917	37974

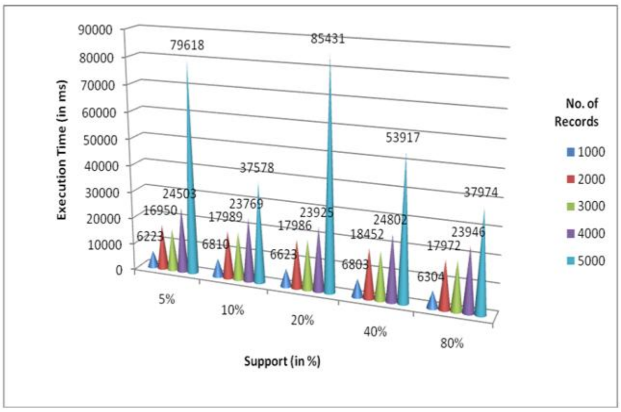


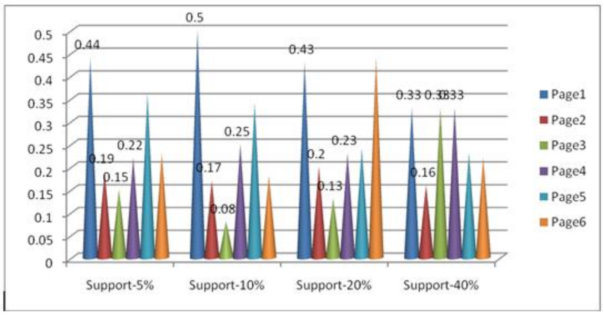
Figure-3: Execution Time (in ms) with different record size and different min_sup

The Figure-3 shows the Running time (in ms) of proposed algorithm under different record size with different support. Running time (in ms) of proposed approach varies from record size 1000 to record size 5000 with support 5%, support 10%, support 20%, support 40%, and support 80% . This Graph shows that while taking record size 1000 with support 5% then running time of proposed algorithm is 6223 ms, similarly with support 5%, 10%, 20%, 40% and 60% running time of proposed algorithm is 6810 ms, 6623 ms, 6803 ms and 6304 ms respectively. Similarly with record size 2000, 3000, 4000 and 5000.

Table-6: Probability of Items in Web Pages with different support

min_sup	P1	P2	P3	P4	P5	P6
5%	0.44	0.19	0.15	0.22	0.36	0.23
10%	0.5	0.17	0.08	0.25	0.34	0.18
20%	0.43	0.2	0.13	0.23	0.24	0.44
40%	0.33	0.16	0.33	0.33	0.23	0.22

Figure-4: Probability of items in Web Pages with different min_sup



The Figure-4 shows the probability of occurrence of each item with different support. This graph show the probability of occurrence the page1, page2, page3, page4, page5 and page6 of Website with different min_sup just like 5%, 10%, 20%, and 40%. The probability of page1 is highest with min_sup 10%, probability of page2 is highest with min_sup 5%, similarly probability of page3, page4, page5 & Item-6 is highest with min_sup 40%, 40%, 5% and 20%.

Table-7: Comparison of FP-Tree and Proposed Approach with different support (By using Record-Size 5000)

Algo.	5%	10%	20%	40%	80%
FP-Tree	72847	74736	78394	81646	94982
Proposed	70383	70384	57116	58737	61391
Efficiency	3%	6%	27%	28%	35%

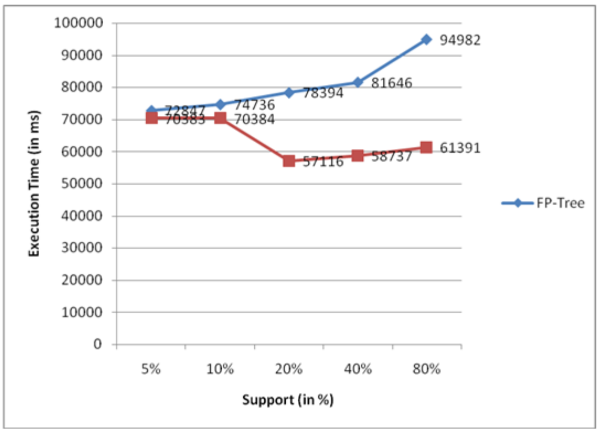


Figure -5: Comparison of FP-Tree and Proposed Algorithm with different support

The Figure-5 show the comparison between execution time (in ms) of FP-Tree and Proposed algorithm using record size 5000 with support 5%, support 10%, support 20%, support 40%, and support 80%. The execution time of proposed algorithm is lowest as compared with FP-Tree algorithm with support 5% to support 80%. This comparison show that proposed algorithm is more efficient.

The results of this algorithm produce good scalability with big data in pattern computing environment. It is extremely important for decision making when studying behaviour of the Internet users. This research improve response time in web usage mining when users having more data in different format.

6. CONCLUSION

In this research a new approach for mining closed frequent item sets from transactional data sets with fuzzy clustering is introduced. It maintain prefix graph data structure with partition to effectively detected and judge the possibility of closed items in very less time. It handles big graph data with different minimum support and number of clusters. Closed frequent pages are clustered by fuzzy clustering as per each user profiles, so that only partial database is scan not whole database.

In future work, other data mining algorithms can be implemented in cloud to efficiency handle big data of many Hospital website in distributed environment for finding any critical diseases.

7. REFERENCES

[1] Ms.N.Preethi1, Dr.T.Devi2,- "New Integrated Case And Relation Based (CARE) Algorithm". International Conference on Computer Communication and Informatics (ICCCI -2013) Jan. 04 – 06, 2013, Coimbatore, INDIA.

[2] Chun-Chieh Chen1, Kuan-Wei Lee2, Chih-Chieh Chang3, De-Nian Yang4, and Ming-Syan Chen5 – "Efficient Large Graph Pattern Mining for Big Data in the Cloud", 2013 IEEE International Conference on Big Data.

[3] Upa Gupta1, Leonidas Fegaras2- "Map-Based Graph Analysis on MapReduce", 2013 IEEE International Conference on Big Data.

[4] Yi Pan, HongYan Du – "A Novel Prefix Graph Based Closed Frequent Itemsets Mining Algorithm", IEEE International Conference on Computational Science and Engineering.

[5] Raut Ms. Anjali B., and Bamnote Dr. G. R., "Clustering Method based on Fuzzy Equivalence Relation," International Conference on Computer & Communication Technology (ICCCT), pp. 666-671, 2011.

[6] Azimpour-Kivi Mozhgan, and Azmi Reza, "A Webpage Similarity Measure for Web Sessions Clustering Using Sequence Alignment," International Symposium on Artificial Intelligence and Signal Processing (AISP), IEEE, pp. 20-24, Jun. 2011.

[7] Zaarour Omar, Nagi Mohamad, "Effective web log mining and online navigational pattern prediction," International Journal of Knowledge Based Systems, Elsevier, Vol. 49, pp. 50-62, Sept. 2013.

[8] Moriwal Rahul and Prakash Vijay, "An Efficient Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs based on Dynamic Weight Constraint," Proceedings of the Third International Conference on Trends in Information, Telecommunication and Computing, Lecture Notes in Electrical Engineering, Springer, Vol. 150, pp. 483-490, 2013.

[9] Varghese Nayana Mariya, John Jomina, "Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic," World Congress on Information and Communication Technologies, IEEE, pp. 948-952, 2012.

[10] Zan Mo, Yanfei Li, "Research of Big Data based on the Views of Technology and Application", American Journal of Industrial and Business Management, pp. 192-197, Apr. 2015.

[11] Yingdi Guo, Kunhong Liu, "A New Spatial Fuzzy C-Means for Spatial Clustering", VSEAS Transactions on Computers, Volume 14 pp. 369-381, 2015.

[12] Jyoti Tyagi, Neeta Verma, "Optimization of Fuzzy C Means Clustering using Genetic Algorithm for an Image", International Journal of Computer Applications, Volume 121, No 17, pp. 29-32, July 2015.

[13] Mohammad Naimur Rahman, Amir Esmailpour, "A Hybrid Data Center Architecture for Big Data", Elsevier, pp. 29-40, 2016.

[14] Xue Yang, Rongxine Lu, "A Secure and Fine-Grained Privacy Preserving Matching Protocol for Mobile Social Networking", Elsevier, pp. 2-9, 2016.