Advances in Intelligent Systems Research, volume 156
7th International Conference on Management, Education and Information (MEICI 2017)

Research on the Integration of Academic Resources Based on Data Mining

Bo Yang[1, a] and Lina Zhang[2, b*]

[1] Changchun University of Finance and Economics, Changchun, Jilin, 130122, China

[2] Jilin Agricultural University, Changchun, Jilin, 130118, China

[a]yangbo@163.com, [b]zhangln@163.com

Keywords: Data Mining; Academic Resources; Resource Integration

Abstract. Academic resources have their unique characteristics, and they are important sources of information for teaching and research workers in universities. Too rich data resources make it easy for people to fall into the "information poverty" situation. In order to effectively promote the development of science and the library and provide services for users, it is necessary to increase the mass of information of digital library development and deep mining, extraction of information out of order relation. Using classification techniques, it can integrate the reader to the relevant literature, and provide data for consulting the literature acquisitioning work; Association analysis, can understand what books are often borrow these books in similar position, optimize the construction of collections.

Introduction

The application of data mining in digital library, from the potential value model and knowledge extraction of unknown collection or a large amount of text in the database, according to the needs of users, readers make more detailed analysis, is conducive to the relevant departments to make a decision at the same time, improve the scientific research or high school staff work efficiency, provide academic value reference, to facilitate the use of academic resources for readers. In a word, data mining plays a very important role in the construction of digital library, and improves the utilization ratio of academic resources.

Data Mining Technology

Data mining is also called knowledge discovery, is from large, incomplete, noisy, fuzzy and random data sets through the algorithm search and identification of effective and potentially useful, and ultimately understandable patterns. It is a wide range of cross disciplines, including mathematical statistics, neural networks, machine learning, databases, rough sets, pattern recognition, fuzzy mathematics and other related technologies.

- Association Analysis

Association analysis refers to the dependence and causality between an event and two or more events, thus describing the rules and patterns of the simultaneous occurrence of certain attributes in a thing. In the excavation of academic resources, this can predict the readers' demand for information, and thus effectively improve the efficiency of information push, and facilitate readers to find the required information.

- Time Series Analysis

Time series pattern analysis is the use of known data to predict future values. It occurred within a certain time interval based events, these events constitute a sequence, through the analysis of these events, find the corresponding rules and trend, but the difference of these events is the attribute value at different time.

- Cluster Analysis

Cluster analysis is called clustering for short, and its analysis objects have not been classified. It is the process of dividing a data object into a series of subsets. Each subset is called a cluster, and the objects in these clusters are similar and are not similar to those in other clusters.

- Classification

Classification is the construction of a classification function (classification model). By analyzing the attributes of samples of the same class, we obtain the rules or methods that determine samples belonging to various categories.

- Forecasting

Data mining automatically searches for predictive information in large databases. It speculates on future trends in data from user history and current data.

Apriori Algorithm

Association rules reflect interdependencies or associations between objects, and their most famous algorithms are Apriori and R.Srikant proposed by Agrawal. The steps are divided into frequent itemsets and association rules.

- Generation of Frequent Itemsets

Apriori algorithm, a frequent itemsets algorithm for mining association rules, uses layer by layer search iterative methods, where k itemsets can derive (k+1) itemsets. First, find the 1- frequent itemsets, denoted as L1, and then through the L1 2- to find frequent itemsets, denoted as L2, and so on, to scan the entire database, collect all itemsets that satisfy the minimum support, until cannot find frequent itemsets K.

Example: Mining Boolean association rules and discovering frequent itemsets of Apriori algorithm

Input: object database: D; minimum support threshold: min_sup

Output: all frequent itemsets D in L

Method：

(1)L1=find_frequent_1_itemset (D);

(2)for (k=2;Lk-1≠null;k++){

(3) Ck=apriori_gen (Lk-1, min_sup);

(4) for each t∈D

(5) {Ct =subset (Ck, t);

(6) for each c∈Ct

(7) c.count++;}

(8) Lk={c∈Ck|c.count>min_sup}}

(9) Return L=UkLk;

Procedure apriori_gen (Lk-1,min_sup)

Method for generating candidate K itemsets according to k-1 frequent itemsets

(1) for each l1∈Lk-1

(2) for each l2∈Lk-2

(3) if (l1[1]= l2[1]∧⋯∧ (l1[k-2]= l2[k-2])∧ (l1[k-1]= l2[k-1] then {

(4) c= l1∞l2;

(5) if has_infrequent_subset (c,Lk-1) then

(6) delete c;

(7) else Ck=Ck∪{C} }

(8) return Ck;

Procedure has_infrequent_subset (c,Lk-1)

Test whether all subsets of candidate sets are frequent itemsets

As the algorithm above, Apriori has two main actions: the connection step and the pruning step.

The first algorithm: apriori_gen get all the candidate frequent itemsets, step 4 is to scan the database D, then step 6 uses the subset function to find all candidate subsets, and the cumulative for all candidate subsets, such as steps 7 and 8.

Finally get the candidate items all meet the minimum support of min_sup in L.

Link step: Lk-1 and Lk-2 links generate possible candidate sets, step 1-4.

Pruning step: removes candidates with non frequent subsets using Apriori related properties

The test processing of infrequent subsets is performed in the has_infrequent_subset process.

- Generation of association rules

When you find all relevant frequent itemsets in the database D, you can use them directly to obtain strong association rules.

Formula is as follows:

Confidence (A⇒B)= P(A | B)=(support_count (A∪B))/(support_count (A))

The support_count (A, B) contains all transaction itemsets A, B, support_count (A) is all itemsets A.

Based on the formula above, the association rules can be generated this way:

(1) First, all the nonempty subsets of each frequent itemset l are generated;

(2) For each nonempty subset of l, s,if:

(support_count (l))/(support_count (s))≥min_conf

Then you can output strong association rules: s⇒ (l-s)

Table 1 Transaction data sheet for a digital library

TID	Loan record ID list	TID	Loan record ID list
T100	I1, I2, I5	T600	I2, I3
T200	I2, I4	T700	I1, I3
T300	I2, I3	T800	I1, I2, I3, I5
T400	I1, I2, I4	T900	I1, I2, I3
T500	I1, I3		

Suppose the minimum support is 2, and the table contains frequent itemsets X={I1, I2, and I5}. Then what association rules can be generated by X? The X set contains the non empty set {I1}, {I2}, {I5}, {I1, I2}, {I1, I5}, {I2, I5}. The association rules and confidence are given as follows:

I1⇒{I2, I5}, confidence=2/6=33%

I2⇒{I1, I5}, confidence=2/7=29%

I5⇒{I1, I2}, confidence=2/2=100%

{I1, I2}⇒I5, confidence=2/4=50%

{I1, I5}⇒I2, confidence=2/2=100%

{I2, I5}⇒I1, confidence=2/2=100%

If the minimum confidence threshold given by the user is 90%, then the output rules can be third, fifth, and sixth, all of which are strong rules.

Summary

The research of academic resources integration based on data mining not only optimizes the allocation of collection resources, but also enhances the interaction with users, and makes knowledge services become high-level information services in the true sense. On the basis of relevant theories, the model designed in this paper is being explored. In the future, with the continuous change of information technology and environment, knowledge service theory and practice will have further development, in view of the knowledge service to create academic information resources integration mode will be more diversified, is not only a new information resource integration way it appears, prompting the library to complete the transformation from

information service to knowledge the service, which play its special role in knowledge economy society.

Acknowledgments

This work was financially supported by the Jilin Province Education Science "The 13th Five-year" planning issues (GH170992, ZD16038) and Jilin Provincial Institute of higher education issues in 2017(JGJX2017D308, JGJX2017C32).

References

[1] Olfa Nasraoui, Mrudula Pavuluri, Accurate Web Recommendations Based on Profile-Specific URL-Predictor Neural Networks. Communications of the ACM, New York, 2014, 22 (10): 300~301
[2] Wenquan Yi,Fei Teng,Jianfeng Xu. Noval Stream Data Mining Framework under the Background of Big Data[J]. Cybernetics and Information Technologies,2016,16(5).
[3] Pek San Tay,Cheng Peng Sik. Data mining and copyright: A bittersweet technology gift for copyright owners and the Malaysian public?[J]. Computer Law & Security Review: The International Journal of Technology Law and Practice,2016,32(6).
[4] Viktor Medvedev,Olga Kurasova,Jolita Bernatavičienė,Povilas Treigys, Virginijus Marcinkevičius, Gintautas Dzemyda. A new web-based solution for modelling data mining processes[J]. Simulation Modelling Practice and Theory,2017.
[5] Duan Xiaohong. Research and implementation of undergraduate teaching quality monitoring platform [D]. Southwest Jiao Tong University, 2012.
[6] Yang Bo, Zhang Lina. Study on the method of intelligence analysis of the competency of teachers in private universities [J/OL]. Heilongjiang animal husbandry and veterinary medicine, 2016, (15): 263-265+299-300.
[7] Li Qiongyuan. Construction and operation of teaching quality monitoring system in private universities. Taking Nanning Institute as an example, [J]. education and teaching forum, 2013, (26): 126-128.
[8] Olfa Nasraoui, Mrudula Pavuluri. Accurate Web Recommendations Based on Profile-Specific URL-Predictor Neural Networks. Communications of the ACM, New York, 2004, 22 (10): 300~301
[9] Farzaneh A. Amani,Adam M. Fadlalla. Data mining applications in accounting: A review of the literature and organizing framework[J]. International Journal of Accounting Information Systems,2017,24.
[10] Agrawal R, Imiclinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C] . Proceedings of the 1993 ACM SIGMOD Conference. . Washington, DC. 1993 (5): 207~216.