Performance Improvement of Web Usage Mining Over Uncertain Data

Anupa Mishra[1], Sneha Soni [2]

[1]Sagar Institute of Research & Technology-Excellence, Bhopal

[2]Sagar Institute of Research & Technology-Excellence, Bhopal

Abstract: ***Customer behavior plays an important role in any organization to understand their future need. Web usage mining is an active research topic where customers session clustering is used to understand the customer's activities. It investigates the problem of mining frequent pattern and especially focuses on reducing the number of rules using the partition of a sequential pattern. It also reduces scan size of the database using SOM clustering technique. In the website, every web pages access by the user having some patterns and these patterns are merging and finding the frequent set of web pages. If the user needs next request page in advance then it searches only partial web data, not in whole web data. So it solves the problem through Partition based Pattern Mining Algorithm to remove undetermined data. This system is based on user's interest with less execution time.***

Keywords: Web Usage Mining, Sequence Tree, Web Database, Web Services, Sequential Pattern, Clustering

1. INTRODUCTION

Today, organizations are depending more and more on their websites to correlate with customers. Holding recent customers and attracting potential ones push these organizations to come across in striking ways to make their websites more useful and efficient. The world wide web (www) is a massive source of information that can come either from the web content, delineated by the billions of web pages openly accessible, or from the Web usage, represented by the registration information collected daily by all the servers around the world.

Web mining is that part of data mining which deals with the extraction of interesting information from the websites. It has various applications i.e., personalization of web substance, design support, recommendation systems, pre-fetching and caching. In e-commerce web usage mining play an important role.

Customers and products can be targeted with appropriate advertisement suggested in real-time while browsing the website. According to web usage mining it may divided into three steps. The first step starts with data cleaning and pre-processing. Second step is the pre-processed data are mined for some unseen and productive information. and the last step of the web log mining process ends by analyzing the mining results.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. You submit your paper print it in two-column format, including figures and tables. In addition, designate one author as the "corresponding author". This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

2. LITERATURE SURVEY

Kaichun et al., [2] proposed frequent route pattern mining from personal trajectory data is the basis of location awareness and location services. However, because personal trajectory data is highly uncertain, most exciting approached are only capable of funding short and incomplete route patterns. In this paper a novel approach is proposed for the discovery of frequent pattern based on trajectory abstraction.

Pazdor et al., [3] proposes a method for Knowledge discovery of big data is one of the most interesting topics in this paper and research pattern mining is a major task. With the rapid growth of modern technologies, high volumes of data which are of different veracities (i.e. may be precise of uncertain).In this paper we design a memory efficient data structure called uncertain data streams.

Badran et al., [4] extend the SAT-based encoding of frequent pattern data mining to interesting uncertain databases with novel declarative mining framework transactions databases.

Sheetal Sahu et al.,[6] proposed neural network based approach for web usage mining in which Web usage mining try to discovers useful knowledge from the secondary data obtained from the connections of the users with the Web. It represents a novel method self organizing map, which is a kind of neural network, in the process of web usage mining to detect user's patterns. It analyzes the traditional K-Means algorithm result with comparison to SOM. The process details the transformations necessaries to modify the data storage in the web servers log files to an input of SOM.

3. PROBLEM DESCRIPTION

Frequent sequence mining is an important part related to web data and now yet a challenging data mining work. The mining frequent sequence has become an important component of many prediction or recommendation systems. So the problem in the current scenario is –

1) It uses uncertain data streams for each item in different transactions because it may have a different existential probability, this uncertainty property of items may make the tree unmanageable and large.
2) To deal with stream property pattern growth does not keep the most recent information.
3) Every time the whole database scans for searching the frequent pattern not partial database.

4. PROBLEM DESCRIPTION

In this research we use novel approach for finding frequent sequential pattern mining using SOM clustering. It is used as a trend analysis to identify customer's patterns in the process of web usage mining. It depends on the performance of the clustering of the amount of requests. Here the proposed approach using SOM clustering for access the partial web data. In the next step the input support gather by user using interface if item support is less than and equal to given support then it produce the frequent item using pruning strategy of the item.

4.1 Proposed Algorithm Description -

We uses following steps to find new value to share a single node in the uncertain stream.

Step 1: Collection of web navigation history of website.

Step 2: Apply pre-processing techniques to remove noise from web log data and also convert into proper format.

Step 3: Now find the frequent pattern using given support threshold value in the model.

Step 4: Now generate rules using partition sequential pattern of web data.

Step 6: For clustering the web data it require to first find the smallest probability value and after that merge these two by taking smallest value in the form of cluster.

Step 7: Now select different size of web data and generate the rules.

Step 8: So put those item in the cache which are having higher frequency.

Step 9: For the next item prediction put some items in the cache which has higher frequency. Sometimes if next item not available in the cache so that it scan the related item from the cluster web data not the whole data.

4.2 Pseudo code of Proposed Algorithm

The proposed Sequential Pattern Mining using SOM Clustering approach is applied for discover frequent sequential patterns by using clustering approach for producing the cluster of web data set. This cluster is used to access the partial web data set not whose web data set. By using closed sequential, it generates fewer candidates set for generating the rules so that response time is increases. The merging method in realism is reconstructing a small Pattern tree. So at this time this tree having the web pages of website in proportional sessions.

```
Procedure: Proposed Algorithm (Support, Web Dataset)
{
  D = find the web data
  int i=0;
   do
   {
      Compute the mean weight of prefix item x
     if (x.item.weight >= support)
     {
     Output prefix;
     }
   i=i+1;
   }
  While (x.item.count>=i);

 //clustering of web data set
  int k=0;
 Generate the cluster of given frequent item
 int totCluster = CountNoOfItem();
 int sdv = search(smallest distance value);
 int nsdv = searchNearstItem(sdv);
do
  {
    if (sdv <= nsdv)
    {
     Allocate-Cluster (sdv);
     Generate-ClosedPatterTree();
    }
   k=k+1;
  }
  While (k<= totCluster);

SetCacheItem(ClosedPatterTree);
```

5. RESULT ANALYSIS

All experiments were conducted on a 2GHz Intel Core2 Duo processor PC with 4GB main memory running Microsoft Windows XP. The algorithms were implemented in Asp.Net with C# and were executed. In this research a real data set is used, which is having click stream data from an e-commerce web store and it has been used widely to assess the performance of frequent pattern mining. This dataset contains click stream sequences by customers with number of item purchases. It concludes that it produce good result for finding next item prediction.

No Of Support	Execution Time (millisecond) WAP Algorithm	Execution Time (millisecond) Partition Based Pattern Mining Algorithm
2	178	169
5	119	112
8	104	96

15	89	78

Table-1: Execution Time of WAP Algorithm and Partition based Pattern Mining Algorithm with Different Number of Supports

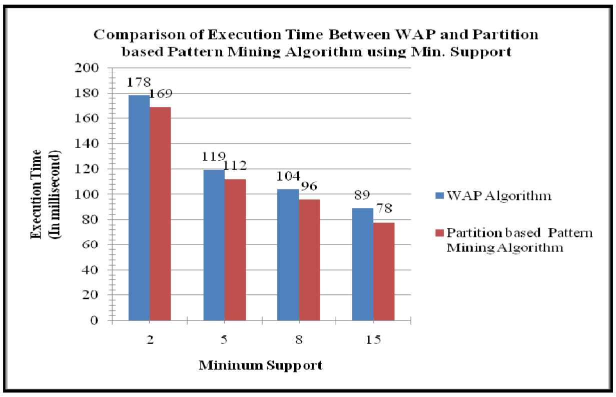

Figure-1: Comparative graph of WAP VS Partition based Pattern Mining Algorithm in terms of Supports

No Of Records	Execution Time (In millisecond) WAP Algorithm	Execution Time (In millisecond) Partition based Pattern Mining Algorithm
200	189	148
300	247	233
400	361	279
500	431	381

Table-1: Execution Time of WAP Algorithm and Partition based Pattern Mining Algorithm with Different Number of Records

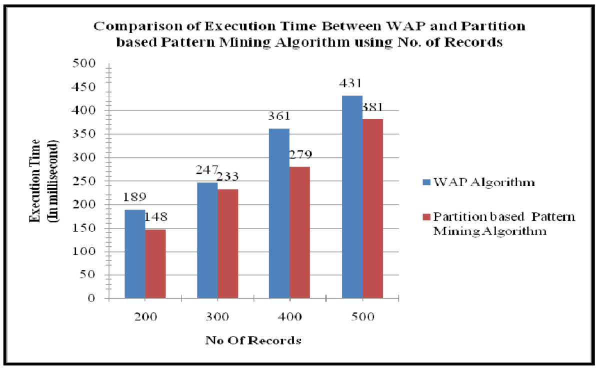

Figure-2: Comparative graph of WAP VS Partition based Pattern Mining Algorithm in terms of Records

6. CONCLUSION

In this research a novel approach is proposed for finding sequential patterns and scan only partial web data for next item prediction. It is having minimum support and initially each item as cluster & merges them in final cluster so that partial web data is scan. Here frequent pages are very less and useful rules in the form of clustered by SOM clustering. Now multiple scanning of database will be reduced. Scanning only partial database not the whole database with improves the response time. It enables effective tracking for the development and improvement of the user interface and software by analyzing user behavior. In future work, other data mining algorithms can be implemented in cloud to efficiency handle large web data of many Hospital website in distributed environment for finding any critical diseases. So there are many areas just like parallel sequential pattern, grouping of similar type of customers, in distributed servers.

References

[1] Miss. Veranda Khairnar and Sonal Patil, "Efficient clustering of data using improved K-means algorithm", Imperial Journal of Interdisciplinary Research (IJIR), Vol.2, Issue-1, 2016.

[2] Fan Muhan, Shao Sujie and Rui Lanlan, “A Mining Algorithm for Frequent Closed Pattern on Data Stream based on Sub Structure Compressed in Prefix Tree”, IEEE Proceedings of CCIS, pp. 434-439, 2016.

[3] Minubhai Chaudhari, Chirag Mehta, “Extension of Prefix Span Approach with GRC Constraints for Sequential Pattern Mining”, International Conference on Electrical, Electronics, and Optimization Techniques, pp. 2496-2498, 2016.

[4] Doddegowda B. J., G. T. Raju and Sunil Kumar, “Extraction of Behavioral Patterns from Pre processed Web Usage Data for Web Personalization”, IEEE International Conference On Recent Trends In Electronics Information Communication Technology, pp. 494-498, 2016.

[5] Jerry Chun, Wensheng Gan, Tzung and Pei Hong, “Efficiently Maintaining the Fast Updated Sequential Pattern Trees With Sequence Deletion”, IEEE Access - The Journal for Rapid open access publishing, Vol. 2, pp. 1374-1383, 2014.

[6] Sheetal Sahu, Praneet Saurabh and Sandeep Rai, “An Enhancement in Clustering for Sequential Pattern Mining Through Neural Algorithm Using Web Logs”. IEEE Proc. of the 2014 Sixth International Conference on Computational Intelligence and Communication Networks, pp 758-764, 2014.

[7] Rahul Moriwal and Vijay Prakash, “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, Vol. 150, 2013.

[8] Zaarour Omar and 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.

[9] Varghese Nayana Mariya and 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] 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.

[11] 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.

[12] Giannikopoulos P., Varlamis I. and Eirinaki M. “Mining Frequent Generalized Patterns for Web Personalization in the Presence of Taxonomies”, IJDWM, 2010.

[13] K. R. Suneetha and Dr. K. R. Krishnamoorthy, “Identifying User Behavior by Analyzing Web Server Access Log File”, IJCSNS International Journal of Computer Science and Network Security, Vol. 9, No.4, pp. 327, 2009.

[14] Krzysztof D., Wojciech K. and Marcin S., “Effective Prediction of Web User Behavior with User-Level Models”, Fundamental Informaticae , IOS Press , Vol. 89, No. 2-3, pp. 189, 2008.

[15] Khan, J. I. Qing ping T., “Web space surfing patterns and their impact on Web prefetching”, Cyber worlds Proceedings, pp 478 – 485, 2003.

[16] Catledge, L. D. and Pitkow, J. E., “Characterizing Browsing Strategies in the World- Wide Web”, In Proceedings of the Third International World-Wide Web Conference, Darmstadt, Germany, pp. 1065–1073, 1995.