ISSN 2348 – 7968

Cognitive Analysis in Web Log using Comparative Study of Apriori and Eclat Algorithm

Naresh Kumar Kar[1], H. R. Sharma[2], Asha Ambhaikar[3]

Rungta College of Engineering and Technology,
Kohka-Kurud Road, Bhilai, India,
[1]nareshkar@gmail.com
[2]hrsharma44@gmail.com
[3]asha.ambhaikar@gmail.com

Abstract

World Wide Web plays a vital role in serving the needs of the user's on web. Interaction between the client and the service provider on web generates web log files. web log file contains lots of hidden important information pertaining to the visitors, we can use it to predict the navigation behavior of the users. However the task of discovering frequent sequence patterns from the web log is challenging. Sequential pattern mining provides a important role in serving a promising approach of the access behavior of the user. This paper focuses on adopting an intelligent technique that can provide personalized web service for accessing related web pages more efficiently and effectively, so that it can be determined which web pages are more likely to be accessed by the user in future. This paper uses two intelligent algorithms for predicting the user behavior's namely Apriori and Eclat and also does the performance comparison of the two algorithms in terms of time and space complexity for the filtered data.

Keywords: *Proposed Architecture, Filtered Data, Algorithm Comparison, Web Usage Mining, Sequential Pattern Mining, Apriori, Eclat, Recommended System.*

1. Introduction

The World Wide Web serves as a vast, widely distributed, global information service center for advertisement, consumer information, e-commerce, education, financial management, government, news and many other information services. So, it has become much more difficult to access relevant information from the web with the explosive growth of information available on the internet. Therefore, further research work needs to be carried out for extracting the appropriate content as per the user's needs. This can be performed using sequential mining as it helps in extracting the common sequences of the user's accessed web pages. The literature review focuses on the study, comparison and contrast of the available preprocessing techniques. Data Cleaning is done to remove the inappropriate records with unsuccessful status[4]. The ability of using the data mining techniques to extract information from the server logs was first introduced by [11], [12], and [13].

This paper includes the comparative study of two sequential mining algorithms.

1.1 Sequential Pattern Mining

Sequential mining is the process of applying data mining techniques to a sequential database for the purpose of discovering the correlation relationships that exist among an ordered list of events. The approach taken by Ciesielski and Lalani [14] focuses on the relevant feature extraction ideas from the dataset rather than developing better algorithms.

With respect to the Internet, web usage mining is a vital problem with wide applications, including the analyses of customer purchase behavior, web access patterns, scientific experiments, disease treatments, natural disaster, and protein fonnations. The algorithm for the sequence pattern mining extract the sequence database looking for repeating patterns (known as frequent sequences) that can be used later by end users to find associations between the different items or events in their data for purposes such as business enterprises, marketing campaigns, planning and prediction. The new user is identified [3], if the extracted page requested by the user doesn't match the sequence page.

Web log mining an exceptional case of sequential pattern mining deals with finding user navigational patterns by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items, with the assumption that a web user can actually access only one web page at any given point in time. Presently, most web usage mining solutions consider that user access one web page at a time, which gives rise to special sequence database with only one item in each sequence's ordered event list. Thus for a set of events E ={a, b, c, d, e, f },which may represent product web pages accessed by the clients in an E-Commerce application.

Algorithm for Sequential Pattern Mining Using Apriori Algorithm:

INPUT:
T = (AI, A2 ,...., An) / / Database from the filtered Data where A1, A2 are the item sets

miniSup / / Support = 20%

OUTPUT: Frequent Sequential Pattern

Algorithm for Sequential Pattern Mining:

T = T sorting on Transaction ID and Find Candidate Sequence table

L = Apriori (T, minSupp, L) Find the sequential data of L

Algorithm for Sequential Pattern Mining Using Eclat Algorithm:

Input:
D = (Tl, T2, ... Tn) / / Database from the filtered Data where Tl, T2 are the Transaction ID

miniSup / / Support = 20%

Output: Frequent Sequential Pattern

Algorithm for Sequential Pattern Mining:

D = D sorting on Item sets

L = ECLAT (D, minSupp, L)

Find the sequential data of L

Apriori Algorithm: Apriori is an algorithm proposed by R. Agrawal and R Srikant in 1994 for mining frequent item sets for Boolean association rule. This algorithm uses prior knowledge of frequent item set properties. Apriori make use of an iterative approach known as Level-wise search, where k item set are used to explore (k+1) item sets.

Each iteration consists of two steps:

- Generates a set of candidate item sets
- Count the occurrence of each candidate set in database and Prunes disqualified.

Pruning Techniques used by Apriori: Apriori uses two pruning techniques;

- First is based on Support Count (Greater than User specified support threshold)
- Item set to be frequent, all its subset should be in last frequent item set.

The iterations process starts with size 2 item sets and the size is incremented after any iteration. According the algorithm if a set of items is frequent, then all its proper subsets is also frequent.

ECLAT Algorithm: This algorithm is based on the concept of depth-first search. It is opposite of Apriori algorithm so it prunes itemsets that have a lower support than the threshold. It calculates the support for itemsets by maintaining a transaction list for each item [6]. In this manner the transaction database is only required once for counting the support. Support for the subsequent.

Recommendation System: The navigation behavior of the client can be predicted with the knowledge of browsing patterns gathered from the previous stage, current user can be recommended links to pages that are similar to the one's is presently viewing. Recently browsed user's pattern is matched with the analyzed data of previous users and based on this comparison; current user is suggested similar web pages of interest not yet visited. The pages requested by the client can be discovered by two methods:

- Highest Confidence
- Last Sequence.

Highest Confidence: It is chosen to predict next page, Highest confidence is gathered from the predecessor as it is based on pattern matching rule.

Last Sequence: It considers the Highest Confidence if different rules are equal. This procedure selects rules where the requested pages are approximately near to the consequent. This process basically considers the distance between pages requested by a user and the consequences of a rule whereas the distance is the number of clicks from one page to another.

User navigation prediction model is built based on the patterns extracted by him-self. The paper hypothesis is that pages accessed in recent times have a great influence on pages that will be accessed in the near future. The prediction for the users' navigation behavior is based on discovered rules matching the current user session.

1.2 Web Usage Mining

Among the first researchers who have put their efforts towards the Web Usage Analysis were Cooley et al. [7], who proposed Web Miner. Web usage mining is an application of data mining in which the meaningful information is extracted from the Web Server Log for the various purpose such as for the business strategies, financial activities etc. Web usage mining is the concept of web mining activity which involves the automatic detection of user access patterns from one or more Web servers. The elements [1] of the web usage access are the users and web pages accessed by the user. The goal of web usage mining is to analyze the user access behavior patterns. Web mining can be practices in three different domains i.e. the content mining, hyper link web structure mining and web usage mining. These approaches effort to extract valuable information from the web which are then applied to some real world problems. The user's surfing behavior analysis follows three phases [6]: data collection and preparation, pattern discovery and content recommendation. Figure 1 Show the web mining applications.

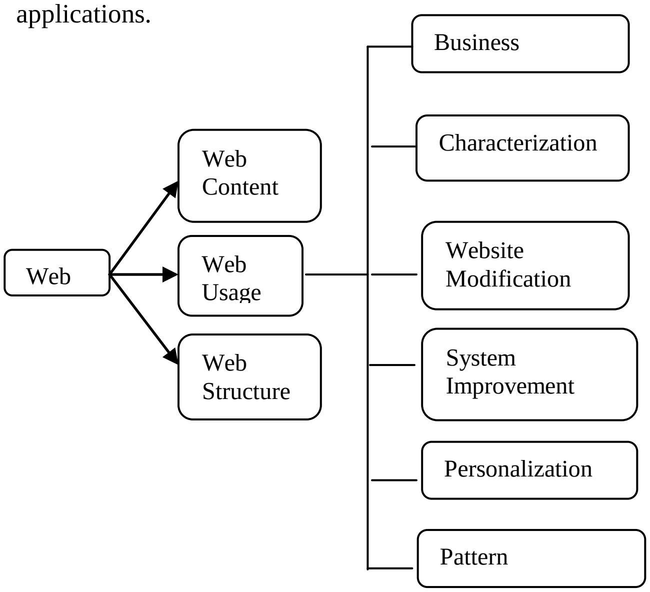


FIG. 1: APPLICATIONS OF WEB MINING

Web usage mining is important for cross marketing strategies, web advertisements and promotion campaigns. It is an application of data mining techniques that extract usage pattern from the click-stream. The extraction of valuable infonnation about users' accesses is obtained from analysis of navigation behavior from the web server logs, where all accesses to web pages are recorded. The access information includes IP Address (Request Originated), Page Requested (URL), Time and Date of the request etc.

The output generated from the pattern analysis consists of sequences of accesses with corresponding probabilities. Frequent Pattern describe how often the pages are accessed together in a sequence. This has been introduced in 1993 by Argawal et al. [4]The algorithms used to mine the usage are association rule mining and sequence analysis. Association rule mining discovers relationships between different web pages within a web site whereas Sequential pattern is used for pre-fetching instead of simple association rules; this approach helps to fmd out the order in which the pages are visited, reduces the bandwidth usage and storage needs, which undoubtedly results in improving the system efficiency and effectiveness i.e. an improved system. The objective of the association rules mining [8], [9] is to discover correlations of association between existing records in a dataset. In [9] fundamental association rules have been described, from their discovery and up to the present moment. These works some classic algorithms like: Apriori, Eclat, Clique, FP-Growth. In [10], more practical and efficient methods are being presented, in order to fmd association rules in the case of less frequent items. Web log [5] can do this only after analyzing data resulted from the users' current and history data.

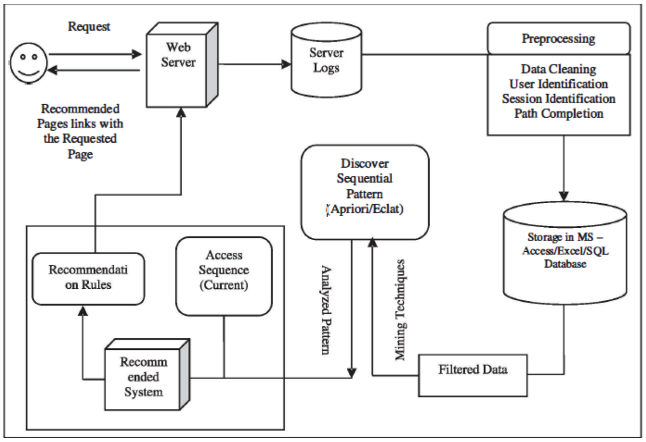
This paper focuses on analyzing the navigation behavior of the web access users from the accessed (filtered) data by applying two algorithm and performing comparison analysis of the two. This paper is organized into following sections.
Section 2 presents the Material and Methodologies. Section 3 consists of Result and Discussion, Section 4 performs comparison and finally Section 5 concludes the paper.

2. MATERIAL AND METHODOLOGIES

In this paper the material is taken from the Learning Management System. An experiment has been conducted on a filtered data of the web log file. This paper applies the concept of data mining onto the filtered data. The proposed architecture working is shown in Figure 2. The methodologies applied in this paper are

Sequential Mining Techniques using the Apriori and Eclat Algorithm.

2.1 Proposed System Architecture



2.2 Working of Proposed System Architecture

Following are the steps to be carried out:

Step 1: Client sends the request to the server for the web pages he/she wants to access on the web.

Step 2: The server log contains the raw data which is generated as of an interaction between the client- server. This information is initially raw, it contains garbage also which needs to be removed. The log data contains the nine attributes for the server log.

Step 3: The raw data is preprocessed. During the preprocessing the unwanted data is removed. After the log files are preprocessed the data is stored in the query language fonnat.

Step 4: This extracted 20% information is stored in the database for applying data mining techniques. Again the data is filtered for applying finding the frequent sequence pattern.

Step 5: We apply the Sequential pattern mining technique based on Apriori rule/Eclat. This algorithm mines the filtered database and it looks for frequent patterns which is also known as frequent sequences which afterwards used by end user for finding the relation between the different events.

Step 6: Recommendation Rule generator take into consideration the currently accessed web pages based on some threshold value defined and the pattern that are discovered after applying sequential pattern mining based on apriori algorithm into consideration and generate the pages that are frequently accessed by the user.

Step 7: The page recommended by the recommendation rule generator system is then sent to the server when a client accesses any web page. The web page contains those links that are of his/her interest.

2.3 Working of Apriori and Eclat

Apriori Algorithm: It is a breadth-first search algorithm, it makes use of two-pass strategy for finding frequent item sets. The lists of candidate item sets are generated in the first pass at each level and item sets are pruned that are supersets of infrequent item sets. In the second pass support values for the remaining item sets are calculated and again performing pruning of those item sets which have a support less than the user-defined threshold. Item sets support can be calculated by either counting the number of transactions in the database for each item set.

Steps for Apriori

Input: Preprocessed Data, Minimum Support (minSup).

Output: Generating Frequent Item Sets From The Preprocessed Data.

```
F1= Frequent Item Set
j=n;   /* Maximum Number Of Elements N */
for  n= MAXLENGTH to 1
{
for i=n to 2
   {
     while each  transaction Fi
        {
          if (Fi Repeated)
             {
                Fi.increment++;
             }
x=0;
for (;i<j-x;)
  {
     if ( Fi is a subset  of each  transaction Fj-x of
          order  j-x)
         Ti.increment++;
         x++;
    }
 }
```

Eclat Algorithm: The Eclat algorithm utilizes the aggregate memory of the system, it portioned the candidates into disjoints sets using the concept of

Equivalence Class Partitioning. It was aim to defeat the shortcomings of the Count and Candidate Distribution algorithms. Eclat uses the concept of vertical database layout which keeps all relevant information in an itemset's tid-list. In Eclat local database partition is scanned only once whereas in contrast Candidate Distribution must scan it once in each iteration. Eclat doesn't search complex data structure, it doesn't generate all the subsets of each transaction thus doesn't pay the extra computation overhead.

Steps for Eclat

```
Begin
// Initialization Step
Mining frequent item set in the database
Then take the containing transaction ID
Support = 20%

// Transformation Step
Mining Frequent 2 -Item sets in Vertical data
format
If (Transaction ID < Support ) then
Left those items sets

// Asynchronous Step
Mining Frequent k -item sets in Vertical data format
If (Transaction ID < Support) then
Left those items sets
k = 1 to n

// Final Reduction
Now Total Results and Output
End
```

Where

Transaction ID	**Item Sets**
T100	{A1, A2} {A1}
T200	{A1, A3, A4} {A1,A2}

3. Result and Discussions

3.1 Working of Apriori

TABLE 3: SHOWING THE USER ID AND CATEGORY

User Name	**Category**
T1	{ Web, IT, SE} { IT } { IT, Web, Edu}
T2	{ Down, Govt, IT} {IT, Govt } { Web, IT, SE}
T3	{Edu, Down} {Edu, IT, Web} {Web, IT, SE, Down}
T4	{ IT, Web, Edu} { IT, Govt, Down}
T5	{Web, IT, SE} {IT, Govt} {Edu, IT,Web } { IT}
T6	{ Down, Govt, IT} { IT, Web, Edu}
T7	{ Web, IT, Down, SE}
T8	{ Edu, Down} { IT, Govt} { IT}

⇩

Category	**Items**	**Supp_Min**
{ Web, IT, SE}	A1	3
{ IT, Web, Edu}	A2	5
{ IT}	A3	2
{ Down, Govt, IT}	A4	3
{IT, Govt}	A5	3
{ Edu, Down}	A6	2
{Web, IT, SE, Down}	A7	2

Minimum support = 20%

The table uses the following alias names as follows:

l. Web: Normal navigation
2. IT: Information technology related websites
3. SE: various search engines navigated
4. EDU: Educational sites
5. Down: downloaded sites
6. Govt: Government organization related sites.

Table 4: SHOWING FREQUENTLY ACCESSED PATTERN

Category	Support
A1A2	2
A1A3	2
A1A5	2
A2A3	2
A3A5	2

Frequent Data Sets

⇩

A1A2A3	2

A1 = {Web, IT, SE}
A2 = {IT, Web,Edu}
A3 = {IT}

3.2 Working of ECLAT

TABLE 5: Working of ECLAT for generating frequent pattern

Category	User Name
A1	T1, T2, T5
A2	T1, T3, T4, T5, T6
A3	T1, T5, T8
A4	T2, T4, T6
A5	T2, T5, T8
A6	T3, T8
A7	T3, T7

Category	TID
A1A2	T1, T5
A1A3	T1, T5
A1A5	T2, T5
A2A3	T1, T5
A3A5	T5, T8

A1 A2 A3	T1, T5

Frequent Item Set
A1 = {Web, IT, SE}
A2 = {IT, Web, Edu}
A3 = {IT}

In this paper we have applied two sequential data mining techniques using Apriori and Eclat. The working of both of the algorithms is shown above. From the above shown working of Apriori we can conclude that Apriori uses more number of candidate sequence sets, it prunes the infrequent pattern from the data. On other hand Eclat generates less number of sequence tables which takes less time for generation of frequent accessed patterns as compared to Apriori. In apriori if massive data is their then it takes huge time to generate the frequent accessed patterns.

4. Algorithm Comparison Based on Discovered Frequent Patterns

TABLE 6: Performance of two Algorithms Applied to Filtered Data

S. No.	Properties	Apriori	Eclat
1	Data base used	College data	College data
2	Size of database	10 MB	10 MB
3	No of transaction	294	294
4	No of items/columns	5	5
5	Type of data	Sparse	sparse
6	Min supp	20 %	20 %
7	Data base scans	5	3
8	Memory consumed	7 MB	5 MB
9	Running time sec	7 Sec	3 Sec

5. Conclusions

The extraction of valuable information about users' accesses is obtained from analysis of navigation behavior from the web server logs, where all accesses to web pages are recorded. This paper adopted an efficient sequential pattern mining techniques using the Apriori and Eclat algorithm for the filtered data set. Both the algorithms helps to find out the navigation behavior of the user based on the previous visits and also shows the comparison of the two techniques adopted for predicting user access behavior. Discovering the frequent itemsets from the two algorithms a also shows that Eclat algorithm serves better for the large databases despite Apriori as it generates less tables and therefore less time it takes to perform the analysis.

References

[1] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining", Communications of the ACM, New York, Volume 43, Issue 8, Aug 2000.

[2] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, IEEE, 2008.

[3] Robert. Cooley, Bamshed Mobasher and Jaideep Srinivastava, "Web mining:Infonnation and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[4] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[5] Perkowitz, M., Etzioni, 0.: Adaptive sites: Automatically learning from user access patterns. In: Proc. of the Sixth International WWW Conference, Santa Clara, CA. (1997).

[6] Mobasher, B. , Cooley, R. , Srivastava, J. : Automatic personalization based on Web usage mining, pp. 142-151. Commun. ACM 43, 8 (2000).

[7] Cooley, R., Mobasher, B. , & Srivastava, J. : Data preparation for mining World Wide Web browsing

patterns. Knowledge Information Systems, 1(1), pp. 5-32. (1999).

[8] Agrawal, R., Irnielinski, T., Swami, A. : Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international Conference on Management of Data, pp. 207-216. Washington, D.C. (1993).

[9] Ceglar, A. , Roddick, J. F.: Association mining. ACM Comput. Surv. 38, 2 (2006).

[10] Ding, 1., Yau, S. S.: TCOM, an innovative data structure for mining association rules among infrequent items, pp. 290-301. Comput. Math. Appl. 57, 2 (2009).

[11] Data mining for path traversal patterns in a web environment. In Proceedings of the 16th international Conference on Distributed Computing Systems (ICDCS '96) (May 27 - 30, 1996). ICDCS. IEEE Computer Society, Washington, DC, 385. [12] H. Mannila, H. Toivonen. Discovering generalized episodes using minimal occurrences. In: Proc.Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.

[13] Yan, T. W., Jacobsen, M. , Garcia-Molina, H., and Dayal, U. 1996. From user access patterns to dynamic hypertext linking. In Proceedings of the Fifth international World Wide Web Conference on Computer Networks and ISDN Systems (Paris, France). P. H. Enslow, Ed. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1007-1014.

[14] Ciesielski, V. and Lalani, A., Data mining of web access logs from an academic web site. In Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03).