International Journal of u- and e- Service, Science and Technology
Vol.8, No.5 (2015), pp.37-46
http://dx.doi.org/10.14257/ijunesst.2015.8.5.04

Semantic-based Keyword Extraction Method for Document

FangJiang[a,b*], GuoheLi[a,b], XueYun[a] and XiangYue[c]

a (College of Geophysics and Information Engineering, China University of Petroleum, Beijing, 102249, China)
b(Beijing Key Lab of Data Mining for Petroleum Data, China University of Petroleum, Beijing, 102249, China)
c(Information & Data Center, CNOOC Research Institute, Beijing, 102249, China)
jiangfangzhang@163.com

Abstract

Keyword extraction is one of the most important contents of information retrieval research. Document keywords extraction on the basis of semantic is an effective way to improve the accuracy of automatic extraction. This paper proposed a semantic-base keywords extraction method with the Chinese document as the processing object. First, semantic distances between words are calculated through the synonyms dictionary. Then theme related classes are obtained by density based clustering of words. Finally, the headwords are selected from topic related classes and regarded as keywords. An artificial contrast experiment, a corpus classification experiment and a scoring experiment were conducted. Results show that the proposed semantic-based keyword extraction method has higher accuracy and recall rate, and the extracted keywords are more related to the topic.

Keywords*: Semantic Distance; Density Clustering; Keyword Extraction*

1. Introduction

Keywords extraction is a research focus in the information retrieval field. At present, in the aspects of study for keywords extraction algorithms, there are heuristic rule-based methods [1,2,3], statistics-based methods[4,5]and methods based on machine learning[6], *etc*. Heuristic rule-based methods extract keywords according to characteristics of document formats and the simple structure of documents. Keywords extracted by these methods have their limits, because they cannot represent very well of the documents theme. Statistics-based methods can be implemented easily, but this method may ignore some keywords which are low-frequent or inconspicuous but essential to a document, and these methods rely on large-scale training samples. Methods based on the machine learning mostly uses algorithms such as supervised-learning[7,8], unsupervised-learning[9], mutual information algorithm[10], maximum entropy method[11]and so on. Precision and recall rates of methods based on supervised-learning have been improved a lot compared with the mentioned methods, but the method could not improve the effect of keywords extraction where training corpus and test corpus are from different fields, because the methods are limited by fields. Methods based on unsupervised-learning have low performance and low rate of accuracy. Keywords extracted by methods based on mutual information have low rate of accuracy. Selection of feature sets and estimation of feature parameters in methods based on maximum entropy are very important issues. If the feature parameters are not estimated accurately, inaccurate keywords will be extracted.

Semantics represents word meanings and relations of theses meanings. Semantic mining is a tool, automatically extracting meaning from any given body of document. This paper

ISSN: 2005-4246 IJUNESST

presents a semantics-based keywords extraction method（SKEM）.When a document is taken as the processing object, semantic distances between words are first calculated; then, the words are clustered based on density; then, theme related classes come into being; finally, the headwords selected from topic related classes are regarded as keywords. The method is not limited by fields, and does not rely on a large scale of training samples. What is more, precision and recall rates are improved.

2. Semantic-based Keyword Extraction Method

2.1 Method Introduction

Currently there have already been so many methods in extracting keyword, but there are mainly two kinds: one is to calculate word weights after segmentation by using a conventional or improved statistical method; the other is that after word segmentation, structure collinear network, semantic network and grammar network are built by using a collinear relationship between words, semantic relationship or grammar relationship. The network is mainly built by using HowNet，‘Little world’ model and HIT Synonyms Dictionary (expanded). Network based on HowNetis to build the word chain through calculating semantic similarity degree, and then choose the keywords based on the word frequency and areal features. Network structured by ‘Little world’ model is to use the word frequency and the word’s collinear probability, and then divide word’s collinear network into ‘clusters’ according to different themes, and identify fragment boundary to make ‘cluster’ correspond with the text fragment.

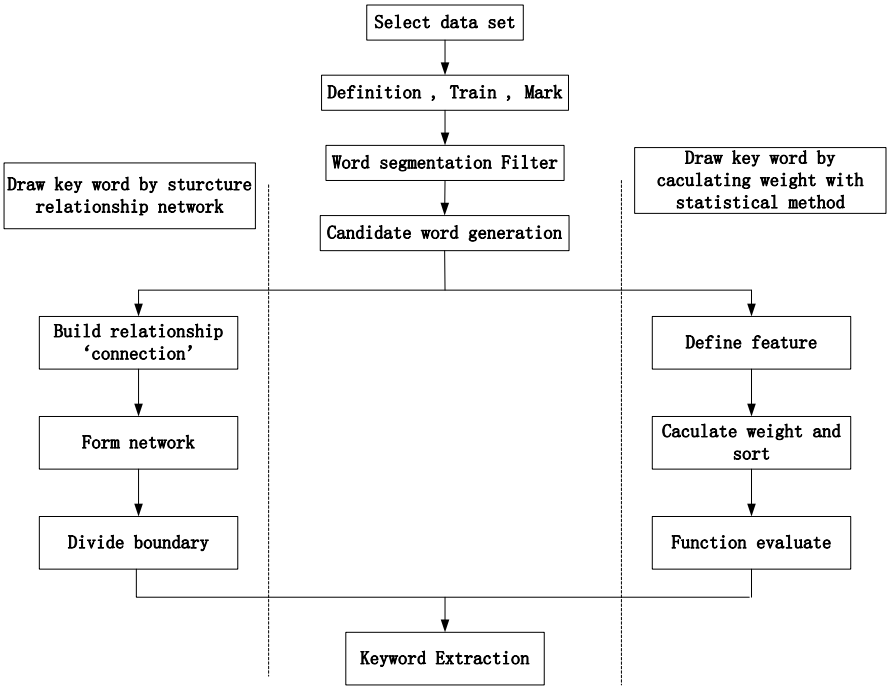

Figure 1. Existing Keyword Extraction Method

According to Figure 1, the above two keyword extraction methods have problems listed below:

1. The statistics based method depends more on corpus, while missing some low frequent words or words in unimportant places but crucial to the text.
2. Network structure method based on HowNet is built by matching synonym and

near-synonym. Due to various languages used by Chinese literature authors, when they express the keyword for the same theme, the used words are not synonym or near-synonym. It results in that the words of the same theme can not be semantically associated and thus makes semanteme cannot play the supposed effective role in extracting keywords. And also, choosing keywords by using the word frequency and areal features will easily ignore those low frequent but important keywords.

3. When 'Little world method' is used to build up network, statistical method and large corpus support is needed. It restricts comprehensiveness of network structure, and also lacks semantic comprehension, which makes that words under the same theme but in different calculating area are unable to be associated.

In order to make up all the disadvantages above, this paper brings forward a keyword extraction method based on the semanteme. It is a different network structure method from HowNet and 'Little world', which is to build network with semantic distance between words, and extracts keyword with density clustering method. The differences are:

1. This method is to measure importance of words from the view of semanteme, and is more close to people's perception logic.
2. The proposed structure network not only includes the match of synonym and near-synonym, but also uses HIT Synonyms Dictionary (expanded), which increases the accuracy of semantic calculation.
3. The proposed method does not depend on corpus.

2.2 Method Principle

Definition 1: $U(D)=\{w_i|1\le i\le n\}$ is a term set of document D, where w_i is a word in document D, and n is the number of term sets in document $D.S(D)=\{ w_t|1\le t\le m \}$ is a stop-word set of document D, where w_t is a stop word in document D, and m is the number of stop-word sets in the document $D.W(D)= U(D)-S(D)=\{ w_k|1\le k\le N \}$, where w_k is a non-stop word of document D, and N the number of non-stop words in the document D. In the following of the paper, all w_k are non-stop words. $SetW(D)=\{w_k|1\le k\le N \}$ is the wordsetw_k for document D.

2.2.1 Semantic Distance Calculation[12]

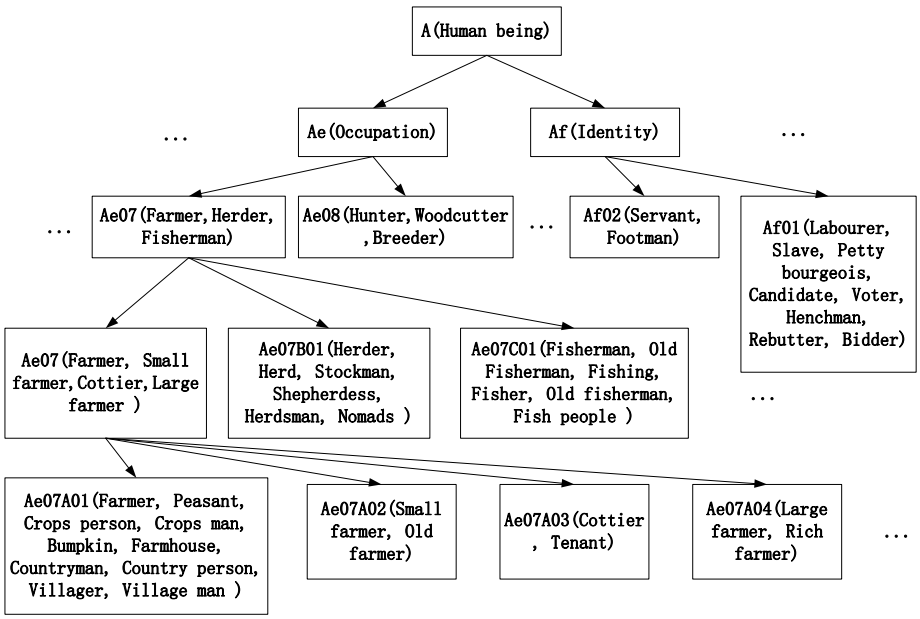

Figure 2. Tree Structure of HIT Synonyms Dictionary (Expanded) Words

Figure 2 describes the tree structure of HIT Synonyms Dictionary(expanded) words. As it shows, synonyms dictionary has a detailed classification, and every word has a corresponding code. $Code_i=X_{i1}X_{i2}X_{i3}X_{i4}X_{i5}F_i$, is described as 'Class', 'Division','Section', 'Clump', 'Group of atoms'. Flag $F_i$has 3 different kinds of marks, which are'=','#', '@', where '='represents 'equal' and 'synonymy', and '#' at the end represents 'unequal' and 'similar', which belongs to related words, and '@' at the end represents 'self-reclusive' and 'independent', which has neither synonymies nor related words. According to Figure 2, every word has a corresponding code. For example, the code for the Small farmer/Old farmer is Ae07A02, and the code for the Cottier/Tenant is Ae07A03. The semantic distance can be calculated through each word's code.

Definition 2: Assuming word $w_1$has m codes in expanded version of HITSynonyms Dictionary, which are: $Code_{11}$, $Code_{12}$, …, $Code_{1m}$, and the word w_2 has n codes: $Code_{21}$, $Code_{22}$, …, $Code_{2n}$, and then the semantic distance $Dis(w_1, w_2)$ between word w_1 and w_2 is defined as:

$$Dis(w_1, w_2) = \min_{i=1,2,\ldots m;\, j=1,2,\ldots n} Dis(code_{1i}, code_{2j}) \quad (1)$$

Definition 3: Assume two codes$Code_1$ and $Code_2$ are different from each other since layer i, and $1 \le i \le 5$. The larger the i is, the larger the semantic distance will be. Different weightsare distributed to each layer. It is defined as: $weights=[W_1,W_2,W_3,W_4,W_5,W_f]$, $W_1>W_2>W_3>W_4>W_5>W_f$. In the paper weights are defined as [1.0,0.5,0.25,0.125,0.06,0.03].

Definition 4:Semantic Distance $Dis(Code_1, Code_2)$ of $Code_1$ and $Code_2$is defined as:

$$Dis(Code_1, Code_2)= \begin{cases} init\_dis & \text{if } F_1=\text{"@" or } F_2=\text{"@"} \\ 0 & \text{if } Code_1=Code_2 \text{ and } F_1=F_2=\text{"="} \\ weights[5]\times init\_disif & Code_1=Code_2 \text{and } F_1=F_2=\text{"\#"} \\ weights[i-1]\times init\_disif & Code_1 \text{ and } Code_2 \text{ become different from layer i} \end{cases}$$

whereinit_dis is the customized initial value of the distance. In the paper, init_dis = 10.

2.2.2 Density-based Clustering[13,14,15]

After getting semantic distance of the words, we start to cluster all the words. As density based clustering algorithm does not need to set the number of clusters, and it can discover random forms of clusters.Density based clusters are also not sensitive to the noise, so the noise can be classified into a single class, and thus we choose a density-based clustering algorithm.Through density clustering, achieved word set $w_k$after word segmentation can be divided into severalclasses. Calculation steps are as below:

Input: ε --- radius
MinPts --- Minimum points of given points to be the core object in field ε
D --- set
Output: The target cluster set
Method: repeat Judge ifthe input point is core object or not
Core object {
Figure out all the reachable points in field ε of core object
}
Non-Core object {
If p is a boundary object, then mark p as the noise
}
untilThe judgment of all the input points hasbeen finished.
repeat Forall the reachable points in field ε of core objects, find out the set of connected

objects of maximum density, and combine density reachable objects.
Until Iterate through the field ε of all core objects.

2.2.3Keywords Determination

Several classes are obtained after density clustering. Threshold N is set to filter irrelevant classes. When n>N, the class can be defined as the theme related classification, where N is a constant and n is the number of relative words in a certain theme. The larger the number is, the more important the class to the theme will be.

Definition 5: Assuming thatj^{th} theme class $C_j(D)=\{w_1,w_2,\ldots,w_k\}$, $1\le k\le n$,and$w_M$is set as $C_j(D)$'s key word, and then

$$w_M = \arg\min_{0\le M\le k} \sum_{i=0}^{k} Dis(w_i, w_M)^2 \quad (2)$$

where$w_i \in C_j(D)$, and Dis(w_i, w_M) is the semantic distance betweenw_i and w_M.

According to Equation (2),all the theme related central words for document D are calculated, and are viewed as the key word to the document D.

The central word of themerelated class is the key word to the document. C (D)={w_1,w_2,…,w_n} is the key words set, and w_n is the central word of theme related class C(D).

3.System Structure and Procedure

This paper integrates semantic features whenextracting the key words, and proposes the keyword extraction algorithm based on semantics. The logical structure of the algorithm is shown in Figure3.It consists 4 modules: the text pre-processing module, the semantic distance calculation module, the words cluster module and the central point data calculation module.

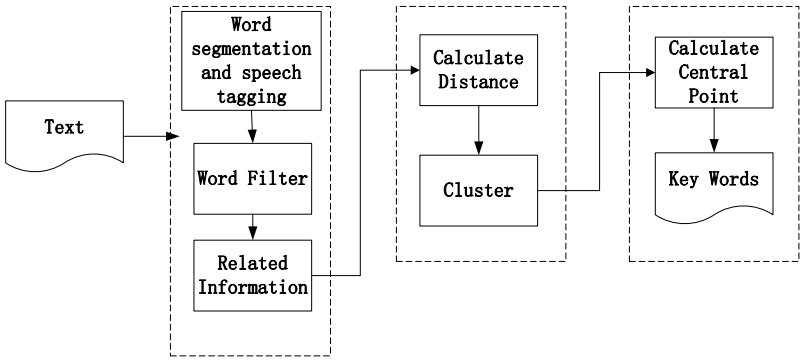

Figure3. Algorithm Logical Structure

First, word segmentation and speech tagging are conducted for the document, and then the the stop words are filtered from the word segmentation result to get the word sets. Afterwards, the semantic distance are calculated according to HITSynonyms Dictionary(expanded)and clustering based on the density of word set is done to get classification results.Then theme related classes are filtered from the results. Finallythe key words are determinedthrough calculating each theme related central words.

Algorithm processing steps is as below:
Input Text D
Output Key word for Text D
(1) Segment and tagspeech for text D to achieve Candidate Words.
(2) Remove the stop words in Candidate Words, andkeep adjective, adverb adjective, adnoun, set phrase, abbreviation, idiom, verb, action morpheme, auxiliary verb, noun verb

and noun, to get word data set W(D).

(3) Calculate semantic distance according to algorithm in 2.2.1

(4) Density clustering for word data set W(D) according to algorithm in 2.2.2, and filter theme related classes according to threshold N.

(5) Calculate central words for theme classes according to algorithm in 2.2.3to get the key words set $C(D)=\{w_1,w_2,\ldots,w_n\}$.

4.Experiments

As the number of key words and key wordsthemselvesarevariable, it is difficult to evaluate the result of key word extraction.Different persons will choose different key words, which makes the difference in evaluating the result of key words extraction. This paper usesartificial contrast test and grading test to evaluate the automatic extraction algorithm.

4.1 Artificial Contrast Experiment

The paper chooses 20 texts as test corpus from People's dailyof Jan, 1988.Then keywordsextracted by 5 different persons areselected as the standard of the tests. We comparekeyword extraction method(SEKM) based on semantics withthe method based on statistic (TF-IDF,the maximum entropy model(MEM) and method based on *HowNet*(WN).The evaluationstandard includes accurate rate(P), recall rate(R) andthe average value of both rates (measure value F1).

$$P=\frac{A}{A+B},\quad R=\frac{A}{A+C}\quad ,\quad F_1=\frac{2PR}{P+R}$$

whereA is the numbers of keywords judged by both manually and automaticallyextraction; B is the number of keywords judged only by automatically extraction; C is the number of keywords judged only by manually extraction. Then 5 different manually extraction results are calculated, andtheir average value is taken as the final test result, shownas Table 1.

Table 1. Artificial Contrast Experiment

Number of keywords	Accurate Rate（P）				Recall Rate（R）				Measure value F_1			
	TF-IDF	MEM	WN	SKEM	TF-IDF	MEM	WN	SKEM	TF-IDF	MEM	WN	SKEM
4	50.24%	51.56%	53.21%	54.76%	44.78%	45.36%	48.72%	49.48%	47.35%	48.26%	50.87%	51.99%
5	53.68%	55.35%	56.23%	58.25%	47.56%	49.35%	53.86%	57.61%	50.44%	52.18%	55.02%	57.93%
6	48.75%	50.25%	54.89%	56.83%	46.35%	50.21%	57.86%	65.39%	47.52%	50.23%	56.34%	60.81%

As the table shows, with respect to accurate rate and recall rate, SKEM algorithm performed obviously much better than statistical based method TF-IDF and MEM, slightly better than *HowNet* method. It has more advantages as the number of keywords increase.

4.2 Corpus Classification Experiment

The experiment adopts text sets provided by Department of Computer Science of Fudan University. The number of classes |C|=20, and the number of texts |D|=19637. Class distribution is shown in Table 2. 15 keyword-marked articles are randomly selected from each of the class, and then the contrast experiment on SEKM and statistical-based method (TF-IDF) is conducted.

4.2.1 Evaluation Standard

The evaluation standard of keywords extraction results is recall rate, accuracy rate and

value F, which are set according to the classification effect. The definitionis shown as below.

x(c): Total number of keywords extracted correctly from c type of text.

y(c): Total number of keywords extracted from c type of text

z(c): Total number of keywords contained in c type of text.

Recall rate: $rec(c) = \frac{x(c)}{z(c)}$

FValue: $F(c) = \frac{2 \times pre(c) \times rec(c)}{pre(c) + rec(c)}$

Marco accuracy rate: $macro\_pre = \frac{\sum_{\forall c \in C} pre(c)}{|C|}$

Marco recall rate: $macro\_rec = \frac{\sum_{\forall c \in C} rec(c)}{|C|}$

Micro recall rate: $micro\_rec = \frac{\sum_{\forall c \in C} x(c)}{\sum_{\forall c \in C} z(c)}$

Marco FValue: $macro\_F = \frac{\sum_{\forall c \in C} F(c)}{|C|}$

Micro accuracy rate: $micro\_pre = \frac{\sum_{\forall c \in C} x(c)}{\sum_{\forall c \in C} y(c)}$

Accuracy rate: $pre(c) = \frac{x(c)}{y(c)}$

Micro FValue: $micro\_F = \frac{2 \times micro\_pre \times micro\_rec}{micro\_pre + micro\_rec}$

Table 2. Experiment Corpus

No.	Type	Number of text
1	Art	1482
2	Literature	67
3	Education	120
4	Philosophy	89
5	History	934
6	Space	1282
7	Energy	65
8	Electronics	55
9	Communication	52
10	Computer	2715
11	Mine	67
12	Transport	116
13	Environment	2435
14	Agriculture	2043
15	Economy	3201
16	Law	103
17	Medical	104
18	Military	150
19	Politics	2050
20	Sports	2507

4.2.2 Experiment Results

Experiment results are shown in Table 3:

Table 3. Contrast Experiment of Corpus

	Marco Accurate rate	Marco Recall rate	Marco F value	Micro Accurate rate	Micro Recall rate	Micro F value
TF-IDF	49.25%	51.35%	50.02%	53.75%	55.21%	54.47%
MEM	50.65%	51.15%	50.39%	53.98%	56.89%	55.40%
WN	51.72%	52.46%	51.23%	54.26%	58.29%	56.20%
SEKM	52.63%	53.32%	51.84%	55.42%	60.57%	57.88%

Table 3 shows the accuracy rate, recall rate and F value of algorithm SEKM is higher than all other algorithms, which means that SEKM presents a better performance in classified corpus.

4.3 Scoring Experiment

The quality of keywords is determined by the reflection degree and coverage degree of themes. Reflection degree of themes describes the degree that keywords actually reflect the theme. The coverage degree means the key words need to cover all the key points instead of giving several synonymous words against only one point. The scoring standard is set as below.

3 points: Each key word accurately reflects the theme of text and is able to cover all the key point;

2 points: Each key word accurately reflects the theme of text but is not able to cover all the key point;

1 point: Part of key words reflects the theme of text but the rest do not

0 point: All the given key words cannot reflect the theme of the text

By using the standard above, the results of TF-IDF and SKEM algorithm are scored to each text, and each score's percentage is calculate as the test result, and then 5 different tests on 5 different readers is used. Finally, the average value is taken as the final test result, shown in Table 4 as below.

Table 4. ScoringTest Result

	TF-IDF	MEM	WN	SKEM
0分	7.20%	3.60%	0.00%	0.00%
1分	43.30%	44.32%	40.98%	35.10%
2分	47%	48.65%	54.86%	59.70%
3分	2.50%	3.43%	4.16%	5.20%

According to Table 4, it is clearly that percentage of 0 point and 1 point forSKEM algorithm is less than TF-IDF, slightly less than WN, and the percentage of 2 points and 3 points for SKEM algorithm is higher than other algorithms. The results validates that SKEM performs better than the other 3 algorithms.

Based on the three experiments above, we can safely come to the conclusion that SKEM has a better performance than TF-IDF and WN, which proves the effectiveness of SKEM algorithm.

5. Conclusion

The paper brings out an extraction algorithm of key word based on semantics. This method is able to reflect features of semantic level of the text. It is not only to count the words with a higher characteristic value, but also is not restricted by the field. It also does not need a large scale of training texts. The main idea of this algorithm is to make density based clustering for words through semantic distance between words, and get the theme related classes, and choose central words as the key words of the text from the theme related classes. Artificial contrast experiment is able to prove the effectiveness of this algorithm while raising the accurate rate and recall rate. Contrast experiment of classification corpus shows this algorithm is not restricted by classification of text. Scoring test makes it clear that the key words extracted by this algorithm are better than other methods with respect to

References

[1] G. Krupka, "SRA: Description of the SRA system as used for MUC-6", Proceedings of the Sixth Message Understanding Conference, **(1995)**; Morgan Kaufmann and California.
[2] D. H. Jang and S. H. Myaeng, "Development of a document summarization system for effective information services", RIAO Conference Proceedings:Computer-Assisted Information Searching on Internet, **(1997)**; Montreal, Canada.
[3] B. Krulwich and C. Burkey, "Learning user information interests through the extraction of semantically significant phrases", Hearst M and Hirsh H, EDS, AAAI Spring Symposium on Machine Learning in Information Access, AAAI Press, **(1996)**; California.
[4] W. F. Yang, "Chinese keyword extraction based on max duplicated strings of the documents", Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(2002)**; Tampere, Finland.
[5] J. Wang, "Updating Thesaurus via Extracting Keywords from Metadata", Journal of Chinese Information Processing, vol. 19, no. 6, **(2005)**, pp, 36-43.
[6] Y. C. Liu, X. L. Wand, Z. M. Xu and B. Q. Liu, "Mining Construction Rules of Chinese Key phrase Based on Rough Set Theory", Electronics Sinica, vol. 35, no. 2, **(2007)**, pp. 371-374.
[7] P. D. Turney, "Learning to extract key phrases from text", National Research Council, Canada, NRC Technical Report ERB-1O57, **(1999)**.
[8] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill, "KEA：Practical automatic key phrase extraction", Proceedings of the 4th ACM conference on Digital libraries，**(l999)**; Berkeley，California，US.
[9] A. MuQnoz, "Compound key word generation from document databases using hierarchical clustering ART model", Intelligent Data Analysis, vol. 1, no. 1, **(1996)**.
[10] A. M. Steier and R. K. Belew, "Exporting phrases: A statistical analysis of topical language", Second Symposium on Document Analysis and Information Retrieval, **(1993)**.
[11] S. J. Li, H. F. Wang, S. W. Yu and C. S. Xin, "Research on Maximum Entropy Model for Keyword Indexing", Chinese Journal of Computers, vol. 27, no. 9, **(2004)**, pp. 1192-1197.
[12] L. X. Wang and X. Y. Huai, "Semantic-based Keyword Extraction Algorithm for Chinese Text", Computer Engineering, vol. 38, no. 1, **(2012)**, pp. 13-17.
[13] L. Ertoz, M. Steinbach and V. Kumar, "Finding clustering of different sizes, shapes, and densities in noisy, high dimensional data", SIAM International Conference on Data Mining (SDM), **(2003)**.
[14] Y. C. Zhan, M. Song, F. Xie and J. Song, "Clustering datasets containing clusters of various densities", Journal of Beijing University of Posts and Telecommunications, vol. 26, no. 2, **(2003)**, pp. 42-47.
[15] G. Karypis, E. H. Han and K. V. Chameleon, "A hierarchical clustering algorithm using dynamic modeling". IEEE Computer, vol. 32, no. 8, **(1999)**, pp. 68-75.

Authors

FangJiang, she is doctoral candidate,her main research topics: Intelligent Information Processing.

