Prediction of Users Behavior through Correlation Rules

Navin Kumar Tyagi
Marathwada Institute of Technology
Bulandshahr, UP, India

A. K. Solanki
Meerut Institute of Engineering and Technology
Meerut, UP, India

***Abstract*— Web usage mining is an application of Web mining which focus on the extraction of useful information from usage data of severs logs. In order to improve the usability of a Web site so that users can more easily find and retrieve information they are looking for, we proposed a recommendation methodology based on correlation rules. A correlation rule is measured not only by its support and confidence but also by the correlation between itemsets. Proposed methodology recommends interesting Web pages to the users on the basis of their behavior discovered from web log data. Association rules are generated using FP growth approach and we used two criteria for selecting interesting rules: Confidence and Cosine measure. We also proposed an algorithm for the recommendation process.**

Keywords- Web usage mining; FPgrowth; Cosine measure; Usability; Association rules.

I. INTRODUCTION

The ease and speed with which information exchange and business transactions can be carried out over the Web has been a key driving force in the rapid growth of the Web. Recommendation systems have become popular among users in World Wide Web environment. Web sites generates huge amount of usage data which consists useful information about the users behaviour.Automatic discovery of user access patterns from server log is known as web usage mining . The term web usage mining was introduced by Cooley in 1997.Data mining techniques such as association rules, sequential patterns, clustering and classification can be used to analyze the web site usage data. Association rules mining is one of the important and widely used data mining technique. It is highly successful technique for extracting useful information from very large databases [1, 2, 3, and 4]. In web environment, HTTP server log contains historical user sessions. Web sessions reflect user behavior while navigating throughout a web site and considered as an important source of information about users. Association rules shows similarities between web pages derived from user behavior, can be utilized in Recommender systems. The main objective of such recommendation is to suggest web pages which are useful for the user. Proposed system generates association rules from web log data and then correlation analysis is performed to obtain interesting rules. Pages visited by a user are matched with the antecedent of the rules and consequents of matching rules become the recommendations. In this way proposed system can enhance the usability of the site.

This paper is organized as follows. In section II association rule mining and correlation analysis are presented. In section III we proposed a Methodology and algorithm to predict web pages for the users. An example is presented in section IV. We evaluated the performance of proposed system through example in section V. Section VI presented some related work and conclusion is given in section VII.

II. ASSOCIATION RULES MINING

Association rules [5] are used to show the relationship between data items. These uncovered relationships are not inherent in the data. Association rules are frequently used by retail stores to assist in marketing, advertising, floor management, and inventory control. An association rule A→B represents a relationship between itemsets A and B and it is characterized by two measures, support and confidence. The support of the rule is the percentage of transactions in the database that contain AUB and confidence or strength of the rule is the ratio of the number of transactions that contain AUB to the number of transactions that contain A.

Association rule mining can be viewed as two-step process. In first step frequent itemsets that satisfies a minimum support are generated from the transactional database and in second step strong association rules that satisfies minimum confidence are generated.Apriori[6] can be used to generate frequent itemsets, but it can suffer from two nontrivial costs[3] . It may need to generate a huge number of candidate sets and it may also need to repeatedly scan the database and check a large set of candidates by pattern matching.

An interesting method FP-growth can be used to generate frequent itemsets without candidate generations. This method works on divide and conquers strategy. It compresses the database representing frequent itemsets into a frequent pattern tree or FP tree, which retains the itemset information. It then divides the compressed database into a set of conditional databases; each associated with one frequent item and mines each such database separately. FP growth algorithm is efficient and scalable for mining long and short frequent patterns and is about an order of magnitude faster than the apriori algorithm. It is also faster than Tree-Projection algorithm, which recursively projects a database into a tree of projected databases. To generate association rules from frequent patterns, following steps are to be performed.

For each frequent itemset l, generate all nonempty subsets of l.

- For every nonempty subset s of l, generate the rule s→ (l-s) if support (l)/support(s) ≥ Min_conf, where Min_conf is the minimum confidence threshold.

A. Correlation analysis

An association rule is interesting or not can be assessed either subjectively or objectively. The user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another. However objective interestingness measures based on the statistics behind the data can be used to extract uninteresting rules. Support and confidence measures are insufficient to filter out uninteresting rules as confidence of rule A→B is only an estimate of the conditional probability of itemset B given itemset A. It does not measure the real strength of

The correlation and implication between A and B. In order to overcome this weakness, a correlation measure can be used to augment the support-confidence framework for association rules. This leads to correlation rules of the following form

A→B [support, confidence, correlation]

A correlation rule is measured not only by its support and confidence but also by the correlation between itemsets A and B.Many different correlation measures [3] such as lift, chi square, cosine and all _confidence can be used to perform correlation analysis. Lift between two itemsets A and B can be given by the following equation.

confidence(A→B)/support (B) (1)

If the resulting value of equation (1) is less than 1, then occurrence of A is negatively correlated with occurrence of B(Fig.3b).If the resulting value is greater than 1,then A and B are positively correlated(Fig.3a).If the resulting value is equal to 1, then A and B are independent (Fig 3c).For two itemsets A and B, the Cosine Measure can be defined by the following equation.

$$support\ (A \cup B)/\sqrt{support(A) \times support(B)} \quad (2)$$

The Cosine Measure can be viewed as a harmonized lift measure. Cosine value is only affected by the support of A,B and AUB not by the total number of Transactions.Moreover,Cosine Measure is null invariant as it is not affected by the number of null trasactions.This property is important for measuring correlations in large transaction Databases. Support-confidence framework can be augmented with a correlation measure to mine correlation rules. It can reduce the number of rules generated and leads to the discovery of more meaningful rules. It is better to augment Cosine measure with left when the result is not conclusive.

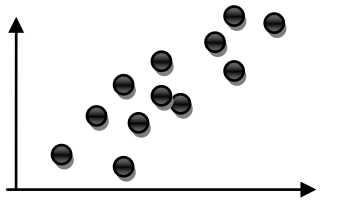

Figure 1a: Positive correlation

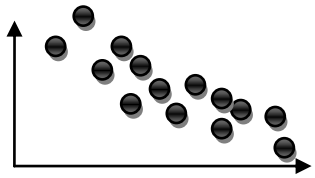

Figure 1b: Negative correlation

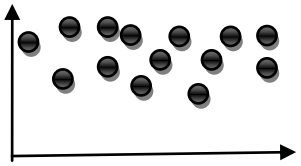

Figure 1c: No correlation

III. METHODOLOGY

Web server stores large volume of data as a result of access to a website. Data may includes date and time of request, URL requested, amount of data, IP address of client, browser and operating system information etc .In proposed methodology (Fig.3) server logs are preprocessed to get sequential list of pages that were visited in the sessions[16]. In Web environment, sessions and pages can be treated as transactions and items respectively. FP growth method [3] is used to generate frequent itemsets and then association rules are generated from frequent itemsets. Cosine measure is used to filter out uninteresting rules. In order to produce better results cosine measure may be augmented with lift measure. We consider dependencies only between 1-page set i.e. single Web pages. Interesting rules are stored in knowledgebase. When a user request for a page, then it is matched with the antecedent part of rules in the knowledgebase and a recommendation list of pages with highest confidence presented to the user [13]. We proposed an algorithm in pseudo codes for overall process of recommendation.

Algorithm
Inputs: Database of Sessions (D)
Minimum Confidence (α)
Minimum support(s)
Threshold cosine value (d)
Output: Recommended web pages
Begin
K_b=ϕ // K_b; knowledgebase
Determine the set of frequent page set L using FP Growth method.
For each l ϵ L // l; frequent page set
Generate association rules $p_i \rightarrow p_j$ and determine Cosine (p_i, p_j). // where i≠ j
If (Confidence ($p_i \rightarrow p_j$) ≥ α && Cosine (p_i, p_j) >d) Then
$K_b = K_b$ U ($p_i \rightarrow p_j$)
Else
Remove the Rule
End If
End For
For each visited page p_i of user u_i
P_i is matched with the antecedent of rules in knowledge base.
End For
Return (consequents from matching rules)
End

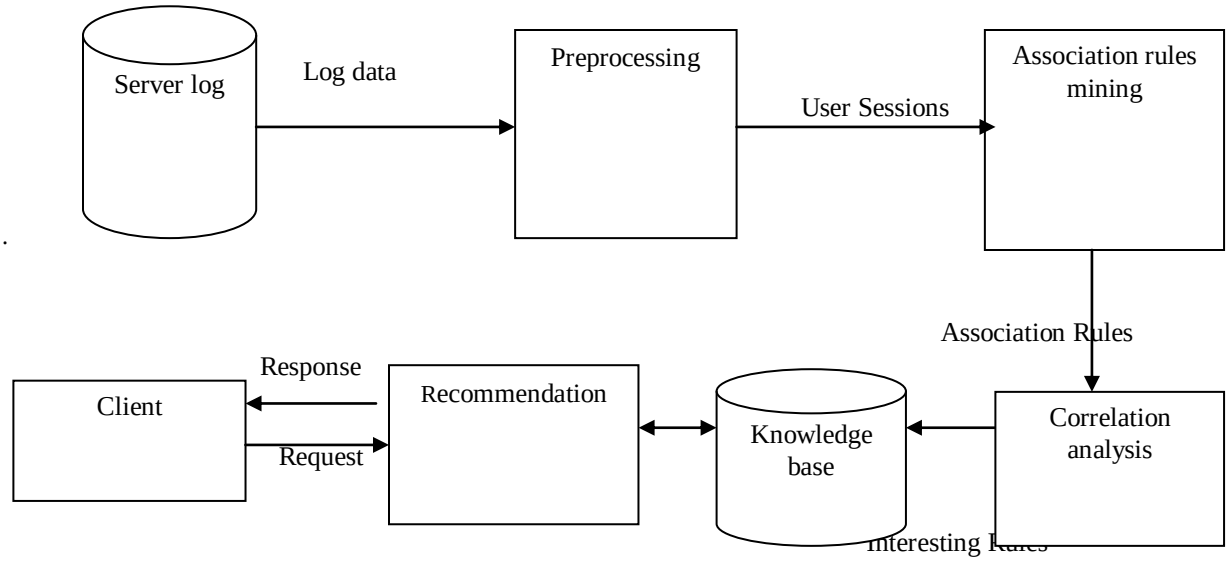

Figure 2.Proposed Methodology

IV. EXAMPLE

Let us consider an example set of nine user sessions within a website which contains five pages (table I), D {A, B, C, D, E}.We used data of Table I to construct FP tree (Fig.3) and then tree is mined [3] to get frequent patterns (Table.II).Association rules generated from frequent patterns are shown in figure 4 and recommendation list for each page is shown in Table III.

TABLE I. DATABASE OF TRANSACTIONS

Session id	Pages
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

TABLEII.FREQUENT PATTERNS (SUPPORT COUNT=2)

Page	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns
E	{{B,A:1},{B,A,C:1}}	<B:2,A:2>	<B,A,E:2> <B,E:2> <A,E:2>
D	{{B,A:1},{B:1}}	<B:2>	<B,D:2>
C	{{B,A:2},{B:2},{A:2}}	<B:4,A:2>, <A:2>	<B,C:4> <A,C:4> <B,A,C:2>
A	{B:4}	<B:4>	<B,A:4>

C→B	E→A	B→C
A→C	D→B	A→E
C→A	A→B	B→E
E→B	B→A	B→D

Figure 4.Interesting Association rules (α=25%, d=0.5)

V. PERFORMANCE EVALUATION

In this study, we used FP Growth method to find frequent item set, which is faster than Apriori method. The execution time for the two algorithms [15] for different support values on a data set is shown in Fig. 5. Cosine measure is used to prune the generated association rules (only positive correlation between page set has been taken in to account).

Performance of a Recommender system can be evaluated on the basis of three measures: Recall, Precision and F1. Precision measures the degree to which the system produces accurate recommendations. It is the number of relevant web pages retrieved divide by the total number of web pages in the recommendation set. On the other hand Recall measures the ability of the system to produce all of the page views which are likely to be visited by the user and it is the number of relevant web pages retrieved divide by the total number of web pages that actually belong to the user sessions.F1 measure attains its maximum value when both precision and recall are maximized.

Recall= Relevant and Retrieved/ Relevant (3)

Precision= Relevant and Retrieved/ Retrieved (4)

F1= 2(Precision. Recall)/ (Precision+ Recall) (5)

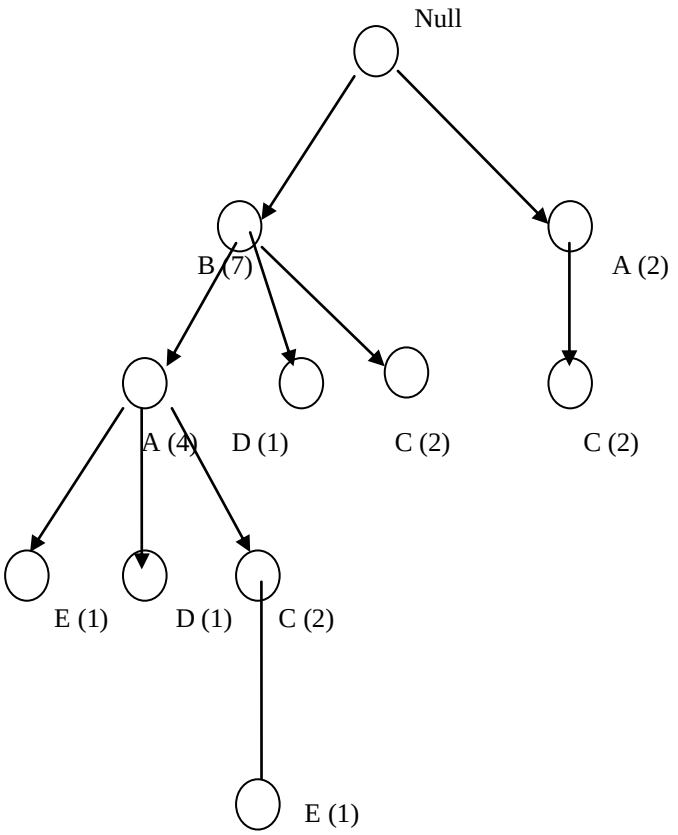

Figure 3.FP Tree

TABLE III.RECOMMENDATION LIST

Page	Recommended Pages
A	{C,B},E
B	{C,A}, {E,D}
C	{A,B}
D	B
E	{B,A}

We obtained the values of precision, recall and F1 using (3), (4) and (5) for above mentioned example as shown in Table IV.

TABLE IV.PERFORMANCE METRICS

Precision	Recall	F1
0.53	0.93	0.67

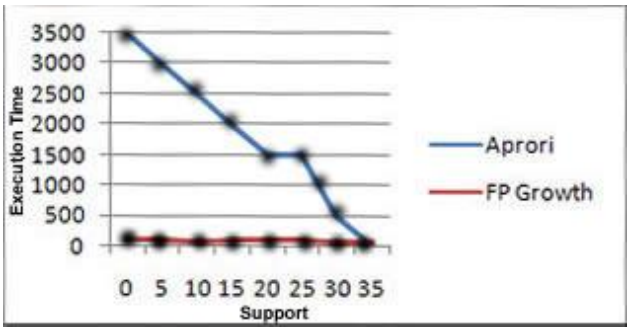

Figure 5.A comparison between FP Growth and Apriori

VI. Related Work

Association Rules Mining is one of the important Data Mining Technique. Association Rules can be used for the recommendation of Web Pages. In [7] Complex association rules have been used for the recommendation of Web pages. [8] Discovered the association rules by using data cube structure and applying OLAP operations. In [9] coordination is achieved between caching and prefetching.Collaborating filtering technique can be used to recommend Web pages within a Web site[10].

This approach uses Association rules mining to form a set of predictive rules, which are further pruned by using minimum reaching distance(MRD) information. Two Rule learning algorithms, Set covering and CN2 to analyze sequences of WWW Pages visits in click stream data are presented in [11].A simplified WWW data model[12] can be used to represent data in the cache of Web browser to mine association rules .These rules are stored in Knowledgebase and prefetched the pages according to user interest. [13] Presented a Recommendation model by generating association rules. An integrated system (Web Tool) for applying Data mining Techniques such as association rules or sequential patterns on access log files is presented in [14].

VII. Conclusion

In this paper we proposed a recommendation methodology based on correlation rules. Association rules are generated from log data by using FP Growth algorithm and then Cosine measure is used for generating correlation rules. We considered only positive correlated rules in our recommendation process and other types of rules (negative and independent) have been pruned. Proposed methodology can recommend web pages to the users which are interesting to them. Moreover negative Correlation may be used to remove the links which are uninteresting to the users.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc.of ACM SIGMOD Conf. on Management of Data*, 1993, pp. 207-216.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. of the Eleventh Int. Conf. on Data Engineering*, 1995,pp. 3-14.

[3] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Generalized Affinity-Based Association Rule Mining for Multimedia Database Queries," *Knowledge and Information Systems (KAIS): An International Journal*, vol. 3, no. 3, August 2001, pp. 319-337.

[4] Margaret H.Dunham and S.Sridhar,"Data Mining: Introdutory and Advanced Topics", Pearson Education Publisher, 2008.

[5] Agrawal, R., R. Srikant,"Fast Algorithms for Mining Association Rules, In Proc. of international conference on very large databases, 1994, pp.487-499.

[6] Przemyslaw Kazienko,"Mining Indirect Association Rules for Web Recommendation", International journal of applied math and Computer Science, 2009, vol.19, No.1, pp.165-186.

[7] Galina Bogdanova and Tsvetanka Georgieva,"Discovering the association rules in OLAP data cube with daily downloads of Folklore Materials",International conference on Computer Systems and Technologies 2005.

[8] A.Nanopoulos,.Katsaros and Y. Manolopoulos,"Exploiting Web Log Mining for Web Cache Enhancement", WEBKDD 2001.

[9] M. Shyu, C.Haruechaiyasak and N Zhao," Collaborative Filtering by Mining Association Rules from User Access Sequences"

[10] Peter Berka,"Click stream data analysis using rule based approach".

[11] W.Xu, W.Song, and H.Yang,"Pre-Fetching Web Pages Through data mining based prediction".

[12] A.Jorge, M.Alves, and P.Azevedo,"Recommendation with association rules: A Web mining Application".

[13] F.Masseglia, P.Poncelet, and M.Teisseire,"Using data mining techniques on Web access logs to dynamically improve hypertext structure".

[14] Cornelia Győrödi, Robert Győrödi, prof. dr. ing. Stefan Holban, "A Comparative Study of Association Rules Mining Algorithms".

[15] Murat Ali Bayır, Ismail H. Toroslu, Ahmet SAR"A Performance Comparison of Pattern Discovery Methods on Web Log Data".

AUTHORS PROFILE

Navin Kumar Tyagi received M.Tech in Computer science and Engineering from Kurukshetra university, Kurukshetra (India) in 1998 and currently pursuing Ph.D in Computer Science and Engineering from Bhagwant University, Ajmer (India).Presently, he is working as Assistant Professor and Head of Computer Science and Information Technology Department in Marathwada Institute of Technology, Bulandshahr (India).He has more than 10 years of teaching experience. His areas of interest include Web usage mining, Software Engineering, Operating Systems and Data Structures.

A.K Solanki received M.E. in Computer Science and Engineering from NIT Allahabad (India) in 1996 and Ph.D in Computer Science and Engineering from Bundelkhand University Jhansi (India) in 2005. Presently, he is working as Professor and Director in Meerut Institute of Engineering and Technology, Meerut (India). He has more than 22 years of teaching experience. Professor Solanki is appointed as an executive committee member of national executive council of Indian society of technical education (ISTE) for three years 2009-2012 for Uttar Pradesh and Uttarakhand Region. He is also the member of selection and inspection committee of AICTE, UPTU and other Universities. Professor Solanki research contribution is in the field of Data warehousing and Web Mining.