Journal of Software Engineering and Applications, **2013, 6, 30-33**
doi:10.4236/jsea.2013.67B006 Published Online July 2013 (http://www.scirp.org/journal/jsea)

The Application of Book Intelligent Recommendation Based on the Association Rule Mining of Clementine

Jia Lina, Mao Zhiyong

Graduated School, Liaoning Technical University, Hu Ludao, China.
Email: jelena1988@sina.cn

Received May, 2013

ABSTRACT

The traditional library can't provide the service of personalized recommendation for users. This paper used Clementine to solve this problem. Firstly, model of K-means clustering analyze the initial data to delete the redundant data. It can avoid scanning the database repeatedly and producing a large number of false rules. Secondly, the paper used clustering results to perform association rule mining. It can obtain valuable information and achieve the service of intelligent recommendation.

Keywords: Data Mining; Association Rules; Clustering; Intelligent Recommendation; Clementine

1. Introduction

The recommended service plays an important role in the process of the digital library gradually toward personalization and intelligent. The system can recommend books to the readers by the relevant information which is found from the readers' lending behavior and preferences from data mining. Relevance information mining is association rules mining[1]. This question has been paid attention and studied by many international researchers after it has been put forward by Rakesh Agrawal and researchers also raise many kinds of algorithms.

Association rules are put forward to break the transaction limit. To find the relationship between different transactions so that to predict events that users interest reasonably. It will be a long time to do the data mining and the rules will be a lot with false rules when transaction analysis is carried out on the large database. And the mining efficiency is reduced. Based on it, this paper uses the data mining software Clementine to clustering analysis on the reader firstly, and cluster the behavior of borrowing books for high frequency, medium frequency and low frequency[2]. To do the association rule mining to the books which is borrowed by readers who borrow by high frequency and medium frequency? Finally, transfer the mining result to the client user by Web service. Choose the books borrowed by users which are the high frequency and medium frequency to have the association rule mining is because the amount of borrowed books is huge and the association rule is strong. So it narrows the amount of data involved in association rule, save scanning time, and then to improve the quality of mining.

2. Clementine Software Introduction

Clementine is data mining software developed by SPSS company. It puts clustering, association rules, decision trees, neural network and many kinds of data mining technology to integrate in the intuitive visual graphic interface. Clementine combine with business technology to build the data model quickly to apply it to business activity and help people to improve the decision making process. The paper applies clustering and association rules mining in Clementine 12.0 to book intelligent recommendation service[3].

2.1. Characteristics of Clementine

1) It provides that visual, strong and easy-to-use data mining platform. The process of user modeling is to connect each node. It can be built the data mining model without programming so that user can be more focused on the solving specific business problems by using data mining rather than the use of tools.

2) Fully follow the CRISP-DM standards to establish. Clementine provides good project management function. And it can manage overall process effectively from business understanding to result release.

3) It provides steady and strong release function. Clementine can release data mining model or the whole flow of data mining to improve efficiency of operations.

4) High flexibility and extensibility. Clementine has open database interface. It provides almost all the rela-

tionship database. Meanwhile, it owns extended function.

2.2. Six Stages of CRISP-DM Process Model

1) Business understanding. It is the most important stage in data mining. It includes that confirm business object, estimate situation, confirm target of data mining and set out engineering plan.

2) Data understanding. It provides materials of data mining to realize data characteristics of data source. It includes that collect initial data, describe data, clean data, and check the quality of data.

3) Data preparation. Classify the data source from data mining. It includes that data selection, cleaning, structure, integration and formatting.

4) Modeling. It is the core part of data mining. It includes that choose modeling technology, generate test design and structure and evaluation model.

5) Model evaluation. It can evaluate result of data mining that can help to realize business target after choosing the model. It includes that result, view the process of data mining and confirm the next step.

6) Result deploys. It can combine the new knowledge with daily business flow to solve initial business problems. It includes that plan deploy, monitoring, maintain, produce final report and review the project[4].

3. Library Data Mining Based Clementine

The information requests and forms of users in library are diversified. It provides personalized recommendation service based on the requests and interests of readers. The paper clusters analysis to the times of readers. It can be divided into three types: high frequency, medium frequency and low frequency. And then association rules analysis to the books which are borrowed by high and medium frequency readers to realize personalized recommendation service[10].

3.1. Data Acquisition

The data in this paper is from library in Liao Ning Technical University. The total amount of reader borrowing books is 62261 from Nov 7th, 2011 to Mar 7th, 2012. And extract 3108 from it to serve as the experimental subject.

3.2. Data Pre-Processing

The paper gets to the Excel table to import SQL Server 2000 database to do the data pre-processing. The data pre-processing mainly reprocess data in previous stage to check the integrity of data and consistency of data. It includes noise immunization, deduce to calculate missing data, remove duplicate record and complete data type transfer. In preprocessing stage, delete "dirty data" which is redundancy vacancy data, not completing, noise information. It establishes the foundation for data mining in next step and improves the digging efficiency and digging quality[7].

3.3. Modeling Based on Clementine

3.3.1. Cluster Modeling

Input the data which is collected after preprocessing into cluster modeling in SPSS Clementine to cluster modeling analysis. The paper uses K-means algorithm to cluster modeling for the reader's borrowing behavior. K-means[15] algorithm is a process of iterating to calculate "centroid" and being based on the distance between sample and centroid to appoint every sample to cluster. The following is the process[5].

1) Make sure initial centroid. Select the first sample as the first centroid. And calculate the distance and Squared Euclidean distance between it and centroid for every sample. Define centroid vector $C\left(c^1, c^2 \cdots c^Q\right)$ and a sample vector $X\left(x^1, x^2 \cdots x^Q\right)$, Q is the amount of properties in data set. x^q is the first q attribute values, $q = 1, 2, \cdots, Q$. So the following is computational formula of Euclidean distance between sample and centroid: $d = \sqrt{\sum_{q=1}^{Q}\left(x_q - c_q\right)^2}$ After the initial K centroids are generated, the algorithm begins to iterate and appoint[14]. Select the biggest sample of Euclidean distance to be as another centroid. And repeat it till K centroids are all identified.

2) Appoint sample. During every iteration, each of the samples is appointed to the cluster which is nearest to itself. The distance is defined by the square of the Euclidean distance so the distance between sample I and centroid j: $d_{ij} = \left\|X_i - C_j\right\|^2 = \sum_{q=1}^{Q}\left(x_{qi} - c_{qj}\right)^2$ X^i is vector which is constituent by attribute values of sample i, C^{qj} is centroid vector of cluster j,Q is the amount of property, x^{qi} is the number q property value of number i sample, c^{qj} is the number of q property value of the centroid in cluster j. Begin to update every centroid of cluster after all the records are all appointed.

3) Update centroid. Some samples in one cluster may be transferred into other clusters in the process of appointing samples. So it needs to recount centroid of every cluster. Establish m^j is the sample amount of number j cluster after appointing sample. So the vector of recount the centroid of cluster is: $\overline{X_j} = \left(x_{1j}, x_{2j}, ..., x_{Qj}\right)$, number $q\left(q = 1, 2, ..., Q\right)$ in vector and component x^{qi} is: $x_{qj} = \frac{\sum_{i=1}^{m_j} x_{qi}(j)}{m_j}$, $x_{qi}(j)$ is the number q property value

in sample i of cluster j[11].

4) Stopping criterion. Firstly, "the max iterations" controls that the algorighm search stable cluster. The algorighm will repeat "appoint sample-update centroid" until "the max iterations"[13]. It will generate final model after it reaches the limitation and the algorighm will stop to update cluster. And "Tolerance of differences" provides another way to control algorighm to be stopped. Calculate distance in centroid space after every iteration finish. Such as, iteration after t times finish, the distance in centroid space in number j cluster is: $\left\|C_j(t)-C_j(t-1)\right\|$, $C_j(t)$ is centroid vector of number j cluster of iteration in t times, $C_j(t-1)$ is the centroid vector of number j cluster when the last iteration. So there are k results that produced by k clusters. Select the max in it: $\max\left\|C_j(t)-C_j(t-1)\right\|$, if the max is less than Tolerance of differences which is predefined. So the algorighm will stop. If not, it will go on.

Through these steps, the following **Figure 1** is view of cluster model.

The result shows that it divides it into three classifies: high frequency (cluster 2), medium frequency (cluster 3), and low frequency (cluster 1). Extract the high and medium users because their borrowing amount is huge and the association rules in the books are strong. The cluster1 is regarded as noisy data to delete so that the association rules are more typical.

3.3.2. Association Rules Mining

Regard the clustering analysis as the pretreatment part of association rules mining. It can find association rules efficiently and avoid generating the false rules[6]. It can make data more illustrative, pertinency, veracity. Extract reader data in Cluster 3 and Cluster 2 are totally 764. Query the 764 students' borrowing information from database to save as data sheet. Use Apriori note in Clementine to do association rules mining. The process is:

1) Generate frequent item set. Based on $(k-1)-$ frequent item sets to make up gather L_{k-1}, and generate all candidate k-item-set C_k, and prune C_k, and calculated support in every item-set w C_k: support = $\frac{N_i}{N}$, N_i is the amount of transaction of including item-set w. N is amount of all the transaction. Put item set of support $\geq \min\_\sup$ into item-set L_k in frequently k-. Find the frequently k- item-set and k is less than $k_{\max}$, which is predefined by user. Repeat above steps and search the frequently item-set $(k+1)-$.

2) After getting all the frequently item-set L, the algorighm will generate association rules based on frequently item-set. Firstly, generate l's all nonvoid subset based on frequently item-set l of L. Secondly, for very nonvoid subset A, if it content valuation criterion ($\frac{\sup port(l)}{\sup port(A)} \geq \min\_conf$, $\sup port(l)$ and $\sup port(A)$ are item-set l and A 's support), and then the output role is "$A \Rightarrow \overline{A}$",and $\overline{A} = l\text{-}A$ [12].

So the association rules is **Figure 2**, **Figure 3**

The call number of library is Chinese Library Classification. From picture **Figure 2**, the reader who borrows B83-09/13(historical pedigree and theoretic finality) also want to borrow B83/20 = 3(aesthetics introduction. revised edition), it can be the reason for reader recommend. From **Figure 3**, it can be clearly shown the association rules among books. And the association rules with thick line are stronger than fine line.

Figure 1. Model view.

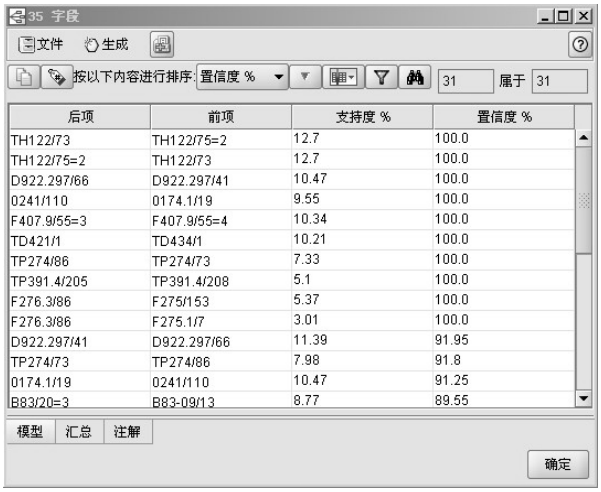

Figure 2. Model view.

4. Realize Intelligent Recommendation

By the data mining process, transfer the association rules to readers through agent. When there is a request from client to Web server, transfer the request to the reader recommended agent to match. And transfer the matching recommended rules to Web server. Finally, transfer it to the user in client[8]. This can give readers more selections, and improve the use ratio of books. **Figure 4** is mode pattern of books intelligent recommendation.

5. Conclusions

It is important to provide flexible and targeted books recommending service by digital library development in the direction of intelligence[9]. The paper views the cluster as the data pre-processing of association rules mining to make the rules more accurate. The paper shows that the subject is effective and viable.

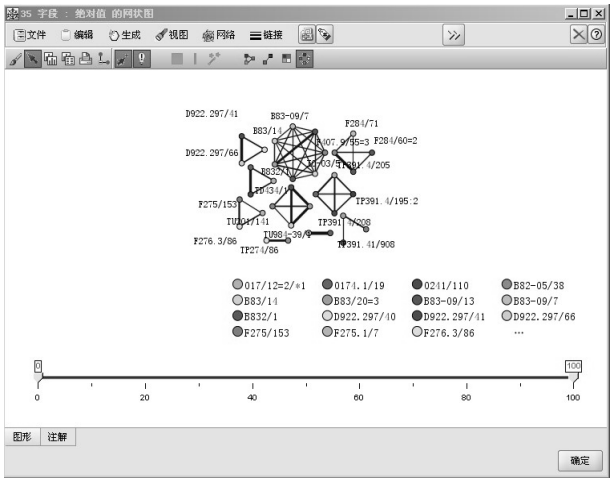

Figure 3. Association rules webs.

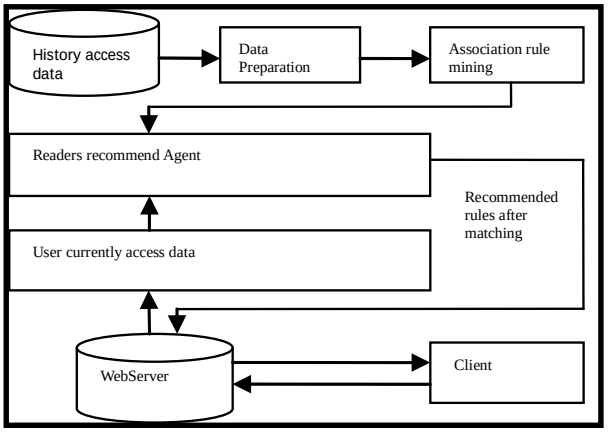

Figure 4. Mode pattren of intelligent recommendation.

REFERENCES

[1] C. G. Yuan, "Data Mining Theory and SPSS Clementine Application," Beijing: Electronic industry publishing, 2009, pp.547-578

[2] F. Y. You,"Data Mining and digital library application "*Office automation magazine*,2007, pp.51－52

[3] C. H. Bao, "Data warehouse and Data Mining," Beijing: Tsinghua University Press, 2006.

[4] J. Han, M. Kamber, "Data Mining and Technology" M. Fan, Translator, Beijing: China Machine Press, 2001, pp.10-33.

[5] H. Y. Chen, "Based on Weighting Association Rules and Browse Behavioral Personality," Chongqing University, 2005.

[6] W. Wang, "Reader Behavior Analysis Based on Data Mining," *Modern Library and information technology*, 2006, pp.51-54.

[7] H. Y. Cai, "The Application in University Library System for Data Mining about Association Rules," *NUT College Journa*l, 2005, pp.85-88.

[8] W. H. Li, "Personality Information Recommend System in Digital Library," 2007, pp.109-110.

[9] W. W. Chen, "Data Mining Research about Reader Behavior," Chongqing Southwest University, 2007.

[10] B. C. Xie, "Data Mining Clementine Application," Beijing: THU press, 2008, pp.213-215.

[11] J. Bao, S. W. Fan, "The Data Pre-processing for Data Mining," *Library and Information Science*, Vol. 26, No. 2, 2008, pp. 31-33.

[12] Z. G. Li, G. Ma, "DW and DM Application," Beijing: Higher Education Press, 2008, pp.150-170.

[13] Q. H. Xiao, "Data Mining Apply in Information Server," *Library forum*, Vol. 24, No. 1, 2004, pp.140-142.

[14] B. H. Wang, "Data Mining and Application," *Statistics and decision*, 2006, pp.122- 123.

[15] X. Li, C. H. Yang, "K-means Cluster Application," *Library and information Science*, Vol. 25, No. 2, 2009, pp.15-17

