Analysis of Web Log Mining Based on Association Rule

Chunya Li, Yan Li, Yang Yang, and Yangyang Deng

Abstract With the development of computer technology and the popularity of the Internet, Web data breaks through the limitations of traditional data formats, and it becomes more and more important, becoming an effective way for Web users to better obtain information. Web log data is data that records user access information to Web sites, stores a large amount of path information, and user access patterns obtained by mining these log information, in personalized information services, improved portal site design and services, and targeted E-commerce, building intelligent Web sites and improving the reputation and effectiveness of the site will play an important role. However, due to the particularity of Web data and applications, traditional mining techniques cannot be directly applied to Web mining. This paper first preprocessing the Web Log data through a real estate Web site, after cleaning and deleting invalid data, by using the method of simple association rule to find the characteristics of the user's search behavior, thereby providing relevant suggestions to the Web site and improving the user experience of the Web site.

Keywords Web log mining · Apriori algorithm · Association rule

1 Introduction

Due to the rapid development of the Internet, the amount of information on the Internet is growing at an exponential rate. The problem is that the amount of information on the Internet has far exceeded the processing power of people, which directly led to Data Mining (DM) technology generation, data mining [1] refers to the process of mining knowledge from a large amount of data stored in a database or other information base, it is a process of discovering and mining useful knowledge from massive data. Web data mining is the application of data mining on the Web. From data content such as Web page content, page structure, and server log files, data mining technology is used to extract potential and implicit information with

C. Li (✉) · Y. Li · Y. Yang · Y. Deng
Business School of Nantong Institute of Technology, Nantong 226002, Jiangsu, China
e-mail: 10880945@qq.com

C.-T. Yang et al. (eds.), *Innovative Computing*,
Lecture Notes in Electrical Engineering 675,
https://doi.org/10.1007/978-981-15-5959-4_81

application value by mining the Web logs, the Web site administrator can access the log files to understand the user's browsing mode and the usage of the Web site, thereby optimizing the structure of the Web site, which can facilitating user browsing, improving access efficiency, and providing personalized recommendations for users.

2 Literature Review

Chen [2] and other scholars discovered in 1996 that data mining methods can be used in Web technologies. In 1997, Cooley [3] first proposed the concept of Web log mining. He analyzed the user's access patterns by analyzing the rules in the Web logs, thus helping the Web site identify potential users and provide better services for users. In 1999, Spiliopoulout [4] proposed an access mode miner Web Utilization Miner (WUM), which can process the log data, display the sequence pattern structure using MINT query language, and provide visual environment and model query function. Developed by Simon Fraser University, Web Log Miner composes log data into data cubes and then performs online analytical processing (OLAP), providing a variety of data mining methods including association rules and sequence models [5]. The WebMiner system at Minnesota University is a general-purpose Web log mining architecture that enables the system to automatically discover serial patterns and association rules for Web logs by converting Web log data into transactional data formats [6].

In the research field of Web data mining, the domestic research on data mining started later than foreign research. In 1999, Ning [7] reviewed the foreign application of data mining technology in the Internet, including the mining of path patterns, how to perform intelligent queries, discovery of user interest patterns, and more. Bin [8] also introduced the E-OEM model, based on customer data and log data mining to find meaningful user access patterns and potential customer groups. Dongshan and Yiyi [9] designed the Web site access matrix and proposed concepts such as support preference to explore user preference browsing paths. As far as the application is concerned, the most influential company in China for data mining research is Finite-Futong [10, 11], which developed the business intelligence platform (BI Office), and based on the platform's decision analysis software packages in retail, banking, power, tax, customs, communications and many other industries, related companies using the platform can make relevant decisions using relevant functions.

But so far, there have been relatively few applications for data mining technology in the real estate industry. The factors that influence customer selection are based only on simple statistical summaries. They are often hidden and undetectable for certain underlying factors, and this paper uses the data mining technology to identify the relationship between various factors through the data log from a real estate Web site to find the rules between various factors and the characteristics of the browsing behavior, thus providing advice on the construction of real estate Web sites and improving user access efficiency as well as access experience.

3 Data Preprocessing

To mine Web log data, it is necessary to preprocess the log. The purpose is to delete the useless data in the log, normalize the data, and convert the log into a format recognizable for data mining. Data preprocessing includes data cleaning, user identification, session identification, etc. The data in this article is the Web log data, and Web log information includes date and time of access, user IP address, requested query, path of click, access method, request method, and user agent.

3.1 Data Cleaning

Data cleaning refers to removing redundant invalid data and erroneous data in Web log data according to requirements, that is, deleting data records unrelated to data mining: data cleaning includes six steps, step 1: deleting files that are automatically downloaded by Web pages like images, audio files, and those files are usually end with ico, css, jpg, gif, png, js, swf, etc., step 2: Delete the record of the access request failure. That is, only the record with the status code 2XX is kept. Step 3: Delete the record which its request method is not "GET" or "POST." A variety of request methods are recorded in the Web log, "GET" means user information acquisition, suitable for most requests; "POST" is used to update the site; "HEAD" identifier only requests the header of the page; "OPTIONS" refers to allowing the client to view server performance. So for data mining, we just keep the record of the request method of "GET" and "POST." Step 4: Delete the Web Robots record. Web Robot, also known as "Web Crawler," is a program that automatically extracts Web pages and is an important component of search engines. Since the access of the "crawler" is purposeless and does not represent the behavior of the user itself, we have removed the common "Web Crawler" in the Web log. Step 5: Delete the local browsing access record. The data used in this article is Web log records from a company's Web site server. As a Web site, employees need to continuously update and maintain the Web site, which results in a large number of local access records, so it needs to be deleted here. Step 6: Delete abnormal access data. If some IDs have more than 10,000 daily visits, this should be the machine automatically collecting data, so it needs to be deleted. Finally, 10,336,000 queries remain after data cleaning.

3.2 User Identification

In reality, due to the existence of caching, proxy servers, firewalls, etc., it is very difficult to determine the uniquely determined user in actual use. For example, multiple users in the same LAN can use the same IP address to access the Web site. So, only IP address cannot accurately identify the user. In general, cookie information can be used to distinguish users. However, not all Web sites use cookies, and some users block cookies because of privacy settings. Therefore, there are problems in using cookies to identify users. Thus, according to the analysis of the Web log file format, this paper uses IP + USER_AGENT to identify the user [12] that means different IP represents different users, and the same IP has different USER_AGENT information in the log, is determined to be a different user. Then, add a [userid] column to the database table.

3.3 Session Identification

A user's effective access to the Web can be defined as a user session. Different users represent different sessions. In the same session, one user accesses different pages. The session is a sequence of all Web pages accessed by the user during the access process. J Pitkow's research shows that when the same user stays on a certain page for more than 25.5 min, it can be judged as a different session of the same user. Currently, 30 min is generally used as the criterion for judgment [13]. In this paper, we use 30 min as the criterion for distinguishing sessions that means the time difference between two adjacent clicks of the same user exceeds 30 min determined to be two sessions. Also, add the [sessionid] column to the table in the database.

4 Simple Association Rules

Association rules are one of the important research directions in Web log mining. Association rules are mainly used to represent the interdependencies between transactions. If there are two transactions in a transaction set or there are associations between multiple transactions, then the attribute value of a transaction can be inferred based on the attribute values of other transactions.

4.1 Measurement Indicators for the Validity of Simple Association Rules

According to the sample data, we can get a lot of association rules, but not all association rules are valid because some rules may not be convincing, and some rules have limited scope which are not valid associated rules, valid association rules should meet certain measurement indicators, and the most common indicators are rule confidence and rule support.

(1) Confidence is a measure of the accuracy of a simple association rule. Describes the probability of including item Y in a thing containing item X, reflecting the

possibility of Y appearing under the condition of X occurrence. For example, if X stands for bread and Y stands for milk, such as bread milk ($S = 85\%$, $C = 90\%$), it means the customer buys bread whose possibility of buying milk is 90%.

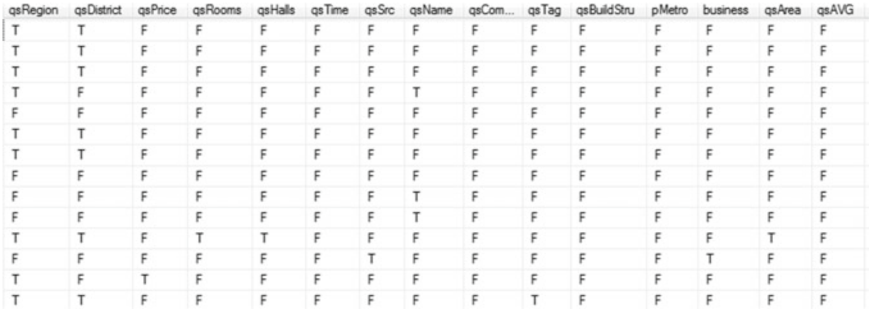
(2) The support of the rule measures the universality of the simple association rule, indicating the probability that the item X and the item Y appear at the same time. For example, the bread and milk example above, $s = 85\%$ means that the probability of purchasing milk and bread at the same time is 85%.

4.2 Measurement Indicators for the Practicality of Simple Association Rules

An ideal simple association rule should have higher confidence and support. If the rule support is higher but the confidence is lower, the rule is less trustworthy. Otherwise, the application of the rule is rare. In the actual situation, a valid association rule does not mean that the rule is practical. In other words, an effective rule does not necessarily have the correct guiding significance. Therefore, you can also consider the following measurement indicators: 1. rule lift. That is, the ratio of confidence to support of the latter. The degree of rule lift reflects the extent to which the occurrence of Project X affects Project Y. It usually makes sense to have a degree of lift, and the greater the degree of rule lift, the better. 2. Confidence Difference. Similar to the rule promotion, it also utilizes the support of the latter, which is the absolute difference between the rule confidence and the support of the latter.

4.3 Application of Simple Association Rules

The most famous algorithm for association rules is the Apriori algorithm, which was proposed by Agrawal and Srikant in 1994. The process consists of two parts, first generating frequent itemset and then generating association rules based on frequent itemset. This time we use the Apriori algorithm in software of SPSS Modeler to discover the association rules of the query in the user access log, the data used in this article is the user access log in the second-hand housing site of a real estate Web site. Query contains the preferences of the user's choice, such as region, room type, price, and area, finally, there are 10,336,000 records, and 14 options have been analyzed. The purpose of this excavation is to find the association between the user's browsing behavior preferences and provide suggestions based on the discovery to improve the user search efficiency.



qsRegion	qsDistrict	qsPrice	qsRooms	qsHalls	qsTime	qsSrc	qsName	qsCom...	qsTag	qsBuildStru	pMetro	business	qsArea	qsAVG
T	T	F	F	F	F	F	F	F	F	F	F	F	F	F
T	T	F	F	F	F	F	F	F	F	F	F	F	F	F
T	T	F	F	F	F	F	F	F	F	F	F	F	F	F
T	F	F	F	F	F	F	T	F	F	F	F	F	F	F
F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
T	T	F	F	F	F	F	F	F	F	F	F	F	F	F
T	T	F	F	F	F	F	F	F	F	F	F	F	F	F
F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
T	T	F	T	T	F	F	F	F	F	F	F	F	T	F
F	F	F	F	F	F	T	F	F	F	F	F	T	F	F
T	F	T	F	F	F	F	F	F	F	F	F	F	F	F
T	T	F	F	F	F	F	F	F	T	F	F	F	F	F

Fig. 1 Query (Boolean matrix type)

4.3.1 Data Conversion

Since the Apriori algorithm in SPSS Modeler can only process data of Boolean matrix type, we first need to convert the data as Fig. 1

Each row in Fig. 1 represents a search behavior for a user, *T* indicates that the user has selected the selection in this search, and F indicates that the user has not selected the option.

4.3.2 Rule Setting and Result Output

Import the processed data into SPSS Modeler, connect the Apriori model in SPSS Modeler, select the minimum confidence level of 80% in the tab, the minimum support is 5%, and set the confidence difference to 5% in order to verify the practicability of the rule. Run the software and get the following results:

As can be seen from Table 1, there are a total of five association rules. Rule support between 5.501% and 34.411%; confidence level between 89.371% and 99.999%; all rules' lift is greater than 1, which means all the rules are acceptable.

Table 1 Result of simple associate rule

Latter	Former	Rule ID	Number	Support%	Confidence Level%	Rule Support%	Lift
Region	Area	1	970,401	9.361	89.371	8.366	1.544
Region	Room type	2	1,015,543	9.797	95.514	9.357	1.65
Region	District	3	3,567,110	34.412	99.997	34.411	1.727
Region	Area and District	4	570,271	5.501	99.999	5.501	1.727
Region	Room type and District	5	656,162	6.33	99.998	6.33	1.727

Rule 1: Area → Region ($S = 8.366\%$, $C = 89.371\%$) This indicates that 89.371% of users who have selected the area option may select the region option, and the probability that the area will appear together with the region is 8.366%.

Rule 2: Room type → Region ($S = 9.357\%$, $C = 95.514\%$) indicates that the user who selected the room type option has 95.514% probability of selecting the region at the same time, and the probability of selecting the room type and the region is 9.357%.

Rule 3: District → Region ($S = 34.411\%$, $C = 99.997\%$) indicates that 99.997% of the regions may be selected after selecting the district, and the probability of selecting region and district is 34.411%.

Rule 4: Area and district → region ($S = 5.501\%$, $C = 99.999\%$) indicate that the probability of selecting the region option after selecting area and district is 99.999%, and the probability that these three are selected at the same time is 5.501%, which reflects the low generality of this rule.

Rule 5: Room type and district → region ($S = 6.33\%$, $C = 99.998\%$), which reveals that the probability of selecting region option after selecting room type and district is 99.998%, and the probability of occurrence of these three is 6.33%.

The five rules generated have a confidence level of more than 89%, which indicates that the accuracy of these five rules is very high, and the rule support degree is not high except for the rule 3. However, since the degree of lift is greater than 1, this shows that although the generality of the rules is low, these rules are within the scope of acceptance.

4.3.3 Drawing and Analysis of Network Diagram

In order to more intuitively reflect the strength of the association between the options, we also use the network module in the SPSS Modeler to draw a network diagram of all links, as shown in Fig. 2:

The thick lines in Fig. 2 represent the strong associations, the thin lines represent medium associations, and the dotted lines represent weak associations. The criterion is that the number of links is more than 25% of the total number of links is a strong link, and less than 10% is a weak link. Therefore, for this sample, a strong link refers to a link of 891,749 or more, and a weak link refers to a link below 356,700. As can be seen from Fig. 2, there are two strong links and ten medium links. Region and District, Region and Rooms are strong links; Region and Price, Region and Area, Region and Housing Source (Src), District and Rooms, District and Area, District and Price, Rooms and Area, Price Area, Price and Rooms, District and Housing Source (Src) are medium links, others are all weak links. This also shows that the user is most concerned about the location of the house during the search process, followed by the size and type of the house and the price of the house and the source of the house. In addition, it is worth noting that the strength of the link between queries is not directly related to the generation of association rules. Although the association rules are generated between the strong links in the data of this sample, it is not said

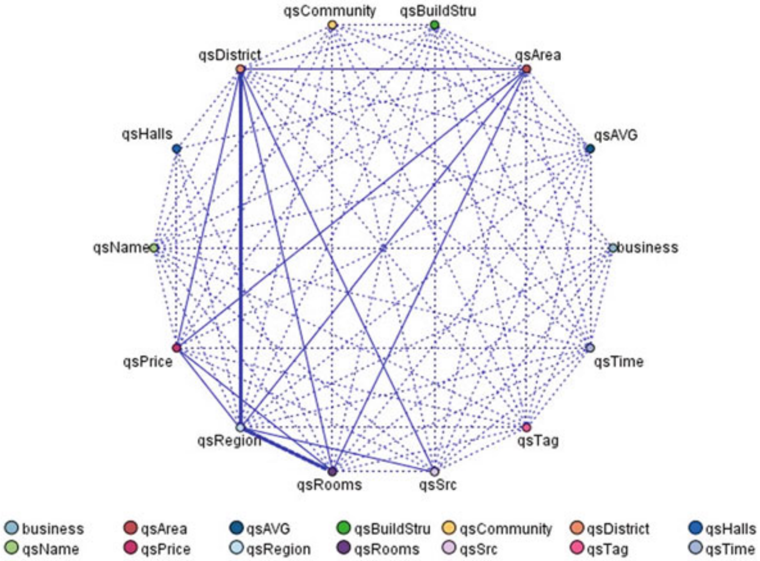


Fig. 2 Query associated mesh

that there is a strong link can generate simple association rules. The generation of simple association rules needs to meet the thresholds of various parameters of simple association rules. In this paper, the minimum confidence should be 80% or more, the minimum support should be greater than or equal to 5%, and the confidence difference should be greater than 5%. Strong links only represent a higher degree of rule support, but it is possible that the confidence level between the two does not meet the set threshold, so a simple association rule cannot be generated. In summary, the confidence of the five rules generated by this sample data is very high, but the rule support degree is generally not high except for rule 3. The rule support of the other four rules is generally 5–10%, this shows that except for rule 3, the other four rules are generally low. This shows that the user pays more attention to the geographical location of the residence and the size of the house, and the attention to the average price, label, transaction type, etc., of the house is relatively low. That is to say, when the user searches for a house, the average price of the house and other factors have little influence on the user's search behavior. Therefore, for the Web site, when setting the search option, the position of the four options of region, district, area, and room type can be placed in front, and the less attention-oriented option can be placed behind or deleted, thereby improving the user' search efficiency and providing users with a better access experience.

5 Conclusion

This paper uses the Web log data of a real estate Web site and uses simple association rule analyzes the association rules of user search behavior.

Firstly, based on the relevant theories, this paper conducts a comprehensive analysis of the research status of Web logs. Secondly, it preprocesses the original log date data, establishes a new database table to meet the requirements of subsequent analysis, and then searches for user search options. The association rules were mined. From the sample data, five rules were found. Those five rules have high confidence, and the degree of lift is greater than 1, but the support of the other four rules except the rule 3 is not high enough, which shows that the five rules are valid and have some practical significance, while the low support indicates that the rules are not universal.

Then, in order to more intuitively reflect the strength of the association between the options, the corresponding mesh diagram is also be drawn, but the strong link does not mean that the association rule will be generated because the strong link only indicates that the rule support is higher, and it does not represent the confidence level in the rules. At the same time, because the rules are generated between the four options of room type, region, district, and area, the options such as average price and source are less concerned. Therefore, in order to improve the search efficiency of users, to improve the user access experience, the Web site can consider setting the location of the former to the front end, while the option with lower attention can reduce the setting or set the position later.

Acknowledgements This work is support by "The Second Batch of Young and Middle-aged Research Scientists" of Nantong Institute of Technology (Grant No. ZQNGG206), the Philosophy and Social Science Fund of Jiangsu Provincial Department of Education (Grant No. 2018SJA1287), and the 13th Five-Year Plan of Jiangsu Province "Key Construction Discipline Project of Business Administration Level 1" (SJY201609).

References

1. Luo, Q. 2008. Advancing knowledge discovery and data mining. In *1st International Workshop On Knowledge Discovery and Data Mining, WKDD*, 3–5.
2. Chen, M., J. S. Park, and P. S. Yu. 1996. Data mining for path traversal patterns in a web environment. In *Proceeding of the 1996 16th International Conference on Distributed Computing Systems*, 385–392.
3. Cooley, R., B. Mobasher, and J. Srivastava. 1997. Web mining: information and pattern discovery on the world wide web. In *Proceedings if the 1997 IEEE 9th IEEE International Conference on Tools with Artificial Intelligence*, 558–567.
4. Spiliopoulou, M., L. C. Faulstich, and K. Winkler. 1999. A data miner analyzing the navigational behaviour of web users. In *Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI 99*, 588–589.
5. Jie, Feng. 2004. *Research on web log mining related algorithms and its original statistical design*, 3. Chengdu: Southwest Jiaotong University.

6. Wu, X., V. Kumar, J.R. Quinlan, et al. 2008. Top10 algorithm in data mining. *Knowledge and Information System* 14: 1–37.
7. Ning, Chen, and Zhou Longzhen. 1999. Application of data mining in the internet. *Computer Science* 26 (7): 44–49.
8. Bin, Zhou and Wu Quanyuan. 1999. Research on model and algorithm of user access pattern data mining. *Computer Research and Development* 36(7): 870–875.
9. Dongshan, Xing, Shen Yiyi, and Song Yubao. 2003. Mining user views and preference paths from web logs. *Chinese Journal of Computers* 26 (11): 1518–1523.
10. Finette's Official Website. Finette-Futong Bank Business Intelligence Solution [EB/OL] 2018.4.12.
11. He, Wang, and Liu Wei. 2011. Network log analysis based on data mining. *Journal of Suzhou University* 27 (2): 43–47.
12. Cooley, R., B. Mobasher, and J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and Information System* 1 (1): 5–23.
13. Srivastava, J., R. Cooley, M. Deshpande, et al. 2000. Web usage mining: discovery and applications of usage patterns from web data. *Proceedings ACM SIGKDD* 1 (2): 12–23.