Keywords Extraction Based on Text Classification

Yang Li-gong[1a,]Zhu Jian[2], Tang Shi-ping[1b]

1.Department of Computer Science and Technology,

Beijing Institute of Technology, Beijing 100081, China
[a]yyllgg@gmail.com；[b]simontangbit@bit.edu.cn

2. China Youth University For political Sciences, Beijing ,100089,China

[2] zhujian07@foxmail.com

Key words: text classification; keywords extraction; candidate word；weight function CLC Number: TP391.1

Abstract: In this paper, we propose new keywords extraction method based on texts classification. We first classify texts to determine their categories. Then determine weights of candidate words according to both their frequency and the relevance between text words and text category. Finally, keywords are extracted by sorting weights of candidate words. We conduct this experiment to show that on the premise of accurate text classification, this method can extract keywords effectively from text without title or with deviated title which can not reflect text's subject. Objective selecting of candidate word weighting function still needs to be further researched.

Introduction

Keywords extraction of articles is widely applied in text summarization, information retrieval and other aspects. It is not only of great value and significance for scientific and technical literature work, but also a very significant work for the information acquisition of abundant articles without subjects on the internet. Artificial method is the common method for keywords extraction; however, it is suitable for texts in professional fields like scientific and technical literature and academic report and is inefficient for abundant non-professional articles without subjects on the internet, so the extraction only could be achieved through machine learning.

In machine learning method, the determination of candidate word's weight is essential for article keywords extraction. Generally, the weights of candidate words are determined by their importance in reflecting the subject or theme of the article. Scientific and technical literatures have distinct subjects which could accurately reflect literature content, so keywords extraction could be conducted by utilizing the relevance between keywords and the subject. However, regular news reports and narrative articles on the internet tend to have indistinct subjects and less prominent themes, and it is a problem to select the keywords that could accurately reflect the subject or narrative content of the article. In order to solve this problem, this paper will propose a keywords extraction method based on text classification: first classify normal texts to determine their categories, then determine the weights of candidate words according to factors such as their relativities with corresponding text categories and their frequency in the text, finally extract keywords according to the sorting result of the candidate words.

Text classification and word segmentation

Normal articles all have emphases in statements of facts and introductions of knowledge, so they usually could be classified into different text categories, such as economics, politics and military, according to the differences in their narrative contents.

There are many text classification methods, and the most common one is to use vector space model to convert texts into high-dimensional vectors, and then use different classifications to classify texts

Published by Atlantis Press, Paris, France.

into different categories, and the classifiers frequently used are Bayes Network, nearest neighbor classifier and support vector machine, etc.

Because Chinese texts are different from English texts and word boundaries could not be judged naturally. Word segmentation should first be conducted on the text for the machine processing of article. Here we use the word segmentation software of ICT (Institute of Computing Technology of the Chinese Academy of Sciences) for the word segmentation processing of normal articles. This software could conduct word segmentation for most texts, and its word segmentation accuracy is relatively high. Different articles have different narrative contents; so many articles may have technical terms or phrases. And these technical terms could not be identified correctly by normal word segmentation software, because these technical terms usually are noun phrases or abbreviations. In order to solve this problem, we selected technical terms in economics, politics and military fields from specialized vocabularies in electronic dictionaries and added them in the specialized vocabularies; then use matching method to compare the aforementioned texts handled by normal word segmentation software with specialized vocabulary, so as to recognize the technical terms and phrases thereinto; finally, convert segmented texts into vectors in vector space, then use support vector machine to classify these texts to determine their categories.

Calculate the relativity of each word in word group with article category

Classification of category word group

During text classification in Literature [1], vocabularies in texts with different categories are sorted into different category vocabulary sets; then words belong to different categories are structured into an intersection model which will be used for classification of new texts. The set model is as figure 1.

During keywords extraction in this paper, we will continue to utilize this set model, i.e., according to the occurrence of each word and technical term in different texts, sort the words and technical terms into different category sets. Then determine the positions of keywords to be extracted according to this category set and words found in different category sets after text classification.

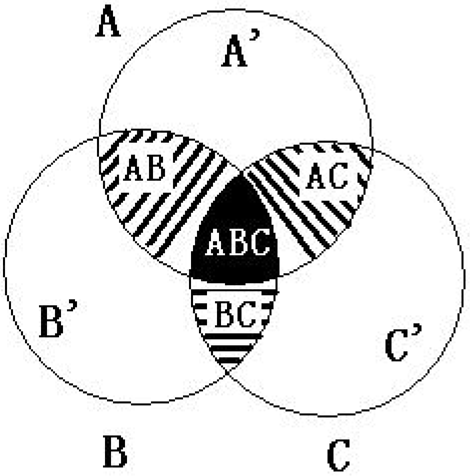

Figure 1 Vocabulary intersection of text classification

Statistics of the relativity between words and article subjects

When use normal keywords extraction methods, we often need to calculate the relativity between words and article subjects. In normal scientific and technical literatures or academic reports, articles' subjects are usually distinctly reflected by their titles. So when considering the relativities between words and article subjects, most methods are to calculate the relativity between words and article titles. This is certainly very rational for normal scientific and technical papers or academic reports, because these academic literatures' titles usually correspond to their contents. However, this method usually is ineffective for texts like normal news reports, narrative articles or brief evaluations. Because these texts' titles tend to be fresh and exaggerated to attract readers, but the titles are usually not enough to reflect article subjects or main contents. So it would be inaccurate if the relativity between words and article titles is still used in keywords extraction.

In this paper, we mainly extract keywords according to the category, word frequency and other relevant information of the text to be extracted. The detailed method is to firstly conduct word segmentation on the text to be extracted and determine the text category, as mentioned in Part 2. Because the general subject range of a text will be determined once its category is determined. Then count different words (including technical terms) occurring in the text and which word category sets the words are in. For example, economics articles will certainly contain some important words and phrases of economics field, and these words and phrases also would be found in the set of economics vocabulary or in vocabulary intersection which closely relate to economics field. Obviously, keywords in economics articles mainly should be selected from vocabularies related to economics, and of course several keywords also could be found in vocabularies of non-economics fields. Under this condition, the word frequency of each word in the text should be considered. With this method, we initially filtered the relativity between keywords and subjects, which is the candidate range of keywords.

Selection of keywords

The above Part 3 initially filtered the source range of keywords, but it doesn't mean that all keywords should be selected from word groups in the same category. Because although some words are found in corresponding word category, they might be only occasional in the text and they are not enough to reflect article subject or theme content. In order to evaluate these words' importance in the text, we also consider their comprehensive weights according to their frequencies in the article and their positions in category word groups. For example, consider a word or a technical term's frequency in the text to be extracted, or consider whether it is found in the category vocabulary set corresponding to text category, and consider whether it is found in independent set or intersection, because the weight calculation methods for words in different sets are different. So we comprehensively evaluate each candidate word's weight according to these 3 factors: category, position in set and frequency, then filter out important keywords according to the sorting for each candidate word's comprehensive weight.

The following formulas are the calculation methods for the comprehensive weight of candidate words in the text to be extracted:

（1）Vocabulary category weight function:

$$L(w)=\begin{cases} 2 & \text{w belongs to the independent vocabulary set in the same category with T} \\ 1 & \text{w belongs to a intersection with the vocabulary set in the same category with T} \\ 1/2 & \text{w belongs to a intersection with the vocabulary set not in the same category with T} \end{cases} \qquad (1)$$

Where, w represents word and T represents text.

（2）Vocabulary position weight function

$$P(w)=\begin{cases} 1 & \text{w} \in \text{ Independent set} \\ 1/2 & \text{w} \in \text{ Intersection of 2 categories} \\ 1/4 & \text{w} \in \text{ Intersection of 3 categories} \end{cases} \qquad (2)$$

(3) Vocabulary frequency function

$$f(w)=x \qquad (3)$$

Where, x represents w (word)'s frequency in T (text).

(4) Candidate word's comprehensive weight function

$$S(w)=L(w)*P(w)*f(w) \qquad (4)$$

According to the four function formulas, we could calculate each word's comprehensive weight as a candidate word.

Test results and analysis

Test materials and procedures

For keywords extraction, texts in relevant categories should be collected firstly. And here we use the texts already classified into politics, economics and sports by Fudan University. Each category has 500 articles, and 400 of those 500 articles are used in word segmentation and category vocabulary set construction, and the residual 100 articles are used as the test texts for keywords extraction. After finishing constructing three category vocabulary sets, we add specialized vocabularies of different fields acquired from terminological dictionaries of electronic dictionaries to construct the final category vocabulary set. Then after word segmentation and identification of technical terms for test texts, we use lineal SVM to classify the texts, and finally we use the aforementioned method to extract the keywords of these texts. In order to evaluate the quality of keywords machine extraction, we manually extract keywords of these texts, 5 keywords in each, and then we contrast the effects of keywords machine extraction and manual extraction. We mainly use coverage rate as the indicator, that is, 5 keywords extracted by machine are found in manually extracted keywords, and we use two people to conduct keywords extraction to guarantee the fairness of evaluation standard, then we contrast the machine result with the two people’s results and take the average of those two coverage rates and use the average as the indicator for machine extraction accuracy.

Test result

Table 1 Text classification accuracy table

	Politics	Economics	Computer
Precision	86％	87％	91％
Recall	82.7％	86.1％	93.8％
F1 value	84.％	86.％	92.4％

Table 2 Correct keywords extraction coverage table

	0％	20％	40％	60％	80％	100％
Manual 1	3.40	9.84	16.29	33.71	28.79	7.95
Manual 2	2.65	8.7	19.32	34.85	26.89	7.58
Average	3.03	9.27	17.81	34.28	27.84	7.77

Text classification is conducted on 300 texts in total, where, 264 texts are classified correctly and 36 texts are classified incorrectly. We use the aforementioned method for the keywords extraction of all texts, and the keywords extraction of texts classified incorrectly is based on their incorrect categories. Now the average coverage rates of correct and incorrect texts are shown in table 3, where, RAO represents the average coverage rate of correct texts and WAO represents the average coverage rate of incorrect texts.

Table 3 Errata of text classification and keywords extraction

	0％	20％	40％	60％	80％	100％
RAO	3.03	9.27	17.81	34.28	27.84	7.77
WAO	77.78	16.67	5.55	0	0	0

Test analysis

It could be seen on the table above that the classification result for texts of support vector machine is good, and the accuracy rates of three categories all are nearly 90%; and our keywords extraction for classified texts is based on this classification. The difference is that we also use this method for the keywords extraction for incorrectly classified texts, so as to evaluate this method's extraction effect on incorrectly classified texts.

In table 2, we manually extract 5 keywords in each text and use these keywords as the comparison standard with keywords extraction through machine. In order to guarantee the objective fairness of comparison as far as possible, two people independently extract the keywords in texts, and then we contrast these two manual extraction results with machine extraction results and use the average as the final indicator for machine extraction accuracy. We use coverage rate as the indicator in comparison. For example, the proportion of machine extracted texts having one keyword in manually extracted keywords in total texts is the proportion of articles with 20% coverage rate, then we get table 2. As shown in table 2, extraction coverage rates are in different degrees, but if we use the coverage equal to or more than 60% as the good effect of keywords extraction, this method's effect is up to 69.89%. Considering the distribution of these coverage rates, most of them are 60% or 80%, and the distribution under 40% is quite scarce. especially there are only quite a little proportion with 0% coverage rate. This shows that this keywords extraction method is relatively effective, especially in consideration that this method is used for keywords extraction of informal texts like news reports or normal narrative texts, so this effect is acceptable because most extracted keywords could effectively reflect the subject contents of texts.

Text classification could not guarantee that all texts are classified into correct categories, so we also have to consider this method's keywords extraction effect for incorrectly classified texts. In this test, there are 36 incorrectly classified texts, accounting for 12% of the total test texts. We conduct keywords extraction for the incorrectly classified texts according to their incorrect categories, and the extraction effect is shown in table 3. For the comparison with correctly classified texts, we also list the extraction effect of correctly classified texts in this table. It could be seen from this table that, most extracted keywords are incorrect if the classification is incorrect namely, those extracted keywords could not faithfully reflect the main content of texts, but a few extracted keywords still have some effect. Through detailed text analysis, we could find that most of the words in these texts are in the intersections of two vocabulary categories. Under this condition, most keywords extracted by this method are in these sets and they still have some effect for these incorrectly classified texts.

Conclusion and Future Work

This paper proposes new keywords extraction method based on classification. This method could effectively extract keywords in texts without subjects or texts with titles which could not reflect their subjects to reflect their contents. It could be seen from the test result that, on the premise of ensuring the text classification accuracy, the keywords extraction with this method is close to those extracted manually. Because this paper considers words' category properties during the calculation of candidate words' comprehensive weights, the selection range of keywords is initially filtered and a foundation is laid for the correct selection of keywords. However, during determining the weight function of candidate words, categories and positions, relevant numerical values are defined artificially, which brings factor that is not subjective to the calculation of word's comprehensive weights. In order to accurately reflect words' category and position weights, the next problem to be solved is how to determine these two functions objectively.

References

[1] Brook Wu Yi-fang，Li Quan-zhi，Razvan Stefan Bot，et a1．KIP：a keyphrase identification program with learning functions[C]. Proceedings of the International Conference on Information Technology：Coding and Computing(ITCC'04)，2004.

[2] Hulth A． Improved automatic keyword extraction given more linguistic knowledge[C]. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing，2003：216--223．

[3] Tumey P． Learning to extract key phrases from text，NRC／ERB－1057[R]，1999,02,17．

[4] Yang Wen-Feng. Chinese keyword extraction based on max2dup licated strings of the documents[A]. In: Proceedings of the 25 th Annual InternationalACM SIGIR Conference on Research and Development in Information Retrieval[C] , Tampere, Finland, 2002: 439 - 440

[5] ZHENG Jia-heng, LU Jiaoli .Study of An Improved Keywords Distillation Method[J].Computer Engineering ,2005，31(18)：194－196 (in Chinese)

[6] LIU Jia-bin,CHEN Chao,SHAO Zheng-rong et al. Automatic extraction of key phrases from scientific articles based on machine learning method[J]. Computer Engineering and Application. 2007,43(14):170-172(in Chinese)