WUJNS
Wuhan University Journal of Natural Sciences

Vol.12 No.5 2007 917-921

Article ID: 1007-1202(2007)05-0917-05
DOI 10.1007/s11859-007-0038-4

Keyword Extraction Based on tf/idf for Chinese News Document

☐ **LI Juanzi, FAN Qi'na, ZHANG Kuo**

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract: Keyword extraction is an important research topic of information retrieval. This paper gave the specification of keywords in Chinese news documents based on analyzing linguistic characteristics of news documents and then proposed a new keyword extraction method based on tf/idf with multi-strategies. The approach selected candidate keywords of uni-, bi- and tri- grams, and then defines the features according to their morphological characters and context information. Moreover, the paper proposed several strategies to amend the incomplete words gotten from the word segmentation and found unknown potential keywords in news documents. Experimental results show that our proposed method can significantly outperform the baseline method. We also applied it to retrospective event detection. Experimental results show that the accuracy and efficiency of news retrospective event detection can be significantly improved.

Key words: keyword extraction; keyphrase extraction; news keyword

CLC number: TP 391.1

Received date: 2007-02-12

Foundation item: Supported by the National Natural Science Foundation of China (90604025)

Biography: LI Juanzi (1964-), female, Associate professor, research direction: data mining, semantic Web. E-mail: ljz@keg.cs.tsinghua.edu.cn

0 Introduction

Keywords of a document are usually several words or phrases that best related to the content of the document. Keywords provide rich semantic information for many text mining applications, for example, document classification, clustering, retrieval, analysis and topic search[1].

Much research have done for the keyword extracting and a number of methods have been proposed, for example, keyword extraction using Naïve Bayes, decision trees, genetic algorithm, SVM (support vector machines) and some string-frequency methods. These methods have achieved satisfying results and have been used widely.

However, there are still some challenges in keyword extraction, especially in Chinese news documents. They are: ① Dependence on word segmentation**.** For Chinese documents, words need to be tokenized by a word segmentation tool and dictionary before the extraction, the quality of word tokenization will affect the keyword extraction result. ② Unknown keywords finding**.** Some new, popular words/phrases that are not contained in the dictionary may not be processed correctly. It may result in the missing of many important keywords. ③ Lack of a standard keyword annotated corpus**.** It is necessary to have a standard corpus with keywords annotated correctly. In Chinese this kind of corpus is unavailable, so machine learning based methods lack of the standard training data. It is difficult to compare these different methods[2,3].

In this paper, we propose multistrategies to deal with these problems. We give the specification of key-

words for Chinese news documents and propose a new keyword extraction method based on tf/idf with multi-strategies. Our experimental results show that the method can significantly outperform the baseline method and also can improve the accuracy and efficiency when it is applied in news retrospective event detection.

1 Related Work

Keywords are widely used as a brief summary and index of documents. Keyword extraction is the task of selecting a small set of words/ phrases from a document that can describe the meaning of the document[4]. Keyword assignment is aimed to assign keywords from a predefined controlled-vocabulary to documents. Index term extraction[5] or keyword indexing[6] is an intermediate between extraction and assignment. In that approach, some keywords are assigned from the predefined tree-structure controlled vocabulary while others are extracted from the text.

We see that the goals of these approaches are different, but some methods are applicable to all these proposed. Supervised machine learning methods and string-frequency method[6,7] are frequently used in these extraction tasks.

Naïve Bayes[5], decision trees[8] and SVM[1] are representative methods in keyword extraction. They made use of machine learning method to learn classifiers from a set of training documents. They achieved satisfying accuracy and have excellent stability, but require a large corpus in which each document is annotated with keywords in advance. It involves great skill and exhaustive labor. Moreover, the unknown keywords are difficult to extract because the model have not seen the words before. tf/idf is a typical method of string-frequency keyword extraction. It is the easiest method to extract keyword, simply and quickly. This method only scans the corpus once, and got the document frequency of all the words. However, there are many factors to consider a keyword except its frequency, such as its part of speech (POS), position in the document and its morphology. In order to improve its performance, people must design many rules and features applied on it.

Although Naïve Bayes, decision trees and SVM work well by using machine learning, string-frequency method also achieved satisfying results by adopting some features and strategies. In news domain, we find that string-frequency method was more suitable than machine learning methods. With the following: ① Machine learning methods cost a lot of time for model training and need a large amount of training documents with keywords. The news contents are manifold, plentiful keywords must be extracted manually by human for data training. ② New events happen every day and new documents about these events contain the newest and hottest unknown keywords which are difficult to extract because the model have not seen the words before. ③ News has clear theme and rich information. The important keywords usually occur many times in a news document.

2 Our Approach

2.1 Definition of Keywords

We know that the news event is usually described by 5W and 1H, they are "When, What, Who, Where, Why and How". The key words of news should be as much concerned to 5W1H as possible. Table 1 shows their POS by observing the documents and addressed the possible POS of them according to 5W1H.

There are four main kinds of POS in Table 1, they are basic noun, verb phrases, noun phrases and modifier. In addition, we think that keywords in news should satisfy following conditions: ① they describe the documents, especially the news events; ② they occur with unexpected frequency in a text; ③ they can be used to find useful results in internet searches; ④ they can be comprehended by human easily. In order to reduce plausible meanings of keywords, we think that single Chinese character can not be a keyword, except some proper noun such as chemical element, general designation of something and other proper names. Moreover, people usually are not interesting in common date except holiday or festival, so temporal noun excludes the particular date unless the text really focuses on it.

Table 1 Possible POS of 5W and 1H

5W1H	Possible POS	Example
Who	Person name(nr)	张晓梅(ZHANG Xiaomei)
Where	Organization name(nt), Place name(ns)	上海(Shanghai), 清华大学(Tsinghua University)
When	Temporal noun(*t*)	中秋节(mid-autumn festival)
What	Basic noun, Noun phrase, Verb phrases	硬币(coin), 蝴蝶效应(butterfly effect), 挪用公款(Embezzlement)
Why	Noun phrase, Verb phrase	讨厌学习(disgusted with study)
How	Modifier	轻松(easy), 紧张(tensional)

2.2 Generation of Keyword Candidates

2.2.1 *N*-gram word from word segmentation

For a news document, we conduct the sentence split, word segmentation, stop-word filtering and POS tagging using ICTCLAS[2,3]. After that, we employ bigram and trigram to create candidate words/phrases and then filter the phrases whose frequencies are below a predefined threshold (we set it to 2). We obtain a set of keyword candidates.

2.2.2 New words finding

News documents always include some new, popular and unknown words/phrases that are not contained in the dictionary, which can not be extracted by word segmentation tool. We propose a new word finding strategy to extract these potential keywords. We find that most of potential keywords often occur in the news' title. There are usually two kinds of potential keywords. The first type is the words/phrases in the quotation, for example 《星球大战》 (Star War). They are always split into several parts by the word segmentation tool. This kind of potential keywords can be extracted from the title easily. The second type is always new abbreviation or proper name. In order to estimate whether a string is a word or not, four measures are used:

① Max-duplicated: Let S be a string in the news. $S \subset S'$ means that string S' has a sub-string S. max-duplicated string is the string S, whose frequency at least 1 and the frequency of S' of all $S' \subset S$ is not bigger than frequency of S[9].

② Completeness: Suppose string S occurs in k distinct positions $p_1, p_2, \cdots, p_k$ in text T, S is "complete" if and only if the (p_i-1)th token in T is different with the (p_j-1)th token for at least one (i, j) pair, $1 \leqslant i < j \leqslant k$, and the $(p_i+|S|)$th token is different with the $(p_j+|S|)$th token for at least one (i, j) pair, $1 \leqslant i < j \leqslant k$ [10].

③ Stability: Suppose $S =$ "$c_1c_2\cdots c_p$", the stability of S is $\mathrm{MI}(S) = \dfrac{f(S)}{f(S_L)+f(S_R)-f(S)}$, where $S_L =$ "$c_1c_2\cdots c_{p-1}$", $S_R =$ "$c_2\cdots c_p$", and $f(S), f(S_L), f(S_R)$ are frequencies of S, S_L, S_R.

④ Relative frequency: The relative frequency of string S is $\dfrac{f(s)}{\sum f(w)}$, where w is the word extracted.

Max-duplicated and completeness are true/false values while stability and relative frequency are numerical value. We define some conditions that the potential keywords must satisfy: ① It is not a single character words; ② It is max-duplicated and completeness; ③ The stability of word must be at least t. For the word length=2, t = 0.38; For the word length=3, t = 0.67; For other, t = 0.8; ④ TF must be at least 3, relative frequency is bigger than 0.021.

The rules above are four basic conditions. Then, we removed some improper strings, such as the strings that are contained by other keyword candidates. In addition, we set some flexible rules to limit potential keywords according to different case. For example, generally the length of keyword is not smaller than 8.

2.3 Features Calculation and Combination

For every keyword candidate w_i, we define $f(w_i)=f$ (w.tf, w.ctf, w.inTitle, w.quo, w.inFirst, w.sign, w.POS, termSum) based on four factors: frequency, position, morphology and POS. Table 2 shows the features and how they are evaluated.

For each candidate keyword, we calculate its score.

$$\mathrm{score}(w) = (w.\mathrm{tf})^{t_1} \times (1 + \sum_{f_i \in F} w.\mathrm{f}_i \times t_{f_i}) \times \ln \frac{\mathrm{termSum}^{t_2}}{w.\mathrm{ctf}^{t_3}} \quad (1)$$

where F = {inTitle, quo, inFirst, sign} is a set of features values,and t_{f_i} are their weights. t_1,t_2 and t_3 are the weights of w.tf, termSum and w.ctf. We set t_1, t_2, t_3, t_{inTitle}, t_{quo}, t_{inFirst}, t_{sign} to 0.99, 1.0, 0.95, 2.3, 2.3, 0.01, 0.85 respectively.

Table 2 Features used in our propose

Item	Features	Comment
Frquency	tf (w.tf)	Term frequency in the news document. Important words/phrases usually have high frequencies in a news document
	Corpus tf (w.ctf)	The term frequency in a large corpus. It indicates how common a word/phrase is
	termSum	The sum of frequency of all uni-gram/bi-gram/ tri-gram (depend on what gram the word w is) words in the corpus
Position	w.inTitle w.inFirst	Whether or not w occurs in the news title and the first paragraph of the text respectively. The words occurring in this position have high possibilities to be keywords. The values are 0/1
POS	w.POS	The part of speech of w
	Quotation (w.quo)	Whether w is bracketed by a pair of specific punctuations such as '《》' and '“ ”'. Its value is 0/1
Morphology	Length (w.sign)	Long words usually contain more concrete information than short words. We adopt the feature *w.sign* for each candidate word according to the word length. It can be estimated as $Se(\mathrm{word}) = g(x)$, where $g(x)$ is a heuristic function about the word length, $g(1)=0$, $g(x)=\log_2 x$ when $2 \leqslant x \leqslant 8$ and $g(x)=3$ when $x > 8$ [10]

2.4 Refinement

After score calculation, every keyword candidate gets a score. In Refinement, we will deal with the problems about overlapped, incomplete words and weak features' word etc in top 5 keyword candidates according to the score. The words in bi-gram and tri-gram, which are composed of two or three single words, we design a measure of stability to show how compact they are. It is defined as:

$$\text{stability}(w) = \text{MI}(S) = \frac{f(S)}{f(S_L) + f(S_R) - f(S)}$$

Suppose "w_1+w_2" and "$w_1+w_2+w_3$" are the patterns of words in bi-gram and tri-gram. For bi-gram $S_L=w_1$, $S_R=w_2$ and for tri-gram, $S_L=w_1+w_2$, $S_R=w_2+w_3$.

We remove the words whose length is more than 5 meanwhile their stability is under 0.8. The word whose stability is under 0.5 meanwhile it had few or no feature in the news is also removed. Then, we check if there is any inclusive relationship between two or more words.

In refinement, we keep checking top 5 keyword candidates, remove the unsuitable words and continue check the new top 5 candidates, until there is no more candidate or no problem with top 5. Finally, we get the keywords as final result.

3 Experimental Results

3.1 Data Set and Evaluation Measure

We collected 27 612 news documents from Xinhua news agency as our corpus that covers different domains. We removed some improper documents such as the documents with no title and news catalog from the corpus. We got 19 947 valid news documents as our data set for keyword extraction, named Test set I. Among 19 947 news, we random selected 400 documents as Test set II and extracted their keywords manually. Most of news documents have 5 manual keywords and there are at least 3 manual keywords in a news document. In the experiments, we evaluate our method in terms of precision (p), recall (r) and F1-measure ($2p\times r/(p+r)$).

3.2 Keywords Distribution

We got the keywords distribution by running the proposed keyword extraction algorithm in Test set I. Experimental result is shown in Table 3. We found that though more than 75% keywords are from uni-gram, which bi-gram and tri-gram also contribute 18.59% keywords. And about 6% keywords are extracted by new word finding procedure. It shows that our strategy of including bi-gram and tri-gram keywords/keyphrases and New Word finding is reasonable.

Table 3 Keywords distribution in test set I

Item	Uni-gram	Bi-gram	Tri-gram	In quotation	New word	Total
Number of keyword	72 601	16 372	1 531	1 069	4 700	96 273
Percent of total keyword/%	75.41	17.00	1.59	1.11	4.88	100.00
Keywords per Document	3.639 7	0.820 8	0.076 8	0.053 6	0.235 6	4.826 4

3.3 Experiments on Performance

We use Eq.(2) as the baseline to extract keywords. Then, we extended the baseline with our proposed four strategies including feature calculation, considering bi-gram and tri-gram, New word finding and refinement. We performed a experiments to show how our approach improved the performance of baseline method. The experiments were run in Test set II and the results are in Table 4.

$$\text{baseline}(w) = (w.\text{tf}) \times \ln \frac{\text{termSum}}{w.\text{ctf}} \quad (2)$$

In Table 4, each block represents the result of each method. In each block, we show the precision and recall at different number (1 to 5 from top to down) of keywords we extracted and F1-measure when the number of keywords is five. From Table 4, our method got the greatest improvement over baseline method from 47.17% and 49.19% to 74.16% and 74.19% with the increase of 26.99% and 25%.

Table 4 The performance of baseline and our method %

Method	Precision	Recall	F1-measure
Baseline	59.00	11.81	—
	57.63	23.09	—
	55.42	33.30	—
	52.08	41.68	—
	49.17	49.19	49.18
Our method	88.50	17.74	—
	84.63	33.90	—
	81.92	49.21	—
	78.45	62.84	—
	74.16	74.19	74.18

3.4 Experiments on NRED

We applied our keyword extraction method to news retrospective event detection (NRED). The task is defined as the discovery of previously unidentified events in historical news corpus. We use keywords extracted to generate the word vector to represent the news document. Experiment of NRED is implemented on Test set I. We first picked some representative words and found the

related documents using our news retrieval system. Then the result of NRED was obtained by a clustering algorithm. These documents have been classified according to the events by experts in news domain in advance.

We used "农业"(agriculture) and "住房"(housing) as query keywords. Figures 1, 2 are their experimental results. We can see that by using the proposed keyword extraction, a significant improvement can be obtained on the task. The results indicate that our proposed keyword extraction is effective in NRED.

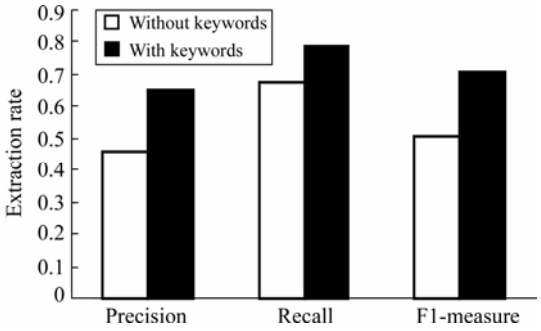

Fig.1 The results of RED for query keyword "NongYe"(agriculture)

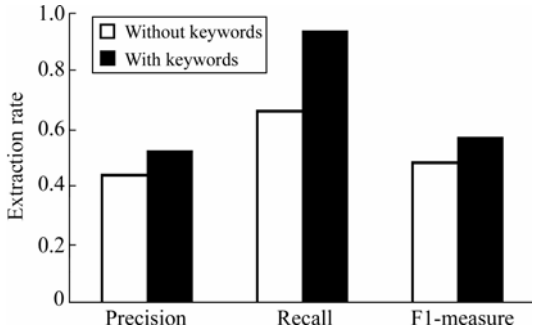

Fig.2 The results of RED for query keyword "ZhuFang"(housing)

4 Conclusion

We have investigated the problem of news keyword extraction in Chinese. We first give a specification to define the keywords/keyphrases in news document according to the news characteristics in both linguistic and semantic level. Then, we propose an approach to keyword extraction methods form Chinese news documents. Experimental results show that our approach can significantly outperform baseline method for keyword extraction. When applying it to NRED, we observed a significant improvement on the clustering accuracy. As the future work, we plan to make further improvement on the accuracy, especially reduce the missing errors caused by intransitive verb and hottest words. We also want to apply the keyword extraction method to other text mining applications.

References

[1] Zhang Kuo, Xu Hui, Tang Jie, *et al*. Keyword Extraction Using Support Vector Machine[C]//*Proceedings of WAIM*' 2006. Hong Kong: Springer-Verlag, 2006:85(Ch).

[2] Zhang Kevin. Online Document about ICTCLAS [EB/OL]. [2002-08-16]. *http://www.nlp.org.cn/project/project.php*? *proj_id* =6.

[3] Liu Qun, Zhang Huaping, YU Hongkui,*et al*. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model[J]. *Chinese Journal of Computer Research and Development*, 2004, **41**(8):1421-1429(Ch).

[4] Anette H, Karlgren J, Jonsson A, *et al*. Automatic Keyword Extraction Using Domain Knowledge[C]//*Proceedings of Second International Conference on Computational Linguistics and Intelligent Text Processing*. Mexico City: Springer-Verlag, 2001:472.

[5] Medelyan O, Witten A I H. Thesaurus-Based Index Term Extraction for Agricultural Documents[C]//*Proc of the* 6*th Agricultural Ontology Service* (*AOS*) *Workshop at EFITA/WCCA*. Vila Real: IEEE Press, 2005:1122.

[6] Li Sujian, Wang Houfeng, Yu Shiwen, *et al*. News-Oriented Automatic Chinese Keyword Indexing[C]//*Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Sapporo: IEEE Press, 2003: 92.

[7] Wang Houfeng, Li Sujian, Yu Shiwen, *et al*. A Combining Approach to Automatic Keyphrases Indexing for Chinese News Documents[C]//*Computational Linguistics and Intelligent Text Processing* (*CICLing*-2004).New York: Springer-Verlag, 2004:435.

[8] Peter T. Learning to Extract Keyphrases from Text [R]. Ottawa: National Research Council, 1999.

[9] Yang Wenfeng, Li Xing. Chinese Keyword Extraction Based on Max-Duplicated Strings of the Documents[C]// *Proceedings of the* 25*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tampere: ACM Press, 2002: 439.

[10] Zhang De, Dong Yisheng. Semantic, Hierarchical, Online Clustering of Web Search Results[C]// *Proceedings of the* 6*th Asia Pacific Web Conference* (*APWEB*). Hangzhou: Springer-Verlag, 2004:69(Ch).

□