Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080329

Research on Data Mining Technologies for Complicated Attributes Relationship in Digital Library Collections

Yumin Zhao, Zhendong Niu and Xueping Peng*

School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Received: 31 May. 2013, Revised: 5 Oct. 2013, Accepted: 6 Oct. 2013
Published online: 1 May. 2014

Abstract: We present here the research work on data mining technologies for complicated attributes relationship in digital library collections. Firstly our work and ideology is introduced as the research background of this paper. Digital library evaluation is an important topic in information systems domain. We creatively import data mining technologies into it to get an intelligent decision support. But traditional data prediction algorithm didn't work well. This is the problem which would be solved in this paper. Secondly related preliminary research is introduced. We researched on attributes of digital library collections, proposed a parallel discretization algorithm based on z-score theory, and by the discretization algorithm discovered a complicated condition attribute relation among attributes, it is the reason why traditional data prediction algorithm didn't work well. At last a stratified decision tree algorithm for value prediction about digital collection is put forward as the ultimate solution to solve the problem. Stratified attribute concept is imported in this algorithm. It can expand the selection of splitting attribute in decision tree from flat information to stereoscopic information, eliminate the influence of complicated condition attribute relationship, nested use existing decision tree algorithms, and solve the bottleneck of data mining application in digital library evaluation.

Keywords: Discretization Algorithm, Stratified Decision Tree Algorithm, Digital Library Collections

1 Introduction

Digital Library is considered an inevitable development trend of computer information service and Internet technology. Data processing and use of digital library has been a hotspot in the study of computer knowledge organization and digitization.

Evaluation research about digital library is a comprehensive topic focusing on all kinds of digital library resources and has been widely concerned in recent years.

There are two evaluation ideologies about digital library presently [1]. One is expert-centric evaluation and the other is user-centric evaluation. Expert-centric evaluation bases on expert's judgment with various professional indexes, and user-centric evaluation bases on the users' experience about digital libraries, as discussed in [2,3]. Establish and calculation of evaluation indexes is adopted as main method in expert-centric evaluation. Design and statistic analysis of user surveys is adopted as main method in user-centric evaluation.

With the rapid development of computer technology both above methods face many problems, such as processing and utilization of mass information, complex data relationship in digital library collections, etc. They lack advanced intelligent information processing technologies and decision support idea, have difficulty to obtain satisfactory results.

2 Research Background

In our research a natural-integration evaluation model for digital library has been developed to reflect our evaluation ideology [4], and the data mining technologies are imported into digital library evaluation for the first time [5]. This model can provide a way to integrate merits of expert-centric ideology and user-centric ideology, also provide a way to utilize the mass information of digital library through intelligent technology. More reasonable evaluation result can be gotten through this way. These concrete evaluation ideologies and work can be found in my other papers.

Furthermore, collections evaluation of digital library is the core of evaluation of digital library. On the one

* Corresponding author e-mail: zhymbit@gmail.com

hand, expert-centric evaluation ideology carries out work around the scale of digital collections. This way can't catch the appropriate score about quality of digital collections, because they can't guess or calculate the numerous users' feeling. On the other hand, user-centric evaluation ideology is a kind of effect evaluation method, it is easy to get the significative score about each digital collection. But this way do not fit the quantity evaluation, because user always has some wrong feel and it is inevitable, as discussed in [6].

So in our research a new evaluation method about digital collections has been put forward. Brief procedures are as follows.

(1) Collect data which expert-centric evaluation ideology concerns about each digital collection to form dataset-A; Collect data which user-centric evaluation ideology concerns about each digital collection to form dataset-B; Integrate dataset-A and dataset-B to form evaluation dataset;

(2) Build the prediction model for digital collection based on the evaluation dataset through a method of data mining; Get the quality value for each digital collection according to the prediction model;

(3) Add the quality value into the scale evaluation process.

In short, the reason why we research on data mining technologies in digital library collections is to build the data prediction algorithm for digital collection. By this way we can merge quality evaluation and quantity evaluation together to get rational score about a collection of digital library.

For this paper we put the research emphasis on the innovations about relevant data mining technologies which would be applied into digital library evaluation.

Our research of digital collections have two types of data. One is objective data corresponding to dataset-A, and the other is subjective data corresponding to dataset-B. The objective data reflect the inherent and unchangeable attributes about a digital collection, such as author, translator, publish time, publisher, price, etc. The subjective data reflect users' experiences and feelings attributes about a digital collection, such as user's review, the number how many people have read, user's score, etc.

The subjective data in our research mainly refer to DOUBAN.com. It provides a open platform for users. Everyone can share personal feelings information about books, movie and music in which people are interesting. Users can not only pose reviews and discussions but also give their own score about each book, movie and music. This information is a kind of meaningful evaluation data for digital library collections. In the study of digital library evaluation these users' feelings information could be taken for the subjective data. These information include user comments about a book, number of people who are reading the book, number of people who want to read the book, number of people who have read the book, score, etc [7].

The main purpose of building the integration dataset containing objective data and subjective data is to get a score prediction model for new collection of digital library. A score can be gotten for a new collection without users' evaluation through prediction model. Traditional work always depend on the collections scale, because the quality of digital collections is very difficult to define. Now we can get evaluation results more reasonably through merging the scale information and quality information into this score prediction model.

But work didn't go well when we directly apply present data mining technologies for evaluation of digital library collections. This paper should research on these problems.

3 Related Preliminary Research

The ultimate purpose of this paper is to put forward a data mining algorithm for score level prediction of digital collection. To get valid algorithm for score level prediction we have some necessary work to do. These related preliminary researches include discretization algorithm for numerical attributes and complicated attributes relationship in digital library collections.

Both studies have been proved very meaningful during the research process of prediction algorithm. Discretization algorithm for numerical attributes can solve the problem of information loss as much as possible. It contributes to improve the effect of prediction algorithm. Discovery of complicated attributes relationship in digital library collections provides the key to find correct prediction algorithm. They are good for getting the valid prediction algorithm.

These topics were studied and discussed as previous research in my other paper [7]. We should give a brief introduction here.

Our research imported the subjective data which express users' feel into collections dataset as a solution for quantizing collection's quality. Then we applied traditional data mining technologies into it to get the intelligent decision support. But traditional data prediction algorithm didn't work well, it confused us. There existed mass data and we had the advance evaluation idea, the application condition looked all right, why it didn't work? We researched the complex relationship among collections' properties, and got two results.

(1) We proposed a parallel discretization algorithm based on z-score theory (PDOZ algorithm). It improves the correlativity between normal attribute and prediction attribute, and can discover attribute relation more efficiently.

(2) Based on PDOZ algorithm we discovered a 'nonlinear conditional attributes relationship'. It is the reason why traditional data prediction algorithm didn't work well. This relationship is detailly described in [7].

Conditional attributes relationship is very similar to the splitting attribute of traditional decision tree. They all have the ability to distinguish different data, split the present dataset according to the value of splitting attribute selected, and would be chosen.

But the key point of this 'Nonlinear conditional attributes relationship' is that the conditional attribute can't be directly recognized by traditional decision tree because it has a low discrimination degree. The splitting attribute of traditional decision tree has a higher discrimination degree, so it can be selected through many ways. For example, J.R.Quinlan proposed the information gain criteria [8,9,10], L.Breiman proposed Gini-Index criteria [11], J.Mingers proposed the x2 statistical criteria [12], K.Kira proposed the relief criteria [13,14], S. J.Hong proposed the CM criteria [15], etc.

For example, attribute C is a class attribute, there exits nonlinear conditional attributes relationship between attribute A and attribute B. Correlation are all weak between A and C, B and C, A and B. Neither of A or B can be the splitting attribute in traditional decision tree algorithm, but when the dataset is divied by A, B immediately have the strong ability to distinguish different data. In fact B should be treated as an important attribute.

In our evaluation study of digital library it is difficult to pick up the conditional attribute with traditional decision tree algorithm. Through collections' attributes research we propose a stratified decision tree algorithm based on PDOZ for value prediction about the score of digital library collections. This algorithm solves the problem that traditional data mining algorithm can't be well applied in the digital library collections. Stratified attribute concept is imported in this algorithm. It expands the select of splitting attribute in decision tree from flat information to stereoscopic information, eliminates the influence of complicated condition attribute relation, can use nested existing decision tree algorithms, solves the bottleneck of data mining application in digital library evaluation.

4 Stratified Decision Tree Algorithm

Decision tree algorithm is one of the most popular data prediction technologies [16,17]. Else prediction technologies include SVM algorithm [18,19], KNN algorithm [20,21], Bayes algorithm [22,23], Genetic algorithm [24] , Neural Networks [25] , etc.

In our evaluation dataset the number of attributes is relatively less, attributes are independent in semantics, data of attributes are not sparse, there exist many nominal attributes, there exist some numerical attributes, and the number of class attribute value is larger than 2 but smaller than 10. All these characteristics of evaluation dataset fit the decision tree algorithm.

That there are many nominal attributes doesn't fit the computation pattern of KNN. The class attribute is users' score and always discretized to nominal type, number of its value is larger than 2 but smaller than 10, this condition doesn't fit the SVM which tends to process the fewer categories. Bayes algorithm very relies on the correctness of prior information, this is difficult to gain in digital collections. Genetic algorithm and Neural Networks belong to the soft computing forecast algorithm which solve problem with highly nonlinear and complexity relationship, they are a kind of compute method to fast search the better solutions, but not to get exact solutions for the target. Although there exists 'nonlinear conditional attributes relationship' listed above, the else relationship among attributes is clear and simple, we only need to transform the 'nonlinear conditional attributes relationship' to the clear and simple relationship.

So we have chosen the decision tree algorithm as the preferred research object.

The key principle of decision tree algorithm is to identify the splitting attribute which has the ability to distinguish different data in present dataset, then to split the present dataset according to the value of splitting attribute selected, this process should be called recursively and finished until present data belongs to one category. In the decision tree algorithm the selection methods of splitting attribute play the most important role.

In our integration dataset there is a complicated relationship and not fit for present decision tree methods. Then a stratified decision tree algorithm based on PDOZ (SDT-Z algorithm) is put forward in our research.

Based on collections' attributes research we propose a stratified decision tree algorithm based on PDOZ for prediction mining about the value of digital library collections.

The key of SDT-Z algorithm is to choose the splitting attribute by circularly judging the change of correlation coefficients between the class attribute and common attribute.

Process of SDT-Z algorithm is presented in Figure 1.

(1) Enumerate all non-class attributes, suppose the sum is n and calculate correlation coefficients between each non-class attribute and the class attribute.

(2) Pick up $m(m \leq n)$ attributes which correlation coefficients are less than the threshold and put them into candidate hierarchical attributes bunch.

(3) Take out an attribute $A_i\,(1 \leq i \leq m)$ from candidate hierarchical attributes bunch, and split dataset into p subsets based on the value number of $A_i\,(1 \leq i \leq m)$. Assuming the value's number of $A_i\,(1 \leq i \leq m)$ is p.

(4) In each data subset recompute correlation coefficients between each non-class attribute and class attribute. Accumulatively record the improvement of each attribute's correlation coefficient compared to the value in origin dataset, certainly there is no $A_i\,(1 \leq i \leq m)$ in non-class attributes. If the improvement is large than a

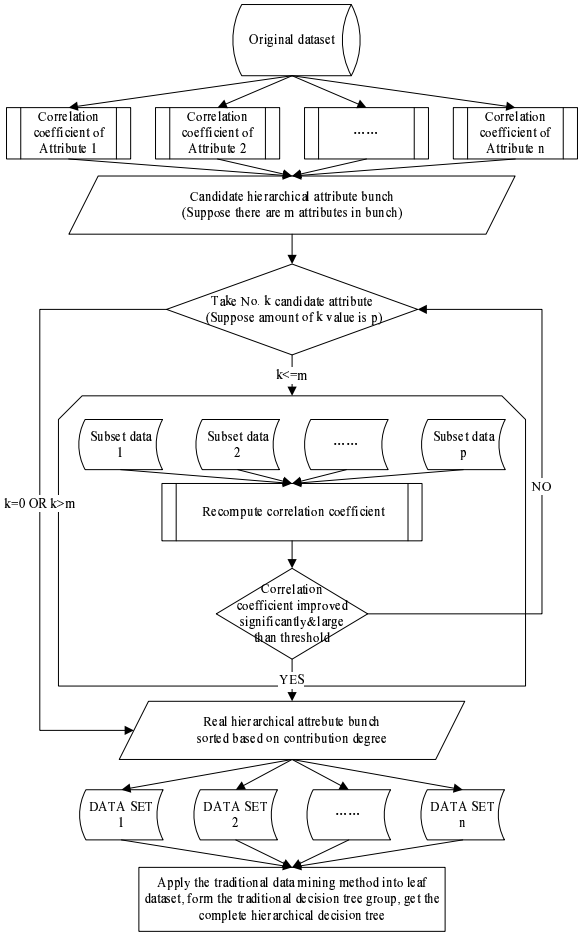

Fig. 1: Model of stratified decision tree algorithm

threshold, we should judge $A_i\,(1 \le i \le m)$ is a real hierarchical attribute.

(5) Repeat the third step and fourth step, and get a real hierarchical attributes bunch. Assuming the number of the hierarchical attributes is k.

(6) In the real hierarchical attributes bunch, according to the descending order of contribution (the improvement of Step (4) get the sort of attributes $\{A_1, A_2, \cdots A_k\}$.

(7) Take out the hierarchical attribute from $\{A_1, A_2, \cdots A_k\}$ in turn and split the data set. At last form the hierarchical decision tree by the hierarchical attributes bunch $\{A_1, A_2, \cdots A_k\}$. The level of the hierarchical decision tree is k.

(8) In the 'leaf-subset' corresponding to the leaf of the hierarchical decision tree, nested apply the existing data mining method, such as J48, to form the traditional decision tree group. At last we merge the hierarchical decision tree and the traditional decision tree group and get the complete stratified decision tree.

This algorithm solves the problem that traditional data mining algorithm can't be well applied in the digital library collections. Stratified attribute concept is imported in this algorithm. It expands the selection of splitting attribute in decision tree from flat information to stereoscopic information, eliminates the influence of complicated condition attribute relation, can use nested existing decision tree algorithms, and solves the bottleneck of data mining application in digital library evaluation.

5 Experiment

To test the stratified decision tree algorithm we constructed a dataset which attributes included *numberAuthor*, *haveornotTranslator*, *price*, *publishTime*, *numberReading*, *numberHaveread*, *numberWantread*, *score*, etc. Attribute *score* is the class attribute, there exists nonlinear conditional attributes relationship between attribute *publishTime* and attribute *price*.

In first experiment we set the conditional attribute (attribute *publishTime*) in two states. One state is presence and the other state is absence. We use the traditional decision tree algorithms for data prediction and compare the results in precision rate and recall rate.

We select 5 traditional decision tree algorithms. Every algorithm is made five times under different data scales which include 2000, 4000, 6000, 8000, 10000, and recorded the average of five results.

Table 1: Precision rate under presence and absence state of conditional attribute

algorithm	ID3	J48	BFTree	LADTree	NBTree
presence	69.3	75.3	65.2	50.9	52.4
absence	78.3	82.7	74.6	60.2	71.7
raise	9	7.4	9.4	9.3	19.3
raise ratio	13.0%	9.8%	14.4%	18.3%	36.8%

Table 2: Recall rate under presence and absence state of conditional attribute

algorithm	ID3	J48	BFTree	LADTree	NBTree
presence	55.2	67.1	38.4	43.1	45.2
absence	69.3	79.2	52.1	55.3	52.1
raise	14.1	12.1	13.7	12.2	6.9
raise ratio	25.5%	18%	35.7%	28.3%	15.3%

From Table 1 and Table 2 we can get precision rate and recall rate both raise when conditional attribute is absent. It proves the influence of conditional attribute. But even absence state of conditional attribute the precision

rates are below 80%. So low rate can not meet the actual application's need.

Form above two tables, we can find the explicit experiment data and get that J48 is the best traditional algorithm for digital collection of DOUBAN.COM relatively.

In the second experiment the precision rate and recall rate is compared in high quality generalization dataset between J48 and stratified decision tree. Result is presented in Figure 2 and Figure 3.

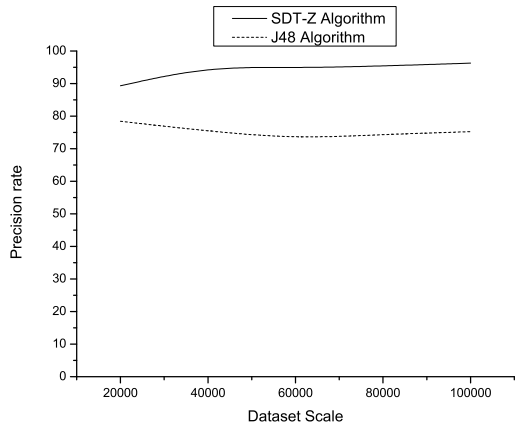

Fig. 2: Comparison of precision rate between SDT and J48

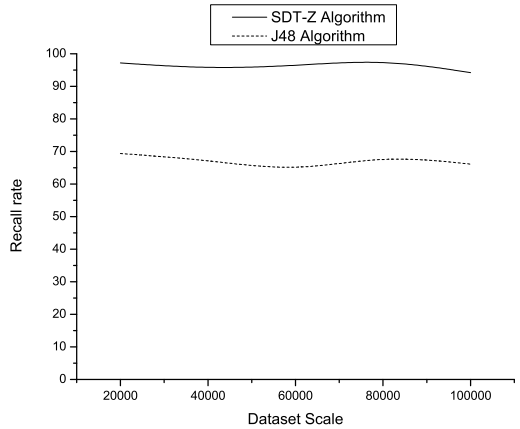

Fig. 3: Comparison of recall rate between SDT and J48

Form Figure 2 and Figure 3 we can find stratified decision tree is better than J48, and the average value is large than 90%. This rate can meet the actual application's need. We think the improvement of rate is attributed to the reason that attribute *price* play a important role in prediction, especially in an appropriate way.

6 Conclusions

This paper presents research on data mining technologies for complicated attributes relationship in digital library collections. We puts forward stratified decision tree algorithm as the prediction technology for digital collections. In research we decided to integrate the quality evaluation and the quantity evaluation, dissatisfactory application effect of data mining technology was met at first. Based on the analysis discretization algorithm was studied in preliminary work. Our algorithm adopts Z value idea to solve the two questions of discretization. During the studies we found the complicated attributes relationship in digital library collections. Based on these two preliminary works we put forward the stratified decision tree algorithm. Through the experiments the algorithm has been proved effective for a better decision-making support.

Acknowledgement

This work is supported by the Program for New Century Excellent Talents in University, China (grant no. NCET-06-0161, 1110012040112), Key Foundation Research Projects of Beijing Institute of Technology (grant no.3070012231001), Beijing Municipal Commission of Education (grant no.1320037010601).
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] I. Hsieh-Yee, Digital Library Evaluation: Progress & Next Steps, Presentation of 2005 Annual Meeting of the American Society for Information Science and Technology, 1-10 (2005).

[2] J. C. Bertot, J. T. Snead, P. T. Jaeger, C. R. McClure, Functionality, usability, and accessibility: Iterative user-centered evaluation strategies for digital libraries, Performance Measurement and Metrics, **7**, 17-28 (2006).

[3] M. Kyrillidou, S. Giersch, Developing the DigiQUAL protocol for digital library evaluation. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 172-173 (2005).

[4] Y. Zhao, Z. Niu, K. Zhao, L. Dai, G. Xu, and W. Wang A Natural-integration Model of Digital Library Evaluation Proceedings of 2010 IEEE International Conference on Computer Engineering and Technology, **2**, 477-479 (2010).

[5] Y. Zhao, Z. Niu, L. Dai, Evaluation algorithm about digital library collections based on data mining technology, Proceedings of Role of Digital Libraries in a Time of Global Change - 12th International Conference on Asia-Pacific Digital Libraries, 266-267 (2010).
[6] Y. Zhao, Z. Niu, K. Zhao, L. Dai, G. Xu, and W. Wang. What and Why about Divarication in Research of Digital Library Evaluation Proceedings of 2010 IEEE International Conference on Computer Engineering and Technology, **2**, 325-328 (2010).
[7] Y. Zhao, Z. Niu, X. Peng, L. Dai, A Discretization Algorithm of Numerical Attributes for Digital library Evaluation based on Data Mining Technology, Proceedings of 13th International Conference on Asia-Pacific Digital Libraries, 70-76 (2011).
[8] J. R. Quinlan, Induction of decision trees. Machine learning, **1**, 81-106 (1986).
[9] J. R. Quinlan, C4.5: programs for machine learning. Morgan Kaufmann, 31-37 (1993).
[10] J. T. Kent, Information gain and a general measure of correlation. Biometrika, **70**, 163-173 (1983).
[11] L. Breiman, Classification and regression trees, Wadsworth International Group, 16-23 (1984).
[12] J. Mingers, Expert systems-rule induction with statistical data, The Journal of the Operational Research Society, **38**, 39-47 (1987).
[13] K. Kira, L. A. Rendell, The feature selection problem: Traditional methods and a new algorithm, Proceedings of the tenth national conference on Artificial intelligence, 129-134 (1992).
[14] K. Kira, L. A. Rendell, A practical approach to feature selection, Proceedings of the ninth international workshop on Machine learning, 249-256 (1992).
[15] S. J. Hong, Use of contextual information for feature ranking and discretization. IEEE Transactions on Knowledge and Data Engineering, **9**, 718-730 (1997).
[16] J. Han, M. Kamber, Data mining: concepts and techniques, third edition. Morgan Kaufmann, 126-141 (2011).
[17] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 76-77 (2005).
[18] M. Kumar, M. M. Gromiha, G. P. S. Raghava, Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins: Structure, Function, and Bioinformatics, **71**, 189-194 (2008).
[19] S. Shalev-Shwartz, N. Srebro, SVM optimization: inverse dependence on training set size, Proceedings of the 25th international conference on Machine learning, 928-935 (2008).
[20] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, **13**, 21-27 (1967).
[21] H. Zhang, A. Berg, M. Maire, J. Malik, SVM-KNN: Discriminative nearest neighbor classification for visual category recognition, Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 2126-2136 (2006).
[22] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, D. Freeman, AutoClass: A Bayesian classification system, Proceeding of the Fifth Intl Workshop on Machine Learning, **27**, 54-64 (1988).
[23] D. Heckerman, Bayesian networks for data mining, Data mining and knowledge discovery, **1**, 79-119 (1997).
[24] D.E.Goldberg, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Professional, 19-21 (1989).
[25] S. Haykin, Neural networks: a comprehensive foundation, The Knowledge Engineering Review, **31**, 409-412 (1999).

Yumin Zhao received the MS degree from Beijing Institute of Technology. He is currently an PhD Student in the School of Computer Science and Technology at Beijing Institute of Technology, China. His research interests are in the areas of digital library, data mining, computer network.

Zhendong Niu received the PhD degree from Beijing Institute of Technology. He is currently a Professor in the School of Computer Science and Technology at Beijing Institute of Technology, China. His research interests are in the areas of computer software architecture, knowledge management, intelligent education software system, digital library, neural information system.

Xueping Peng received the MS degree from Beijing Institute of Technology. He is currently an PhD Student in the School of Computer Science and Technology at Beijing Institute of Technology, China. His research interests are in the areas of information retrieval, personalized service, web log mining.