ISSN:0975-9646

Web Usage Mining: An Incremental Positive and Negative Association Rule Mining Approach

Anuradha veleti[#], T.Nagalakshmi[*]

[#]*Department of computer science and engineering*
Aurora's Technological and Research Institute
Hyderabad, A.P, India
[*] *Sr.Associate.Prof, Dept. of Computer Science & Engineering*
Department of computer science and engineering
Aurora's Technological and Research Institute
Hyderabad, A.P, India

***Abstract*— Web usage mining (WUM) integrates the techniques of two popular research fields – Datamining and the Internet. Web usage mining is the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular website. At present usage of Association Rules in web usage mining claiming a wide research scope. Mining positive and negative association rules in web usage data become a hot spot. In this paper we proposed an incremental algorithm (IPNAR) that mines positive and negative association rules in web usage data. The incremental based algorithm incrementally update web log association rules by utilizing the metadata of old database transactions as well as old mined rules, performs single scan over the dataset , and it overcomes the limitations of other mining methods. When comparing with the other existing algorithms, Incremental algorithm is highly efficient, reduced number of passes over the database, reduce the number of non-interesting negative rules and it will find all the association rules quickly .**

***Keywords*—Web mining, Web usage mining, pattern discovery, Positive association rules, Negative association rules, support-confidence framework, IPNAR Algorithm.**

I INTRODUCTION

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: Web Contents Mining and Web Usage Mining and Web Structure Mining. Web Contents Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. Web Structure Mining is used to examine data related to the structure of a particular Web site, i.e., the topology present in the web site structure and hidden relations present between pages in the tree like structures of web sites.

Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by web servers. I web mining applies the data mining to the web data and traces users visiting behavior, and then extracts the users' using pattern. In Agrawal et al.(1993) the problem of mining association rules was first outlined. The concept was applied to the supermarket data Implement association rules to on-line shopper can generally find out his/her spending habits on some related products. For example, if a transaction of an on-line shopper consists of a set of items, while each item has a separate URL. Then the shopper's buying pattern will be recorded in the log file, and the knowledge mined from which, can be the form like the following:

- 30% of clients who accessed the web page with URL/company/products/bread.html, also accessed /company/products/milk.htm.
- 40% of clients who accessed /company/announcements/special.html, placed an online order in/company/products/products1.html

After that the association rules was also adopted for the analysis of web site traffic. The resulting association rules indicate which pages are often requested together. With this information it is possible to forecast the next pages a visitor will frequent. we can easily trace the behavior of users visit by combination of the positive and negative association rules in practical applications. In previous work many of the researches find that Negative association rules A=>¬B (or ¬A=>B, ¬A=>¬B) plays an important role on e-commerce and reconstruction of web site.. For mining positive and negative association rules in web log usage data the existing algorithms need extra scan i.e. multiple scans are required which would consume extra resources and the accuracy of the results is of a lower degree. In this paper we introduced IPNAR algorithm that incrementally updates web log association rules by utilizing the metadata of old database transactions as well as old mined rules and it performs single scan, consuming lesser resources and the accuracy of the results is of a higher degree.

The organization of this paper is given as follows. Section (II) describes the main objective of this paper. Section (III) revisits the overview of web usage mining, Section (IV) describes the positive and negative association rules. Section(V) IPNAR algorithm and experimental data .Conclusion and future work in Section (VI).

II OBJECTIVE

Our proposed model should incrementally update web log association rules by utilizing the metadata of old database transactions as well as old mined rules. By implementing the incremental approach over the dataset the association rules incrementally updated and single scanning takes place over the dataset in order to generate the rules, so no need of multiscanning over the dataset in order to generate the rules. One major contribution of work should be the technique for efficiently using position codes of small items in database itemsets to restore information about previous small items that were not stored in earlier scan, when the database is updated and these items should become frequent, without re-scanning old database. In order to avoid multiple scans and multiple passes over to the database an incremental approach is implemented

III WEB USAGE MINING

Web usage mining system should be able to gather useful usage data thoroughly, Filter out irrelevant usage data, establish the actual usage data, Discover interesting navigation patterns, Display the navigation patterns clearly, Analyze and interpret the navigation patterns correctly, and apply the mining results effectively. Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data. Web server log files were used initially by the webmasters and system administrators for the purposes of "how much traffic they are getting, how many requests fail, and what kind of errors are being generated", etc. However, Web server log files can also record and trace the visitors' on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as "from what search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors?"

Web log file is one way to collect Web traffic data. After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining techniques such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted

The web usage mining generally includes the following several steps: data pretreatment and knowledge discovery and pattern analysis.

A. Data Preprocessing

The portions of Web usage data will exist in serverlogs, referral logs etc.The information that contains in these logs needs to be integrated in order to form a dataset for mining. Before the integration of data files need to be cleaned/filtered, using techniques like filtering the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units. The data preprocessingperforms

B. Pattern Discovery

After pre-processing phase, Data Mining methods and algorithms should be applied to user sessions and transactions identified before. These methods must be sometimes slightly modified to adapt themselves to the particularities of Web data. There are many types of analysis to be performed on this data: simple statistical analysis, traversal path analysis, association rules discovering, sequential patterns finding, clustering and classification of pages or paths, etc.

C. Pattern Analysis

Pattern Analysis is the third phase of Web Usage Mining .The task of this stage is to remove irrelevant rules or patterns and extract the interesting rules or patterns from results of pattern discovery. The current patternanalysis. Methods and tools include SQL query mechanism, OLAP and visualization.

IV DESCRIPTION OF POSITIVE AND NEGATIVE ASSOCIATION RULES

Let D be a database of transactions. Each transaction consists of a transaction identifier and a set of items {i1, i2 , ...,in} selected from the universe I of all possible descriptive items Let D be a database of transactions.

Transaction	Items
1	Gmail,Yahoo,Facebook
2	Gmail,Facebook
3	Gmail,Rediff
4	Yahoo,Twitter,Orkut

Table 1: *Database with 4 transactions*

In the above table the items represents the users visited Websites. In web usage mining each transaction represent the user visited web pages. A positive association rule is an expression of the form: X =>Y where X□ I, Y□I and X ∩ Y = Ø. Each association rule is characterized by means of its support and its confidence defined as follows:

$$\text{Supp}(X \Rightarrow Y) = \frac{\text{Number of transactions containing } (X \cup Y)}{\text{Total number of transactions}}$$

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)}$$

From the above example, rule Gmail=> Facebook has support 50% and confidence 66.7%. According to the above measures, the support measure can be considered as the percentage of database transactions for which (X*U*Y) evaluates to true. The confidence measure is understood to be the conditional probability of the consequent given the antecedent. Association rule mining essentially boils down to discovering all association rules having support and confidence above user-specified thresholds, minsup and minconf, for respectively the support and the confidence of the rules. For example, from the 100% confidence of the rule Gmail=>Orkut.We

can conclude that page Gmail is always visited in combination with page Orkut. As the forward, confidence of this rule is 0% we can deduct that page Orkut is always visited before page Gmail.

In the dataset their exists other association rule: $X \rightarrow \neg Y, \neg X \rightarrow Y, \neg X \rightarrow \neg Y$ The rule $X \rightarrow \neg Y$ means the data objects which have itemsets X do not have the itemsets Y. The rule $\neg X \rightarrow Y$ means the data objects which do not have itemsets X have the itemsets Y. The rule $\neg X \rightarrow \neg Y$ means the data objects which do not have itemsets X do not have the itemsets Y. These rules can be called negative association rules. The rule $X \rightarrow Y$ can be called positive association rule. In the existing paper researchers expressed their views on negative association rule in web usage mining is that negative association rule is very useful to the web site administrator to adjust the page structure and business decision making from the Web usage Log, resolve the lack of past only researching positive association rules, makes the user's access pattern that is mined more objective and comprehensive. In order to calculate the support and confidence for negative association, we can compute the measures through those of positive rules.

1) supp ($\neg A$) = 1-supp(A);
2) supp(A U $\neg B$) = supp(A)-supp(A U B);
3) supp($\neg A$ U B) = supp(B)- supp(A U B);
4) supp($\neg A$ U $\neg B$) = 1- supp(A)- supp(B) + supp(A UB)
5) conf(A => $\neg B$)

$$= \frac{\text{supp(A) - supp(A U B)}}{\text{1-supp (A)}} = 1 - \text{conf(A=>B)};$$

6)conf($\neg A$=>B)

$$= \frac{\text{supp(A) - supp(A U B)}}{\text{1-supp (A)}}$$

$$= \frac{\text{supp (B)- supp(A) * conf (A => B) ;}}{\text{1 - supp(A)}}$$

7) conf($\neg A$=>$\neg B$)=

$$\frac{\text{1 - supp(A) - supp(B) + supp(A U B)}}{\text{1 - supp (A)}} = \text{conf}(\neg A \Rightarrow B)$$

V IPNAR ALGORITHM AND EXPERIMENTAL DATA

A. IPNAR algorithm for mining positive and negative association rules:

IPNAR algorithm consists of two phases:

1) Phase 1 discusses about the generation of frequent item sets

2) Phase 2 discusses about the IPNAR algorithm for mining association rules (both positive and negative) and for incrementally updating the association rules

B. Generation of frequent itemsets

By using the apriori model, support and confidence framework we can generate the frequent itemsets. A frequent itemset is an itemset that meets the user-specified ms. Accordingly we define an infrequent itemset as an itemset that does not meet the user-specified ms. we can find the frequent itemsets from candidate itemsets, Generate all itemsets that support is greater than the minimum support.

C. Experimental data and generating frequent itemsets

Let's take a record database the record consists of 6 related record users visit pages on e-commerce site.

ID	UsersClickingSequence
1	Cosmetics.,books,jewellery,computer,toys
2	Cosmetics,books,computers,toys
3	Cosmetics,jewellery,computer
4	Footwear,jewellery,computer
5	Cosmetics,books,jewellery,toys
6	Cosmetics,footwear,books,jewellery,computer

From the above table we need generate the frequent itemset by using the Apriori algorithm, for that we need set minsupp=0.3 minconf=0.4.

L1

Itemsets	Support
Cosmetics	0.83
Footwear	0.33
books	0.83
Jewellery	0.83
computers	0.83
toys	0.5

L2

Itemsets	support
Cosmetics, Books	0.83
Cosmetics, Jewellery	0.67
Cosmetics, Computers	0.67
Books, Jewellery	0.67
Books, Computers	0.67
Books, Toys	0.5
Jewellery, Computers	0.67
Cosmetics, Toys	0.5
Footwear, Jewellery	0.33
Footwear, Computers	0.33
Jewellery, Toys	0.33
Computers, Toys	0.33

L3

Itemsets	support
Cosmetics, Books, Jewellery	0.67
Cosmetics, Books, Computers	0.67
Cosmetics, Books, Toys	0.5
Cosmetics, Jewellery, Computers	0.5
Books, Jewellery, Computers	0.5
Cosmetics, Jewellery, Toys	0.33
Cosmetics, Computers, Toys	0.33
Books, Jewellery, Computers	0.33
Books, Computers, Toys	0.33
Jewellery, Computers, Toys	0.33
Footwear, Jewellery, Computers	0.33

L4

Itemsets	Support
Cosmetics, Books, Jewellery, Computers	0.5
Cosmetics, Books, Jewellery, Toys	0.33
Cosmetics, Books, Computers, Toys	0.33

D. IPNAR Algorithm

In the algorithm we have assumed that the frequent items obtained and stored in sets L. After generating the frequent itemsets we can implement the IPNAR algorithm for mining positive and negative association rules. In this algorithm we can mine the positive and negative association rules by using the PNARC model algorithm and made some modifications. By using the incremental approach for mining positive and negative association rules we can reduce multiple passes to the database.

E. Algorithm Design

Algorithm 1: Incremental Positive and Negative Association rules(IPNAR)

/Input: L,frequent itemsets,DB, ms, mc, Ims, Imc, respectively a set of transactions, minimum support, minimum confidence, Increment minimum support, Increment minimum confidence

//Output: IAR: Incremental Positive and Negative Association Rules.

(1) positiveAR ←Ø ; negativeAR ←Ø ;
(2) scan the database and find the set of frequent 1-itemset (L1)
(3) PositiveAR←PositiveAR U L1
(4) for (k = 2; Lk-1←Ø; k + +){
(5) Ck= Lk-1 ∞ L k-1
(6) for each i∈Ck {
(7) s=supp(i)
8) if s □ms then{
9) Lk←Lk U{i}
10) }
11) PositiveAR←PositiveAR U Lk
12) ELSE{
13) negativeARk←negativeARk U{i}
14) negativeAR←negativeAR U negativeARk
/Generate positive association rules
15) for each expression A U B= i and A∩B= Ø {
16) corrA,B=supp(A U B)/(supp(A) * supp(B))
17) if corrA,B >1 then
18) if conf(A → B) ≥ mc then
19) positiveAR ← positiveAR U { A → B }
20) }
// Generate negative association rules
21) for each expression A U B= i and A∩B= Ø{
22) corrA,B=supp(A U B)/(supp(A) * supp(B))
23) if corrA,B <1 then
24) negativeAR ← negativeAR U { A=>¬B }
25) negativeAR ← negativeAR U {¬ A=>B }
26) }
27) if supp(A U B) □Ims and conf(A □¬B) □Imc then
28) negativeAR ← negativeAR U { A □B }
29) if supp(¬A U B) □Ims and conf(¬A □ B) □Imc then
30) negativeAR ←negativeAR U { ¬A □B }
31) }
32) }
33) }
34) IPNAR ←positiveAR U negativeAR
35) Return IPNAR

F. Experimental Results

Based on the incremental approach we will get the positive association rule Cosmetics, Books, Jewellery→ Computer, conf(Cosmetics, Books, Jewellery→ Computer) =0.75 Is an effective positive association rule, so then the page jewellery is directly linked to page computers.

IPAR,INAR	Confidence
Cosmetics→ Books	1.0
Footwear → Jewellery	0.8
Cosmetics, Books, Jewellery →Computers	0.75
Jewellery→ ¬Toys	0.6
Computers→ ¬Toys	0.6

From the above table supp(Jewellery→¬Toys) =0.5>0.3 and supp(computer→ ¬toys)=0.6 >0.3 meets the minsupp requirements and conf(Jewellery→¬Toys)=0.6>0.4, conf(Computer→¬Toys)=0.6>0.4 meets the minconf requirements. So these are the effective negative association rules, so we directly delete the link from page Jewellery to page Toys and from page Computer to Toys. By using the IPNAR algorithm we incrementally mine the positive and negative association rules.

VI CONCLUSION AND FUTURE WORK

In this paper we proposed an incremental based approach algorithm (IPNAR) for positive and negative association rule mining in web usage data .These algorithm not only generating positive and negative association rules but also incrementally updating the association rules. The main advantage of incrementally updating association rules is it avoids multiple scans over the dataset i.e. whenever the dataset is updated ,the whole transaction also participates in Rule mining, So multiples scans are required for rule mining. In order to avoid multiple scans through the dataset we recommended an incremental approach.. By implementing Incremental approach no extra scan is required; within a single scan it is going to mine the positive and negative association rules and multiple passes to database is reduced. This IPNAR algorithm is very effective, efficient and it will improve the search space when compared to the other existing algorithms, giving quick results. In future work we will present improved algorithm by using different techniques and measures for mining positive and negative association rules in web usage data

ACKNOWLEDGMENT

I would like to thank , Sr. Associate.Prof. Nagalakshmi and Prof.(Head of Department) Sujatha (Aurora's Technological and Research Institute Hyderabad, India) for their careful reading and valuable suggestions for the improvement of the paper's presentation. I would also like to thank the anonymous referees for their helpful comments, correction and suggestions to improve this work.

REFERENCES

[1] Huang hao, Wang jianjun. Research of Web usage mining [J] Computer Systems Applications. 2008, 01.

[2] Siriporn Chimphlee, Naomie Salim,etc. Using Association Rules and Markov Model for predict next access on web usage mining T.Sobh and K.Elleithy (eds.), Advances in Systems, Computing Sciences and Software Engineering, 371–376.

[3] Dong, X..,Niu,Z..,Shi,X..,Zhang, X., Zhu, D.., Mining Both Positive and Negative Association Rules from Frequent and Infrequent Itemsets[J] .

[4] Chen M SPark J SYu P S. Efficient data mining for path traversal patterns in a Web environment[J].IEEE Trans on Knowledge and Data Eng1998,10(2):385-390.

[5] Antonie, M-L., Zaiane, O.: Mining Positive and Negative Association Rules: An Approach for Confined Rules, Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD04), LNCS 3202, Springer-Verlag Berlin Heidelberg, Pisa, Italy (2004) 27-38.

[6] Kotsiantis S, Kanellopoulos D., Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[7] Dong, X., Sun, F., Han, X., Hou, R. Study of Positive and Negative LNAI 4093, Springer-Verl ag Berlin Heidelberg, 2006: 100-109

[8] Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer- Aided Industrial Design and Conceptual Design, 2008.CAID/CD 2008. 9th International Conference on 22-25 Nov.2008

[9] O. Daly and D. Taniar, "Exception Rules Mining Based On Negative Association Rules", *Lecture Notes in Computer Science*, Vol. 3046, 2004,pp 543–552.

[10] X. Wu, C. Zhang and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules", *ACM Trans. on Information Systems*, vol.22(3), 2004, pp 381–405.

[11] Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference on 22-25 Nov.2008

[12] Jaideep Srivastava and Robert Cooley Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD Explorations Newsletter .2000,1(2):12-23.