Study of Keyword Correlation based on Information Analysis

Maoguo Chen[1, a], Shuxin Zheng[2, b, *], Zhongguo Zhang[1, c] and Weiwei Xu[1, d]

[1]School of Computer Science and Technology of Taizhou University, Taizhou 225300 China

[2]School of Business Administration of Taizhou University, Taizhou 225300 China

[a]114552248@qq.com, [b, *] 1121561926@qq.com, [c]1142568841@qq.com, [d]834390273@qq.com

Abstract. This paper extracts keywords, takes keyword frequency statistics and analyzes correlation by collecting index texts from student books in the library and questionnaires to understand college students' preference interests, value orientations and hotspots. Thereby, they can further understand the reading tendency of college students, summarize their reading characteristics and then sort and process them, and provide students with accurate reading information services and book recommendations, and provide reading guidance for students.

Keywords: support degree; correlation analysis; keywords co-occurrence.

1. Introduction

As an important part of the strength of and universities, university libraries are mainly reflected in its information function. In the information age, it is an important task for universities to take advantage of muti-functional information network technology to improve the work of libraries, to create a digital, efficient university library. The information revolution has fundamentally promoted the development of libraries. The library uses various automated integration systems to establish its own internal network environment, showing the characteristics of network, Information, intelligence and socialization, which has laid a good foundation for providing intelligence service for students and teachers. It makes it possible to make use of a certain relationship between users, between users and resources to conduct resource recommendations by analyzing correlation of indexed keywords for correlation analysis in a library management system with a large number of users and multiple types of resources. In the same recommendation process, not only the friends but also the relevant types of resources are recommended to the user. In particular, a stable interpersonal network can be established through constantly updated, original and rich contents. However, in today's era of wisdom education, in order to accelerate the pace of national development and improve the culture and quality of the whole nation, college students, as the central force of national development, must first improve their own reading ability and ability to acquire knowledge. But it's pitied that college students are now trapped in a "reading crisis". College students read not in purpose of personal interests and needs, but for exams or entertainment. This kind of utilitarian, entertaining and shallow reading hinders the improvement of college students' reading ability, and it is also not conducive to the development and progress of the society, which has a negative impact on the improvement of the overall quality of the nation. Therefore, it is very necessary to study the reading interest of college students, to provide books according to their preferences, to promote college students' reading and even guide the deep reading of college students. However, in order to provide students with effective reading services, they must study their hobbies and interests as well as value orientation and understanding of learning so as to provide accurate reading guidance.

2. Research Methods

Firstly, the library texts of the students' books and the title texts of the library collections in a certain period of time are obtained from the library. Afterwards, the natural language word segmentation tools are used to process the title texts of the collections of the library, and the keywords are extracted from the texts after the word segmentation. The data of book title is used to calculate

the word frequency to obtain the high-frequency words. Finally, relevant books and friends are recommended to provide personalized services through the correlation analysis.

2.1 Keyword Extraction

In order to avoid the impact of synonymous keywords on the analysis results, the synonym keywords are combined, and the representative of the statistical results is combined to determine the threshold of the high frequency words and the low frequency words.

This paper uses the TF-IWF algorithm, which is an improvement on the TF-IDF algorithm. The TF-IDF algorithm calculates the word weight by counting the word frequency of the document. The basic idea is that the more times a word appears in a particular document, the stronger ability to distinguish the content attributes of the document (TF) is, the broader the scope of a word appears in the document, the lower attributes to distinguish the content of the document are (IDF). Its classic formula is

$$\omega_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{n_j}\right) \tag{1}$$

tf_{ij} ——The times of occurrences of the word " t_i " in the document" d_i "

idf_i —— The countdown of the document in which the word" t_i " appears

N ——Total number of documents

n_i ————Number of documents with words t_i

w_{ij} ——Word weight in the document

The greater the weight is, the stronger the ability of the word to represent the subject matter of the document contents is. TF-IDF algorithm is widely used for word weight calculation because of simplicity and effectiveness. However, the word frequency in the algorithm has a greater impact on the calculation results. The problems with dataset skew and inter-class and intra-class distribution bias exist in text categorization [1]. In response to the deficiencies above, scholars have proposed some improved algorithms, among which the TF-IWF algorithm is more influential [2]. It uses the square of IDF to balance the weight of the word frequency. The formula is as follows

$$\omega_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{n_j}\right) \times \log\left(\frac{N}{n_i}\right) \tag{2}$$

The feature items extracted by the algorithm achieve better results in text classification, and the speed is faster. Document keywords are extracted automatically. In information retrieval, keywords as a basic unit for summarizing the subject content of a document. The number of keywords in the document is generally selected according to the application. The appropriate number is generally within nine words. It is a general user who does not need special efforts. The number that can be remembered, at the same time, in the scientific literature, the number of keywords manually labeled is generally no more than five. The number of document keywords is generally selected according to the application. It is generally suitable for less than nine words. The number is the one that a general user does not need to take special efforts to remember. At the same time, in the scientific literature, the number of keywords manually labeled is generally no more than five.

Since the number of words in each document and the degree of richness of the content included are both different, the number of document keywords should be dynamically determined according to the length of the document. That is, the number of keywords is determined according to the number of words in the document. In the specific implementation, the number of keywords of 300 words and below is set to 3, and for each 300 words added, one keyword is added. The specific calculation formula is as follows

$$keywordNum_i = \left\lfloor \frac{WordCount_i}{300} \right\rfloor + 3 \tag{3}$$

In the formula, $KeywordNum_i$ ——The number of keywords in the document is up to 9.
$WordCount_i$ ——The number of words in the document d_i .

2.2 Calculating Word Frequency

In the TFIDF model, the term frequency (TF) refers to the frequency at which a given word appears in the text. Inverse Document Frequency (IDF) is used to measure the universal importance of a word. It calculates the number of documents containing a specified word in the document set. The smaller the number is, the more the word represents the specified book. The TF value for the bibliographic keyword ti specified in the document set can be expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

$n_{i,j}$ in the numerator of the above formula is the number of occurrences of the keyword t_i in the specified bibliographic name d_j , and the denominator is the sum of the occurrences of all the words in the d_j . Divide the total number of books by the number of books containing the keyword, and then take the logarithm of the result to get the IDF value of the word t_i . That is:

$$idf_i = \log \frac{|D|}{|\{j : t_i \varepsilon d_j\}|} \tag{5}$$

Where | D | is the number of texts in the document collection, $|\{j : t_i \varepsilon d_i\}|$ is the number of texts containing the word t_i . The formula for calculating $tf_i df$ is as follows:

$$tf_i df_{i,j} = tf_{i,j} \times idf_i \tag{6}$$

2.3 Correlation Analysis

As a commonly used data processing method, correlation analysis has been widely used in data analysis. The intrinsic relationship between things can be found quickly and efficiently by means of correlation analysis.

Association rules. The association rules were originally proposed by researchers to solve the "shopping basket" analysis problem. This discovery of association can help sellers understand which products are frequently purchased by consumers, thus helping them to develop better sales strategies [3]. Nowadays, the main carrier of information has been transformed from traditional digital data to structure-specific text data. When faced with massive text data, the association rules proposed for the "shopping basket" problem can be used to analyze the correlation of specific words in text data. In the association rules, support and confidence [4-7] are two important indicators used to describe the results, which measure the validity and certainty of the rules found in the analysis results. The minimum support value and the minimum confidence value are set autonomously according to

different research needs. Suppose D is the total number of records in the database, t is the element of D, and I is the true subset of t then the support of I is as follows,

$$Suppot(I) = \frac{\|\{t \in DI \subseteq t\}\|}{\|D\|} \tag{7}$$

The association rule is a rule form similar to $X \to Y$, ie the rule $X \to Y$ is established in D. $Suppot(X \to Y) = P(X \bigcup Y)$

$P(X \bigcup Y)$ represents that the rule $X \to Y$ is established in database D and possess the support. Where $Suppot(X \to Y)$ is the percentage of $Suppot(X \to Y)$ contained in D, which is the probability.

$$Confidence(X \to Y) = \frac{Support(I_x \bigcup I_y)}{Support(I_x)} = P(Y / X) \tag{8}$$

$P(X \to Y)$ indicates that the rule $X \to Y$ has a confidence level $Suppot(X \to Y)$ in the database D. This confidence is the percentage of data in D that contains both X and Y, which is the conditional probability $P(X / Y)$.

The study of text association rules using keywords can be roughly divided into two steps: one is to use text mining algorithms to obtain keywords in text; the other is to use association rules in frequent items consisting of keywords to derive potential rules.

3. Case Analysis

This paper uses the method of association rule to analyze data of keywords. In order to verify the above keyword automatic extraction model and conduct the correlation excavation, the data collected by the Taizhou University Library and the corresponding questionnaires were used to extract the keywords, and the documents were preprocessed, including: Clear/Labels, word segmentation, stop words, etc., and then use Apriori algorithm to calculate, according to the support threshold value to get some items set Table 1.

Table 1. Frequent items of keyword data

Frequency	Support	Frequency	Support
Prospects	0.21	Pressure	0.05
Schoolwork	0.19	Online games	0.03
Quality	0.17	Postgraduate	0.02
Love	0.14	Classmate relation	0.016
Graduate	0.14	Security	0.016
boringness	0.10	Partner	0.016
Academic warning	0.08	Encouragement	0.016
Efforts	0.06	No regrets youth	0.016

Table 1 shows that college students are very concerned about their academic and future, while focusing on their academic and graduation status. Therefore, according to the learning situation of college students, especially the academic achievement, the teacher should provide the book recommendation in a targeted manner and give corresponding content guidance. At the same time, it can be seen that college students have low support for ‘classmate relationship’, ‘encourage’, ‘partner’, ‘security’ and other issues, which indicates that there is still a lot of space for college students to raise awareness of teamwork, mutual support and safety awareness. So, the ideological and political

teachers, class teachers and counselors should consciously recommend relevant books and give them reading instructions.

Association rule analysis. In general, the larger the frequent item sets are, the more relevant the association rules are. Table 2 is the partial association rule obtained in this paper.

Table 2. Association rules in keyword data (partial)

Association rules	Confidence	Association rules	Confidence
Postgraduate → Prospects	0.51	boringness →Online games	1.00
Prospects → Schoolwork	0.71	Quality → Efforts	1.00
Schoolwork → Resources	0.43	Study → Pressure	0.88
Study → Boringness	0.89	Postgraduate → Family	0.73
Postgraduate→ Teacher guidance	0.43	Prospects → Efforts	0.75

From the data in Table 2, we can analyze the basic rules and learning status of the students in Taizhou University. It's very common that Students passively learn under pressure and have a feeling of helplessness. It's certain that they will create ideas that want to relax, and the easy way to release stress is to play the game. However, they have very clear consciousness. ——Their own future is closely related to their own efforts. In particular, although students are not very active in learning, they firmly believe that their own quality must be improved by their own efforts. They have realized that their own future is inseparable from your own efforts. This indicates that the students have upward consciousness, but they have poor self-control ability. What is interesting is that the intention or idea of a student's postgraduate entrance examination is more influenced by the family than the teacher's guidance, which is verified in the student work of the School of Computer Science. Table 2 also shows that students have a general recognition of the future of admission to graduate students, which also shows the diversity of modern college students' values and employment destinations. This phenomenon is evident in science and engineering majors, especially computer-related majors.

4. Conclusion

The increasing collection capacity of the library has led to explosive growth of library bibliographic data. Therefore, how to provide effective and accurate book recommendation in the process of student retrieval has become an urgent problem in current research. Based on the student's book retrieval information and questionnaire survey information in the library, this paper extracts the key words. On this basis, the word frequency analysis method is used to statistically sort out the high-frequency keywords of students' searches. After correlation analysis, the potential problems of college students are analyzed, then their ideas and needs are refined, to achieve high efficiency, purposeful, accurate book recommendation, so as to provide students with an effective reading recommendation method.

Acknowledgments

This study was funded by the Provincial Key Construction Disciplines in jiangsu province (Management Science and Engineering) during the 13th five-year plan.

References

[1]. Zhang Yufang, Peng Shiming, Lv Jia. Improvement and Application of TFIDF Method Based on Text Classification, Computer Engineering[J], 2006,32(19):76-78.

[2]. Liu Hua. Clustering Field Words by Character Extraction in Text Classification, Applied Linguistics[J], 2007(1):139-144.

[3]. Liu Hua. Knowledge Repository Acquire for Keywords Auto-Indexing System Based on Labeled and Classed Corpus, Library and Information Service[J], 2007,51(7): 41-43.

[4]. Zhang Hongying. Chinese Text Keywords Extraction Based on Fuzzy Processing, New Technology of Library and Information Service[J], 2009(5):39-43.

[5]. Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, Second Edition[M], Beijing, Mechanical Industry Publishing House, 2012.

[6]. Dmitry D, Oren T, Ari R. Enhanced sentiment learning using Twitter hash tags and smileys[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg: Association for Computational Linguistics, 2010:241-249.

[7]. CASTRO J, RODRIGUEZ RM, BARRANCO M J. Weighting of features in content -based filtering with entropy and dependence measures[J]. International journal of computational intelligence systems, 2014,7 (1): 80 -89.