ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Web Usage Mining: Web log Pre-processing and Online Visitor's frequent Pattern Discovery

Aruna Kumari G K, Sudheer Shetty

M. Tech Scholar, Dept of CSE, SCEM, visveswaraiah Technological University, Adyar, Mangalore, India

Associate Professor, Dept of CSE, SCEM, visveswaraiah Technological University, Adyar, Mangalore, India

ABSTRACT: Web Usage Mining (WUM) is one of the data mining techniques to discover the knowledge or information apparent in the web log file, such as user access patterns from web log data and for analyzing behavioural patterns of users. Association rules are used as a main technique to find the frequent item set in data mining. Apriori algorithm can be required to produce large number of candidate sets. To procreate the candidate sets, it requires several scans over the database. Apriori gains more memory space during the candidate generation process. This paper introduces a new Modified Reverse Apriori algorithm in which an Apriori algorithm can be enhanced .The Modified Reverse Apriori algorithm is one of the new approach for frequent pattern generation. It generates large frequent item sets which are to be started by considering a maximum number of total attributes in the dataset. It is efficiently considering a maximum combination of all the itemsets in pairs and then it generates a huge frequently mined set of items based on the condition; also satisfy the user defined support. If it satisfies, then it decreases gradually the number of items in the item sets unless it obtains the largest set of frequent items.

KEYWORDS: Web usage mining, web log, web log processing, association rule, frequent pattern.

I. INTRODUCTION

Now a day a Web is a distributed, very largest dynamic data repository, global information service [1] to provide the services such as advertisements, news, financial management, education, government, e-commerce etc. It contains web pages information, user accessing hyperlink information and usage information. Web Usage mining (WUM)[3] is a web mining technique which is used to discovery and analysis of web usage pattern from web logs. It is also called web log mining. Web Usage mining (WUM) is one of the method of identifying user's browsing patterns, with the means of knowledge gained from web logs over web. The results of the WUM can be used in[5] improving the speed of the system, modification of the site, web personalization, usage characterization, business intelligence etc.
Web mining has three important categories:
They are
1. Web content mining
2. Web structure mining and
3. Web-usage mining.

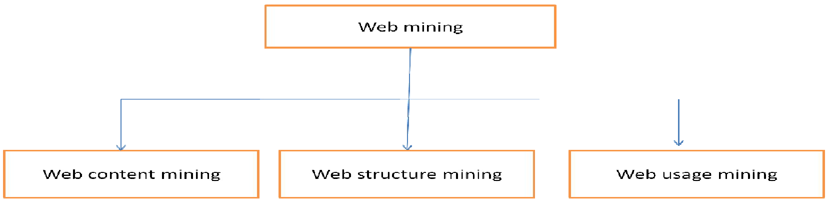


Fig 1: Web mining categories

- Web content mining is used to mine semi-structured and structured data and extract the useful information/knowledge from the contents of web page. Applications include, automated e-mail routing and reply, customer support and knowledge management, such as content categorization, document clustering and keyword extraction and associations.
- Web structure mining focuses on the interior document structure in which discovering the hyperlink's link structure at the inter-document level pages. Web-structure mining focuses on analyze the link structure of the Web to identify interesting associations and patterns describes the connectivity of documents in the Web. Such associations are then used to retrieve relevant documents in response to user requests. and
- Web usage mining is used to discover the user access patterns from web server log files [7], which record the activities of the user while they are browsing and navigating through the web.

1.2 Web Usage Mining

Web usage mining is one of the methods of extracting useful information from server logs. Web Usage Mining [4] is one of the applications of data mining techniques to discover the pattern and analysis of interesting usage patterns from Web log data, in order to improve web based applications. Web Usage mining consist of four phases such as Data collection, Pre-processing of web log data, Pattern Discovery and Pattern analysis.

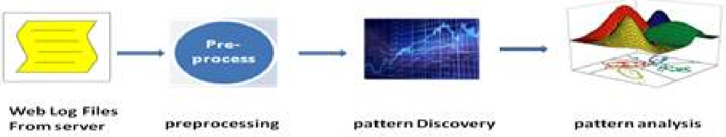


Fig 2: Web Usage Mining phases

In **first Phase** the data is collected form Web Log files which is semi-structured. In Web Usage Mining, data can be collected from different sources like server logs, browser logs, proxy logs, or obtained from an organization's database.

The **Second Phase** [7] Data Pre-processing is required to eliminate irrelevant information from original log file and to make the web log file easy for Session and user identification process. The main purpose of preprocessing is to improve the quality and accuracy of data.

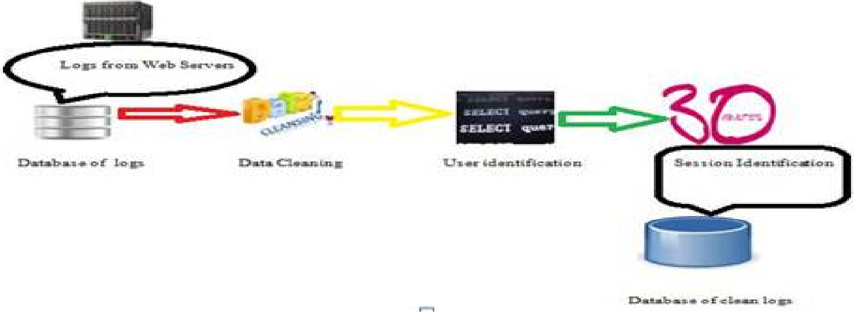


Fig 3: Data pre-processing steps

Next and **Third Phase** of [5]Web usage mining is Pattern Discovery which means discovering patterns from pre-processed data using various data mining techniques like association, clustering , Statistical analysis and so on.

In the **last phase** of WUM, Pattern analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

The proposed system is used to discover useful pattern from servers weblog file. The web log file contains the information about website visitor's activity. By implementing the apriori [11] and Modified reverse apriori algorithm on web log file which gives frequently accessed web pages and unique users. The application of the project is working as world wide .This usage pattern extracted from web log file can be applicable to wide range of applications like web patronization, system improvements, site modification, business intelligent discovery etc.

II. RELATED WORK

Fayyad et. al [7] Focused on the feature and structure of server web log Files. They discussed on Log file Structure, Format, location and Type. Log file contains information on user's request while surfing through the web site. It contains noisy, irrelevant, ambiguous data. Pre processing removes the irrelevant, noisy data and resolves inconsistencies. Data pre-processing is a step that performs filter and organize information before applying to web discovery algorithms. They proposed algorithms for field extraction and weblog data cleaning.

S.Prakash and R.M.S. Parvathi [9] focused Association rule with scaling Apriori mining that defines a new effective protocol suite. It gives normal prediction sequences compared with equal to those consider by the set up association rule with the priority of confidence.

The new protocol suite is smaller than the set up association rule, specifically considering smaller minsup.

Kumar et. al [10] focused on collection of log data at a web server on discover the usage patterns of website from log files . They implements important three faces of web usage mining apriori algorithm generate association rule that interrelates the client's usage pattern for particular website. They implements FP-Growth algorithm. Apriori is implement to operate the database contains set off transaction.

Sheila A. Abaya et. al [11] proposed algorithm the improved Apriori that introduces different factors such as set size and size frequency which are used to remove non significant candidate keys. It is implemented to management of the memory efficiency and apriori intersection algorithm complexity. This algorithm improves execution efficiency.

Agrawal et al. [5] proposed Apriori algorithm. Apriori is more efficient algorithm during the process candidate generation. This algorithm uses a breadth-first search schemes to count the itemsets support and also uses a candidate generation module. Apriori had pruning techniques to avoid from measuring particular item sets, while guaranteed completeness. Apriori having merit in which any subset of a frequent item sets is also a particular frequent item set. It is taking more time and memory for the process of candidate generation. While generating the candidate set it needs multiple scanning over the database.

Association rule mining is used to find out association rules in which that satisfy the previously defined minimum support and a confidence from a given transactional database.

Association rule mining is to be used to discovery and correlation analysis that can find web pages types groups that are commonly accessed together (to discover correlation or relationship between pages types found in a server web log).

III. PROPOSED ALGORITHM

The proposed research system concentrates on web usage mining to track the user's behavioural patterns searching either a website or web page. The information on navigation paths, which is available in log files. It depicts how the people navigate the internet. It is a dynamic environment, where a variation can found in exactly any point of time. The browsed information on navigation paths is updated in log file .The raw data contains noise, missing values and inconsistency. So Log file must be pre-processes the data for the purpose of enhancing the efficiency and quality of data. The Association rules have been used to discover the usage pattern of web visitor. It helps in searching interesting pattern and associations between the weblinks attributes.

A. *Design Considerations:*

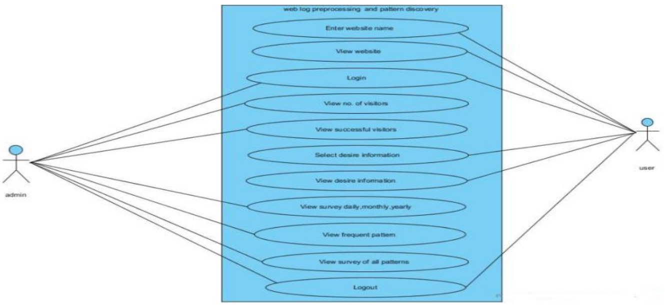


Fig 4: Use case diagram

Fig 4 shows the use case diagram. There are two entities admin and user.

B. *Description of the Proposed Algorithm:*

Implementation details

Link pages in a particular site

pid	**userid**	**pages**
p1	192.168.90.5	h.jsp
p2	192.168.90.2	c.jsp
p3	192.168.90.3	b.php
p4	192.168.90.4	s.jsp
p5	192.168.90.5	X.jsp

Tid	**Pages accessed**				
	1	2	3	4	5
1	P1		P3	P4	
2		P2	P3		P5
3	P1	P2	P3		P5
4		P2			P5

Linking pages

C	L
TID	Links in a page
1	1 , 3 , 4
2	2 , 3 , 5
3	1 , 2 , 3 , 5
4	2 , 5
Min_sup=2	

Link pages in a particular itemset

1-itemset

items	links
1	2
2	3
3	3
4	1
5	3

Apriori result

C3s	L3s
Itemset	links
2 3 5	2

Frequent 3-ItemSet L2
{2 3 5}

Reverse Apriori

- **Generating 5 combinations for minimum required support from 1, 2,3,4,5.**

5-Itemsets Support		transactions
1,2,3,4,5	0	0

Outcome: 5-Itemsets does not satisfy the minimum support min sup, so 4-Itemset combinations should be generated.

- **Generating 4 combinations for minimum required support from 1,2,3,4,5**

4-Itemset Support		transactions
1,2,3,4	0	0
1,2,3,5	1	T3
1,2,4,5	0	0
1,3,4,5	0	0
2,3,4,5	0	0

Outcome: 4-Itemsets does not satisfy the minimum support, so 3-Itemset combinations should be generated.

- **Generating 3 combinations for minimum required support from 1,2,3,4,5**

3-Itemset Support		transactions
1,2,3	1	T3
1,2,4	0	0
1,2,5	1	T3
1,3,4	1	T1
1,3,5	1	T3
1,4,5	0	0
2,3,4	0	0
2,3,5	2	T2,T3
2,3,5	0	0
3,4,5	0	0

Outcome:

3-Itemset Support		
2,3,5	2	T2,T3

- **Generating 2 combinations for minimum required support from 1,2,3,4,5**

3-Itemset Support		transactions
1,2	1	T3
1,3	2	T1,T2
1,4	1	T3
1,5	1	T1
2,3	2	T2,T3
2,4	0	0
2,5	3	T2,T3,T5
3,4	4	T1
3,5	2	T2,T3
4,5	0	0

- **Generating 1 combinations for minimum required support from 1,2,3,4,5**

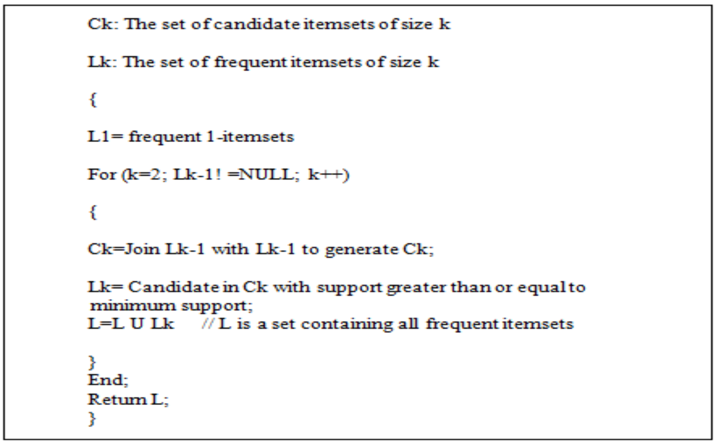
3-Itemset Support		transactions
1	2	T1,T3
2	3	T2,T3,T4
3	3	T1,T2,T3
4	1	T1
5	3	T2,T3,T5

Frequent 3-ItemSet is {2 3 5}

III. Pseudo code

A. *Apriori Algorithm*

Apriori Algorithm:
Input: Transactional Database, min-supp-Minimum Support Threshold
Output: L-Large item set



```
Ck: The set of candidate itemsets of size k

Lk: The set of frequent itemsets of size k

{

L1= frequent 1-itemsets

For (k=2; Lk-1! =NULL; k++)

{

Ck=Join Lk-1 with Lk-1 to generate Ck;

Lk= Candidate in Ck with support greater than or equal to
 minimum support;
L=L U Lk     // L is a set containing all frequent itemsets

}
End;
Return L;
}
```

B. *Finding Frequent Itemsets using Modified Reverse-Apriori algorithm*

The modified Reverse apriori algorithm

The proposed approach checks the user specified minimum item support and then only it generates large frequent item sets. It then decreases one by one the number of items in the item sets, so it gets the largest frequent item sets. Reverse approach begins with the largest occurrences of collective total attributes from a database.

These collective attributes are tested against the minimum support for the associated rule and is thereby selected for the next level or pruned from the subsequent levels.

Input:
Database D, Minimum Support
Output: Large Frequent Item sets
Method:

1. x= Total number of attributes
2. a=0 ,cti=0
3. For all combinations of (x-a) number of attributes
4. Do
5. Generates Candidate key CD(x-a) and Frequent item FI(x-a) combined and find set of TID for each individual sets(by scanning only x-a set)
6. where, support count of generated itemsets >=min_support
7. Put in Frequent item by making union operation according to their TID and delete that candidate itemset(x-a)
8. If successful, then go to step 11
9. Else
10. a=a+1 and repeat step 3 till 1-itemset is true
11. Return sets of large frequent itemsets
12. End

IV. Simulation Results

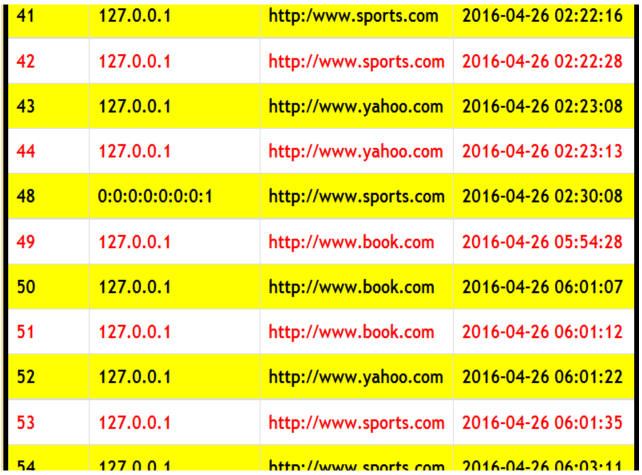
1) Browsing and visiting web sites and stores browsed data to web log files.
Web server log is a file which is created and maintained by the web server.

2) Preprocessing the logs:
The weblog created by the web server contains details of different users' requests. It contains a lot of irrelevant, Noisy and incomplete data. Pre-processing involves removing such data.

3) Convert the of log file in to database table:
The weblog is not directly used for data mining. The whole dataset is converted to a database. This creating a Database and then import the log file to the MySQL database table.

6) Association rules:
Apriori and Modified reverse apriori algorithms
Association rules show correlation among different items. In case of Web mining, an example of an association rule is the correlation among accesses to various web pages on a server by a given client.

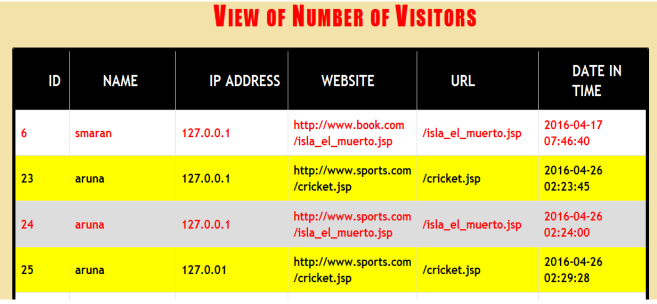
The fig 5 shows the user viewing different site and log into the web site .Which information has to be stored in the log file. By accessing the log file contents to find the number of visitors visited to the different web sites is considered as number of hits.



41	127.0.0.1	http:/www.sports.com	2016-04-26 02:22:16
42	127.0.0.1	http://www.sports.com	2016-04-26 02:22:28
43	127.0.0.1	http://www.yahoo.com	2016-04-26 02:23:08
44	127.0.0.1	http://www.yahoo.com	2016-04-26 02:23:13
48	0:0:0:0:0:0:0:1	http://www.sports.com	2016-04-26 02:30:08
49	127.0.0.1	http://www.book.com	2016-04-26 05:54:28
50	127.0.0.1	http://www.book.com	2016-04-26 06:01:07
51	127.0.0.1	http://www.book.com	2016-04-26 06:01:12
52	127.0.0.1	http://www.yahoo.com	2016-04-26 06:01:22
53	127.0.0.1	http://www.sports.com	2016-04-26 06:01:35
54	127.0.0.1	http://www.sports.com	2016-04-26 06:03:11

Fig5: Number of hits

The fig 6 shows the users those who are login into the particular site and viewing different information means he/she is the number visitor visit in to the particular site.

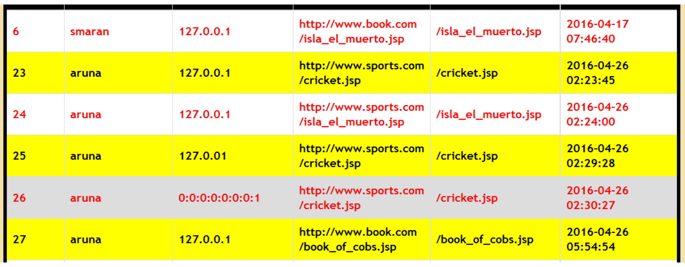


VIEW OF NUMBER OF VISITORS

ID	NAME	IP ADDRESS	WEBSITE	URL	DATE IN TIME
6	smaran	127.0.0.1	http://www.book.com /isla_el_muerto.jsp	/isla_el_muerto.jsp	2016-04-17 07:46:40
23	aruna	127.0.0.1	http://www.sports.com /cricket.jsp	/cricket.jsp	2016-04-26 02:23:45
24	aruna	127.0.0.1	http://www.sports.com /isla_el_muerto.jsp	/isla_el_muerto.jsp	2016-04-26 02:24:00
25	aruna	127.0.01	http://www.sports.com /cricket.jsp	/cricket.jsp	2016-04-26 02:29:28

Fig 6:Number of visitors

The fig 7 shows the users those are login into the particular site means he/she is the successful visitor visit in to the particular site.



6	smaran	127.0.0.1	http://www.book.com /isla_el_muerto.jsp	/isla_el_muerto.jsp	2016-04-17 07:46:40
23	aruna	127.0.0.1	http://www.sports.com /cricket.jsp	/cricket.jsp	2016-04-26 02:23:45
24	aruna	127.0.0.1	http://www.sports.com /isla_el_muerto.jsp	/isla_el_muerto.jsp	2016-04-26 02:24:00
25	aruna	127.0.01	http://www.sports.com /cricket.jsp	/cricket.jsp	2016-04-26 02:29:28
26	aruna	0:0:0:0:0:0:0:1	http://www.sports.com /cricket.jsp	/cricket.jsp	2016-04-26 02:30:27
27	aruna	127.0.0.1	http://www.book.com /book_of_cobs.jsp	/book_of_cobs.jsp	2016-04-26 05:54:54

Fig7: Successful visitors

The result of pattern discovery and analysis helps to improve the system performance and to modify the web site. It helps to attract the visitors and to give the personalized services to regular user. The result of such analysis might include: most recent visit per page, who is visiting which page, the frequency of use of each hyperlink, and most recent use of hyperlinks.

V. Conclusion and Future Work

Preprocessing include data cleansing, user identification, session identification. Data cleaning which is very useful and reduces the size of web log file and also improves the quality of contents in the log file. Association rule mining is a one of the data mining algorithm and used for extracting knowledge and updating the information. The Apriori algorithm, it generates the item sets by using large item sets of previous scan without considering the transactions in the database. The main drawback of Apriori algorithm like requires more time and takes large number of scans which are required to mine the frequent item sets and generation of candidate set is very costly. The drawbacks are overcome by proposing Modified reverse apriori algorithm in such a way that it takes less time and less number of scans than the apriori algorithm. In future work, classification algorithms can be used for finding frequent itemsets. It is recommended for further research work to use enhanced algorithms in order to produce association rules mining. It and also make use of evolutionary algorithms for clustering the sessions, and use neural networks in order to learn the navigation patterns of users.

REFERENCES

[1] L. Wang, J. Fan, L. Liu, H. Zhao," Mining Data Association Based on a Revised FP-Growth Algorithm," *Proc. of the 2012 InternationalConference on Machine Learning and Cybernetics, Xian*, July 2012.

[2] Bamshad Mobasher, "Data Mining for Web Personalization" The Adaptive Web, LNCS] 6. Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

[3] V.Chitraa, Dr. Antony Selvdoss Davamani,2010 "A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No. 3.

[4] R.Agrawal, T.Imielinski, and A.Swami, 1993. "Mining association rules between sets of items in large databases", in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216.

[5] R. Agrawal and R. Srikant. *Fast algorithms for mining association rules.* IBM Research Report RJ9839, IBM Almaden Research Center, San Jose, California, June 1994.

[6] R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.

[7]Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.

[8] Goswami D.N., Chaturvedi Anshu.,Raghuvanshi C.S.," *An Algorithm for Frequent Pattern Mining Based On Apriori*", In: Goswami D.N. et. al. / (IJCSE) International Journal on Computer Science and Engineering ,,Vol. 02, No. 04, 2010, 942-947, ISSN : 0975-3397

[9] S. Praksh, R.M.S. Parvathi 2010. "An enhanced Scalling Apriori for Association Rule Mining Efficiency", European Journal of Scientific Research, vol. 39, pp.257-264, ISSN: 1450-216X.

[10] Kumar, B.S. and Rukmani, K.V., 2010, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms" *International Journal of Advanced Networking and Applications*, Vol. 1, Issue 6, pp. 400-404

[11] Sheila A. Abaya, "*Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation*", In: International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012