Importance of web mining in developing the recommender system for e-business

D. Santhi Jeslet
Asst. Prof, Department of Computer Science,
M.G.R.College,Hosur,TN, India

Abstract: Through the development of computer and internet (web), business has stepped into online. This connected both the company and customer very closely without their physical presence. So, the web is becoming an important part of people's life. The web is a very good place to run successful businesses. This turned the business into e-business (electronic business) or e-commerce (electronic commerce). So it is necessary to maintain the website to improve the e-business. To maintain the website, it is required to identify the frequently visited and associated pages. This is found by applying one of the data mining techniques called the apriori association rule mining algorithm. To apply the algorithm web usage data is collected from the server log, which is preprocessed and converted into a database. The database is given as input for the apriori algorithm to generate interesting rules. Through the rules, one can identify the frequently visited and associated pages which play a vital role in increasing the number of visitors/customers. So recommendations are given to the website administrator to redesign or restructure the website in order to retain the existing customer and to attract new customers. The webpage redesigning and the website restructuring, not only resulted in increasing the number of visitors/customers it has also increased the profit of e-business tremendously.

Key words: *Web mining, web usage mining, data mining, Association rule, apriori algorithm*

1. INTRODUCTION

With the rapid development in the field of computer networks and multimedia technology, World Wide Web (WWW) has swell up [1]. The most dominant application of Internet is e-business [2].The explosive growth of information sources available on the WWW has increased the necessity of automated tools in order to find, extract, filter, and evaluate the desired information and resources for the users. For addressing the above said tool, a new concept called "**Web mining"** is introduced**.** This is a technique that can be used to study the behavior of the user of the website.

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

Web mining can be 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. By mining web usage data, more complete knowledge about the user's behavior can be obtained which helps in improving the e-business.

a) Web content mining

Web content mining mainly deals with mining, extraction and integration of useful data, information and knowledge from Web page content. It describes the discovery of useful information from the web documents. The content of the web page may be text, image, audio, video, metadata and hyperlinks etc.

Web content mining is related as well as different from data mining and text mining. It is related to data mining because many of the data mining techniques can be applied on web content mining. It is related to text mining because much of the contents of the web are texts [3]. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because the content of the web page are semi structured nature, while text mining focuses on unstructured texts. Web content mining thus needs creative applications of data mining and/or text mining techniques and also its own unique approaches.

In the past few years, there was a rapid expansion of activities in the Web content mining area. This is because of the phenomenal growth of the Web contents. The major area where the web content mining is used is resource discovery from the web, document categorization and clustering, and information extraction from web pages.

b) Web Structure Mining

World Wide Web can disclose more information than just the information contained in documents. Web structure mining helps to minimize the problems in accessing the relevant information of the World Wide Web due to its vast amount of information. It extracts previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Website to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining.

The structure of the web consists of web pages as nodes and hyperlinks as edges connecting between two related pages [4]. For example, links pointing to a document indicate the specialty or popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document.

Web Structure mining is classified into two types namely intra-page structure and inter-page structure. Intra-page structure means the existence of links within a page. No separate page will be opened in this case. Inter-page structure involves the connection of one page with the other page.

To truly utilize your website as a business tool web structure mining is a must.

c) Web Usage Mining

Web usage mining involves mining the usage characteristics of the users of Web Applications. It is a part of Web Mining, which, in turn, is a part of Data Mining. Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications.

Web usage mining mainly deals with the discovery of user access patterns from the web usage logs. It focuses on various data mining techniques to understand and analyze search patterns. It improves the navigation of information on the web which provides productive marketing and produces a higher quality of information to the user. Understanding customer behavior, evaluate effectiveness of a particular website and helps to quantify the success of a marketing campaign. Business intelligence, Competitive intelligence and Pricing analysis are also the advantages which can be gained through the application of web usage mining.

2. APPLICATION AREAS OF WEB MINING IN E-BUSINESS

Web mining can be viewed as a key enabler for the success of e-business. In e-business web mining can be used in the following areas: [5]

- ✓ Web mining can provide companies managerial insight into visitor profiles, which help the top management to take strategic actions accordingly.
- ✓ The company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will helps the e-business to improve and align their marketing strategies timely.
- ✓ In the e-business world, structure mining can be quite useful in determining the connection between two or more business Websites.
- ✓ The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.

3. STEPS IN WEB MINING

The major modules in the Web mining are data collection, data preprocessing, Pattern Discovery, Personalization and Recommendations. In these modules first three are offline modules and remaining are online modules.

3.1. Data collection

Data collection is the first step of web mining. This is an important and most difficult step. The main sources of data for web mining are server side data, client side data and proxy server data.

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, session identification and path completion.

3.2.1 Data Cleaning

The purpose of data cleaning is to eliminate irrelevant items. Since the target of web usage mining is to get the user’s travel patterns, following two kinds of records are unnecessary and should be removed:

1. The records of graphics, videos and the format information are irrelevant. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URL field of every record
2. The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

3.2.2 User Identification

The task of identification of a single user is fundamental to distinguish her/his behavior in web mining. Various methods have been proposed to automatically recognize the user. Some of the methods are using cookies, based on IP address, path analysis based on the network topology etc.

3.2.3 User Session Identification

This is related with identifying the users behavior with a particular website. The user session is identified by the set of URLs corresponding to the pages visited by a user from the time he/she enters a website to the time he/she leaves it. Heuristic method based on time and reference can be used for user session identification.

3.3 Pattern Discovery

After completing the preprocessing of web data, discover patterns of usage of website by using statistical, data mining, machine learning and pattern recognition techniques. In particular association rule and clustering techniques of data mining are very frequently used for pattern discovery.

3.4 Personalization and Recommendations

After the identification of patterns, personalization and recommendations can be given to the website administrator or owner of e-business.

4. IMPORTANCE OF WUM IN THE RECOMMENDER SYSTEM

In order to understand the importance of Web usage mining in the recommender system, a study was performed with the website “crescentexcel.com”. The study was performed with the aim to improve the effectiveness of the website by giving the recommendations to improve the e-business.

4.1 Application of Association rule mining in the recommender system

Association rule mining [5] helps us to find the interesting relationships among the pages in the website. 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.

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

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

After finalizing the associated pages, recommendation will be given to the website administrator to improve the website. He/She will 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.

4.2 Apriori Algorithm

Well known association rule mining algorithm is **Apriori** algorithm[6]. It uses prior knowledge of the frequent itemset properties. K-itemsets are used to explore K+1 itemsets. First, the set of frequent 1-itemset is found. This set is denoted as L_1 . 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 $C_{k.}$

4.3 Implementation of Apriori Algorithm

The various page of crescentexcel.com are crescent, product, solar water heater, RO water purifier, UPS etc. The pages are coded as Crescent –CRE, Products-PRD, Solar Water Heater – SWH, UPS- UPS, RO water purifier- WP etc. The server log data is collected, preprocessed and converted into the database format so that the algorithm it can be directly applied on it. The database consists of the sets of transactions and the pages visited (Table 1).

Table 1. List of visited pages

TID	LIST OF PAGES
t1	CRE, PRD,WP
t2	PRD, SWH
t3	PRD, UPS
t4	CRE, PRD, SWH
::: :::	:::: :::: :::: :::: :::: ::::

The preprocessed data is in the database, D, consisting of 82683 records. Fixing the minimum support count as 30% and minimum confidence required as 70%, the frequent itemset is found using Apriori algorithm. Then, Association rules are generated accordingly.

Generate frequent itemsets by applying the Apriori algorithm to the database D. The 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 strong association rules satisfy both minimum support & minimum confidence).

5. RESULT AND ANALYSIS

Considering the candidate set C_4, the association rules are framed which is shown below, with its confidence:

RULE 1:
CRE ^ PRD ^ UPS → WP, *Confidence* = 310 / 460 = 67%

RULE 2:
CRE ^ PRD ^ WP → UPS, *Confidence* = 310 / 579 = 53%

RULE 3:
CRE ^ UPS ^ WP → PRD, *Confidence* = 310 / 310 = 100%

RULE 4:
PRD ^ UPS ^ WP → CRE, *Confidence* = 310 / 330 = 93%

RULE 5:
UPS ^ WP → CRE ^ PRD, *Confidence* = 310 / 450 = 68%

RULE 6:
PRD ^ WP → CRE ^ UPS, *Confidence* = 310 / 590 = 52%

RULE 7:
PRD ^ UPS → CRE ^ WP, *Confidence* = 310 / 330 = 93%

RULE 8:
CRE ^ WP → PRD ^ UPS, *Confidence* = 310 / 564 = 54%

RULE 9:
CRE ^ UPS → PRD ^ WP, *Confidence* = 310 / 410 = 75%

RULE 10:
CRE ^ PRD → UPS ^ WP, *Confidence* = 310 / 355 = 87%

RULE 11:

CRE → PRD ^ UPS ^ WP, *Confidence* = 310 / 21179 = 1.46%

RULE 12:

PRD → CRE ^ UPS ^ WP, *Confidence* = 310 / 21050 = 1.47%

RULE 13:

UPS → CRE ^ PRD ^ WP, *Confidence* = 310 / 1697 = 18.26%

RULE 14:

WP → CRE ^ PRD ^ UPS, *Confidence* = 310 / 1049 = 29.55%

If the minimum confidence threshold is 70%, then only the first, fourth and fifth rules above are output, i.e., **R3**, **R4, R7, R9** and **R10** are **selected** and considered as interesting strong rules and the remaining rule are rejected due to less confidence.

Now recommendations are given to crescentexcel.com website administrator to improve the quality of the pages CRE, UPS, WP and PRD. At the same time the rearrangement of pages are quite necessary here. For example those who visit CRE and UPS also visited PRD and WP. So the administrator should change website design in such a way that if the visitor visits CRE and UPS simultaneously, PRD should be the next page to visit in order to avoid unnecessary crawling through other pages.

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 to increase the number of visitors/customers. It also helps to retain its existing visitors/customers from moving away to some other websites. The webpage redesigning and the website restructuring, resulted in increasing the number of visitors/customers, which in turn increased the success rate of e-business i.e the profit of e-business has increased tremendously.

From the above study it is understood that web mining is the back bone to improve the quality of the website based on the recommendations. So it can be concluded that “web mining is a key enabler of recommender system in e-business”.

7. References

1) Mingyu Lu, Shuying Pang, Yan Wang, Yuchang Lu, Lizhu Zhou, “WebMe – Web Mining Environment”, 2002 IEEE,SMC,WPIRS.
2) Nivedita Roy, Tapas Mahapaatra, “Web Mining – A Key Enabler in E-Business”, IEEE 2005
3) Ramesh Yevale, Mayuri Dhage, Tejali Nalawade,.Trupti Kaule. Unauthorized Terror Attack Tracking Using Web Usage Mining, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1210-1212
4) Sonia Gupta and Neha Singh,” Web Miining: Summary”,International Journal of Computational Engineering Research,Volume 03,Issue 4,April 2013,Page 149
5) B.Santhosh Kumar, K.V.Rukmani (2010), “Implementation of Web Usage Mining using APRIORI and FP Growth Algorithms”, International Journal of Advanced Networking and Applications, Vol: 01, Issue: 06, Pages: 400-404.
6) Jiawei Han & Micheline Kamber, “Data Mining Concepts and techniques”(2005).