Journal of Computational Information Systems4:3(2008) 1169-1180
Available at http://www.JofCI.org

Automatic Keyword Extraction from Documents Using Conditional Random Fields

Chengzhi ZHANG [1,2,†], Huilin WANG[1], Yao LIU[1], Dan WU[1,3], Yi LIAO[4], Bo WANG[5]

[1] *Institute of Scientific & Technical Information of China, Beijing 100038, China*

[2] *Department of Information Management, Nanjing University of Science & Technology, Nanjing 210093, China*

[3] *Department of Information Management, Peking University, Beijing 100871, China*

[4] *Department of Management, Lingnan University, Hong Kong*

[5] *Department of Computer Science and Technolog, Peking University, Beijing, 100871, China*

Abstract

Keywords are subset of words or phrases from a document that can describe the meaning of the document. Many text mining applications can take advantage from it. Unfortunately, a large portion of documents still do not have keywords assigned. On the other hand, manual assignment of high quality keywords is expensive, time-consuming, and error prone. Therefore, most algorithms and systems aimed to help people perform automatic keywords extraction have been proposed. Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labeling. In this paper, keywords extraction based on CRF is proposed and implemented. As far as we know, using CRF model in keyword extraction has not been investigated previously. Experimental results show that the CRF model outperforms other machine learning methods such as support vector machine, multiple linear regression model etc. in the task of keywords extraction.

Keywords: Keywords Extraction; Conditional Random Fields; Automatic Indexing; Machine Learning

1. Introduction

Automatic keyword extraction (AKE) is the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document [1]. Since keyword is the smallest unit which can express the meaning of document, many text mining applications can take advantage of it, e.g. automatic indexing, automatic summarization, automatic classification, automatic clustering, automatic filtering, topic detection and tracking, information visualization, etc. Therefore, keywords extraction can be considered as the core technology of all automatic processing for documents.

However, a large number of documents do not have keywords. At the same time, manual assignment of

† Corresponding author.

Email addresses: zhangchz@istic.ac.cn (Chengzhi ZHANG), wanghl@istic.ac.cn (Huilin WANG), liuyao@istic.ac.cn (Yao LIU), woodan@pku.edu.cn (Dan WU), yliao@ln.edu.hk (Yi LIAO), wangbo@pku.edu.cn, (Bo WANG).

March, 2008

high quality keywords is costly and time-consuming, and error prone. Therefore, most algorithms and systems to help people perform automatic keywords extraction have been proposed.

Existing methods on keyword extraction can not use most of the features in the document. Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, and it can utilize most of the features for efficient keyword extraction more sufficiently and effectively. At the same time, keyword extraction can be considered as the string labeling. In this paper, keywords extraction based on CRF is proposed and implemented. In the research community, no previous study has investigated similar method, to the best of our knowledge. Experimental results indicate that the CRF model can enhance keyword extraction, and it outperforms the other machine learning methods such as support vector machine, multiple linear regression model etc.

The rest of this paper is organized as follows. The next section reviews some related work on keyword extraction. In section 3, a detailed description of the proposed approach is presented. Subsequently in section 4, the authors report experiments results that evaluate the proposed approach. The paper is concluded with summary and future work directions.

2. Related Work

There are two existing approaches to automatic keyword indexing: keyword extraction and keyword assignment. In keyword extraction, words occurred in the document are analyzed to identify apparently significant ones, on the basis of properties such as frequency and length. In keyword assignment, keywords are chosen from a controlled vocabulary of terms, and documents are classified according to their content into classes that correspond to elements of the vocabulary [2]. Existing methods about automatic keyword extraction can be divided into four categories, i.e. simple statistics, linguistics, machine learning and other approaches.

2.1. Simple Statistics Approaches

These methods are simple and do not need the training data. The statistics information of the words can be used to identify the keywords in the document. Cohen uses N-Gram statistical information to automatic index the document [3]. N-Gram is language and domain-independent. Other statistics methods include word frequency [4], TF*IDF [5], word co-occurrences [6], and PAT-tree [7], etc.

2.2. Linguistics Approaches

These approaches use the linguistics feature of the words mainly, sentences and document. The linguistics approach includes the lexical analysis [8], syntactic analysis [1], discourse analysis [9] [10] and so on.

2.3. Machine Learning Approaches

Keyword extraction can be seen as supervised learning from the examples. Machine learning approach employs the extracted keywords from training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes [11], SVM [12], Bagging [1], etc. Some keyword extraction tools, e.g. KEA [13], GenEx [14], have been developed.

2.4. Other Approaches

Other approaches about keyword extraction mainly combine the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of the words, html tags around of the words, etc [15].

3. Keyword Extraction Based on CRF

3.1. Why Do We Use CRF?

3.1.1 Introduction to CRF

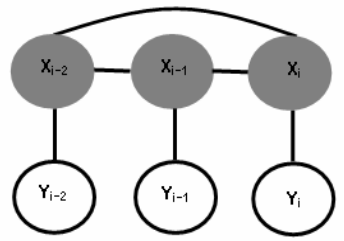

Fig.1 Graphical Structure of a Chain-structured CRF for Sequences

Conditional Random Fields (CRF) model is a new probabilistic model for segmenting and labeling sequence data [16]. CRF is an undirected graphical model that encodes a conditional probability distribution with a given set of features. Figure 1 shows the graphical structure of a chain-structured CRF.

For the given observation sequential data X(X1X2…Xn), and their corresponding status labels Y(Y1Y2…Yn), a linear chain structure CRF defines the conditional probability as follows:

$$P(Y|X)=\frac{1}{Z_X}\exp\left(\sum_i\sum_j \lambda_j f_j(y_{i-1},y_i,X,i)\right) \quad (1)$$

Where, ZX is a normalization and it makes the probability of all state sequences sum to 1. $f_j(y_{i-1},y_i,X,i)$ is a feature function, and λ_j is a learnt weight associated with feature f_j.

Maximum entropy learning algorithm can be used to train CRF. For the given observation sequential data, the most probable label sequence can be determined by

$$Y^*=\arg\max_y P(Y|X) \quad (2)$$

Y* can be efficiently determined using the Viterbi algorithm. An N-best list of labeling sequences can also be obtained by using modified Viterbi algorithm and A* search [17].

The main advantage of CRF comes from that it can relax the assumption of conditional independence of the observed data often used in generative approaches, an assumption that might be too restrictive for a considerable number of object classes. Additionally, CRF avoids the label bias problem.

3.1.2. Keyword Extraction is a Typical Labeling Problem

In process of manual assignment keyword to a document, the content of the document will be analyzed and comprehended firstly. Keywords which can express the meaning of document are then determined. Content analysis is the process that most of the units of a document, such as the title, abstract, full text, references and so on, be analyzed and comprehended.

Usually, we can extract 3~5 keywords from a document in process of manual assignment keyword. These keywords may be in the title, abstract, first section or headings of the document, first or last sentence of paragraph, etc. Sometimes, we may read entire document, then summarize the content of the document, and give the keyword finally.

According to the process of manual assignment keyword to a document, we can transfer this process to labeling task of the text sequences. In other words, we can annotate a word or phrase with a label by a large number of features of them. Therefore, keyword extraction algorithm Based on CRF is proposed and implemented in this paper. We use CRF++ tool [18] to extract keywords.

3.2. Keyword Extraction using CRF

3.2.1 Features in the CRF Model

Because we extract the keywords from the Chinese documents, we must segment the sentence into word and tag the POS of the word. We use SegTag tool [19], which is available at http://www.nlp.org.cn. For automatically processing the labels by computers, the manually labeled data should be formatted as follows:

The sentence '*贸易投资/n 一体化/vn 与/c 就业增长/n ——/w 以/p 江苏省/ns 为/p 案例/n 的/uj 实证分析/n*' is formatted to '*贸易投资/KW_B 一体化/KW_I 与/KW_S 就业增长/KW_N ——/KW_S 以/KW_S 江苏省/KW_N 为/KW_S 案例/KW_N 的/KW_S 实证分析/KW_Y*', in which 'KW_B' represents this word is at the beginning of a keyword, 'KW_I' means this word is one part, but not at the begging of a keyword,, 'KW_S' means this word is not a word in the StopList, 'KW_N' represents this word is neither a keyword nor a word in the StopList, and 'KW_Y' means this word is a keyword.

After the tagging we can find that keyword extraction is a typical labeling problem. We obtain the keyword from the tagging results using CRF model. To utilize the flexibility of CRF and considering the keyword extraction problem, we use the features in table 1.

In table 1, there are three kinds of features, i.e. (1) local context features: Word-2, Word-1, Word, Word+1, Word+2, Len, POS, t, a, c, TF*IDF, DEP, (2): global context features: T, A, H, F, L, R, (3) hybrid features: Word-2 Word-1, Word-1 Word, WordWord+1, Word+1 Word+2. According to the features, we can process the documents collection and transfer these documents into training data for CRF model. Table 2 shows a sample of training data from one of documents for CRF model.

Table.1 Features in the CRF Model

No.	Features	Explanations	Normalization Method
1	Word	current word	-
2	Len	length of the word	$\frac{Len(Word)}{Max\_Len}$

3	POS	Part-Of-Speech of word or phrase，if one of word in a phrase is n, then POS=1, otherwise, POS=0.	{0，1}
4	t	whether the word is in the title	{0，1}
5	a	whether the word is in the abstract	{0，1}
6	c	whether the word is in the full-text	{0，1}
7	TF*IDF	Term Frequency * Inverse Document Frequency of the word	$\frac{Freq(Word)}{Max\_Freq} \times \log_2 \frac{N+1}{n+1}$
8	DEP	the position of the first appearance of the word	$\#(Word) / \sum word_i$
9	T	whether the word has appeared in the title	{0，1}
10	A	whether the word has appeared in the abstract	{0，1}
11	H	whether the word has appeared in the heading	{0，1}
12	F	whether the word has appeared in the first paragraph	{0，1}
13	L	whether the word has appeared in the last paragraph	{0，1}
14	R	whether the word has appeared in the references	{0，1}
15	$Word_{-2}$	second previous word	-
16	$Word_{-1}$	previous word	-
17	$Word_{+1}$	next word	-
18	$Word_{+2}$	second next word	-
19	$Word_{-2}$ $Word_{-1}$	second previous word and previous word	-
20	$Word_{-1}$ Word	previous word and current word	-
21	$WordWord_{+1}$	Current word and next word	-
22	$Word_{+1}$ $Word_{+2}$	next word and second next word	-

Table. 2 Sample of Training Data for CRF Model

Word	**POS**	**t**	**a**	**c**	**TF*IDF**	**Len**	**DEP**	**T**	**A**	**H**	**F**	**L**	**R**	**Lable**
贸易投资	1	1	0	0	0.0915	0.5714	0.0387	1	1	1	1	1	1	KW_B
一体化	1	1	0	0	0.0541	0.4286	0.0548	1	1	1	1	1	1	KW_I
与	0	1	0	0	0.0002	0.1429	0.0671	1	0	1	1	0	1	KW_S
就业增长[*]	1	1	0	0	0.0265	0.5714	0.0775	1	0	1	0	0	0	KW_N
——	0	1	0	0	0.0022	0.2857	0.0866	1	0	0	0	0	0	KW_S
以	0	1	0	0	0.0006	0.1429	0.0949	1	0	0	0	0	0	KW_S
江苏省	1	1	0	0	0.0325	0.4286	0.1025	1	1	1	0	0	1	KW_N
为	0	1	0	0	0.0001	0.1429	0.1096	1	1	0	0	1	0	KW_S
案例	1	1	0	0	0.0077	0.2857	0.1162	1	0	0	0	0	0	KW_N
的	0	1	0	0	0.0000	0.1429	0.1225	1	1	1	1	1	1	KW_S
实证分析	1	1	0	0	0.0128	0.5714	0.1285	1	1	1	0	0	1	KW_Y

3.2.2. Process of the CRF-based Keyword Extratcion

Figure 2 shows the process of the CRF-based keyword extraction. The implementation carries out keyword

[*] We collect many domain-specific words like '*就业增长*' and annotate their POS, then add these information into the lexicon dictionary of SegTag tool.

extraction in the following steps.

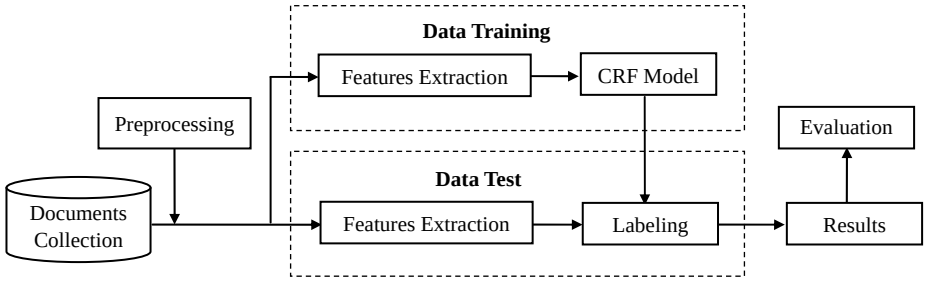

Fig. 2 Process of the CRF-based Keyword Extraction

(1) Preprocessing and features extraction

The input is a document. Before CRF model training, we must transfer the document into the tagging sequences, i.e a bag of words or phrases of the document. For a new document, we conduct the sentence segment, POS tagging. Then, the feautures mentioned above are automatic extracted. The output is the feautures vectors, and each vector corresponds to a word or phrase.

(2) CRF model training

The input is a set of feature vectors by step above. We train a CRF model that can label the keyword type. In the CRF model, a word or phrase could be regarded as an example, and the keyword is annotated by one kind of labels, such as 'KW_B', 'KW_I', 'KW_S', 'KW_N', and 'KW_Y'. The tagged data are used to training the CRF model in advance. In the CRF++, the output is a CRF model file.

(3) CRF labeling and keyword extraction

The input is a new document. The document is preprocessed and its feautures are extracted. Then, we predict the keyword type by using the CRF model. According to the keyword type, the keywords of the document are extrated. For example, '*实证分析/KW_Y*'-> keyword: *实证分析*, '*贸易投资/KW_B 一体化 KW_I*' -> keyword*: 贸易投资一体化.*

(4) Results evaluation

We can evalutate the results of keyword extraction by comparing these results with the manual assignment results. A detailed evalulation mehtod is presented in following section.

4. Experimental Results

4.1. Data Sets

In this study, we collect documents from database of 'Information Center for Social Sciences of RUC', which is available at http://art.zlzx.org/. We randomly chose 600 academic documents in the field of economics from the database. These Chinese documents are divided into 10 data sets and used 10-fold cross-validation for the CRF model. Each document includes the title, abstract, keywords, full-text, heading of paragraph or sections, boundaries information of paragraphs or sections, references, etc. These documents have abundant rich linguistics features and are suitable to perform keywords labeling well. Therefore, this is a very interesting work of keywords extraction from documents using CRF model. The number of the annotated keywords of 600 documents ranges from 5 to 10 and the average of annotated keywords is 7.83 per document.

4.2. Evaluation Measures

Table. 3 Contingence table on Results of Extraction and Manual Assignment

	Keywords assigned by humans	**Non-keywords assigned by humans**
Keywords extracted by system	a	b
Non-keywords extracted by system	c	d

In the evaluation, there are two types of words or phrases in manual assignment of keywords, which are keywords and non-keywords assigned by humans. On the other hand, there are two types of words or phrases in automatic keyword extraction, i.e. keywords and non-keywords extracted by keyword extraction system. Table 3 shows the contingence table on the result of keywords extraction and manual assignment keywords.

From all experiments on keyword extraction, we conducted evaluations according to the general measuring method used in the Information retrieval evaluation, i.e. precision (P), recall (R) and F1-Measure. The evaluation measures are defined as follows:

$$P = \frac{a}{a+b} \tag{3}$$

$$R = \frac{a}{a+c} \tag{4}$$

$$F_1(P,R) = \frac{2PR}{P+R} \tag{5}$$

Where, a, b, c and d denote number of instances. In this paper, we get the evaluation results by using 10-fold cross-validation.

4.3. Other Keyword Extraction Approaches

Automatic keyword extraction can be viewed as a classification problem. In this study, we use some other approaches as the baseline to extract keyword. We carried out the comparison of CRF-based method and these approaches. These approaches include support vector machines (SVM) [20][21][12], multiple linear regression (denoted as MLR) [21], logistic regression (denoted as Logit）[22][21], BasaLine1, BaseLine2.

We give two heuristic baseline approaches to extract keyword, namely, BaseLine1 and BaseLine2. We use TF*IDF and Len as the features of a document in the BaseLine1. The score of word or phrase is defined as follows:

$$Score = TF*IDF * Len \tag{6}$$

Because the average number of keywords annotated manually is six, six words or phrases with the higher score are selected as the keywords of the document. TF*IDF, Len and DEP are used as the features of a document in the BaseLine2. The score of word or phrase is defined as follows:

$$Score = TF*IDF * Len* DEP \tag{7}$$

We select eight words or phrases with the higher score as the keywords of the document.

4.4. Experimental Results and Discussions

4.4.1. Performance Evaluation of Six Models

We evaluate the performance of the six keyword extraction models, i.e. CRF, SVM, MLR, Logit, BaseLine1 and BaseLine2, by using the 10-fold cross-validation.

Table. 4 Performance Evaluation of Keyword Extraction

Model	P	R	F_1
BaseLine1	0.2343	0.4508	0.3083
BaseLine2	0.2778	0.5287	0.3656
MLR	0.3174	0.5233	0.3951
Logit	0.3248	0.5388	0.4067
SVM	0.8017	0.3327	0.4653
CRF	0.6637	0.4196	**0.5125**

Table 4 shows the 10-fold cross-validation results of six keyword extraction models. According to F1-Measure in table 4, we can see that CRF-based approach outperforms the other five models, and the result of the F1-Measure comparison is: CRF > Logit > MLR > Baseline2 > Baselin1. According to the precision in table 4, we know that SVM and CRF model significantly outperform the other four models, and the result of the precision comparison is: SVM > CRF > Logit > MLR > BaseLine2 > BaseLine1. At the same time, we also can see that the result of recall comparison is : Logit > BaseLine2 > MLR > BaseLine1 > CRF > SVM.

According to the precision in table 4, we know that Logit model outperforms MLR model. It shows that logistic regression model outperforms multiple linear regression model in the task of keyword extraction which can be viewed as a typical binary classification problem. We can also know that BaseLine2 model is about 4.35 percentage points higher than BaseLine1. This shows that DEP is a useful feature in the task of keyword extraction.

It is noteworthy that keywords of manual assignment do not appear in the document sometimes. Therefore, it is very necessary to automatic assign keywords for document [2].

4.4.2. Lexicon Dictionary Size Influence on Keyword Extraction

Word segment result has great influence on precision of Chinese keyword extraction model. The lexicon dictionary is required when we use SegTag tool. Yang & Li's experiment shows that precision increased from 50.7% to 59.3% by using a lexicon dictionary including 50, 000 words [23]. In this section, we focus on the lexicon dictionary size impacting on keyword extraction. According to the size of lexicon dictionary, the dictionary can be divided into three types, i.e. full, general, general and domain-specific. The general dictionary includes most of common words and phrases. The full dictionary combines the general dictionary with all keywords from the annotated keywords in the training documents. The general and domain-specific dictionary combines the general dictionary with a large number of domain-specific words and phrases.

If the lexicon dictionary includes all keywords, the task of keyword extraction task will be to identify these words or phrases after preprocessing and features extraction. Therefore, we can efficiently evaluate

the performance of keyword extraction model in this ideal circumstance.

We use three kinds of lexicon dictionary mentioned above to segment documents, and six models to extract keywords from documents. Table 5 shows the 10-fold cross-validation results.

Table. 5 Lexicon Dictionary Size Impact on Keyword Extraction

Model	Full			General			General and Domain-specific		
	P	R	F_1	P	R	F_1	P	R	F_1
BaseLine1	0.2343	0.4508	0.3083	0.0578	0.1111	0.0760	0.1188	0.2284	0.1563
BaseLine2	0.2778	0.5287	0.2656	0.0690	0.1327	0.0908	0.1717	0.3302	0.2260
MLR	0.3174	0.5233	0.3951	0.1105	0.1821	0.1375	0.2135	0.3519	0.2657
Logit	0.3248	0.5388	0.4067	0.1067	0.1759	0.1329	0.2116	0.3488	0.2634
SVM	0.8017	0.3327	0.4653	0.1538	0.0926	0.1156	0.5140	0.1698	0.2552
CRF	0.6637	0.4196	0.5125	0.1734	0.1564	0.1645	0.4702	0.2438	0.3209

According to table 5, we can conclude that lexicon dictionary size can significantly impact on keyword extraction. The result of the F1-Measure comparison is: Full > General and Domain-specific > General. No matter what type lexicon dictionary is selected, CRF-based approach can outperform the other five models.

4.4.3. Training Set Size Influence on Keyword Extraction

Table. 6 Training Set Size Impact on Keyword Extraction

Size	p				R				F_1			
	MLR	Logit	SVM	CRF	MLR	Logit	SVM	CRF	MLR	Logit	SVM	CRF
50	0.3056	0.3183	0.7975	0.6321	0.5123	0.5255	0.3121	0.3793	0.3828	0.3965	0.4486	0.4732
100	0.3067	0.3213	0.7933	0.6526	0.5135	0.5322	0.3236	0.4195	0.3840	0.4007	0.4597	0.5102
200	0.3092	0.3223	0.7911	0.6592	0.5170	0.5330	0.3163	0.4023	0.3870	0.4017	0.4519	0.4991
300	0.3049	0.3257	0.7949	0.6608	0.5185	0.5341	0.3278	0.3908	0.3840	0.4046	0.4642	0.4902
400	0.3109	0.3207	0.7984	0.6606	0.5193	0.5359	0.3236	0.4138	0.3889	0.4013	0.4605	0.5075
500	0.3123	0.3216	0.7971	0.6624	0.5204	0.5381	0.3293	0.4023	0.3903	0.4026	0.4661	0.5098
540	0.3174	0.3248	0.8017	0.6637	0.5233	0.5388	0.3327	0.4196	0.3951	0.4067	0.4653	0.5125

In the experiment, we change the size of training set and use these training sets to train CRF model. Table 6 shows the 10-fold cross-validation results according to the training set size respectively. Because BaseLine1 and BaseLine2 model are dependent to training set size, we only give the result of the others four model in table 6. According to table 6, we can see that the performance of these four keyword extraction model is related with the training set size obviously.

With the size increasing, the effect becomes smaller. CRF-based keyword extraction model can use a small training set, e.g. 50, to extract keywords from documents effectively.

4.4.4. Error Analysis

We conducted error analysis on the results of CRF-based keyword extraction. There are two kinds of errors detailed as follows.

(1) Errors in the Training Set

In the training set, the keywords have some synonym or similar words, for example, '*牧民(herdsman)*' and '*牧户 (makido)*' are similar words. In the evaluation process, we ignored this problem. It can affect the precision of these six models.

(2) Ambiguity of the extracted keywords

Some of errors occurred due to the ambiguity of the extracted keywords. These ambiguity words have several meanings and are difficult to identify whether they are keywords or not. This problem can also affect the performance of CRF-based keyword extraction. Therefore, in the future work, we should take account of the ambiguity of the extracted keywords and adjust the result of CRF-based keyword extraction.

5. Conclusion and Future Work

Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labeling. In this paper, we have proposed and implement the CRF-based keyword extraction approach. Experimental results show that the CRF model outperforms the other machine learning methods such as support vector machine, multiple linear regression model etc. in the task of keywords extraction from academic documents.

CRF model is a promising method in labeling the sequence, and it can take full advantage of all the features of document. As future work, we plan to make further improvement on the precision and recall of CRF-based keyword extraction model. For example, we will use the semantic relations between the keywords. We also plan to apply the keyword extraction approach on Web pages, E-mail and others non-academic documents. Meanwhile, we will apply this method on some standard documents corpus, e.g. LDC corpus. It will also be interesting to apply the CRF-base keyword extraction model to a large number of text mining applications, such as text classification, clustering, summarization, filtering and so on.

Acknowledgements

The work described in this paper has been supported in part by supported by National Key Project of Scientific and Technical Supporting Programs funded by Ministry of Science & Technology of China (NO. 2006BAH03B02), Youth Research Support Fund funded by Nanjing University of Science & Technology (NO. JGQN0701), Project of the Education Ministry's Humanities and Social Science funded by Ministry of Education of China (06JC870001), and Innovation Project of Graduate Education of Jiangsu Province (2006) funded by Jiangsu Education Commission.

References

[1] A. Hulth. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216-223.

[2] O. Medelyan, I. H Witten. Thesaurus Based Automatic Keyphrase Indexing. In: Proceedings of the Joint Conference on Digital Libraries 2006, Chapel Hill, NC, USA, 2006: 296-297.

[3] J. D. Cohen. Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting. Journal of the American Society for Information Science, 1995, 46(3): 162-174.

[4] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[5] G. Salton, C. S. Yang, C. T. Yu. A Theory of Term Importance in Automatic Text Analysis, Journal of the

American society for Information Science, 1975, 26(1): 33-44.
[6] Y. Matsuo, M. Ishizuka. Keyword Extraction from a Single Document Using Word Co-ocuurrence Statistical Information. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.
[7] L F. Chien. PAT-tree-based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR1997), Philadelphia, PA, USA, 1997: 50-59.
[8] G. Ercan, I. Cicekli. Using Lexical Chains for Keyword Extraction. Information Processing and Management, 2007, 43(6): 1705-1714.
[9] S. F. Dennis. The Design and Testing of a Fully Automatic Indexing-searching System for Documents Consisting of Expository Text. In: G. Schecter eds. Information Retrieval: a Critical Review, Washington D. C.: Thompson Book Company, 1967: 67-94.
[10] G. Salton, C. Buckley. Automatic Text Structuring and Retrieval –Experiments in Automatic Encyclopaedia Searching. In: Proceedings of the Fourteenth SIGIR Conference, New York: ACM, 1991: 21-30.
[11] E. Frank, G. W. Paynter, I. H. Witten. Domain-Specific Keyphrase Extraction. In: Proceedings of the 16th International Joint Conference on Aritifcal Intelligence, Stockholm, Sweden, Morgan Kaufmann, 1999: 668-673.
[12] K. Zhang, H. Xu, J. Tang, J. Z. Li. Keyword Extraction Using Support Vector Machine. In: Proceedings of the Seventh International Conference on Web-Age Information Management (WAIM2006), Hong Kong, China, 2006: 85-96.
[13] I. H. Witten, G. W. Paynte, E. Frank, C. Gutwin, C. G. Nevill-Manning. KEA: Practical Automatic Keyphrase Extraction. In: Proceedings of the 4th ACM Conference on Digital Library (DL'99) , Berkeley, CA, USA, 1999: 254-26.
[14] P. D. Turney. Learning to Extract Keyphrases from Text. NRC Technical Report ERB-1057, National Research Council, Canada. 1999: 1-43.
[15] J. B. Keith Humphreys. Phraserate: An Html Keyphrase Extractor. Technical Report, University of California, Riverside, 2002: 1-16.
[16] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conference on Machine Learning (ICML01), Williamstown, MA, USA, 2001: 282-289.
[17] H. Kang, W. J. Liu. Prosodic Words Prediction from Lexicon Words with CRF and TBL Joint Method. In: Proceedings of 2006 International Symposium on Chinese Spoken Language Processing (ISCSLP-2006), Kent-Ridge, Singapore, 2006: 161-168.
[18] CRF++: Yet Another CRF toolkit. http://chasen.org/~taku/software/CRF++. Accessed: 2006.12.20.
[19] CNLP Platform. http://www.nlp.org.cn. Accessed: 2006.12. 25.
[20] V. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
[21] H. J. Zeng, Q. He, Z. Chen, W. Y. Ma, J. Ma. Learning to Cluster Web Search Results. In: Proceedings of 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR'04), Sheffield, 2004: 210-217.
[22] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. New York: Springer-Verlag, 2001.
[23] W. F. Yang, X. Li. Chinese Keyword Extraction Based on Max-dupliated Strings of the Documents. In: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR02), Tampere, Finland, 2002: 439-440.