International Institution for Technological Research and Development
Volume 1, Issue 3, 2016

Survey on Methods for Finding Association rule for Web Mining

Kamalben Patel [1], Dr. Shyamal Tanna [2]

[1] Information & Technology Department, LJIET
Ahmedabad, Gujarat Technological University, India
patelkamal093@gmail.com

[2]Ass.Prof.PG -Department, LJIET
Ahmedabad, Gujarat Technological University, India
drsmtanna@gmail.com

Abstract

In this paper, we examine varius methods for frequent item set sequences for Web Usage data like of Apriori and FP-growth algorithm**.**The frequent item set mining has been performed with the help of Apriori algorithm. It gives us different frequent item set mining results with their support count. Apriori reads file once for every iteration. Where FP-Growth is the fact that the algorithm only needs to read the file twice. In FP-Growth resulting FP-tree is not Unique for the Same "logical" database the process needs two complete scans of the database. That's why use DynFP-Growth for finding frequent item set is more efficient

Keywords: web usage mining, apriori algorithm, FP-growth algorithm, minimum support, association rule.

1. Introduction

Today growth of World Wide Web increases the complexity for users to browse effectively. To increase the performance of web sites better web site design, web server activities are changed as per users' interests decide which products to purchase. User's behavior is used in different applications such as Personalization, e-commerce, to improve the system and to improve the system design as per their interest etc.

Web personalization offers many functions such as simple user salutation to more complicate such as content delivery as per users interests. Content delivery is very important since non- expert users are overwhelmed by the quantity of information available online. It is possible to anticipate the user behavior by analyzing the current navigation patterns with patterns which were extracted from past web log. Usage mining techniques are very useful to focus customer attraction, customer retention, cross sales and customer departure. System Improvement is done by understanding the web traffic behavior by mining log data so that policies are developed for Web caching, load balancing, network transmission and data distribution.

1.1 Web Mining

It is the application of data mining techniques to discover patterns from the World Wide Web. Web mining has mainly three types.

• **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 captures the identity of Web users along with their browsing behavior at a Web site.

• **Web Structure Mining:** Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

• **Web Content Mining:** Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization.

2.Related Works

Many authors presented different pre-processing techniques to improve performance for finding data. This section comprises a brief mention about some methods proposed by different authors.

Neha Goel , Dr. C.K.Jha [1] discussed Web usage mining consists of the following phases: Data extraction or Data Collection, Preprocessing, Pattern Discovery and Pattern Analysis .The objective of this paper is to discuss the preprocessing involved in Web usage mining. A tool has been designed for pre-processing.

Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya [2]discussed an approach to identify web link patterns which has been developed from web log and analysis of patterns is presented .The frequent item set mining has been performed with the help of Apriori algorithm. It gives us different frequent item set mining results with their support count. Web link sequences below support threshold are pruned. We found different frequent item set mining results by varying minimum support (2%-3%).Association rules miner give all the possible rules with their confidence and Lift. Using the knowledge of the web site structure and the behavior of the site's visitors, we analyzed the pruned rule set from the user 's point of view and proposed actions that a webmaster may decide to take based on knowledge extracted from rules in order to enhance a website and improve visitor's browsing experience.

- **Web log Extractor Module:** They developed this web extractor to extract the IPs and Web links from a web log file.
- **Data Pre-processing Module**: Data pre-processing module produces an input file for APRIORI which contains entry of different navigational profile. Here each IP works as user id and web links as item sets.
- **Apriori or FP Growth Algorithm Module**: If a set cannot pass a test, all of its superset will fail the same test as well
- **Association Rule Generation Module**: This module finds out the association rules in between frequent mining pattern results.
- **Information**: Knowledge is derived with the help of extraction of rule those satisfy minimum confidence, below this minimum confidence other rules has been ruled out.

The process for finding frequent item should be done following way.

(a) It assigns each different IP a different transaction number and add all different item sets (web links) accessed by this IP to the particular transaction.

(b) Check for every entry that if IP already exists in list then simply add web link to that navigational profile otherwise declare it a new navigational profile. For a particular navigational profile, if a web link already exists in profile then no need to add this repeated web link to navigational profile.

(c) Write each different transaction in a new line with its accessed item sets (web links). So, now it produces File (2) containing entry of transactions.

(d) We have got File (2) as an output of Data pre- processing module. Fed this File (2) to Apriori produces the different item set mining results. It calculates Support count for every item set mining result. It counts support for 1-itemset, 2-itemset…k- item set. Value of k depends on max number of item sets those may occur together in a transaction.

(e) Write these all different web links (item set) mining result to a File (3).

(f) Now the next task is to find out patterns and analysis of patterns with the help of different association rules. Rule miner generates rule according to minimum confidence given. Our purpose here to find some strong patterns those are occurring as an output of rule miner. Finally, get some strongly occurring patterns on the basis of knowledge we got.

Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya[3] in second paper they are compare Apiori and FP-Growth algorithm for finding frequent itemsets.

Apriori algorithm based on Hash-based items counting, when a k-item set whose corresponding hashing count is below the threshold cannot be frequent .It uses bottom-up search approaches that in step generates a frequent sequences of length n, all 2n subsequence's have to be produces. Apriori algorithm implies that in any item-set that is potentially frequent in database must be frequent in at least one of the partition of database (DB). FP-Growth firstly creates the root of the tree, labeled with "null". FP-Growth scans the database D a second time (First time when scanned, it crate l-itemset and then Ll), whenever the same node is encountered in another transaction, we only increment the support count of the common node. This transforms the problem of mining frequent patterns in database to that of mining the FP-tree. Table I. gives a comparison between the two algorithms based on various factors.

S.Thangarasu , D.Sasikala[4] In this paper extracting intentional knowledge from both the structure and content of the XML document. This knowledge will improve the query answering time and easy access of the content in the XML instead of scanning entire documents. In this paper, the intentional knowledge is generated from the frequent structure of the XML document. Here is chance that the interested information may sometimes be eliminated. So the future work includes to generate the knowledge, the documents are classified according to their type and the knowledge is generated from the categorized documents. As the XML documents are classified, it includes all the information without eliminating any information as in the existing system. This will also reduce the time required to get the intentional knowledge.

Wang Yan,Le Jiajin, Huang Dongmei[5] Privacy preserving in web data mining has arisen worldwide concerns with the promotion of network technology and the demand of application. But there are many drawbacks and open questions. In this paper, they presented a method for privacy preserving mining of association rules based on web usage mining. First, they gave the general framework for mining association rules in the web usage mining, generated session sets by exploring user sessions and transfer session sets to relation two-dimension table. Second, they proposed secondary random response column replacement (SRRCR), a simple and effective privacy preserving algorithm, and achieve privacy protection association rule mining based on SRRCR. Finally, they presented experimental results that validated the algorithm (SRRCR) in practice by simulation. In the future, we will enhance the efficiency of mining algorithm further by parallelization and other methods, and combine SRRCR with other privacy preserving ways to achieve more significant improvements in terms of privacy, accuracy, efficiency, and applicability.

3.Figures

3.1 Figures

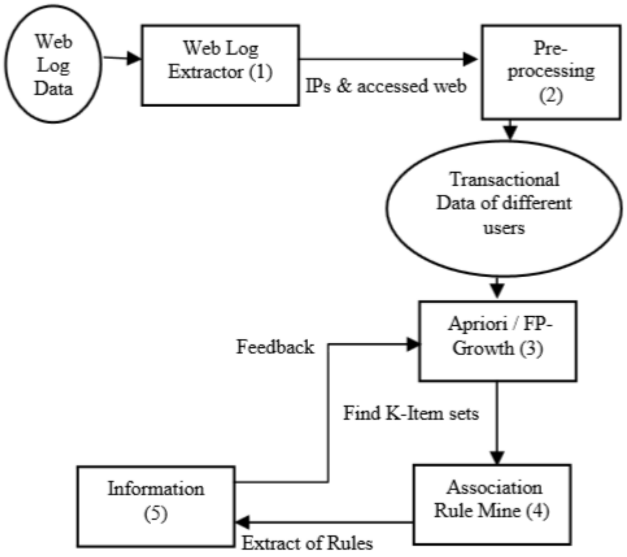

Figure 1 proposed model [2]

4.Conclusion

The frequent item set mining has been performed with the help of Apriori algorithm. It gives us different frequent item set mining results with their support count. Apriori reads file once for every iteration. Where FP-Growth is the fact that the algorithm only needs to read the file twice. In FP-Growth resulting FP-tree is not Unique for the same “logical” database The process Needs two comple te scans of the database . That’s why use DynFP-Growth for finding frequent item set is more efficient.

References

Examples follow:

Proceedings Papers:

[1] Neha Goel , Dr. C.K.Jha "Preprocessing Web logs: A Critical phase in Web Usage Mining ", International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, ISSN- 1530 9866,pp. 672 – 676,2015

[2] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya ,"Association Rule Mining for Web Usage Data to Improve Websites", International Conference on Advances in Engineering & Technology Research (ICAETR),2014

[3] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya"An Empirical Analysis and Comparison of Apriori and FP- Growth Algorithm for Frequent Pattern Mining ", International Conference on Advanced Communication Control and Computing Technologies (lCACCCT) ,2014

[4] S.Thangarasu , D.Sasikala "Extracting Knowledge from XML Document Using Tree-based Association Rules", International Conference on Intelligent Computing Applications(ICICA),2014

[5] Wang Yan,Le Jiajin, Huang Dongmei, "A Method for Privacy Preserving Mining of Association Rules based on Web Usage Mining", International Conference on Web Information Systems and Mining,2010

[6] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases" In Proceeding of the ACM SIGMOD International Conference on Management of Data (ICMD), USA, pp. 207–216, 1993.

[7] Cooley, R., Mobasher, B., and J. Srivastava, "Grouping web page references into transactions for mining World Wide Web browsing patterns" Proceeding of the IEEE Knowledge and Data Engineering Exchange Workshop (KDEEW), Newport Beach, CA, pp 2-9,1997.

[8] E.R. Omiecinski, "Alternative interest measures for mining associations in databases" IEEE Transactions on Knowledge and Data Engineering, vol.15, Issue 1, pp. 57-69, 2003.

[9] L.Cristofor and D.Simovici, "Generating an informative cover for association rules" Proceeding of the IEEE International Conference on Data Mining (ICDM), Boston, USA, pp. 597-600, 2002.

[10] A.Saleem Raja, E.George and Dharma Prakash Raj "MAD- ARM: Mobile Agent based Distributed Association Rule Mining" IEEE International Conference on Computer Communication and Informatics (ICCCI) Coimbatore, pp. 1- 5, 2013.