Special Issue Published in International Journal of Trend in Research and Development (IJTRD),
ISSN: 2394-9333, www.ijtrd.com

Review of Web Usage Mining using Apriori Algorithm

Sai Sudha.C.
Faculty, Dept of Computer Science, M.E.S.Degree College, Malleswaram, Bangalore, India

Abstract: WWW enriches us with enormous amount of widely dispersed, interconnected, beneficial and dynamic hypertext information. Data Mining is also referred as KDD (Knowledge discovery in databases). It is a process of discovering useful patterns or knowledge from data sources. Web mining is the application of data mining techniques for knowledge drawing out from web data. Web mining can be broadly defined as discovery and analysis of useful information from the website topology, Web page content and web usage data. With the rapidly increasing pace of adoption to e-commerce, Market basket analysis might help a retailer to analyze purchase behavior of the customers. Apriori is used to find all the frequent itemsets in a given database.In this paper implementation of Web usage mining is discussed using Apriori algorithm. The limitations of algorithm are also highlighted.

Keyword:s *Apriori,Web Usage Mining,Web Mining*

I. INTRODUCTION

www is a vast and dynamic data repository that consists of mostly raw data.The information excavated from this repository is used by users, web service providers, business analysts, thus making it even complex to be dealt with. The web users hence, want to have the effective search tools to find relevant information easily and precisely. Data mining is the process of excavation for finding out knowledge from data. Web mining is the process of excavating information and patterns from web. It is used to understand customer behavior, evaluate the effectiveness of a particular web site, and help quantify the success of a marketing campaign. It also allows looking for patterns in data through content mining, structure mining, and usage mining[5].Web mining has many advantages which makes this technology to corporations including the government agencies and enables e-commerce to personalize individuals, which results in higher trade volumes. Many Government sectors are using this technology to find threats and fight against terrorism. Companies can understand the customers' actual need and they can react to the customer needs faster.

There are three methods which are applicable for web mining- (1) Web content mining (2) Web structure mining (3) Web usage mining.

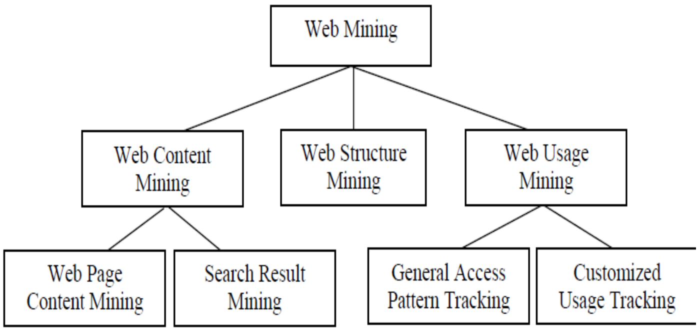

Figure-1 : Taxonomy of Web Mining [1]

II. WEB USAGE MINING

Web usage Mining is the application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of Web-based applications. Usage data encapsulates the identity or origin of web users.[6] Web Usage mining processes are:

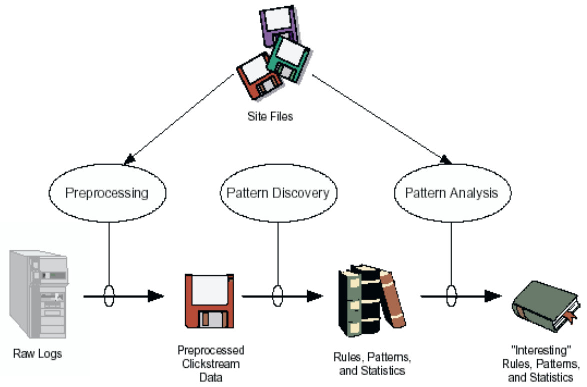
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure -2: Web Usage Mining Processes.

- Preprocessing: Conversion of the raw data into the data abstraction (users, sessions, episodes, click stream and page views) necessary for further applying the data mining algorithm.
- Pattern Discovery: Is the key component of Usage Mining, which converges the algorithms and techniques from data mining, machine learning, statistics and pattern recognition etc.
- Pattern Analysis: Validation and interpretation of the mined patterns

The discovery and analysis of patterns focuses on data accessed by the user. Web browsing behavior of users is captured by Web usage data through web site. User activities are stored in web logs. Due to more usage, the files in log are increasing at higher rate in size. The Preprocessing plays an important role in efficient mining process as Log data is normally noisy and not distinct. In the pattern analysis phase, interesting knowledge is extracted from frequent patterns and these results are used for website modifications. For finding out the information that is hidden in web logs, several data mining techniques are applied on web server logs. Web usage mining is applied to many real world problems to discover interesting user navigation patterns for improvement of web site design by making additional topic or recommendations observing user or customer behavior.

The primary data sources used in Web usage mining are :The web server data, application server data .The log data collected automatically by the Web and application servers represents the fine-grained navigational behavior of visitors. It is the primary source of data in Web usage mining. Each hit against the server, corresponding to an HTTP request, generates a single entry in the server access logs. Each log entry (depending on the log format) may contain fields identifying the time and date of the request, the IP address of the client, the resource requested, possible parameters used in invoking a Web application, status of the request, HTTP method used, the user agent (browser and operating system type and version), the referring Web re-source, and, if available, client-side cookies which uniquely identify a re-peat visitor. The crucial information extracted is discovered with the application of association rules about users' behaviors.

III. ASSOCIATION RULES AND APRIORI ALGORITHM

Association Rules help to discover correlations among pages**.** Deriving Association Rules from data was first formulated in (Agrawal, Imielinski and Swami, 1993) and is called the "market-basket problem". Given a set of items and a large collection of transactions which are sets (baskets) of items.We can find relationships between the containments of various items within those baskets. Apart from supermarket scenario ,other examples where Association Rules have been used, are users visits of WWW pages, in which the structure and its content can be optimized. Xue et al ., (2001) have used re-ranking method and generalized Association Rules to extract access patterns of the Web sites pattern usage. Mannila et al(1999) use page accesses from a Web server log as events for discovering frequent episodes. Batista and Silva, (2001) perform mining process for online newspaper Web access logs by using Apriori algorithm.

The association rules can be formally defined as:

- If the support of item-sets X is greater than or equal to minimum support threshold, X is called frequent item-sets.
- If the support of item-sets X is smaller than the minimum support threshold, then X iscalled infrequent item-sets.

IV. APRIORI ALGORITHM

Apriori is a classic algorithm for learning association rules developed by Agrawal and Srikant (1994). Apriori operate on databases containing transactions (for example collection of items brought by customers, or details of a web site frequentation) Apriori algorithm captures large data sets during its initial database passes and uses this result as the base for discovering other large datasets during subsequent passes. Item sets having a support level above the minimum are called large or frequent item sets and those below are called small item sets. The algorithm is based on the large item set property which states "Any subset of a large item set is large and any subset of frequent item set must be frequent". Apriori algorithm is, the most supervised and important algorithm for mining frequent itemsets. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item set among the candidates. Apriori is a supervised algorithm fc mining frequent itemsets for Boolean association rules. Sinc the Algorithm uses prior knowledge of frequent item set it ha been given the name Apriori. It is an iterative level wise searc Algorithm, where k itemsets are used to explore (k+1)-itemset First, the set of frequents 1- itemsets is found. This set is denoted by L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3 and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of database.

The Apriori Algorithm Pseudo code [8]:

- Join Step: Ckis generated by joining Lk-1with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code:

```
Ck: Candidate itemset of size k
Lk: frequent itemset of size k
L1= {frequent items};
for(k= 1; Lk!=∅; k++) do begin
Ck+1= candidates generated from Lk;
for eachtransaction tin database do
increment the count of all candidates in Ck+1that are contained in t
Lk+1= candidates in Ck+1with min_support
end
return∪kLk;
```

The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. Market basket analysis may provide the retailer with information to understand the purchase behavior of a buyer. This information will enable the retailer to understand the buyer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new buyers (much like the cross-selling concept.A widely used example of cross selling on the web with market basket analysis is Amazon.com's use of "customers who bought book A also bought book B", .Market basket analysis can be used to divide customers into groups.A company could look at what products are most frequently sold together and align their category management around these cliques .

Advantages:

1. It is very easy and simple algorithm.
2. Its implementation is easy.

Disadvantages:

1. The Apriori algorithm requires many scans of the database. If n is the length of the longest itemset, then (n+1) scans are required to generate candidate set for calculating frequent item.
2. Generation of candidate item-sets and support counting are expensive
3. This algorithm only defines the presence and absence of an item.
4. Apriori algorithm is not good for large database
5. This algorithm is allowed uniform minimum support threshold.

V. METHODS TO IMPROVE EFFICIENCY OF APRIORI ALGORITHM

- Hash-based itemset counting: A *k*-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent.
- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans.
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB.
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness.
- Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent.

VI. LITERATURE SURVEY

A. An Efficient Algorithm for Mining Association Rules using Confident Frequent

Itemsets

B. Al-Maqeleh et.al**,** introduced a problem with such a process is that the solution of interesting

patterns has to be performed only on frequent item sets. An efficient algorithm is proposed to integrate confidence measure during the process of mining frequent item sets, may substantially improve the performance of association rules mining by reducing the search space. The experimental results show the effectiveness of the proposed algorithm in reducing the number of discovered rules comparing with the Apriori algorithm.[11]

B. Apriori: A Modified Apriori Algorithm Based on Checkpoint

M. Patel et.al, a proposed of many algorithms to mine association rule that uses support and confidence as constraint. We proposed a method based on support value that increase the performance of Apriori algorithm and minimizes the number of candidate generated and removed candidate at checkpoint which is infrequent which interns reduces storage and time required to calculate support of candidate.[12]

C. Applying Correlation Threshold on Apriori Algorithm

A.H.S *et.al***,** introduced an Apriori algorithm, a classical rule mining algorithm finds its application in areas of data mining, finding association between attributes and in prediction systems. Performance of the redesigned algorithm is evaluated and is compared with the traditional Apriori algorithm. To increase the efficiency of the Apriori algorithm and reduce the time complexity of the proposed algorithm into O (n). [13]

D. Utility of Association Rule Mining: a Case Study using Weka Tool

A.Lekha *et.al* **,** introduced a few case studies pertaining to breast cancer, mushroom, larynx cancer and other datasets are studied to find the utility of association rule mining using Weka tool. They are three association algorithms - Apriori, PredictiveApriori and Tertius Algorithms and comparative study of the three algorithms is also made and also the implementation of the three algorithms gives the strong association rules they have problems with the number of cycles taken to generate the frequent item-sets, minimum support needed, memory utilized and non-numeric data.[14]

E. An Improved Apriori Algorithm Based on Association Analysis

Y. Jia *et.al* , proposed an improved algorithm based on a combination of data division and dynamic item-sets counting. The proposed algorithm has improved the two main problems which are faced by classical apriori algorithm. First is the repeatedly scanning of transactional database and second is the generation of large number of candidate sets. In data division, the transactional database is divided into n parts that don’t intersect each other. In first scan, all the frequent sets of each division are mined which is called local frequent sets. In second scan, the whole database is scanned again, getting support degree of all candidate item-sets and then deciding the global frequent item-sets. After data division, dynamic item-sets counting are used to decide candidate item-sets before scanning database every time. So, the whole process needs only twice the entire database scan. [15]

F. Search complexity of apriorihashtree and apriori direct search hash [7]

The factors affecting the efficiency of algorithms are the parameters with regard to item set factors(such as number of candidate sets, number of the itemsets remaining after pruning and number of frequent sets)and number of rows and columns. As far as AprioriDirectSearchHash algorithm is concerned if the parameters relating to itemsets factors are high it is very much affecting the execution time as the factors getting less in number it drastically comes down than that of AprioriHashTree.

It is usually the case that as support increases and length of the frequent set size increases, the number of frequent sets and candidate sets decreases. So, as support and frequent set size increase, the algorithm AprioriHashTree takes more execution time than the algorithm AprioriDirectSearchHash.

If the element to be searched is over the sorted set of n elements, the search complexity will be O(log2n). If it is not sorted,it will be O(n). when the maximal frequent size is small and there are very large number of 2-itemsets and 3-itemsets, we have tremendous improvement in execution time over other methods.

The, influencing factors for complexities of algorithms are row size ,column size, maximal frequent set size and item size factors.By applying these hash functions on 2-itemsets and 3-itemsets, initially thereby reducing the itemsize factors and then applying the direct search on the candidate set and since the maximal frequent item size is small, AprioriDirectSearchHash algorithm will give us good results. When row size is large, it may adversely affect the vertical mining algorithms since the complexities involve multiplication of item size factors by row size. If item size factors reduced to a threshold minimum then row size will not have much impact. As far as column size is concerned, the AprioriHashTree algorithm will be very much affected because it has the complexity of O((|Column|)Ck) as k increases, this value will increase affecting the algorithm. This problem is mitigated by AprioriDirectSearchHash, but it depends on the item set factors produced in the pass.Once item set factors reached a threshold the AprioriDirectSearch algorithm will perform well.Maximum frequent size affects all the algorithms because it decides the number of passes the algorithm to go through. If item size factors reduced during the later part of the passes, the algorithm is not further affected by the number of passes of the algorithm.

These functions can be used as index for arbitrary combinations of 2-items and 3-items. In such a case it will have time complexity of O(1) instead of having O(log2n) or O(n) depending on whether they are sorted or not.

Suppose the number of input transactions is N, the threshold is M, number of unique elements is R. The complexity for generating set of size i is O(R^i) and the time for calculating support for each set can be done in O(n), if using HashMap. Therefore, time complexity would be O[(R + N) + (R^2 + N) + (R^3 + N) …] = O[MN + (R^1+R^2+ … R^M)] = O(MN+ (1-R^M)/(1-R)).

CONCLUSION

This paper discusses the Web Usage Mining processes, and reviews the association rule based algorithm namely, Apriori algorithm, for mining frequent itemsets. The Apriori association algorithm is built upon pregauges recurrent item sets and it has to browse the entire transaction log/dataset or database which

National Conference on “Digital Transformation – Challenges and Outcomes” (ASAT in CS'17) organized by Department of Computer Science, St.Anne's First Grade College For Women, Bangalore on 3rd Mar 2017 34 | Page

will become a conflict with huge item sets .The methods to improve the efficiency of this algorithm have been discussed. Apriori has many drawbacks which can be minimized using different approaches.The factors affecting the search complexities of AprioriHashTree, AprioriDirectSearchHash are highlighted. In the future work, the problem of large number of candidate sets generated can still be improved and new improved apriori algorithm can be developed which can be used in various fields.

References

[1] IJETAE (ISSN 2250-2459, Volume 2, Issue 1, January 2012) Mahendra Pratap Singh Dohare1, Premnarayan Arya, Aruna Bajpai

[2] www2.ims.nus.edu.sg/preprints/2005-29.pdf

[3] www.interscience.in/IJIC_Vol1Iss1/paper6.pdf

[4] Analysis of Complexities for finding efficient Association Rule Mining Algorithms,. Rathinasabapathy, R.Bhaskaran, International Journal of Internet Computing, Volume-I, Issue-1, 2011.

[5] Data and web mining-Salvatore Orlando

[6] Web Usage Mining-Jinguang Liu &Roopa Datla

[7] www.interscience.in/IJIC_Vol1Iss1/paper6.pdf

[8] APRIORI Algorithm-notes of Prof.Anita Wasilewska

[9] International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016

[10] Review of Web Usage Mining , Mr. AnkitRathi, Prof. Abhijeet Raipurkar

[11] Basheer Mohamad Al-Maqaleh and Saleem Khalid Shaab," An Efficient Algorithm for Mining Association

[12] Rules using Confident Frequent Itemsets," 2012 Third International Conference on Advanced Computing & Communication Technologies.

[13] Mihir R. Patel, Dipti P. Rana, and Rupa G. Mehta," FApriori: A Modified Apriori Algorithm Based on Checkpoint,"IEEE 2013.

[14] Anand H.S. and Vinodchandra S.S.," Applying Correlation Threshold on Apriori Algorithm," 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013).

[15] A. Lekha, Dr. C V Srikrishna and Dr. Viji Vinod," Utility of Association Rule Mining: a Case Study using Weka Tool,"2013 IEEE.

[16] Yubo Jia, Guanghu Xia, Hongdan Fan, Qian Zhang and Xu Li, "An Improved Apriori Algorithm Based on Association Analysis," ICNDC 2012, 3rd IEEE International Conference, pp208-211.