Implementation of the Association Rule Mining Algorithm for the Recommender System of E-Business

D. Santhi Jeslet
Department of Computer Science
M.G.R. College, Hosur, TN, India
santhijeslet@gmail.com

M. Lilly Florence
Department of Computer Applications
Adhiyamaan College of Engineering (Autonomous)
Hosur, TN, India
lilly_swamy@yahoo.co.in

Abstract **- Through the dramatic development of computers and internet (web), business has moved into online. This connected the customer and company very close without their presence. The internet has become a part of individual human being to run their day-to-day life. So, to run an effective e-business web is playing a vital role. To run a successful e-business, it is necessary to uphold the websites. To uphold the website, it is required to identify the frequently visited and associated pages. This can be found by applying the association rule mining algorithms of data mining. In this work, the apriori association rule mining is applied on the web usage data. The web usage data is collected from the server log. Then, it is preprocessed in such a way that it is suitable for the algorithm. Through the association rules generated by the algorithm, one can identify the frequently visited and associated pages. These rules recommend the website administrator to restructure or redesign the website which in turn results in increasing the number of visitors/customers and retains the existing customer.**

Keywords - Data mining, Web mining, Web usage mining, Association rule, Apriori algorithm

1. INTRODUCTION

The rapid development in the field of computer networks and multimedia technology, World Wide Web (WWW) has swell up (Mingyu Lu, 2002). The most dominant application of Internet is e-business (Nivedita Roy, 2005). The information sources in WWW has grown tremendously which increased the necessity of automated tools to find, extract, filter and evaluate the preferred information for the users. For addressing the above said tool, a new concept called “Web mining” is introduced. This technique can be used to study the behavior of the users of the website.

Web mining is the major application of data mining techniques. It is the integration of information gathered by traditional data mining methodologies and the information from the World Wide Web. It automatically extracts information from the web documents and services. Classical data mining techniques that are used in web mining includes classification, clustering and association rule.

Web mining is divided into three categories: content mining, structure mining, and usage mining. Out of these categories of web mining, web usage mining is the suitable category that can be used for studying the previous behavior of the users. Through mining web usage data, the complete knowledge about the user’s behavior can be obtained which helps in improving the e-business.

1.1. Web content mining

Web content mining mainly deals with the extraction and integration of useful data, information as well as knowledge from web page. It also discovers useful information from the web documents. The web page may contain text, audio, image, video, metadata, hyperlinks etc.

Web content mining relates as well as differs from data mining and text mining. Many of the data mining techniques can be applied on web content mining. So it is related to data mining. It also differs from data mining in which the web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data only. It is related to text mining because most of the contents of the web are texts (Ramesh Yevale, 2014). The content of the web page are semi structured nature, while text mining focuses on unstructured texts. In this way it differs from text mining. Thus, web content mining needs innovative application of data mining and/or text mining techniques, in addition to its own unique approaches.

Due to the phenomenal growth of the web content, there was a great explosion in the activities of web content mining in the last few years. Web content mining is used to discover the resources from the web, categorization of documents, clustering and extracting information from the web pages.

1.2. Web Structure mining

World Wide Web can reveal additional information than the available information in the documents. Due to the vast amount of information, Web structure mining helps to minimize the problems in accessing the relevant information of the World Wide Web. The previous unknown relationships between Web pages are extracted. To enable the navigation and cluster information into site maps, the structure data mining helps to link the information of its own Website to the business. This increases the ability of the users to access the desired information through content mining and keyword association.

The web structure consists of web pages as nodes and hyperlinks as edges that connect two related pages (Sonia Gupta, 2013). For example, links pointing to a document indicate the popularity / specialty of the document and links coming out of a document indicate the richness of various topics wrapped in the document.

Web Structure mining is categorized into two types. They are intra-page structure and inter-page structure. Intra-page structure deals with the existence of links within a page. In this case, separate page will not be opened. Inter-page structure entails the link with one page to another page.

Web structure mining is must to accurately utilize the website as a business tool.

1.3. Web Usage Mining

Web usage mining deals with, mining the characteristics of the web users' usage. It is a part of Web Mining, and also a part of Data Mining. Rather than the technical aspect, web usage mining is regarded as an element of the business intelligence in an organization. It is highly needed for deciding business strategies through the proficient usage of web applications.

Web usage mining mainly focuses on finding the user access patterns from the web logs. It concentrates on different data mining techniques to analyze and understand the search patterns. It develops the navigation of information on the web. In addition, it produces a higher quality of information to the user and endows with productive marketing. It helps to quantify the success of a marketing campaign through understanding the customer behavior and evaluating the effectiveness of a particular website.

Through the application of web usage mining, some of the advantages like pricing analysis, Business intelligence and competitive intelligence are also gained.

2. WEB MINING APPLICATIONS IN E-BUSINESS

Web mining can be viewed as a key enabler for the success of e-business. In e-business web mining are used in these subsequent areas: (Santhosh Kumar B, 2010)

- Web mining can help the top management to take strategic actions through companies managerial insight into visitor profiles.
- Web mining assist the e-business in improving and aligning their marketing strategies by obtaining some subjective measurements on effectiveness of their marketing research and campaign.
- Structure mining can be fairly useful in finding the relationships between two or more business websites in the e-business world.

- Through web mining the company can make strategic adjustments by identifying the strength and weakness of the web marketing campaign. It can get the feedback again in order to spot the improvements in e-business.

3. STEPS IN WEB MINING

The major modules in the web mining are namely data collection, data preprocessing, pattern discovery, personalization and recommendations. In these modules first three are offline modules and remaining are online modules.

Preprocessing of data is necessary which helps to improve the quality of data and consequently the mining results. As the first step data has to be extracted from the web server.

3.1 Data collection

In web mining, data collection is the first step which can also be called as "data selection". This is an important, difficult and most time consuming step. For web mining, the main sources of data are from server side, client side and proxy server.

First, the web log file of the portal name **proview.com** has been collected over a period of 4 months with more than **2, 34,000** records. This web site consists of **16** numbers of pages. It has more than **1,40,000** records. The server log consists of 19 attributes. All the 19 attributes are not required for the recommendation system. So it is necessary to identify the needed attributes. The main attributes that are needed for the work are client IP, date, time, URI-Stem, and status. After identifying the relevant attributes, the attributes in the text file need to be separated using blank space as the delimiter.

3.2 Data preprocessing

The data collected from various sources are sometimes insufficient, inconsistent and may include noise. The data preprocessing has to be performed in order to make the data to be clean so that application of data mining algorithm becomes easy. The data preprocessing work mainly include data cleaning, user identification and data transformation.

3.3 Data Cleaning

The main purpose of data cleaning is in web mining is to remove or eliminate irrelevant data. Since the aim of web usage mining is to obtain the user's move patterns, the following two kinds of records are unnecessary and should be eliminated.

a. Graphics, videos and the format information are irrelevant to web mining. These records have their filenames with the suffix of GIF, JPEG, CSS etc., which can be found in the URI_STEM field.

b. The records that have the failed HTTP status code are also not needed for web mining. This can be found by examining the status field of each record in the web log. The records with the status codes above 299 or lesser than 200 are eliminated.

An algorithm for cleaning the entries of server logs is given below. The algorithm is applied on the log files of proview.com.

```
Algorithm : Data Cleaning
Input: Web Server Log File
Output: Cleaned Log Database
While (!eof(Web Server Log file)){
        Read a record in Web server Log file
        Read field (URI-Stem)
        If URI-Stem !=(*.gif / *.jpg/ *.css/ *.png/*.ico)
                Read field (Status)
        If (status < 300 AND status>=200) then
                Save records into a database    }
```

3.4 User Identification

The main job of identification of a single user is essential to differentiate her/his behavior in web mining. In most cases, the log file provides only the computer address (name or IP) and the user agent. For Web sites requiring user registration, the log file also contains the user login that can be used for the user identification. This work assumes the following in order to identify the different users:

- The different IP addresses distinguish different users.
- If the IP addresses are same, the different browsers and operating systems indicate different users which can be found by client IP address and user agent who gives information of user's browsers and operating system.
- In addition, if the time spent by the visitor in the web site is more than 10 minutes that particular visitor is considered to be the important visitor or reliable visitor.

The algorithm for user identification is given below. The outcome of this algorithm is unique users database which gives information about the total number of individual users, users IP address, date and time etc.

```
Algorithm : User Identification
Input: Cleaned Log Database
Output: Unique Users Database
Initialize IPL=0; UList=0; BList=0;
OSL=0; NOU=0;
    While (!eof(Log Database))
    {
        Read a record in cleaned Log Database
        Read field (client IP address, time stamp)
        If (client IP address is not in IPL) and
            (time-stamp>10minutes) then
        {
            Add new client IP address into IPL
            Add Browser details into BList
            Add OS details into OSL
            Increment count of NOU
            Insert new user in to UserList
        }Else
        {
            If( client IP address is present in IPL OR Browser details not in BList
            OR OS details not in OSL and time-stamp>10minutes) then
                Add new client IP address into IPL
                Insert new user in to UserList
                Increment count of NOU
        }
  }
```

3.5 Data Transformation

After identifying the unique users and also the visited pages, it is necessary to transform it into the form applicable for the data mining algorithm. The cleaned data is available in the Web Log database. The unique users are identified and stored in the Unique User database. Here, these two databases are compared and transformed into a database called the Rating Matrix Database.

```
Algorithm : Data transformation
Input : Unique User database, Web Log database
While (!eof(Unique User database))
{   Read a record
    While(!eof(Web Log database)
   {        Read a record
            If (client ip of Unique User database = client ip of
                Web Log database)
            {
            While( not end of predefined page list)
                     {
                              If (uri-stem of Unique User database =  predefined page list)
                                       Insert 1 into the Rating matrix database
                              Else
                                       Insert 0 into the Rating matrix database
                     }
            }
```

Each unique user in the Unique User database is compared with the records in the Web Log database to find out the pages visited by the users. In the resulting Rating Matrix database each transaction is represented as a numeric number starting with 0. The pages are represented as P0, P1,….Pn. Rather than listing the visited pages of a web site in a transaction table, it can be represented in terms of 1 and 0. The intersection of the transaction and the page is marked as 1 if the particular page is visited by the user and 0 if it is not visited. The sample Rating Matrix is shown in Table 1.

3.6 Pattern Discovery

After completing the preprocessing of web data, discover patterns of usage of website by using any of the statistical methods, data mining techniques, machine learning techniques and pattern recognition techniques. In particular, association rule and clustering techniques of data mining techniques are very frequently used for pattern discovery. In this work, the association rule mining technique is used for finding the frequently visited and associated pages to find out the patterns (rules).

Table 1. Sample records in the Rating Matrix database of proview.com

Trans-ID	P1	P2	P3	P4	P5	…..
0	1	1	0	0	1	
1	0	1	0	1	0	
2	0	1	1	0	0	
3	1	1	0	1	0	
4	1	0	1	0	0	
5	0	1	1	0	0	
6	-	-	-	-	-	
-						
-						

3.7 Personalization and Recommendations

After the identification of patterns (rules), personalization and recommendations can be given to the website administrator or owner of e-business inorder to redesign and restructure the website to attract more number of visitors who promotes the profit of e-business.

4. ASSOCIATION RULE MINING TECHNIQUE

Association rule mining (Jiawei Han, 2005) helps us to find the interesting relationships among the pages in the website. This technique can be used to indicate pages that are most often referred together and to discover the direct or indirect relationships between web pages in users browsing behavior (Pierrakos d, 2003).

The problem of finding web pages visited together is similar to finding associations among itemsets in transaction databases.

In the recommender system, it is necessary to find out the frequently visited associated pages. After identifying the frequently visited associated pages, it will be compared with the two thresholds called Support and the Confidence level evaluated using the equations (1) and (2).

$$\text{Support (A->B)} = \frac{\text{No. of tuples containing both A and B}}{\text{Total no. of tuples}} \tag{1}$$

$$\text{Confidence (A->B)} = \frac{\text{No. of tuples containing both A and B}}{\text{No. of tuples containing A}} \tag{2}$$

For example an association rule in the Web Usage Mining area could take the form “the visitor who view web page *product_Detail.html* also view *technical_core.html* with the support=50% and the confidence=60%. It generates huge number of e-association patterns. But constraints on confidence and support threshold, build association rules with page sets of *n* pages from rules with *n-1* page, page sets. This reduces the effective search space. Not all the patterns are good, select only the good e-association patterns. This attempt to predict which web page or document can be most useful to a user.

4.1 Application of Apriori algorithm in the recommender system

Apriori algorithm is one of the well known association rule mining algorithm. It uses previous knowledge of the frequent itemset properties. In this algorithm, the K- itemsets are used to discover K+1 itemsets. First, find the frequent 1-itemset which can be represented as L1. Now a two step process is followed in the algorithm namely:

1. Join step: To find L_k a set of candidate K- itemset is generated by joining L_{k-1} with itself i.e $L_{k-1} \bowtie L_{k-1.}$ This set of candidates is denoted as $C_{k.}$

2. Prune Step: The size of C_k is huge. So to reduce the size of this Apriori property is used. According to this property, any (K-1) itemset that is not frequent cannot be a subset of a frequent K- itemset. Hence, if any (K-1) subset of a candidate K- itemset is not in $L_{k-1,}$ then the candidate cannot be frequent either and so can be removed from $C_{k.}$

4.2 Implementation of Apriori Algorithm

The server log data for the website proview.com is collected, preprocessed and converted into the database format so that the algorithm can be directly applied on it. The preprocessed data is in the database, D, consist of 116758 records. Fixing the minimum support count as 50% and minimum confidence required as 75%, the frequent itemset is found by using Apriori algorithm, which is an association rule mining algorithm. Then, the association rules are generated accordingly.

The Apriori algorithm generates the candidate sets until all the frequent pages are found. The candidates generated are $C_{1,}$ $C_{2,}$ C_3 and C_4. Since $C_5 = \varphi$, the algorithm terminates. These frequent pages are used to generate strong association rules (where the strong association rules satisfy both minimum support and minimum confidence).

5. RESULT AND ANALYSIS

From the results, it is understood that the visitors of the web site “proview.com” have visited the pages P1, P2, P3, P5, P6, P7, P8 and P15 more number of times than the other pages. After finalizing the associated pages, recommendations are given to the website administrator to improve the website. He/She can do the necessary changes in the design of the website so that the visitors visit the website without spending more time in crawling through the pages of the website.

Name of the website	: proview.com
No. of records from server log	: 234000
No. of records after preprocessing	: 116758
No. of unique users	: 4200
No. of Pages	: 16

The maximum size of a frequent pageset is 4 pages. There were 16 out of 213 rules with confidence greater than 75% and with the minimum support of 50%. The following are the rules generated:

1) P6 ^ P15 ====> P7 ^ P8
2) P7 ^ P15 ====> P6 ^ P8
3) P8 ^ P15 ====> P6 ^ P7
4) P6 ^ P7 ^ P15 ====> P8
5) P6 ^ P8 ^ P15 ====> P7
6) P7 ^ P8 ^ P15 ====> P6
7) P2 ^ P5 ====> P1 ^ P3
8) P3 ^ P5 ====> P1 ^ P2
9) P1 ^ P2 ^ P5 ====> P3
10) P1 ^ P3 ^ P5 ====> P2
11) P2 ^ P3 ^ P5 ====> P1
12) P2 ^ P15 ====> P1 ^ P3
13) P3 ^ P5 ====> P1 ^ P2
14) P1 ^ P2 ^ P15 ====> P3
15) P1 ^ P3 ^ P15 ====> P2
16) P2 ^ P3 ^ P15 ====> P1

The following recommendations are given to the web site administrator for enhancing the web site:

- Page P6, P7 and P8 as well as Page P1, P2 and P3 are most associatively and frequently visited. So the order of the pages should not be altered at any cause.
- Many of the visitors after visiting page P6, P7 and P8 also visited page P15. It is obvious that P15 is an important page which has been visited by the visitors who have visited P6, P7 and P8. Page 15 is also visited in association with the pages P1, P2 and P3. But it is recommended to shift page P15 from its present location, in such a way that it appears after the page P8. This is because P15 is more associated with the pages P6, P7 and P8 rather than P1, P2 and P3. This shifting of page improves the efficiency of the web site by attracting huge number of visitors.
- Page P4 is not visited more and it can be shifted. Since most of the visitors who have visited pages P1, P2, P3 have visited P5 and therefore P5 should be shifted in the place of P4.
- In order to retain the existing visitors and also to attract more number of new visitors more importance should be given to the pages P1, P2, P3, P5, P6, P7, P8 and P15 in their designing.
- Pages P9 to Page P14 is not visited more. The draw backs in these pages have to be identified and decision has to be taken whether to retain the pages or not.

6. CONCLUSION

Based on the interesting association rules, recommendations are given to the Website Administrator to launch the website by redesigning or restructuring the pages of the website according to the association rules generated. This helps the e-business website “proview.com” to increase the number of visitors in addition to retaining its existing visitors from moving away to some other websites. This in turn increases the success rate of e-business. From the above study it is understood that web usage mining helps tremendously to improve the quality of the website based on the recommendations. Therefore, it can be concluded that “web usage mining is a key enabler of recommender system to improve the e-business”.

REFERENCES

Jiawei Han, M. K. (2005). Data Mining Concepts and Techniques.

Mingyu Lu, S. P. (2002). WebMe-Web Mining Environment. IEEE, SMC,WPIRS.

Nivedita Roy, T. M. (2005). Web Mining - A Key Enabler in E-Buiness. IEEE, 1121-1125.

Pierrakos d, P. G. (2003). Web Usage Mining as a tool for personalizations: A Survey. User Modelling and user adapted Interactions, 311-372.

Ramesh Yevale, M. D. (2014). Unauthorized Terror Attack Tracking using Web Usage Mining. International Journal of Computer Science and Information Technologies, 1210-1212.

Santhosh Kumar B, R. K. (2010). Implementation of Web Usage Mining using APRIORI and FP Growth Algorithms. International Journal of Advanced Networking and Applications, 400-404.

Sonia Gupta, N. S. (2013 April). Web Mining: Summary. International Journal of Computational Engineering Research, 149-154.