ATLANTIS PRESS
Advances in Computer Science Research, volume 50
4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)

The Method to Determine Bibliographic Types Based on Decision Tree

Si Geng[a], Ning Li[b], Lin Zhao[c], Ying'ai Tian[d]

Computer School, Beijing Information Science and Technology University, Beijing, China

[a]email: gengsi123@163.com, [b]email: ningli.ok@163.com, [*]corresponding author, [c]email: zhaolin_0124@126.com, [d]email: tianyingai@bistu.edu.cn

Keywords: Bibliographic type determination, Normative checking of bibliography, Decision tree.

Abstract. The key to normative checking of bibliography is to find out bibliographic types. In order to solve the problem to check the correctness of bibliography in the absence of bibliographic types, a novel method to determine bibliographic types based on decision tree is proposed in this paper. This method extracts bibliographic items from the references and utilizes machine learning method to construct a decision tree, and then to determine the bibliographic types according to the decision rules. The performance is compared with the Naive Bayes classifier. Experimental results show that the proposed method can effectively determine the bibliographic types, which lays the foundation for the normative checking of bibliography and the effective use of bibliography.

Introduction

With the dramatic expansion of academic papers, the normalization and standardization of academic papers have caused the extensive concern in the editing circle. The correctness of bibliographies is usually essential to academic standardization, as it is important to the inheritance of academic research legacy. A normative bibliography is much helpful to the reviewers, editors as well as the readers to understand the relevant research results, meanwhile it lays the foundation for manuscript reviewing, bibliometrics analysis, digital publishing and so forth. Therefore, it is of great realistic significance to check whether the reference is correct.

At present, the auxiliary reference writing tools mainly divides into two categories, reference editing tools and reference normative checking tools. The former, such as Endnote [1], NoteFirst [2], can help users to typeset the references automatically based on the requirements of the different periodicals. The latter can check the correctness of existing references. Some significant achievements include the checking method mentioned by Zhang (2011) and Liu (2015). Zhang proposed the method checks the references based on XML documents parsing and regular expression matching, while the premise condition is that the bibliographic type symbol must be there in the reference [3]. Liu proposed the method determines the references categories by modeling the existing references and building a reference model database, in which VSM is used to calculate the similarity of the measured references model and each model in the database, but it is difficult to carry on the work when some bibliographic items are absent [4]. Generally, normative checking tools of bibliography are mainly dependent on the symbol of bibliographic type, which determines the bibliographic format that a reference should be followed.

In sum, the bibliographic type symbol is very crucial for the normative checking of bibliography and automatic processing. Unfortunately, the bibliographic type symbol is an optional item in the standard Descriptive rules for bibliography (GB/T 7714) [5]. When the bibliographic type is not given, bibliographic types determination is the first problem to be solved before the normative checking of bibliography. Each reference is composed of many description items, bibliographic type can be inferred by analyzing the features of description items, such as author, publisher and so on. For a reference, we can clearly know the description item of author corresponding to what content in the reference, but how to make the computer know it clearly. It can learn from the named entity recognition method in the natural language processing to extract description items, which is based on the analysis of bibliography description format. After the extraction of description items, the

bibliographic types can be predicted according to the bibliographic rules stipulated in GB/T 7714, which is similar to multiple classification problem in machine learning.

Current classification methods in machine learning are commonly used support vector machine (SVM) [6] [7] [8], artificial neural network (ANN) [9] [10] [11], Naive Bayes [12] [13], decision tree [14] [15] and so forth. SVM is applied to solve the problems such as high-dimensional, non-linear and sparse data, and achieves satisfying classification result. However, the process of dealing with multiple classification is much more complex. ANN has some merits in non-linear self-learning and associative storage, but the unobservable learning process sometime leads to the decline in the credibility of the final result. Naive Bayes has stable classification efficiency and is based on the assumption of independence between attributes. In this paper, there is no guarantee that bibliographic items are independent. Decision tree can deal with the irrelevant features and various types of data. The decision tree model is easy to interpret and understand, which help to make accurate predictions. Based on the above analysis, a method to determine the bibliographic type based on decision tree is proposed in this paper. The method constructs a decision tree via the extracted description items, and uses decision rules to predict the type of reference.

Extraction of Reference Description Items

Since the data used in this paper are large numbers of references, the description items need to be extracted from each reference before determining the bibliographic type. However, different types of references have different description items. The description items can be extracted on the basis of analyzing the reference description features.

The items of the description, sequence of the description and symbols are stipulated in GB/T 7714, which are used to describe the bibliography of various publications. Take the monographs as an example, its description form is shown as follows.

<Description form> ::= primary author“.” <Title items>“.” [other author] “.” edition item“.” <Published items>.
<Title items> ::= title“.” other title information “[” bibliographic type symbol“]”
<Published items> ::= publication place “:”publisher “,”publication year“:” citation page

Through analyzing the reference description form, it can be found that the description items include two types. One is an entity with a fixed form, such as publication year and citation pages etc. The other is named entity which is similar to person name, place name and organization name, such as author and publication place etc. According to the features of the description items, this paper will combine the use of rules and statistics to extract the description items. In order to be able to clarify the meaning of each description item, description items extraction is divided into two major step. The first step is to segment the reference and the second step is to identify the description items. In a reference, the description symbols are used in the connection between each description item. Different symbols have their own usage rules, for example, “.” is used before the title, edition item etc. GB/T 7714 clearly points out that these symbols are used as a prefix-symbol, and do not use any symbol before the first description item. So different prefix-symbols can be used as split points to segment the reference.

After the reference segmentation, different types of description items can be identified respectively. The fixed form description items are identified by regular expressions. The description items, which are similar to named entity, can be identified by Stanford NER method. Unfortunately, the Stanford NER training model cannot meet our needs, we need to build a named entity corpus and train it to obtain a new model, so as to realize the description items identification. The corpus of “People’s Daily” owns a large proportion of nouns, such as the name of person and place etc., which is helpful to identify the description item of author and publication place etc., so the “People’s Daily” tagged corpus in January 1998 is extracted as training corpus. Apart from that, the publisher of journal references is usually the journal name, the paper titles contain some common keywords, such as “based on”, “research methods” and so forth. In order to accurately identify the description items, the journal names, which appeared in the 2015 Peking University version of “A Guide to the Core

Chinese Periodical", and the statistical keywords commonly used in the titles are added as the final training corpus.

Preprocessing of Reference Description Items

Each reference is composed of many description items, and different types of references have various description items. The description items are extracted according to the above method, and they can be seen as the attributes of each record, Table 1 is a fragment selected from the records.

Table 1. Part of records

description item	the value of description item				
author	Zong Chengqing	Liu Baochao	Chen Guoguang	Xun Xiangjun	Li Ning
title	Statistical Natural Language Processing	Design and implementation of dissertation standardization evaluation system	A Rule Based Book Document Logical Structure Extraction Method	Research of Data Mining Model Based on Decision Tree	The Method of Bibliographic Reference Format Checking
patent country	----	----	----	----	China
patent number	----	----	----	----	CN105824791A
report number	----	----	----	----	----
publication place	Beijing	----	----	----	----
publisher	Tsinghua University Press	Yanbian University	Computer Engineering and Applications	2006 System Simulation Technology and Application Academic Exchange Conference Proceedings	----
publication year	2013	2015	2002	2006	----
publication date	----	----	----	----	2016-08-03
volume period	----	----	38(19)	----	----
page number	150-164	----	53-57	----	----
reference type	M	D	J	C	P

As can be seen from the Table 1, some values of description items are blank, and some description items can be ignored. So the description items should be preprocessed before classification. It can be carried out from the following three steps.

Data Cleaning. Data cleaning mainly deals with the description items with null values, such as the type of journal references do not contain "patent" and "report number" and so on. And these attribute values should be filled according to the type of value that already exists.

Correlation Processing. Correlation processing principally deals with the description items with irrelevant information to the results prediction, which should be ignored. For example, “publication place” and “publication year”, no matter when and where it was published, will not affect the final bibliographic type, so they can be neglected.

Data Transformation. Data transformation mostly deals with inconsistent description items which need to be unified, namely data generalization. For instance, the description item of “author” can be generalized to individual and institution. Predicting the bibliographic type is based on whether the “author” is a person name or an organization name, regardless of the specific name. In the same way, the description items of “publisher” can be summarized into four categories: journal (e.g. Computer Engineering and Applications), education (e.g. Peking University School of Mathematics), press (e.g. Xinhua Publishing House) and other (e.g. Beijing Institute of Aerodynamics).

According to the steps mentioned above, the data set is generated by processing 10000 references collected from academic papers, which is used for constructing the decision tree model, as shown in Table 2.

Table 2. Preprocessed references

Id	Author Type (T1)	Report Number (T2)	Patent Number (T3)	Publisher Type (T4)	Volume Period (T5)	Page Number (T6)	Reference Type	Count
1	individual	no	no	journal	yes	yes	J	2527
2	individual	no	no	education	no	no	D	2260
3	individual	yes	no	other	no	no	R	328
4	individual	no	no	press	no	no	M	1135
5	institution	no	no	other	no	no	C	1389
6	individual	no	no	journal	yes	no	J	182
7	individual	no	no	education	no	yes	D	78
8	individual	no	no	press	no	yes	M	1344
9	individual	no	yes	other	no	no	P	268
10	institution	no	no	press	no	yes	S	489

Construction Of Decision Tree Model

The meaning of bibliographic type symbol is explicitly stated in GB/T 7714, such as journal (J), dissertation (D), monograph (M), collected papers (C), standardization (S), report (R) and patent (P) etc. These bibliographic type symbols are the basis for the normative checking of bibliography and automatic processing. Nevertheless, the symbol of bibliographic type is optional in the GB/T 7714. The purpose of using decision tree is to predict the bibliographic type by analyzing the features of description items.

Two key issues need to be considered in the process of decision tree construction. The first is how to choose a current optimal grouping variable from a large number of input variables? That is, which type can be selected as the root of decision tree? The second is how to find an optimal split point from the numerous values of the grouping variable? How to select the best division is usually decided by the impure degree of divided node. The lower impure degree is, the better division will be. There are three methods commonly used to measure purity, Gini index, entropy and error rate. Entropy formula is popularly used, and information gain and information gain ratio are derived from it. The definitions are as follows:

$$Gain(U,V) = Entropy(U) - Entropy(U,V) \quad (1)$$

$$GainRatio(U,V) = Gain(U,V) / Entropy(V) \quad (2)$$

In this paper, information gain ratio is used to address the above two key issues. First of all, choose an optimal grouping variable according to the data in table 2. Respectively calculate Entropy (U), Entropy (U | Ti), Gains (U, Ti) and Gain Ratio (U, Ti). The information gain ratios of description items are shown in Table 3.

Table 3. The information gain ratio of description items

Description item	Author Type (T1)	Report Number (T2)	Patent Number (T3)	Publisher Type (T4)	Volume Period (T5)	Page Number (T6)
Gain Ratio	0.8068	0.3535	0.0793	0.9321	0.8402	0.4683

From Table 3, it can be seen that the publisher type (T4) has the maximum information gain ratio. Consequently, the publisher type will be selected as the optimal grouping variable, namely the root of decision tree. Whereas, the publisher type has four attribute values, “journal”, “education”, “other” and “press”. It needs to choose an optimal split point. The calculation process is actually similar. After calculating, “education” can be selected as the best split point. Now, “education” is assigned as the first node. Then, each subtree is recursively calculated in accordance with the method described above, and the decision tree is finally constructed as shown in Fig 1.

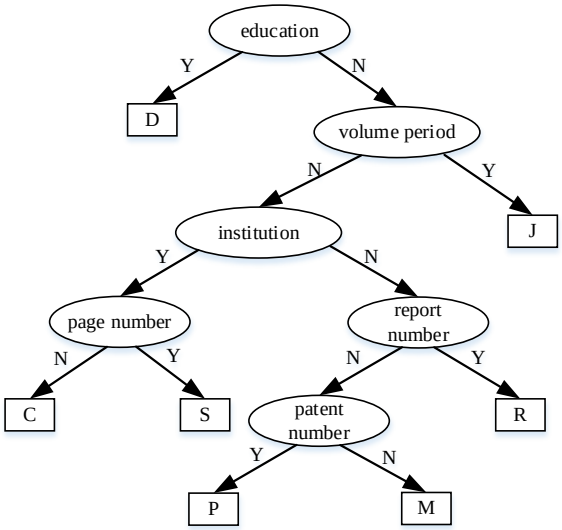


Fig. 1. Decision tree of predicting bibliographic type symbol

Now we can predict the unknown type of bibliography according to the decision tree in Fig 1. As an illustration, a reference is expressed as following:

> Zhu Gang. New fluid finite element method and the positive and negative mixing problem of impeller machinery. Beijing: Tsinghua University, 1996.

Reference checking according to the test conditions is started from the root of the tree, and appropriate branch is selected according to the test results until reaching the leaf node. As the publisher type belongs to education by analyzing the reference, it can be predicted that this reference belongs to academic paper (D) based on the decision tree model.

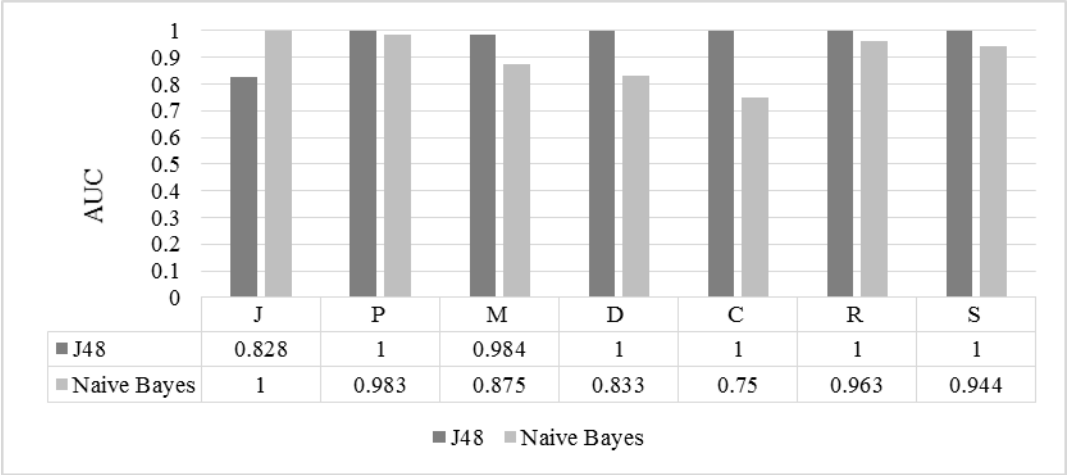
Experimental Results Analysis

The experiment is carried on the open source data mining tool of WEKA. The decision tree algorithm outlines in the previous section is WEKA’s J48 decision tree implementation based on C4.5 algorithm. And the Naive Bayes algorithm is used for comparison. The experimental data set contains 17760 references, which collected from 342 undergraduate theses and 150 graduate theses of computer specialty. We select 7 categories of J, D, M, C, R, P and S. Each category respectively selects 2709, 2388, 2479, 1389, 328, 268 and 489, a total of 10000 references to experiment. This paper uses 10 fold cross-validation to evaluate the classification results. The whole dataset is divided into 10 parts on average, and each time there are nine training sets and one testing set. The performances are evaluated using popular accuracy.

Table 4. Results comparing J48 with Naive Bayes algorithm in terms of accuracy

	J	P	M	D	C	R	S
J48	0.9444	1	0.8943	1	1	1	1
Naive Bayes	1	0.9841	0.8865	0.8749	0.8472	0.9653	0.9589

Table 4 shows the accuracy values comparing J48 with Naive Bayes algorithm. It can be observed that the accuracy of J48 classification algorithm is higher than Naive Bayes classification in most of the classification categories, indicating that J48 algorithm can be effectively classify the bibliography types. Another measure used to evaluate the performances is the ROC curves. With the adjustment of the classifier threshold parameters, not only ROC curves can directly reflect the classifier performance, but also the area under the ROC curve (AUC) can quantify the tendency of classifier to accept the positive examples. As can be seen from Fig 2, only the AUC of "J" and "M" do not reach 1, and the rest of the prediction types reach 1 in the J48 method. While in the Naive Bayes method, the AUC of "J" can reach 1 and the AUC of "C" is minimum, only 0.75. So the average performance of J48 outperforms the Naive Bayes. The overall conclusions that can be drawn are that decision tree algorithm has distinct superiority in the determination of bibliographic type.



	J	P	M	D	C	R	S
J48	0.828	1	0.984	1	1	1	1
Naive Bayes	1	0.983	0.875	0.833	0.75	0.963	0.944

Fig. 2. Comparison of AUC values for J48 and Naive Bayes

Based on those experiment results, a reference format normative checking system (RFNC system) is developed as the application of the determination method. RFNC system can check up whether the description items of references are missing or out-of-order and so on, prompting the specific error location information and giving revision suggestions. As an example, some checking results are shown in Fig 3.

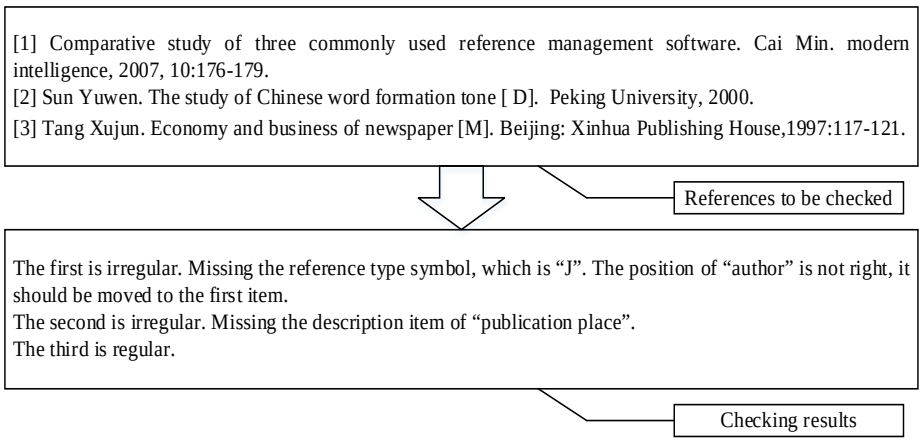


Fig. 3. The results of reference format normative checking

Conclusion

In this paper, we propose an approach based on decision tree to determine bibliographic type. After the segmentation and identification of reference description items, the decision tree is constructed according to the selected features, solving the problem of bibliographic type determination, thus realizing the normative checking of bibliography. The study has great significance for improving the quality of academic publications and the effective use of information, and saving the labor cost of typesetting.

Currently, this paper focuses on the Chinese bibliography, we are not able to pay enough attention on the bibliography of English and other languages. There are some difference in the extraction of description items. In addition, we only check the correctness of bibliography itself, whereas the citations in the text are not checked. All these can be further improved in the future.

Acknowledgement

In this paper, the research was supported by National Natural Science Foundation of China (Project No.61672105); the National High-tech R&D Program (863 Program No. 2015AA015403) and the General Program of Science and Technology Development Project of Beijing Municipal Education Commission (Project No. KM 201511232013).

References

[1] Information on http://endnote.com/product-details.

[2] Information on http://www.notefir-st.com/product/default.aspx?id=0px&defalut.

[3] Chunling Zhang. The XML-based solution for automatically checking the references in the electronic articles of academic journals [D]. Jilin University, 2011.

[4] Baochao Liu. Design and implementation of dissertation standardization evaluation system [D]. Yanbian University, 2015.

[5] The standard Descriptive rules for bibliography (GB/T 7714-2005) [S]. Beijing: China Standard Press, 2005.

[6] Wang H, Ye P, Deng S. The Application of Machine-Learning in the Research on Automatic Categorization of Chinese Periodical Articles[J]. New Technology of Library & Information Service, 2014.

[7] Abdul-Rahman, S., Mutalib, S., Khanafi, N. A., Ali, A. M. Exploring Feature Selection and Support Vector Machine in Text Categorization [C]. Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering. 2013:1101-1104.

[8] Deris A M, Zain A M, Sallehuddin R. Overview of Support Vector Machine in Modeling Machining Performances[J]. Procedia Engineering, 2011, 24(8):308–312.

[9] Siniscalchi S M, Svendsen T, Lee C H. An artificial neural network approach to automatic speech processing[J]. Neurocomputing, 2014, 140(140):326-338.

[10] Hemanth D J, Vijila C K S, Selvakumar A I, et al. Performance Improved Iteration-Free Artificial Neural Networks for Abnormal Magnetic Resonance Brain Image Classification[J]. Neurocomputing, 2014, 130(3):98-107.

[11] Zhang G P. Neural networks for classification: a survey [J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2000, 30(4):451-462.

[12] Ren J, Lee S D, Chen X, et al. Naive Bayes Classification of Uncertain Data[J]. 2009:944-949.

[13] Zhang M L, Peña J M, Robles V. Feature selection for multi-label naive Bayes classification[J]. Information Sciences An International Journal, 2009, 179(19):3218-3229.

[14] Das S, Dahiya S, Bharadwaj A. An online software for decision tree classification and visualization using c4.5 algorithm (ODTC)[C]. International Conference on Computing for Sustainable Global Development. 2014:962-965.

[15] Navada A, Ansari A N, Patil S, et al. Overview of use of decision tree algorithms in machine learning[C]. Control and System Graduate Research Colloquium. 2011:37-42.