Research on the Model and Algorithm of User Access Pattern Data Mining

Junfeng Cheng

Longnan Teachers College, 742500, China

Keywords: Data Mining and Knowledge Discovery, User Access Pattern, E-Commerce.

Abstract: Developing e-commerce from a modern enterprise strategy - almost all the data recorded by customers is of great significance to the potential customer groups related to user access patterns. Among them, the key technical support of E - OEM model is based on the same user, page, pattern and lead mining as well as server topology and user considerations of multiple data sources. The validity of the experiment and the model is proved.

1. Introduction

The maturity of technology makes it possible for technology to penetrate into all aspects of social life in an amazing proportion. The information sharing, exchange and division of labor among educational and academic research institutions, the cooperation within, among and among enterprises, the transformation of e-commerce to the traditional business model, and the information of human interaction must be quantified electronically. Take the web server log as an example, the log data of several web hotspots increases by tens of megabytes per day. It contains useful and important knowledge (patterns, rules, visual structure, etc)[1]. which is another important research field of data mining and knowledge discovery. One of the most important application areas of DMKD is business purpose, which is to discover the shopping mode of customers and support the business decision of shopping center. Digitalization and informatization of life, internationalization and competition of business, DM - KD technology are widely used and developed in many fields. One of the important research topics in the field of information decision-making. Based on the research background of webmdkd, which is the background of e-commerce application, several problems encountered in this field are studied. In a word, housing is not ideal. Discuss the solutions to these problems.

2. Related Work

From the point of view of research objectives, the existing research based on Web server log data is to analyze system performance, improve system design, and understand user intention. The main technologies used are different[2]. The research on the performance analysis of this system is mainly from the statistical point of view. The data items recorded by frequently visited web pages, the number of visits per unit time and the amount of data accessed are simply counted. Most of the existing commercial and free web log analysis tools are in this category. The users of these tools are generally web server managers (web master, etc.)[3]. The design and construction of web server is mainly complex, designers and users need to constantly adjust themselves to change. Perkowitz is a research on how to automatically or semi automatically adjust the expression of organization or web server log data. Chen and DMKD overlap. Chen proposed an algorithm to discover the operation mode of document path. The algorithm finds the path of frequent users from the web server log. In addition, it is also meaningful to protect the data of log analysis and use OLAP technology. In the above research, specific patterns and rules are found from a large number of Web log data, but the current research results are not enough[4]. No patterns and rules were found in these. Ideals are one of the main problems. In this case, the information of the pattern or rule area is not enough, which is difficult for users to understand. For example, pages a, B, C, and D can find the next pattern and often access the path. Other information, such as the characteristics of the user and the potential customers in the path, depends on the user's further analysis. On the other hand, the algorithm is too

  DOI: 10.25236/ictmic.2020.114

small, resulting in too small patterns and too sensitive data.

Table 1 Confusion matrix

	Pre_yes	Pre_no	Total
Act_yes	TP	FN	P=TP+FN
Act_no	FP	TN	N=FP+TN
Total	TP+FP	FN+TN	

3. Model and Algorithm of User Access Pattern Data Mining

This research is based on the application of online shopping in e-commerce. Business people build their own online product catalog on the Internet. You can browse your customer directory (ie, user), or your order can be paid online through your browser[5]. It can store process information (including user's login information and user's reading resume) and user's personal information in the form of record file or customer database. It is very important to find the regularity of commercial marketing. In this article, we investigated how to mine meaningful user access patterns and potential customer groups from a large number of customer and log data. Such knowledge can develop professional promotion strategies.

3.1. E-Oem Data Model

OEM (object exchange model) model is a kind of data model that describes semi-structured data. In order to mine more meaningful knowledge from data, we comprehensively consider multiple data sources and domain knowledge, such as application logic design of server, page topology and user's browsing path. Therefore, we propose an extended OEM model, E-OEM, to describe the problems discussed in this paper

Definition 1. Object o consists of object $id(o)$ and object value $val(o)$. $id(o)$ uniquely identifies object o in object space. Object value $val(o)$ can be in the following two forms.

Atomic form. $val(o)=\{l_1:d_1,\cdots l_m:d_m\}$

Citation form. $val(o)=\{l_1.id(o_1),\cdots l_m:id(o_m)\}$

Definition 2. All the objects (including document objects and application objects) that can be accessed by users on the web site constitute the object space[6]. All the objects in the object space are described by the E-OEM model. On this basis, we construct the basic information that reflects the objects in the object space as domain knowledge to guide the data mining process.

3.2. Application Related Issues

Based on the E-OEM data model described in the previous section, this section describes several related issues.

The object identification of document object and application object is defined as the function of its URL (Universal Resource Locator). The function value is unique in the object space. In addition, we also obtain the following basic information of the object: meta information. For document object, including URL, file size, latest modification time, etc[7]. For application object, record its CGI program name and topology information, including in the object space For document objects, including several attribute names and attribute values in the object; for application objects, we record the effective parameter combination of CGI program as its function description.

Since all users' browsing on the server is registered in the log table, the problem that must be solved before data mining is the demarcation of transactions between users and merchants

Definition 3. Let n be the natural number set, $t(O_i)\in N$ be the access time of object O_i, $C(O_i)\in N$ be the user ID of object O_i ,If P satisfies the following conditions: $C(O_i)=C(O_j),i,j=0,1\cdots n;T(O_i)\le T(O_j),0\le i\le j\le n$ then P is an access path of user C. definition 4. Let t be a set of access paths of user C, for any object O_i,O_{i+1} contained in path P1,

any object O_j contained in path $P_1, P_2, P_2 \in T$ and $P_1 \neq P_2$, if the following conditions are true: $T(O_i+1)-T(O_i) \leq \max time; |T(O_i)-T(O_j)| > \max time$, then t is a transaction of user C, where max time is the maximum time interval defined by user. Define 5 to calibrate the transaction based on page browsing time. In the implementation, we also consider the shortest path principle (path with too small filtering length). The transaction database records all transaction information of customers and merchants, and the data items about browsing path can be directly from the web server log In the research of this paper, we divide transaction database d into two parts: positive transaction database $D+$ and negative transaction database $D-$. $D+$ contain all customer transactions transacted with merchants; $D-$ contain browsing but not transacted transactions. Generally, $D-$ is much greater than $D+$.

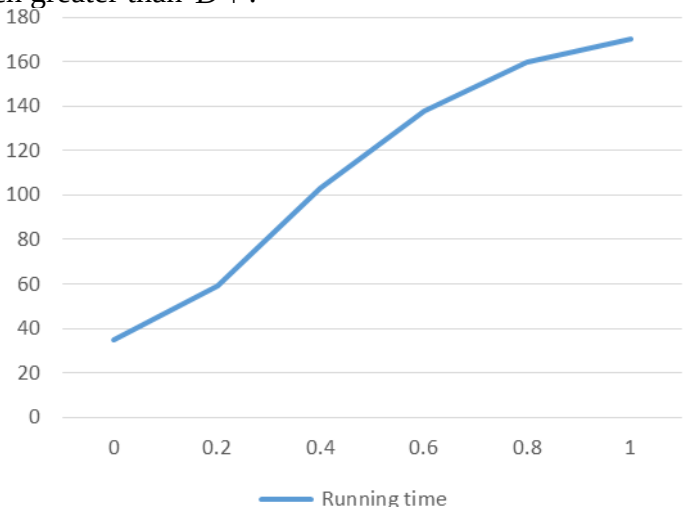

Figure 1 Logistic ROC curve

3.3. Algorithm

Customer's shopping mode and potential customer group can be expressed in many forms and methods. In this paper, customer's shopping mode is expressed as access mode (in this paper, it is expressed in capital letters)[8]. It consists of multiple frequent access paths in a transaction (in this paper, it is expressed in lowercase letters). On this basis, potential customer group is defined as user group based on frequent access mode

Definition 5. For a given positive transaction database $D+$, user-defined minimum support degree $\min\sup \in [0,1]$, if access mode $P = \langle P_1, P_2 \cdots P_K \rangle$ satisfied conditionat least $\min\sup \times |D+|$ transaction set of users contains pattern P, then p is called frequent access pattern of customers, where $|D+|$ represents the number of users in $D+$

Definition 6. For a given negative transaction database $D-$, customer frequent access pattern set $\{P_1, P_2, \cdots, P_K\}$ The equivalent class $C_i \in (D-/R(D-, P_i))$ is called the ith potential customer group in $i = 1,2 \cdots k$.

In the implementation, we use two-layer hash tree to disperse the access patterns to different matching trees. On the one hand, it can avoid the over width of the matching tree, accelerate the matching process, on the other hand, it is conducive to the parallel processing of data blocks. The nodes of the matching tree are the similarity measurement function, and the labels on the edge represent the path. The algorithm works as follows[9]: when the new pattern reaches the root of the tree, the measurement function calculates the If the similarity is less than min SIM, a new sub node and corresponding edge will be created. The first path on the edge is the first path of the new mode, and the sub node counter is set to 1. Otherwise, the sub node with the largest similarity will be taken as the descending node, and the current node counter will be increased by 1. The new mode will delete the first path and descend to the lower node Point, measure the similarity between the current path of the new mode and the path of its lower level nodes, and repeat the process until the new mode is empty. The corresponding access mode of each node is calibrated from the tree root to each

side label of the node in turn. The support of the access mode is the ratio of the current node counter and the root node counter.

4. Performance Simulation and Analysis

In order to evaluate the performance of the algorithm, we have carried out the following simulation experiments under the environment of Pentium 266 / 64M RAM / Windows 95 / MS Visual C + + 5.0: Based on the five pre-determined access modes, using Markov chain model, according to the parameters , respectively generate $D+$ and $D-$; apply the access mode discovery algorithm to $D+$, adjust min SIM, The time cost of the algorithm is examined, the algorithm of lead group discovery is applied to $D-$, min SIM is adjusted, and the classification accuracy of the algorithm is examined.

The running time of the algorithm is longer than that of the corresponding sequential pattern mining algorithm. This is mainly because we consider the concept of similar path, which leads to the increase of search space. The running time of the algorithm increases approximately linearly with the path similarity. The classification accuracy of the algorithm in Figure 4 is up to 66%. It shows the effectiveness of the algorithm.

5. Conclusion

Mining meaningful user access patterns and related potential customers from a large number of customer data and log data will play an important role in business decision-making. Based on the E - OEM model, the logic design of potential customers and application servers, paging topology, user proximity and other comprehensive considerations proves that the algorithm is effective. In addition, it is applied to the actual data and optimized according to the data.

References

[1] James, Xue., Stephen, Jarvis. Mining association rules for admission control and service differentiation in e-commerce applications. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, vol. 8, no. 11, pp. e1241, 2018.

[2] Antonis, Matakos., Evimaria, Terzi., Panayiotis, Tsaparas. Measuring and moderating opinion polarization in social networks. Data Mining & Knowledge Discovery, vol. 31, no. 5, pp. 1480-1505, 2017.

[3] Mohammad, Al Hasan., Vachik, S. Dave. Triangle counting in large networks: a review. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, vol. 8, no. 2, 2017.

[4] Kun, He., Wu, Wang., Xiaosen, Wang. A New Anchor Word Selection Method for the Separable Topic Discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019.

[5] Kun, He., Wu, Wang., Xiaosen, Wang. A New Anchor Word Selection Method for the Separable Topic Discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019.

[6] Gao, Y., Yue, K., Hao, W U. Construction and inference of latent variable model oriented to user preference discovery, 2017.

[7] Chen, Chen., Hanghang, Tong., Lei, Xie. Cross-Dependency Inference in Multi-Layered Networks: A Collaborative Filtering Perspective. Acm Transactions on Knowledge Discovery from Data, vol. 11, no. 4, pp. 1-26, 2017.

[8] Kauffman, R.J., Kim, K., Lee, S.Y.T. Combining machine-based and econometrics methods for policy analytics insights, no. 25, 2017.

[9] Huang, Hongzhan., Arighi, Cecilia, N., Ross, Karen. E. iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Research, no. D1, pp. D1, 2017.