International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 03 (MAY-JUNE, 2016)

A Novel & Efficient Methodology for Web Data Mining

Anil Sharma, Kailash Patidar, Jitendra Rai, Manoj Yadav
Department of Computer Science & Engineering
Sri Satya Sai Institute of Science & Technology, India

Abstract: - In this paper, we present an overview of modern web mining algorithms. Frequent web item set mining is a heart favorite topic of research for many researchers over the years. Also web mining is a computationally expensive task. So still there is a need to update and enhance the existing web mining techniques so that we can get the more efficient methods for the same task. In this paper, we have developed a method to discover frequent web item sets from the web transaction database. The proposed method is fast in comparison to older algorithms. Also it takes les main memory space for computation purpose.

Keywords: - *Apriori, Web Usage mining, Itemsets, Loss counting,* Defer Counting.

I. INTRODUCTION

The web mining [1, 2] is used to extract the useful information from the World Wide Web by using data mining techniques. The overall tasks under web mining are generally divided in to three main categories. These are web content mining, web structure mining and web usage mining. The first one that is the web content mining is used to search the web pages by using the content of the web pages as search words.

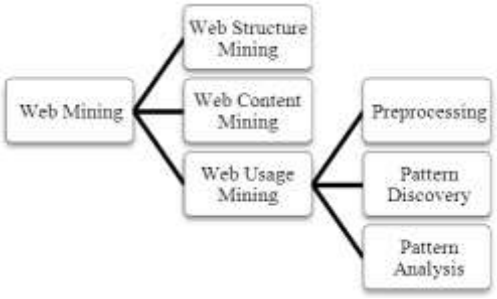

Figure 1 Web Mining Categorization

The second web structure mining is the collection of methods which are used for mining or extracting the structure or hierarchy or the links of a web site. The web usage mining is related to the application of data mining tools and techniques on the web to discover the web user patterns. It helps organizations in increasing the user satisfaction. The WUM or the web usage mining consists of three major steps [3, 4, 5]. These steps are preprocessing, pattern discovery, Pattern analysis. In preprocessing step of web usage mining, the data is extracted from the web data set & then this data is preprocessed. In preprocessing, the noise is removed from the data. The output of preprocessing phase contains information like, how many pages accessed, which page is accessed how many times, which user accessed which page, access time, access date, access duration etc.

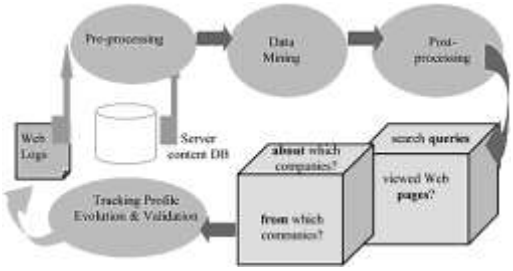

Figure 2 web usage mining process

Data mining [6] represents the integration of several fields, including machine learning, database systems, data visualization, statistics and information theory. Data mining can be defined as a non-trivial process of identifying Valid, Novel, and potentially useful, ultimately understandable Patterns in data. It employs techniques from machine learning statistics databases Knowledge discovery in databases is a complex process, which covers many interrelated steps. Key steps in the knowledge discovery process are:

1. Data Cleaning: remove noise and inconsistent data.
2. Data Integration: combine multiple data sources.
3. Data Selection: select the parts of the data that are relevant for the problem.

4. Data Transformation: transform the data into a suitable format.
5. Data Mining: apply data mining algorithms and techniques.
6. Pattern Evaluation: evaluate whether they found patterns meet the requirements
7. Knowledge Presentation: present the mined knowledge to the user (e.g. Visualization).

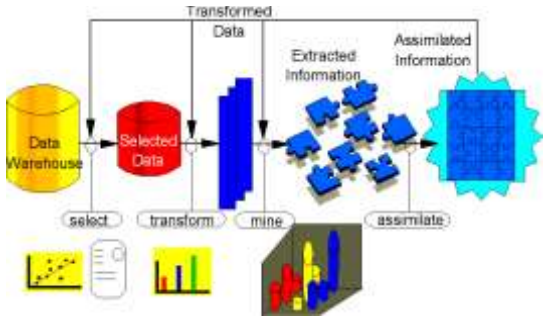

Figure 3 key steps in data mining

II. LITERATURE SURVEY

In 2009 the authors Ling Chen [7] proposed that the fading factor model can be used to compute the frequent itemsets. This fading factor lm contributes more to the recent items than the older. The fading factor ranges between 0<lm<1, where lm is frequency. Value near to 1 is considered to be most frequent item. This technique has two major advantages. Firstly, It takes the all old data items based on the frequency and the other is changes in frequency varies by a small values.

In 2009 the work done by Cai-xia Meng [8] proposed the efficient algorithm for mining frequent itemsets over a high speed data streams. The frequent pattern mining algorithms pose two steps. This involves calculations behind the arrival of every frequency of new item sets and formatting them into the output. In this algorithm these two steps are blended together to reduce too much of time that arrives in LossyCounting (LC) and FDPM. That loss of data occurs in other algorithms is sorted out here. This method uses Defer Counting (DC) method which delays the frequency calculation and provides gap between step 1 and step 2 to avoid the transaction missing problem. In step 1the DC includes the information that achieved the threshold value. In step 2 it frequent pattern from the stored information, the data structures used are List and Trie. The List is for the frequent items and Trie for the frequent item sets. Data structure List is used with three-fields stating, the id of each item, the frequency of that item and error between the actual and the estimated frequency. The Trie composed of two fields in which one points to counter in the list. The other field is to compute the frequency of the itemsets.

In 2010 author Varun Kumar [9] proposed this algorithm which has an ability to hold the various sizes of the batch rather than the fixed in other. The time has been fixed for segregating the Batches. In the previous algorithms the infrequent items were removed. Later if those items become frequent then the data cannot be bagged again. Also they concentrated only on the frequent item sets, but not on the extracting knowledge from it. Such kind of problems is clearly solved by this paper. Proposed work uses an extension of trie structure with the log-time window as its data structure. Method constitute three columns namely tilted-time, frequency and size of the batch. Recent data holds big space whereas the old one holds the less only. The work follows two different types of tail pruning in examining whether the superset needs to be dropped or not based on the different batch sizes and time.

In 2009 work of Sonali Shukla [10] proposed this algorithm with the regression based methodology to find out the frequent item sets continuously that are streaming regularly. In this method the 2-Dimensional stream data is preprocessed and converted into sampling value. Regression analysis is carried out with these values. Method bags the data using sliding window model and then it applies FIM-2DS algorithm

to compute with the item set. It is processed to the sampling value for the further process with least square method. Every data is paired (mi, ni) to find the arrival time difference between them. Data is calculated like t, t-1, t-2, t-3… tn. if the pair is (m,n) then it is mean that m is an independent variable of n and n is dependent variable on m. By taking the help of the pairs of Data Sets dependent and independent variable values are calculated. After that the regression line is drawn from fragment slope values. Also the regression analysis is also used to find out the functional relationships between the paired data items.

In 2010 the author ZHOU Jun [11] proposed this algorithm by considering the space as an important factor. Authors used an improved LRU (Least Recently Used) based algorithm. Proposed algorithm omits the infrequent items before taken for the processing. Method increases the stability and the performance. Method is used to find out the frequent items as well as the frequency of those items.

In 2012 work of Yong-gong Ren [12] proposed this algorithm in order to predict the future data based on the new method called AMFP-Stream known as Associated Matrix Frequent Pattern-Stream. It predicts the frequently occurred item sets over data streams efficiently. Proposed work also has a capability to predict that which item set will be frequent with high potential. Method takes the data in the form of 0-1 matrix and then it updates the values by doing logical bit operations. Then on this it will find out the item sets that will frequently occur in the future. Method uses the associated matrix for the further manipulation. Experimental results says that this algorithm is how much feasible.

In 2011 author Mahmood Deypir [13] proposed this algorithm based on the different kind of sliding window based model. This Method doesn't need entire data that are in streaming. Method takes an advantage of the already existing item sets. To enhance the feature of sliding window concept Also it reduces the amount of space occupying and time taken to calculate based on the fixed size of the window.

III. PROPOSED ALGORITHM

Step 1: *Transaction data set & minimum support threshold.*

Step 2: *First the algorithm scans the transaction data base and calculates the support of each single size item.*

Step 3: *Compare the support of each item found in step 2 with the minimum support threshold.*

Step 4: *If support (item) > = minimum support threshold then add the item in frequent item list. Otherwise add the item in infrequent item list*

Step 5: *In this step, the transaction data base is transformed into a new compressed data structure based table by pruning of all those items which are in infrequent item list (generated in step 4) because they will not appear in any frequent patterns.*

Step 6: *Call algorithm recursively to generate bigger frequent patterns by using the union of lower size items.*

IV. CONCLUSION

In this paper, we surveyed the list of existing web mining techniques. We restricted ourselves to the classic web mining problem. It is the generation of all frequent item sets that exists in market basket like data with respect to minimal thresholds for support & confidence. we presented a novel algorithm for mining web log data sets. Frequent mining of data mining is used for that purpose. Frequent item set mining is crucial for association rule mining. We have evaluated the performance of our proposed algorithm. It is fast.

Also it is taking less main memory for computation in comparison to previous algorithm.

REFERENCES

[1]. Ashok Kumar D. Loraine Charlet Annie M.C., "web log mining using K-Apriori Algorithm", volume 41, March -2012,

[2]. Jian Pei, Jiawei Han, Behzad Mortazaviasl, Hua Zhu, "Mining Access Pattern Efficient from Web Logs"

[3]. B. Santhosh Kumar, K.V. Rukmani," Implementation of Web Usage Mining Using Apriori and FP-Growth Algorithms", volume: 01, Issue: 06, Pages: 400-404(2010)

[4]. J. Han and Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, 2000

[5]. Jiawei Han, Ian Pei, Yiwen Tin, Runying Mao, "Mining Frequent Pattern without Candidate Generation: A Frequent Pattern Tree Approach", Volume-8

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

[7]. Ling Chen, Shan Zhang, Li Tu, "An Algorithm for Mining Frequent Items on Data Stream Using Fading Factor".33rd Annual IEEE International Computer Software and Applications Conference.172-179,2009.

[8]. Cai-xia Meng, An Efficient Algorithm for Mining Frequent Patterns over High Speed Data Streams. World Congress on Software Engineering, IEEE 2009, 319-323.

[9]. Varun Kumar, Rajanish Dass .Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010 IEEE, 978-0-7695-3869-3.

[10]. Sonali Shukla, Sushil Kumar, Bhupendra Verma, A Linear Regression-Based Frequent Itemset Forecast Algorithm for Stream Data. International Conference on Methods and Models in Computer Science, 2009.

[11]. ZHOU Jun, CHEN Ming, XIONG Huan A More Accurate Space Saving Algorithm for Finding the Frequent Items, IEEE-2010.

[12]. Yong-gong Ren, Zhi-dong Hu, Jian Wang. An Algorithm for Predicting Frequent Patterns over Data Streams Based on Associated Matrix. Ninth Web Information Systems and Applications Conference, 2012. 95-98.

[13]. Mahmood Deypir, Mohammad Hadi Sadreddini,A New Adaptive Algorithm for Frequent Pattern Mining over Data Streams, ICCKE,2011, 230-235 FLEX Chip Signal Processor (MC68175/D), Motorola, 1996.