Discovering Generalized Association Rule in Web Usage Mining by FP Tree

Han Ni Ni Myint Thu, Khine Khine Oo
University of Computer Studies, Yangon, Myanmar
hanninimyintthu1993@gmail.com, k2khine.@gmail.com

Abstract

Web mining techniques can use to search for web access patterns, web structures, regularity and dynamics of web contents. Web usage mining analyzes Web log files to discover user accessing patterns of Web pages. Log file data can offer valuable insight into web site usage. It reflects actual usage in natural working condition, compared to the artificial setting of a usability lab. This paper presents web log mining based on hierarchy of web usage data by generalized association rule. Multi-level association rule will be used for implementation of generalized association rule. In this system, Web log database is used to store web log records of log files collected from web server. And web log database are constructed via a process of data cleaning, data transformation. By using FP tree, the system generates rules from web log data, will reduce counting phase of association rule since it stores the pre-computed count values. Frequent patterns are generated instead of page item-sets. The generated frequent patterns can later be applied to improve web site management, decision making process.

KEYWORDS: *Web Log, Association rules, FP tree*

1. Introduction

Web plays an important role and medium of information dissemination. There is a need for data log to track any transaction of the communications. Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Web usage mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web.

Web servers register a (web) log entry for every single access they get in which they save the URL requested, the IP address from which the request originated. Web log data are extracted from web log files and need to be cleaned and transformed. And then these web log data are loaded into a data warehouse as a data cube format for mining multilevel based patterns.

Association rule is applied to get relationship between each dimension of web log data. In this system, mining association rules in hierarchical database will be applied in web usage mining.

2. Related Works

Agrawal and Srikant [4] observed an interesting downward closure property, called Apriori, among frequent k-itemsets: A k-itemset is frequent only if all of its sub-itemsets are frequent. This implies that frequent itemsets can be mined by first scanning the database to find the frequent 1-itemsets, then using the frequent 1-itemsets to generate candidate frequent 2-itemsets, and check against the database to obtain the frequent 2-itemsets. This process iterates until no more frequent k-itemsets can be generated for some k. This is the essence of the Apriori algorithm.

In [1], discovering usage pattern from web log data by association rule mining is presented. SOTrieIT algorithm and Apriori algorithm are used for mining association rule. In their paper, association rule is generated for browsing patterns of web page. Those rules can later be applied recommender system and predicting user behaviors. Weakness of Apriori algorithm is time consuming.

In this paper, Apriori algorithm is applied to data cube so it reduce the counting step of the algorithm since the data cube store pre-computed value.

In [2], discovering association rules from OLAP data cube with daily downloads of Folklore Materials is presented. In their system, association rules based on dimensions instead of browsing page are generated. It is fast in processing since data cube stores pre-computing counts. Analysis is emphasized for all types of documents rather than browsing pages. Besides, only four dimensions of documents are analyzed in their system.

In this thesis, web server log files are used to discover the association rules and present the relationship of the dimension of these log file.

In [3], association rule mining method on OLAP Cube is presented. In their system, different types of OLAP database, ROLAP, MOLAP and HOLAP are presented. Then student information database are used to discover association rule from data cube. It will give the frequent items and rules from the data cube.

In this thesis, web mining is described and emphasize on web usage mining. Raw log files are preprocessing and stored in the warehouse database. Data cube is created from the warehouse database and find the association rules from it so overcomes the problem of Apriori's algorithm.

3. Background Theory

3.1. Web Mining

Web Mining is the application of applying data mining techniques on web data to discover knowledge. Web mining also known as web data mining. There are three kinds of web mining.

3.1.1 Web content mining

Web content mining, also known as text mining, is generally the second step in web data mining. Content mining is the scanning and mining of text, pictures and graphs of a web page to determine the relevance of the content to the search query.

3.1.2 Web structure mining

Web structure mining focuses on the hyperlink structure of the web. The challenge for web structure mining is to deal with the structure of the hyperlinks within web itself. Web structure can be treated as a part of web content so that web mining can instead be simply classified into web content mining and web usage mining.

3.1.3 Web usage mining

Web usage mining – web log mining, reveal the knowledge hidden in the log files of a web server. Application of data mining techniques on large web log repositories to discover useful knowledge about user's behavioral patterns and website usage statistics that can be used for various website design tasks. Web usage mining has mainly focused on the analysis of usage patterns recorded in the web usage log of web servers.

3.2 Process of Web Usage Mining

Pre-processing: The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. Since the origin web logs data sources are blended with irrelevant information, data pre processing acts as an important steps to filter and organize only appropriate information before presenting to any web mining algorithm. The inputs to the preprocessing phase are the server logs. The outputs are the user session file, transaction file. Preprocessing includes following steps:

- Data Cleaning
- User Identification
- Session Identification

Pattern discovery: Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns. In this system pattern discovery, is performed by generating frequent of web usage by association rule algorithm.

The preprocessed log is then passed to a mining application for pattern discovery. Usage mining algorithm is used to discover patterns. In this system multi-level association rule mining by FP-Growth algorithm will be used. The importance of discovered patterns depends on statistical measures like support and confidence which are usually computed by the mining application. These statistical measures provide a first interestingness filter because they allow the setting of thresholds that must be met or exceeded for pattern discovery.

Pattern Analysis: This process targets to understand visualize and give interpretation to these patterns. Pattern Analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of Web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily.

3.2.1. Data Cleaning

Elimination of the items deemed irrelevant by checking the suffix of the URL name such as gif, jpeg, GIF, JPEG, jpg, JPG. Since every time a Web browser downloads a HTML document on the Internet, several log entries such as graphics and script are downloaded too. In general, a user does not explicitly request all the graphics that are in the web page, they are automatically down-loaded due to the HTML tags. Since web usage mining is interested in studying the user's behavior, it does not make sense to include file requests that a user does not explicitly request. The HTTP status code returned in unsuccessful requests because there may be bad links, missing or temporality inaccessible pages, or unauthorized request etc: 3xx, 4xx, and 5xx.

3.2.2. User Identification

Unique users must be identified using heuristic methods that can be used to help identify unique users. If the IP address is the same, a reasonable assumption to make is that each different agent type for an IP address represents a different user. User's IP addresses of two consecutive entries are compared. If the IP address is the same, both the records are considered from the same user.

Input: N records of web log file
Output: User sets identified
Algorithm:
Repeat steps

1. Compare ip address of first log entry with ip address of second log entry.
2. If both are same compare the user agent of both entries else assume as different users.
3. If both user agents are same identify both entries are from same user.

3.2.3. Session Identification

After the preprocessing, the log data are partitioned into user sessions based on IP and duration. Most users visit the web site more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. The individual pages are grouped into semantically similar groups. A user session is defined as a relatively independent sequence of web requests accessed by the same user. If a user stays inactive for a period longer than the max_idle_time, subsequent page requests are considered to be in another episode, thus identificated as another session. Web access sessions are identified based on IP address and time-out does not exceeding 30 minutes for the same IP Address. A new session is created when a new IP address is encountered after a timeout.

4. The System Overview

This thesis will present web usage mining by generalized association rules. Generalized Association Rule – Rules based on Taxonomy / Hierarchy of product or page in this web usage mining Taxonomy and hierarchies are defined based on attributes of web log records such as browsers, referrer, status, URL and so on. Web server logs are preprocessed including following steps: Cleaning, User Identification and Session Identification. Then those preprocessed log records are stored in the transaction database. Then multi-level association rule is applied based on taxonomy of web logs. Frequent patterns are generated as output of the system as shown in figure1.

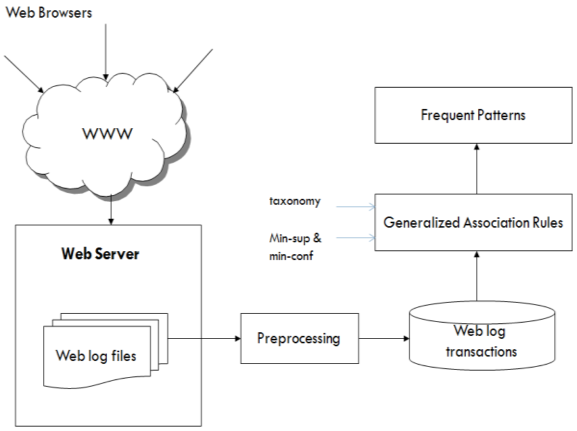


Figure 1. System Overview

5.Implementation of the System

At the start of the system, web log must be loaded to be analyzed. In the **cleaning** phase, image url is removed. Image url is detected using its file extension whether it is .gif, .jpg, .bmp, .png, etc. for example, If (url. ends with(".gif"), remove the sentence. Web log records (with date, access time, client ip, url, method, status) are stored in the memory.

In the **user identification** phase, records with same client ip are considered to be one user and those records are grouped together. For example, log records with client ip address 192.168.0.1 are assumed to be access by a single user and those records are grouped together.

The main idea of **session identification** is that, access records with different times may be different sessions. Hence, in the session identification phase, records with same time group from user group are grouped together again to get the sessions. For example, client ip 192.168.0.1 with different access time records are assumed to be different sessions.

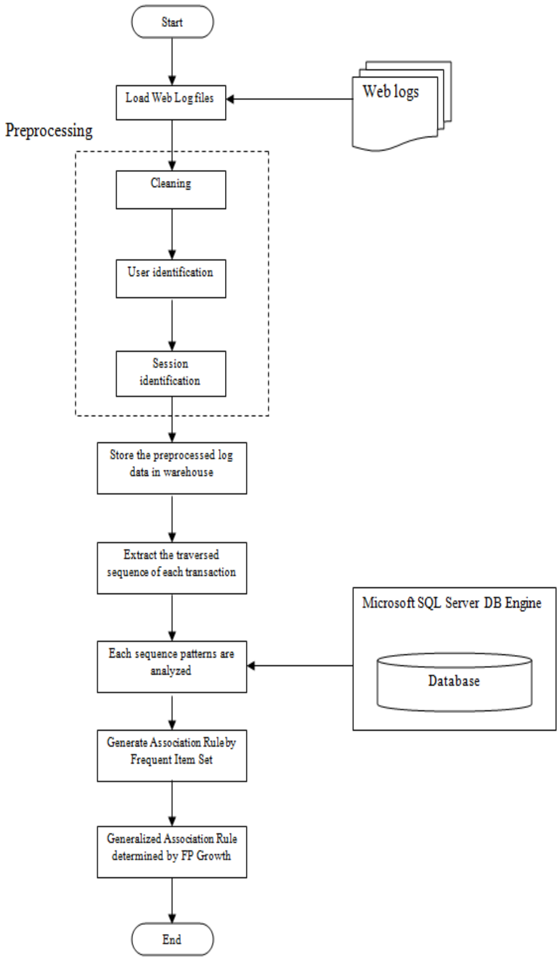
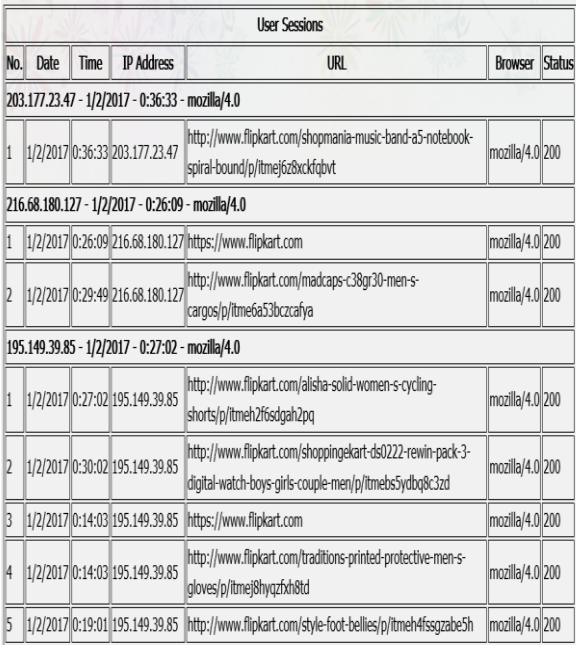

Start
Load Web Log files
Web logs
Preprocessing
Cleaning
User identification
Session identification
Store the preprocessed log data in warehouse
Extract the traversed sequence of each transaction
Microsoft SQL Server DB Engine
Database
Each sequence patterns are analyzed
Generate Association Rule by Frequent Item Set
Generalized Association Rule determined by FP Growth
End

Figure 2. The Process Flow Diagram

5.1. Preprocessed Web Log Sessions

After the preprocessing steps, the web log records are extracted from web log file as follows. Example web log records are shown in table 1.

Table 1. Weblog table



User Sessions						
No.	Date	Time	IP Address	URL	Browser	Status
203.177.23.47 - 1/2/2017 - 0:36:33 - mozilla/4.0						
1	1/2/2017	0:36:33	203.177.23.47	http://www.flipkart.com/shopmania-music-band-a5-notebook-spiral-bound/p/itmej6z8xckfqbvt	mozilla/4.0	200
216.68.180.127 - 1/2/2017 - 0:26:09 - mozilla/4.0						
1	1/2/2017	0:26:09	216.68.180.127	https://www.flipkart.com	mozilla/4.0	200
2	1/2/2017	0:29:49	216.68.180.127	http://www.flipkart.com/madcaps-c38gr30-men-s-cargos/p/itme6a53bczcafya	mozilla/4.0	200
195.149.39.85 - 1/2/2017 - 0:27:02 - mozilla/4.0						
1	1/2/2017	0:27:02	195.149.39.85	http://www.flipkart.com/alisha-solid-women-s-cycling-shorts/p/itmeh2f6sdgah2pq	mozilla/4.0	200
2	1/2/2017	0:30:02	195.149.39.85	http://www.flipkart.com/shoppingekart-ds0222-rewin-pack-3-digital-watch-boys-girls-couple-men/p/itmebs5ydbq8c3zd	mozilla/4.0	200
3	1/2/2017	0:14:03	195.149.39.85	https://www.flipkart.com	mozilla/4.0	200
4	1/2/2017	0:14:03	195.149.39.85	http://www.flipkart.com/traditions-printed-protective-men-s-gloves/p/itmej8hyqzfxh8td	mozilla/4.0	200
5	1/2/2017	0:19:01	195.149.39.85	http://www.flipkart.com/style-foot-bellies/p/itmeh4fssgzabe5h	mozilla/4.0	200

5.2. Generalized Association Rule (GAR)

Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. For each level, any algorithm for discovering frequent item-sets may be used. In this system, FP-Growth algorithm is used. Generalized association rule is one of the commonly used web usage mining technique. Concept hierarchy is used to illustrate the relationship between options provided by technician user (site admin or the one who knows their domain very well). Concept hierarchy shows the set of relationship between different items, generalized association rules allow rules at different levels. Generalized association rules were applied to mine the useful patterns counting. From the server logs, hierarchy of the websites is determined. Comparing with the standard association rules, generalized association rules allow rules at different levels.

5.3. GAS in Web Usage Mining

As more organizations view the Web as an integral part of their operations and external communications, interest in the measurement and evaluation of Web site usage is increasing. Server logs can be used to glean a certain amount of quantitative usage information. Compiled and interpreted properly, log information provides a baseline of statistics that indicate usage levels and support and/or growth comparisons among parts of a site or over time. Such analysis also provides some technical information regarding server load, unusual activity, or unsuccessful requests, as well as assisting in marketing and site development and management activities.

5.4. FP-Growth in GAR

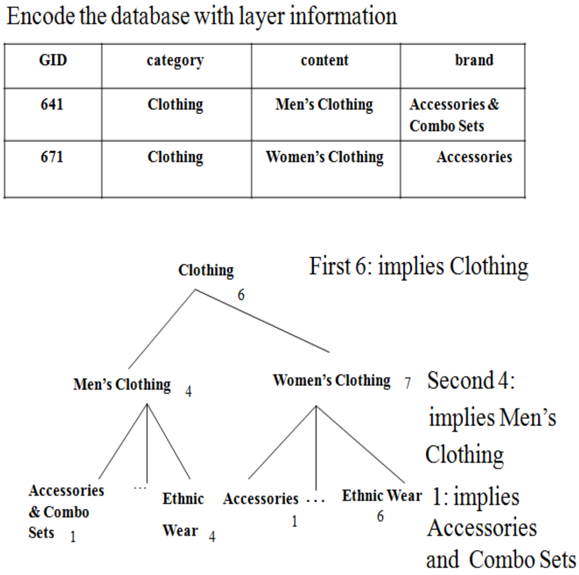
Generalized Association Rule by FP-Growth has following processes –

1. **Encoding Transaction Table** – The first step of GAS is the encoding process which assigns the number to each item in the taxonomy. Pages are encoded with 1, 2, 3, … according to level they stand in the preprocessing steps of web usage mining.
2. **Building Primitive Level FP-Tree**
3. **Generate candidate high-level frequent item sets** through merging and grouping the frequent item sets at primitive level
4. **Calculate the support of each candidate**, and select only ones satisfying the minimum support;

5.4.1. FP-Growth in GAR – Step 1 (Encoding)

Generalized database stores hierarchy information encoded transaction table instead of the original transaction table. Therefore, each item is encoded as a sequence of digits in transaction table In GAR, Pages are at the Primitive level in the hierarchy and level of taxonomy is determined by user input. In generalized association rules, FP-Growth is utilized to mine from primitive items to get frequent item-sets at primitive level. The encoding scheme is used to encode the items according to the concept hierarchy, which uses encoded string to represent a position in a hierarchy. In this system, encoded string is used in building FP-tree and the description is only needed for final display. Encoding process is done according to Encode Tables [Table 2].

Table 2. Tables for Encoding



Encode the database with layer information

GID	category	content	brand
641	Clothing	Men's Clothing	Accessories & Combo Sets
671	Clothing	Women's Clothing	Accessories

Clothing
6
First 6: implies Clothing
Men's Clothing 4
Women's Clothing 7
Second 4: implies Men's Clothing
Accessories & Combo Sets 1
…
Ethnic Wear 4
Accessories …
1
Ethnic Wear
6
1: implies Accessories and Combo Sets

5.4.2. FP-Growth in GAR – Step 2 (Building FP-Tree for Primitive Level)

Normal FP-Growth: In building header table and tree, concepts are not considered; only how many times that page is navigated (count of visit); regardless of browser, method, status and so on.

FP-Growth with Concept (GAS): In case of taxonomy, web pages including taxonomies are counted and saved in header table and FP tree.

Figure 3 shows an FP-tree at the atomic level (level 0), and suppose the support threshold of level 1 is 3. Let's show the process to generate FP (1)-tree.

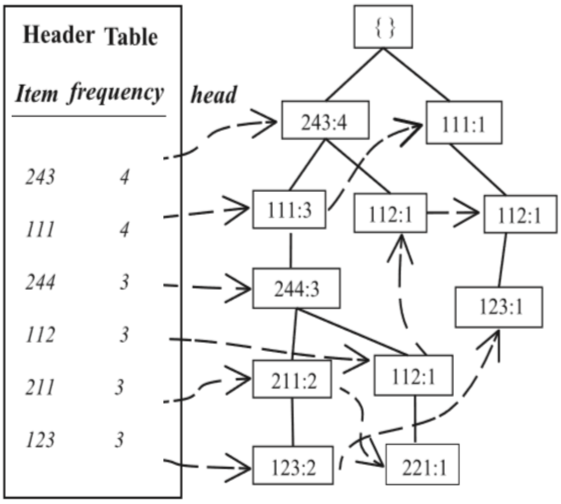


Figure 3. Building FP Tree

First, transform the form of items to concept level 1, for example, both items “243” and “244” in Figure 2 should be changed to “24*”; and then remove the repeated nodes, such as the nodes “24*:4” and “24*:3” are repeated in the same path of FP-tree and the node “24*:4” is the top one, we will keep the node “24*:4” and remove the other one “24*:3”. After that, the changed FP-tree is shown below:

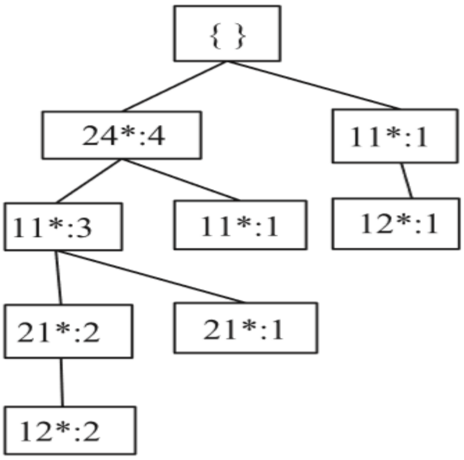


Figure 4. Uncompleted FP (1)-tree (step1-2)

Obviously, some inclusive paths need to be merged as shown in Figure 4. Such as the two paths derived from node “24*:4”, and the two paths derived from node “11*:3”. We merge them by merging the corresponding nodes. The identical nodes in different paths are merged through summing the support count of the corresponding items. For example, the nodes “11*:3” and “11*:1” can be merged as “11*:4”. So after step 3 the changed FP-tree is illustrated in Figure 5:

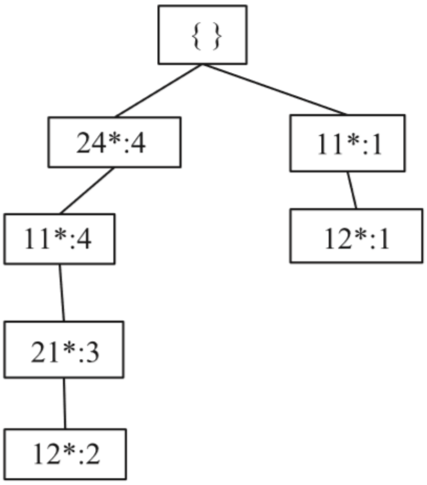


Figure 5. Uncompleted FP(1)-tree (step 3)

The next step is to update the header table as described in steps 4 and 5. And then the nodes in the FP-tree are sorted as the item order in new header table shown in Figure 5. After sorting, the inclusive paths are generated again, such as nodes “11*:4” and “11*:1” can be merged to “11*:5”. All items satisfy the support threshold, so there is no need to remove any items in header table or nodes in FP-tree. The final FP(1)-tree is shown in Figure 6.

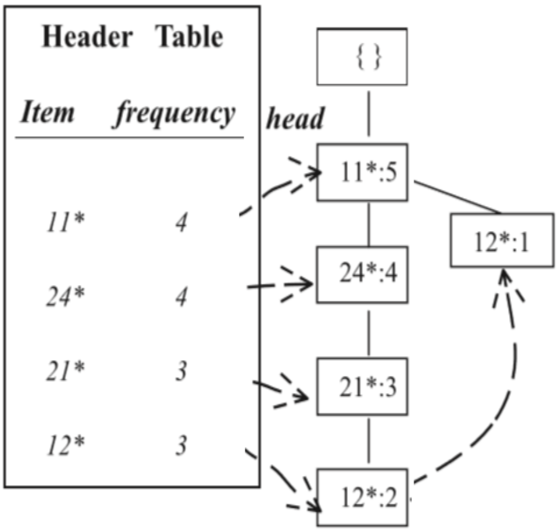


Figure 6. FP (1)-tree (constructed by algorithm CFP)

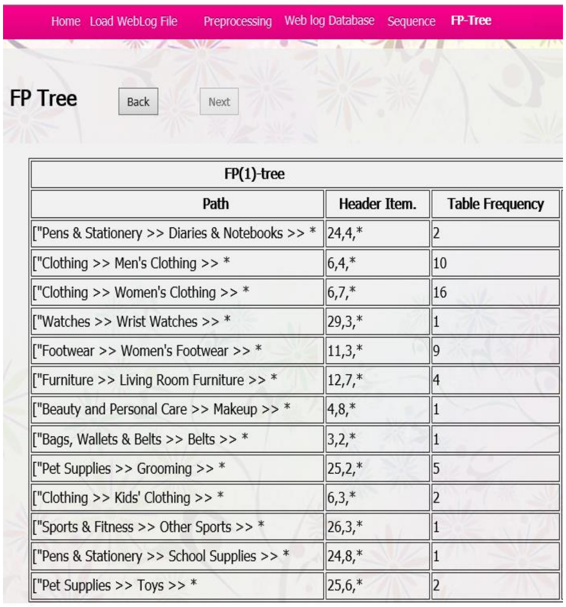
5.4.3. FP-Growth in GAS – Step 3 Detail Explanation (Generating High level frequent item-sets)

Step 3 is processed by the following steps:

1. Change all the items in the header table and all the nodes in FP-tree of primitive level into the form of level l.
2. For each path in FP(l)-tree, if there are repeated nodes, then the top one will be maintained and the others will be removed.
3. Merge the inclusive paths by accumulating support counts of the identical nodes.
4. Remove the duplicate items in header table, and set the frequency of the remaining items by accumulating the support count of relative nodes in the FP(l)-tree.
5. Sort the items in the header table in descending order of frequency.
6. Rearrange the nodes in FP(l)-tree with the same order of the items as appeared in the

header table, and merge the inclusive paths as done in step 3.

7. For the item whose frequency does not satisfy the support threshold of level l, not only remove it from header table, but also remove the relative nodes from FP (l)-tree.
8. Adjust the node-links.



FP Tree

FP(1)-tree		
Path	Header Item.	Table Frequency
["Pens & Stationery >> Diaries & Notebooks >> *	24,4,*	2
["Clothing >> Men's Clothing >> *	6,4,*	10
["Clothing >> Women's Clothing >> *	6,7,*	16
["Watches >> Wrist Watches >> *	29,3,*	1
["Footwear >> Women's Footwear >> *	11,3,*	9
["Furniture >> Living Room Furniture >> *	12,7,*	4
["Beauty and Personal Care >> Makeup >> *	4,8,*	1
["Bags, Wallets & Belts >> Belts >> *	3,2,*	1
["Pet Supplies >> Grooming >> *	25,2,*	5
["Clothing >> Kids' Clothing >> *	6,3,*	2
["Sports & Fitness >> Other Sports >> *	26,3,*	1
["Pens & Stationery >> School Supplies >> *	24,8,*	1
["Pet Supplies >> Toys >> *	25,6,*	2

Figure 7. System Generated Association Rules

8. Conclusion

Discovering the association rule is an important data mining function. This system presents generating association rule patterns from weblog data. In order to analyze the web usage patterns, association rule mining algorithm is applied to weblog. Relationships of web server log are displayed as output in this system. Provide information to better accommodate the website based on user's needs. By the study of the result, the web system can be reorganized the web site structure to access the associated link without delay. Current trend of the web user can be analysed.

References

[1] Hsu Mon Thet Wai, "Discovery of user access pattern from weblog based on Association Rule Mining", 2010

[2] Galina Bogdanova, Tsvetanka Georgieva, "Discovering the Association Rules in OLAP Data Cube with Daily Downloads of Folklore Materials", International Conference on Computer Systems and Technologies, 2005.

[3] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques". Simon Fraser University.

[4] Jigna J. Jadav and Mahesh Panchal, "Association Rule Mining Method On OLAP Cube", In proceeding of International Journal of Engineering Research and Applications, Vol. 2, Issue 2, pp.1147-1151, Mar-Apr 2012.

[5] Agrawal R, Srikant R, "Fast algorithms for mining association rules", in Proceedings of the 1994 international conference on very large data bases (VLDB'94), Santiago, Chile, pp 487–499, 1994.

[6] Brin S, Motwani R, Ullman JD and Tsur S, "Dynamic itemset counting and implication rules for market basket analysis", in Proceeding of the 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 255–264, 1997.

[7] Cheung DW, Han J, Ng V, Fu A and Fu Y, "A fast distributed algorithm for mining association rules", In: Proceeding of the 1996 international conference on parallel and distributed information systems, Miami Beach, FL, pp 31–44, 1996

[8] Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, "Frequent pattern mining: current status and future directions", Data Min Knowl Disc (2007) 15:55–86, DOI 10.1007/s10618-006-0059-1, 2007.

[9] Park JS, Chen MS and Yu PS, "An effective hash-based algorithm for mining association rules", in Proceeding of the 1995 ACM-SIGMOD international conference on management of data (SIGMOD'95), San Jose, CA, pp 175–186, 1995.

[10] Sarawagi S, Thomas S, Agrawal R (1998) Integrating association rule mining with relational database systems: alternatives and implications. In: Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD'98), Seattle, WA, pp 343–354, 1998.

[11] Savasere A, Omiecinski E and Navathe S, "An efficient algorithm for mining association rules in large databases", in Proceeding of the 1995 international conference on very large data bases (VLDB'95), Zurich, Switzerland, pp 432–443, 1995.

[12] Toivonen H, "Sampling large databases for association rules", In: Proceeding of the 1996 international conference on very large data bases (VLDB'96), Bombay, India, pp 134–145, 1996.

[13] Yinbo WAN, Yong LIANG and Liya DING, "Mining Multilevel Association Rules From Primitive Frequent Itemsets", Journal of Macau University of Science and Technology, Vol.3 No.1, 2009