The Design and Research on Management Classification of College Textbooks

Zhang Genlian[1, a]

[1] Jilin Business and Technology College，Educational administration department,Changchun, Jilin 130062, China

[a]zhanggenlian555@163.com

Keywords: College textbooks, classification, system, self-adaptive learning algorithm

Abstract. This paper proposes a method that uses the inter-category and intra-category distribution variance of feature words to describe inter-class and intra-class distribution of feature words, uses the inter-category and intra-category distribution variance of feature words to amend TFIDF value of entries, establishes a booklist hierarchical classification system model based on machine learning facing to Chinese Library Classification, puts forward the design thought constructed by shallow hierarchical classification system, and verifies feasibility and rationality of the model by using a large-scale experiment. The experiment proves that features extracted from the improved TFIDF method are correct and feasible to use for KNN classification

Introduction

With the extensive application of computers in various fields [1, 2], domestic college textbook management is gradually realizing microcomputer management. In terms of business schools of new category, it is still kept in the exploration stage. Lots of colleges develop management software [3, 4] that is suitable for their school in succession and form a kind of situation that fights the enemy separately, because management emphasis of different colleges has a larger difference. However, as for general characters, first of all, all of them are facing to a problem that they should classify these textbooks scientifically, namely how to establish a set of scientific textbook classification. Facing to increasing book publication, cataloguers of libraries are powerless to classify booklist manually. How to realize computers' auto-completion of book classification becomes a key problem to be solved in digital library construction.

The algorithm on text feature extraction is widely studied in the text classification and is an important issue. The traditional TFIDF algorithm can't distinguish these two situations and gives no consideration to distribution of intra-class document of feature words. In the intra-category document, if feature words can be distributed evenly, while the feature word can represent characteristics of this category. If it appears in several documents, this category won't appear in other documents. Obviously, these feature words can't represent characteristics of this category. This paper tries to introduce the adaptive learning classification algorithm to the study of automatic library classification, constructs a multilevel automatic library classification system model based on feature weighting, and makes it provide classification codes of Chinese Library Classification automatically in line with content feature of Chinese books(including title, keyword and abstract), so as to solve some problems, such as larger investment of books' manual classification, low efficiency, strong subjectivity and one-sidedness, as well as prove accuracy and rationality of the model through the experiment.

The Textbook Classification Model Based on the Adaptive Learning Algorithm

The Analysis and Design of the System Model

The paper adopts the adaptive learning technology to realize a basic method of automatic book classification, analyzes problems to be solved in the operation process, and proposes a multilevel automatic book classification system model based on a level classifier by aiming at abundant characteristics of Chinese books and complicated hierarchical structure of Chinese Library Classification.

The basic thinking of realizing automatic book classification based on adaptive learning: first of all, it analyzes booklist data of Chinese books, extracts basic features that describe book content and category codes of Chinese Library Classification, endows different weights in line with importance degree of features, regards feature vectors as booklist expression of books, and constructs binary feature matrix of books. Here row vector expresses booklist, column vector expresses features, and matrix value coij expresses the relevancy between booklist i and feature j. Category information of booklist added in the feature matrix is formed objects of self-adaptive learning, namely a category column vector catai is increased on the basis of feature matrix in Figure 1. Then, self-adaptive learning algorithm is adopted, frequently-common including decision-making tree, neural network and support vector machine, etc., to learn features+ category matrix and acquire the classifier; Ultimately, the classifier acts on feature matrix of books to be classified. The classification situation of books can be obtained through automatic data analysis. The entire process is shown in Figure 1. The entire automatic book classification process is divided into two stages: learning firstly and analyzing later. This adopts a self-adaptive learning method to realize basic methods of specific applications. However, category codes of books' Chinese Library Classification differ from text classification in general meaning, involving in problem solving of specific applications in feature extraction, feature weight setting, adaptive learning algorithm selection and confirmation of classification methods, etc.

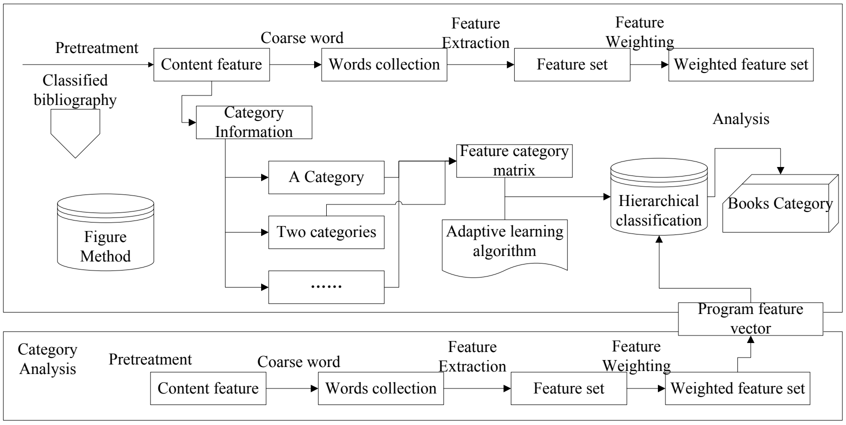

Figure 1: Book Catalogue Classification Design of Based on the Adaptive Learning Algorithm

Book Classification of TFIDF Based on the Improved Self-Adaptive Feature Extraction

Support Vector Machine (SVM) is a machine learning technology developed by the group in Bell Laboratory with the leader of V. Vapnik. The theoretical basis of SVM comes from the statistical learning theory proposed by Vapnik. Its basic idea refers to a given task with limited quantity samples, and how to compromise in accuracy (for given training set) and machine capacity (machine can learn the ability of arbitrary training set wrongly), so as to obtain the optimal promotion performance. SVM is used for text classification and compares with other classification algorithms. The result displays that SVM is superior to other methods. Hyperplane divides two categories. For multi-class text classifications that are greater than two categories, the hyperplane should be constructed on every category, which is separated from other categories. The number of categories can construct the same number of hyperplane, which can be selected the most suitable testing sample in testing.

The common feature extraction includes document frequency (DF), information gain (IG), mutual information (MI), and x statistical magnitude (CHI), TFIDF algorithm and term frequency (TF), inverse document frequency (IDF). The common-used computing method of IDF in this method is as follows:

$$idf(t) = long(\frac{N}{n}) \quad (1)$$

Here, N is the total documents in document set, and n is the number of documents appearing feature item t. The core idea of IDF algorithm means that feature items appeared in most of documents are less important to than feature items appeared in a small part of documents. IDF algorithm can weaken the importance degree of high-frequency feature items appeared in most of documents and also increase importance degree of low-frequency feature items appeared in a small part of documents. In the actual application, TF and IDF generally can be used together. The united formula of TF and IDF is as follows:

$$Weight_{TFIDF}(t) = tf(t) \times idf(t) \quad (2)$$

In many cases, vector should be normalized. The normalization computational formula of TFIDF is as follows:

$$Weight_{TFIDF}(t_i) = \frac{tfidf_i}{\sqrt{\sum_{j=1}^{n}(tfidf_j)^2}} \quad (3)$$

The main idea of TFIDF means that if a word or phrase has higher frequency TF in a document and seldom appears in other documents, it thinks that this word or phrase has the better ability to distinguish categories and is suitable for classification. The main idea of IDF means that if documents contain entry t are fewer, n will be smaller, and IDF is larger, indicating that entry t has the better ability to distinguish categories. IDF gives no consideration to one aspect of feature words in inter-category; on the other hand, though the number of documents containing t is smaller, it distributes in each category evenly. Such the feature word is not suitable for classification, but should endow with smaller weight. According to traditional TFIDF algorithm, the calculated IDF value is larger. The reason for existing these disadvantages is mainly because traditional feature weight algorithm TFIDF regards document set as the entirety to consider, especially for calculation of IDF and giving no consideration to inter-category distribution of feature items.

This paper regards entry t as a random variable. The value of t between categories can be expressed in inter-category word frequency (namely, the times of appearing the word in each category). Through the definition of variance: the distribution variance D (t) of t in each category embodies the dispersion degree of t in each category. If D (t) is smaller, distribution of t will be more even in each category. If t distributes in each category evenly, D (t) is 0, making no contribution to classification. By using the feature of variance, this paper uses D(t) to amend TFIDF formula to make up for the advantage that TFIDF gives no consideration to entry distribution in each category.

Through the above analysis, value of t in each category can be expressed in word frequency, but distribution probability of entry t in each category is hard to be calculated. If variance needs to be calculated, it will be difficult, so this paper uses average mean variation of entry t to replace D(t) appropriately.

Set up a total of n categories, $tf_i(t)$ represents the frequency of occurrence in Ci category, $\overline{tf(t)}$ represents average word frequency of entry t in each category, and computational formula is:

$$\overline{tf(t)} = \frac{1}{n}\sum_{i=1}^{n} tf_i(t) \quad (4)$$

Make t use mean variation square in each category as D_g, and the computational formula of t's mean variation square is:

$$D_e = \frac{1}{n}\sum_{i=1}^{n}(tf_i(t) - \overline{tft})^2 \quad (5)$$

Use D_e to amend TFIDF formula:

$$Weight_{TFIDF}(t) = tf(t) \times idf(t) \times D_e \quad (6)$$

Obviously, when t distributes in each category evenly, because D_e is equal to 0, so $Weight_{TFIDF}$ (t) =0, entry t makes no contributions to classification.

If entry t distributes evenly in the document of C_i category, D_{ii} will be smaller, but t can represent C_i category. The corresponding $1D_{-ii}$ will be larger. Therefore, 1-D_{ii} can be used for amending TFIDF formula:

$$D_{ii} = \frac{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t) - \overline{tf_i(t)})^2}{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t))^2} \quad (7)$$

$$Weight_{TFIDF}(t_{ik}) = tf(t_{ik}) \times idf(t_k) \times D_e \times (1 - D_{ii}) \quad (8)$$

Self-Adaptive Learning Algorithm Classification

Generally speaking, two methods are adopted to realize text classification, namely flat classification and hierarchical classification. Flat classification refers to regard all categories appeared in data as independent categories, but neglects the mutual relation between categories. When classifying, text is classified into the related category with the highest confidence coefficient. Such the classification method is relatively suitable for categories with less quantity. The relation between categories is relatively single data environment. Once the number of categories in classification system can reach a larger scale or mutual categories have the hierarchical relation, the calculation complexity will be sharply increased, while distinguished capacity of categories will be reduced significantly. The analysis accuracy will be decreased rapidly. Therefore, this method is not suitable for the classification environment with a larger scale of categories. Hierarchical classification refers to decompose a complicated classification task into classification with smaller scale for several levels in line with the hierarchical relation between categories and classifies documents to be classified step by step. This method can simplify a complicated problem, not only reduces computation complexity greatly, but also is fit for the automatic classification with large-scale classification system. Hierarchical classification translates the entire classification system into the tree structure, as shown in Figure 4, while the classification problem is translated into to the process of looking for leaf nodes from the root node. First of all, through data training, the m multiple classifiers are established in every internal node. M value is the number of child node of the current node. When carrying out data analysis, it begins with the root node, uses m multiple classifiers to distribute test cases to the subtree of the corresponding child node, analyzes step by step, until test cases reach a certain leaf node, marking the end of the entire classification process. The dotted arrow in the Figure 4 marks the direction of test cases' hierarchical classification.

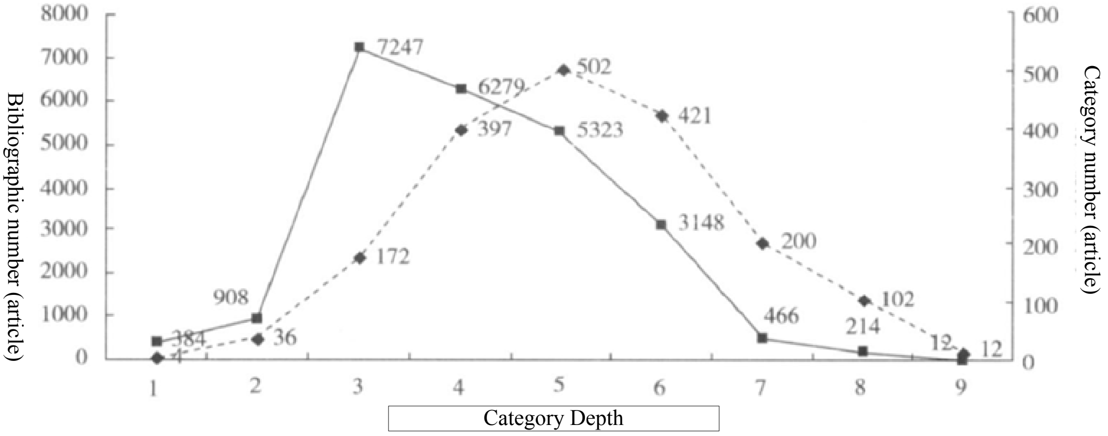

Figure 2: Booklist and Classification Distribution Based on Hierarchical Depth

Booklist distribution (solid line) and category distribution (dotted line) of Chinese Library Classification's different hierarchical depth are shown in Figure 2. It can be observed from the figure that middle-level booklist and category quantity are further higher than high level and low level, resulting in booklist data present a kind of sparse phenomenon with unbalanced distribution, high-level and low-level data. This also indicates that when category allocates automatically, it should concentrate on 3-7 levels as much as possible. The solid line in Figure 6 presents average book contain of each category. The 6th level has already reduced to below 10. Dotted line describes the proportion of sparse categories (booklist quantity is less than 4) of each level. The proportion of 6th level is increased to above 70%. Thus, it can be observed that in terms of DB ACFNX corpus, booklist data from level 1-5 may have an ideal learning effect. In level 6-9, due to reduction of categories' training corpus and increase of sparse categories, it will result in the dramatic decline of these levels' category discrimination.

Experimental Results and Analysis

Two experiments of this paper adopt the data set as shown in Table 1. Here training samples and testing samples have 10 categories, respectively. Training samples have a total of 1882 documents, while testing samples have 934 documents. The detailed information is shown in Table 5.1. It can be observed from the table that the proportion of training documents and testing documents in each category is about 2:1.

Table 1: Experimental Data

Category type	Environment	Computer	Transportation	Education	Economy	Military	Sports	Medicine	Art	Politic
Training set	134	134	143	147	217	166	301	136	166	338
Testing set	67	66	71	73	108	83	149	68	82	167

Through repeated experiment, it can be found that for traditional TFIDF, when k is equal to 8 and 10, the classification effect is the best, while for the improved algorithm, when k is equal to 18, the classification effect is the best. The specific experiment effect is shown in Table 2:

Table 2: The Classification effect in k-value of Different Situations

K Evaluation index	K=8	K=10	K=16	K=18
Check rate(traditional)	88.865%	88.865%	88.544%	87.9015%
Check rate(improvement)	90.578%	90.257%	90.685%	90.792%

The classification effect of traditional TFIDF algorithm and improved TFIDF algorithm can be observed from the Table 2. For traditional algorithm, when k is equal to 8 and 10, and for improved algorithm, when k is equal to 18, it reaches the optimal. However, in order to increase strength of persuading, k value which makes traditional algorithm reach the best classification effect is compared, namely it compares the classification effect of improved algorithm and traditional algorithm, when k is equal to 8 and 10.

X-coordinate is the category code of testing documents, corresponding to environment, medicine, military, economy, education, sports, art, politic, computer, and transportation, respectively. Figure 3 is the classification result of each category and means that feature is extracted from traditional and improved feature extraction algorithms and uses for KNN (when k=8). In the check rate, values of two categories in the tradition are higher than the improved one. Values of two categories are equal to the improved one. Values of other categories are lower than the improved. In the correct rate, traditional algorithm has four categories higher than the improved one, and others are lower than the improved one.

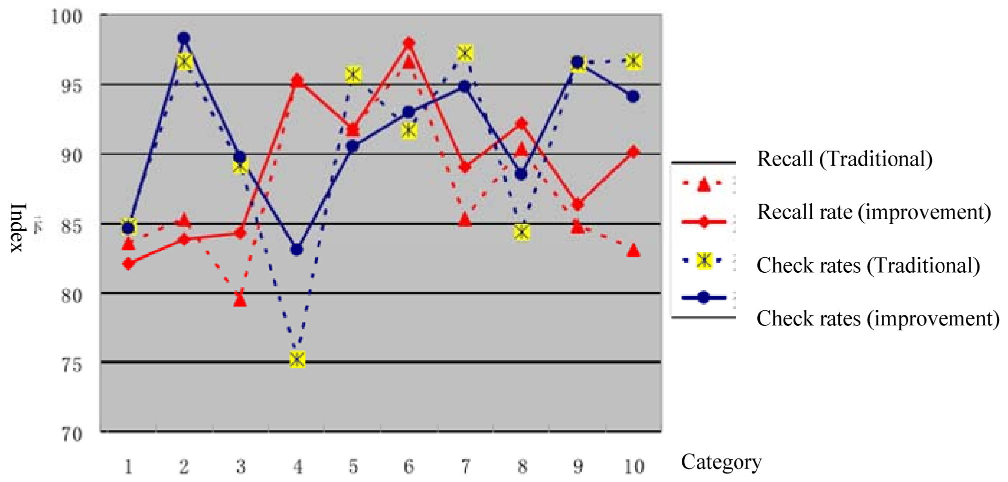

Figure 3: Classification Result Comparison Diagram of Traditional and Improved TFIDF Methods (k=8)

Figure 4 is the classification result of each category and means that feature is extracted from traditional and improved feature extraction algorithms and uses for KNN (when k=10). Values of two categories in the tradition are higher than the improved one. Values of two categories are equal to the improved one. Values of other categories are lower than the improved. In the correct rate, traditional algorithm has four categories higher than the improved one, and others are lower than the improved one.

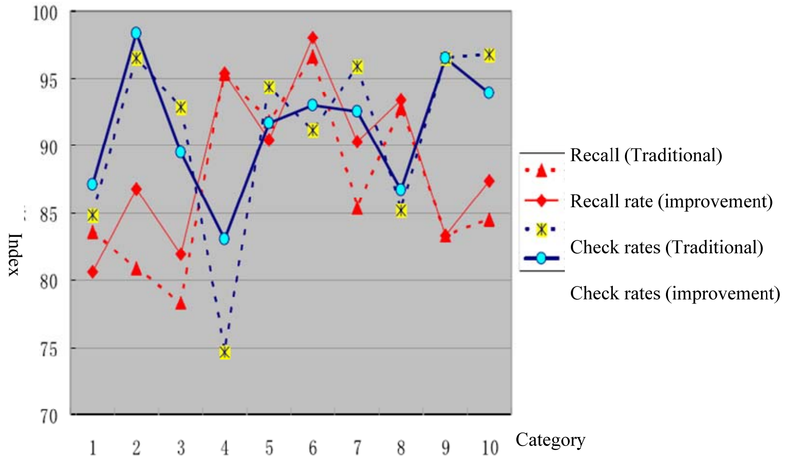

Figure 3: Classification Result Comparison Diagram of Traditional and Improved TFIDF Methods (k=10)

Summary

This paper adopts a machine learning method based on the realization of the shallow hierarchical booklist classification model, regards parts of category books as research objects, conducts experimental demonstration and analysis of the model, establishes practical n multiple classifiers by aiming at different categories or sub-categories in Chinese Library Classification, uses for automatic catalogue of libraries, and promotes the development of digital library economy. The machine learning method is regarded as an important direction of artificial intelligence, not only can be used for the study of automatic booklist classification. The selective work of libraries can give consideration to adopt a machine learning method. The improved TFIDF feature extraction algorithm, correct document quantity of total classification, total check rate, correct rate, and F1 value should be better used for traditional feature extraction algorithm.

References

[1] He Lin, Hou Hanqing and Bai Zhentian, et al., *Multi-level Automatic Classification Based on Index Experience and Combination of Machine Learning*, Journal of Intelligence, Vol. 25, 2006, p.725-729;
[2] Gu Baohua and Liu Zhenyi, *The Analysis and Design of Textbook Management Information System in Institutions of Higher Learning*, Journal of Liaoning University of Technology, Vol.4, 2005 p.134-136;
[3] Guo Shouku and Zhou Zuotao, *The Design and Realization of College Textbook Management System under the C/S Pattern*, Journal of Shaanxi University of Technology(natural science), Vol.3, 2005, p.97-98;
[4] Cheng Ying and Shi Jiulin, *the Current Situation and Prospect of Automatic Classification,* Journal of Intelligence, Vol.18, 2008;
[5] Zhou Yonggeng, Yu Hongqi and Hu Yunfa, et al., *Chinese Text Classification Study Based on N-Gram Information*, Journal of Chinese Information Processing, Vol.15, 2005, 13-15;
[6] Abrizio, Sebastiani, *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, Vol.34 (I), 2002, p.1-47;
[7] Bao Riyuan, *Several Reflections on Speeding up College Textbook Informationalization Management*, Journal of Fujian University of Technology, Vol.8, 2008, 113-115.