Chinese Text Summarization Based on Thematic Area Detection

Po Hu
Department of Computer Science
Central China Normal University
Wuhan, China, 430079
geminihupo@163.com

Tingting He
Department of Computer Science
Central China Normal University
Wuhan, China, 430079
hett@163.net

Donghong Ji
Institute for Infocomm Research
Heng Mui Keng Terrace,
Singapore, 119613
dhji@i2r.a-star.edu.sg

Abstract

Automatic summarization is an active research area in natural language processing. This paper has proposed a special method that produces text summary by detecting thematic areas in Chinese document. The specificity of the method is that the produced summary can both cover many different themes and reduce its redundancy obviously at the same time. In this method, the detection of latent thematic areas is realized by adopting K-medoids clustering method as well as a novel clustering analysis method, which can be used to determine automatically K, the number of clusters.. In addition, a novel parameter, which is known as representation entropy, is used for summarization redundancy evaluation. Experimental results indicate a clear superiority of the proposed method over the traditional non-thematic-area-detection method under the proposed evaluation scheme when dealing with different genres of text documents with free style and flexible theme distribution.

1 Introduction

With the approaching information explosion, people begin to feel at a loss about the mass of information. Because the effectiveness of the existing information retrieval technology is still unsatisfactory, it becomes a problem to efficiently find the information mostly related to the needs of customers retrieval results so that customers can easily accept or reject the retrieved information without needing to look at the original retrieval results. This paper has proposed a new summarization method, where K-medoid clustering method is applied to detect all possible partitions of thematic areas, and a novel clustering analysis method, which is based on a self-defined objective function, is applied to automatically determine K, the number of latent thematic areas in a document

This method consists of three main stages: 1) Find out the thematic areas in the document by adopting the K-medoid clustering method (Kaufmann and Rousseeuw, 1987as well as a novel clustering analysis method. 2) From each thematic area, find a sentence which has the maximum semantic similarity value with this area as the representation. 3) Output the selected sentences to form the final summary according to their positions in the original document.

To validate the effectiveness of the proposed method, use this method as well as the traditional non-thematic-areas-detection method on our experimental samples to generate two groups of summaries. Next, make a comparison between them. The final results show a clear superiority of our method over the traditional one in the scores of the evaluation parameters.

The remainder of this paper is organized as follows. In the next section, we review related methods that are commonly discussed in the automatic summarization literature. Section 3 describes our method in detail. The evaluation methodology and experimental results are presented in Section 4. Finally, we conclude with a discussion and future work.

2 Related Work

The research of automatic summarization begins with H.P.Luhn's work. By far, a large number of scholars have taken part in the research and had many achievements. Most of the researchers have concentrated on the sentence-extraction summarization method (the so-called shallower approach) (Wang et al., 2003; Nomoto and Matsumoto, 2001; Gong and Liu, 2001), but not the sentence-generation method (the so-called deeper approach)(Yang and Zhong., 1998). On the one hand, it is caused by the high complexity and the severe limitation of practical fields of rational natural

language processing technology and knowledge engineering technology. On the other hand, it is closely associated with the great achievements in many fields of natural language processing by statistical research methods, machine learning methods and pattern recognition methods in recent years (Mani, 2001).

The summarization method of sentence-extraction can roughly be divided into two kinds: supervised and unsupervised (Nomoto and Matsumoto, 2001). Generally, the realization of the former relies on plenty of manual summaries, that is so-called "Gold Standards" which help determining the relevant parameters of the statistical model for summarization. However, not all people believe that manual summaries are reliable, so the researchers have begun to investigate the general unsupervised method, which can avoid the requirement of support of manual summaries. Nevertheless it is soon discovered that the summaries produced by this method can't cover all the themes and have great redundancy at the same time. Usually, it can only cover those intensively distributed themes while neglects others. So researchers in Nanjing University proposed a summarization method based on the analysis of the discourse structure to overcome these problems (Wang et al., 2003). By making statistics of the reduplicated words in the adjacent paragraphs of the document, the semantic distances among them can be worked out. Then analyse the thematic structure of the document and extract sentences from each theme to form a summary. It is ideal to employ this method while dealing with those documents with standard discourse structure, because it can effectively avoid the problems caused by the summarization method without discourse structure analysis. Yet when the writing style of a document is rather free and the distribution of the themes is variable, that is the same theme can be distributed in several paragraphs not adjacent to each other, then the use of this method can't be equally effective.

To deal with a lot of Chinese documents which have free style of writing and flexible themes, a sentence-extraction summarization method created by detecting thematic areas is tried following such work as (Nomoto and Matsumoto, 2001; Salton et al., 1996; Salton et al., 1997; Carbonell and Goldstein, 1998; Lin and Hovy, 2000). The thematic areas detection in a document is obtained through the adaptive clustering of paragraphs (cf. Moens et al. 1999), so it can overcome in a certain degree the defects of the above methods in dealing with the documents with rather flexible theme distribution.

3 The Algorithm

In this section, the proposed method will be introduced in detail. The method consists of the following three main stages:

Stage 1: Find the different thematic areas in the document through paragraph clustering and clustering analysis.

Stage 2: Select the most suitable sentence from each thematic area as the representative one.

Stage 3: Make the representative sentences form the final summary according to certain requirements.

3.1 Stage 1: Thematic Area Detection

The process of thematic area detection is displayed in Figure 1.

The each step of Figure 1 is explained in the following subsections.

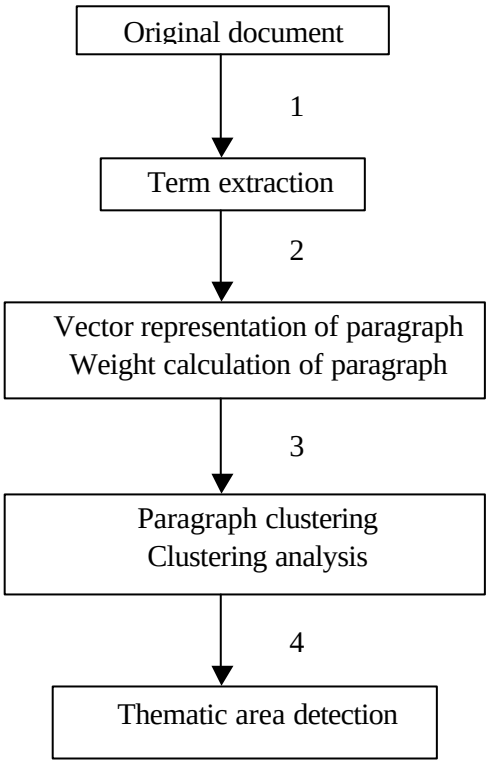

Figure 1: The process of thematic area detection (4 steps in all)

3.1.1 Step 1: Term Extraction

Different from the general word segmentation operation adopted in the traditional Chinese automatic summarization research, we do not take the general operation when pre-processing the original

document, but make use of the method introduced by (Liu et al., 2003) to extract terms
from the document and then express its content by such metadata elements as terms.

The greatest advantage of term extraction technology is that it needs no support of fixed thesaurus, only through the continuous updating and making statistics of a real corpus. We can dynamically establish and update a term bank and improve the extraction quality through continuous correcting of the parameters for extraction. Thus it is of wide practical prospects for natural language processing. In addition, the terms can represent a relative specific meaning, because most of them are phrases, which consist of multi-characters.

3.1.2 Step 2: Vector Representation and Weight Calculation of Paragraph

The advantage of the vector space model (VSM) is that it successfully makes the unstructured documents structured which makes it possible to handle the massive real documents by adopting the existing mathematical instruments. All the terms extracted from the document are considered as the features of a vector, while the values of the features are statistics of the terms. According to this, we can set up the VSM of paragraphs, that is each paragraph P_i (i:1~M,M is the number of all paragraphs in a document) is represented as the vector of weights of terms, VP_i, $VP_i = (WP_{i1}, WP_{i2}, \ldots, WP_{iN})$

Where N is the total number of terms, WP_{ij} denotes the weight of the j-th term in the i-th paragraph. There are many methods of calculating WP_{ij}, such as tf, tf*idf, mutual information (Patrick Pantel and Lin, 2002), etc. The method adopted here (Gong and Liu, 2001) is shown as follows:

$$\mathbf{WP_{ij} = \log(1+TF(T_{ij}))*\log(M/M_j)} \quad (1)$$

Where $TF(T_{ij})$ denotes the number of occurrence of the j-th term in the i-th paragraph, M/M_j denotes the inverse paragraph frequency of term j, and M_j denotes the number of paragraphs in which term j occurs. In accordance, on the basis of defining WP_{ij}, we can further define the weight of paragraph P_i, $W(P_i)$, by the follwing formula:

$$W(P_i) = \frac{\sum_{j=1}^{n} WP_{ij}}{n} \quad (2)$$

In formula (2), *n* represents the total number of different terms occurring in the *i*-th paragraph.

3.1.3 Step 3: Paragraph Clustering and Clustering Analysis

1) Paragraph clustering

The existing clustering algorithms can be categorized as hierarchical (e.g. agglomerative etc) and partitional (e.g. K-means, K-medoids, etc) (Pantel and Lin, 2002).

The complexity of the hierarchical clustering algorithm is $O(n^2 Log(n))$, where n is the number of elements to be clustered, which is usually greater than that of the partitional method. For example, the complexity of K-means is linear in n. So in order to achieve high efficiency of algorithm, we choose the latter to cluster paragraphs.

K-means clustering algorithm is a fine choice in many circumstances, because it is simple and effective. But in the process of clustering by means of K-means, the quality of clustering is greatly affected by the elements that marginally belong to the cluster, and the centroid can't represent the real element in the cluster, So while choosing the paragraphs clustering algorithm, we adopt K-medoids (Kaufmann and Rousseeuw, 1987; Moens et al. 1999) which is less sensitive to the effect of marginal elements than K-means.

Suppose that every sample point in the N-dimensional sample space respectively represent a paragraph vector, and the clustering of paragraphs can be visualized as that of the M sample points in the sample space. Here N is the number of terms in the document and M is the number of paragraphs. Table 1 shows the formal description of the paragraph clustering process based on K-medoids method.

2) Clustering analysis

A classical problem when adopting K-medoid clustering method and many other clustering methods is the determination of K, the number of clusters. In traditional K-medoid method, K must be offered by the user in advance. In many cases, it's impractical. As to clustering of paragraphs, customers can't predict the latent thematic number in the document, so it's impossible to offer K correctly.

In view of the problem, the authors put forward a new clustering analysis method to automatically determine the value of K according to the distribution of values of the self-defined objective function. The basic idea is that if K, the number of clusters, is determined with each value of K, and

Input: <a, b>, they respectively denote the paragraph matrix composed by all the paragraph vectors in the document and the number of clusters, k (the range of k is set to 2~M).

Step 1: randomly select k paragraph vectors as the initial medoids of the clusters (here, the medoids denote the representative paragraphs of k clusters).

Step 2: assign each paragraph vector to a cluster according to the medoid X closest to it.

Step 3: calculate the Euclidean distance between all the paragraph vectors and their closest medoids.

Step 4: randomly select a paragraph vector Y.

Step 5: to all the X, if it can reduce the Euclidean distance between all the paragraph vectors and their closest medoids by interchanging X and Y, then change their positions, otherwise keep as the original.

Step 6: repeat from step 2 to 5 until no changes take place.

Output: <A, B, C>, they respectively denote the cluster id, the representative paragraph vector and all the paragraph vectors of each cluster under the k clusters.

Table 1: Paragraph clustering process based on K-medoid method

suitably, then the corresponding clustering results can well distinguish the different themes in the document, and correspondingly the average of the sum of the weight of the representative paragraph under each theme will tend to maximize. We call this the maximum property of the objective function.

Correspondingly, we define the following objective function Objf(K) to reflect clustering quality and determine the number of clusters, K.

$$Objf(K) = \frac{\sum_{j=1}^{K} W(P_j)}{K} \quad (3)$$

Where $W(P_j)$ denotes the weight of the selected representative paragraph in the j-th cluster, here the selected representative paragraph P_j can be regarded as the medoid in the j-th cluster which is determined by the final output of the presented K-medoid paragraph clustering process, and the weight of P_j is calculated by formula (2). Put the objective function in K clustering results corresponding then make good use of the maximum property of the objective function to adaptively determine the final number of clusters, K.

Figure 2 shows the concrete distribution of the values of objective function obtained in the example document "On the Situation and Measures That Face Fishing in the Sea in Da Lian City" when adopting the proposed clustering analysis method. According to the maximum property of objective function, that is take the value of K when the values of the objective function take maximum as the final number of clusters. From the results in Figure 2, we can know that K equals to six, that is we find six latent thematic areas from nine paragraphs in the document with this method.

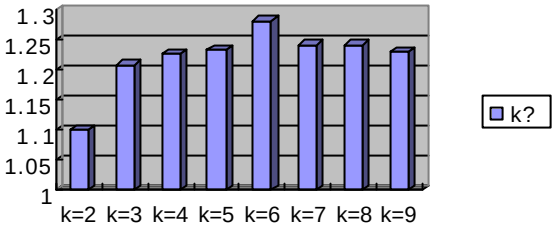

Figure 2: The distribution of the values of the objective function when K takes different values

Figure 3 displays the paragraph clustering results when K equals six in the process of adopting K-medoid clustering method on the example document.

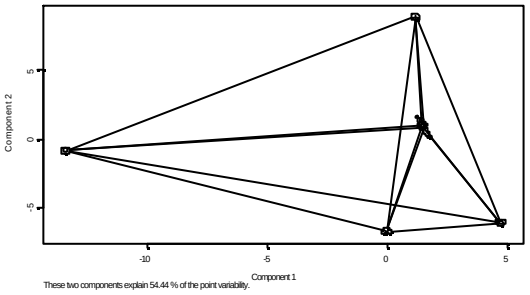

Figure 3: The paragraphs clustering result when K equals to six

3.1.4 Step 4: Thematic Area Detection

Output the complete information table of each thematic area in the form of the representative paragraph and all the paragraphs and sentences covered by the thematic area.

3.2 Stage 2: Selection of the Thematic Representative Sentences

To select a most suitable representative sentence from each thematic area, the author proposes the following method. This is in contrast with a method proposed by Radev (Radev et al., 2000), where the centroid of a cluster is selected as the representative one.

Method: select the sentence which is most similar to the thematic area semantically as representative one.

Before carrying out the method in detail, there are two problems to be solved:

1) The vector representation of sentence and thematic area

The vector representation of sentence and thematic area is similar to that of paragraph introduced before. We only need to change the weight calculation field of the terms from the interior of paragraph to the interior of sentence or thematic area. Accordingly, we can describe the sentence vector and thematic area vector as follows

$VS_j = (WS_{j1}, WS_{j2}, \ldots, WS_{jN})$

$VA_k = (WA_{k1}, WA_{k2}, \ldots, WA_{kN})$

2) The semantic similarity calculation between sentence and thematic area

The calculation of semantic similarity of sentence and thematic area can be achieved by calculating the vector distance between sentence vector and thematic area vector. Here we adopt the traditional cosine method for vector distance calculation. Correspondingly, the distance between the sentence vector VS_j and the thematic area vector VA_k is calculated by the following formula:

$$Cos(VS_j, VA_k) = \sum_{i=1}^{N} (WS_{ji} \times WA_{ki}) \Big/ \sqrt{\left(\sum_{i=1}^{N} WS_{ji}^2\right) \times \left(\sum_{i=1}^{N} WA_{ki}^2\right)} \quad (4)$$

Principles of evaluating summarization redundancy

At the premise of the same number of summarization sentences selected out by different summarization methods:

The ***higher*** the value of RE calculated by the covariance matrix of the summarization sentence vectors.

The ***lower*** the summarization redundancy.

Table 2: The evaluation principles of the summarization redundancy based on RE

3.3 Stage 3: The Creation of the Summary

Ouput the selected representative sentences from each thematic area according to their postions in the original document to form the final summary.

4 Experimental Results and Performance Evaluation

4.1 Evaluation Methodology

It is challenging to objectively evaluate the quality of different automatic summarization methods. Methods for evaluation can be broadly classified into two categories: intrinsic and extrinsic (Mani, 2001). We adopt the former to evaluate the quality of summarization by defining the following parameters for evaluation.

1) Theme coverage (TC)

The definition of TC is the percentage of the thematic contents covered by the selected summarization sentences. The value of the parameter can be got by means of the works of some experts.

2) Representation entropy (RE)

In order to effectively and objectively evaluate the redundancy of the produced summary, we refer to the parameter which was initially proposed by (Mitra et al., 2002) for evaluating the feature redundancy in the process of feature selection and transform it into the novel parameter to evaluate the summarization redundancy.

According to this, some important notations are defined as follows:

N	Number of terms in the original document ;
Nz	Number of sentences in the produced summary ;
Lz	Nz-by-N matrix composed by all the sentence vectors in the produced summary ;
$\sum z$	Nz-by-Nz covariance matrix composed by all the sentence vectors in the produced summary ;
λi	Eigenvalues of $\sum z$ i:1~Nz ;
Υi	$\Upsilon i = \lambda i / \sum_{i=1}^{Nz} \lambda_i$;

Genre	Sample ID	Number of characters	Number of paragraphs	Number of detected thematic areas	Theme coverage (TC)		Representation entropy (RE)	
					Method1	Method2	Method1	Method2
Economy	d10000801	1461	11	5	0.6	0.56	1.44	1.25
	d10000901	1192	7	5	0.64	0.6	1.36	1.35
	d10100101	1936	14	9	0.66	0.64	2.14	2.06
	d10100201	1778	12	6	0.8	0.5	1.62	1.54
	d10100301	2472	4	3	0.64	0.4	0.81	1.05
	d10100601	1553	11	7	0.9	0.64	1.79	1.83
	d29600501	2400	6	4	0.7	0.56	1.33	1.01
	d29800101	670	4	3	0.64	0.6	1.06	1.01
	d40000301	2026	8	5	0.56	0.52	1.45	1.54
	d40100101	1529	7	4	0.6	0.58	1.19	1.31
Art	e10000101	907	4	2	0.72	0.56	0.64	0.24
	e10000201	845	5	3	0.9	0.6	1.06	0.89
	e29600201	2035	5	4	0.72	0.5	1.36	1.21
	e29800201	1831	7	2	0.56	0.52	0.67	0.57
Prose	f20000101	2354	12	7	0.58	0.5	1.92	1.79
	f20000201	1769	9	6	0.64	0.52	1.72	1.50
Military	g00000201	1163	5	4	0.84	0.56	1.34	1.21
	g00000501	790	6	4	0.64	0.54	1.31	1.26
	g00001201	425	5	5	0.92	0.62	1.45	1.49
	g00100101	1629	10	3	0.84	0.6	0.93	0.82
	g00100301	817	6	4	0.76	0.7	1.32	1.26
	g00100501	1355	4	4	0.84	0.5	1.31	1.12
	g09600901	2179	7	6	0.72	0.62	1.75	1.73
	g09601601	1271	5	3	0.7	0.52	1.03	0.98
Life	h00000401	1224	6	6	0.72	0.54	1.75	1.60
	h00000601	1331	15	7	0.6	0.5	1.88	1.80
	h00000901	1507	7	3	0.64	0.68	1.05	0.83
	h00001801	1604	8	6	0.68	0.64	1.73	1.66
	h00100301	960	6	3	0.9	0.4	1.04	1.05
	h00100601	1228	6	3	0.8	0.6	1.06	0.89

Table 3: Experimental data

Genre	Number of samples	Mean of theme coverage ($\overline{TC}$)		Mean of representation entropy ($\overline{RE}$)		Ratio of information and noise (F)	
		Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Economy	10	0.68	0.56	1.42	1.40	2.81	2.27
Art	4	0.72	0.54	0.93	0.73	1.82	1.12
Prose	2	0.62	0.52	1.82	1.65	3.83	2.71
Military	8	0.78	0.58	1.31	1.23	2.89	1.98
Life	6	0.72	0.56	1.42	1.31	2.98	2.08

Table 4: Evaluation results of parameters

The value of RE (Mitra et al., 2002) is calculated as follows:

$$RE = -\sum_{i=1}^{Nz} \Upsilon_i * \log \Upsilon_i \quad (5)$$

The evaluation principles of the summarization redundancy based on RE are demonstrated in Table 2.

3) Ratio of information and noise (F)

$$F = TC / e^{-RE} \quad (6)$$

The novel evaluation parameter proposed by us can objectively evaluate the quality of the produced summary by effectively combining the above two parameters. The more the value of F, the better the quality of the produced summary.

4.2 Experimental Results

We randomly extract 200 documents of different genres from the Modern Chinese Corpus of State Language Commission to form the experimental corpus. Because summarizing short documents doesn't make much sense in real applications (Gong and Liu, 2001), we select 30 documents of more than 400 characters from the corpus as the samples which are summarized by the proposed summarization method (method 1 for abbreviation) and the traditional non-thematic-area-detection method (method 2 for abbreviation), that is the method of determining the weights of sentences in a document, sorting them in a decreasing order, and selecting the top sentences in the end. The specific experimental data and evaluation results of parameters are given in table 3 and table 4.

The synthetic evaluation of the 30 samples proves that our method under the above evaluation parameters is superior to the traditional non-thematic-area-detection summarization method when dealing with different genres of text documents with free style and flexible theme distribution, and the results we have achieved are encouraging.

5 Conclusions

In this paper, we have proposed a new summarization method based on thematic areas detection. By adopting a novel clustering analysis method, it can adaptively detect the different thematic areas in the document, and automatically determine K, the number of thematic areas. So the produced summary can both cover as many as different themes and reduce its redundancy obviously at the same time.

For our experiment, we used three different parameters to evaluate the quality of the produced summaries in theme coverage and summarization redundancy. We achieved a better performance than the traditional non-thematic-areas-detection method in the proposed evaluation scheme. As a future work , we need the additional research for testing the proposed method on larger-scale real corpora , and have the further comparison with earlier similar works such as MMR, etc. In addition, we'll improve our summarization system by considering the structure of thematic areas and user's requirement.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *In Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York.

Yihong Gong, Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. *In Proceedings of ACM SIGIR'01*, pages 19-25, ACM, New York.

L. Kaufmann and P.J. Rousseeuw. 1987. Clustering by means of medoids. *In Statistical Data Analysis Based on the L1 Norm,Y.Dodge,Ed,Amsterdam,* 405-416.

Chin-Yew Lin and Eduard Hovy. 2000. The automatic acquisition of topic signatures for text summarization. *In Proceedings of the 18th International Conference of Computational Linguistics* (*COLING 2000*).

Jian-Zhou Liu, Ting-Ting He, and Dong-Hong Ji. 2003. Extracting Chinese term based on open corpus. *In Proceedings of the 20th International Conference on Computer Processing of Oriental Languages,*pages 43-49. ACM, New York.

Inderjeet Mani. 2001. Summarization evaluation: an overview. *In Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization.*

Inderjeet Mani. 2001. Recent developments in text summarization. *In Proceedings of CIKM'01*, 529-531.

Pabitra Mitra, C.A. Murthy, Sankar and K.Pal. 2002. Unsupervised feature selection using

feature similarity. *IEEE Transactions of Pattern Analysis and Machine Intelligence*: 1-13.

Marie-Francine Moens, Caroline Uyttendaele and Jos Dumortier. 1999. Abstracting of legal cases: The potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50 (2): 151-161.

Tadashi Nomoto, Yuji Matsumoto. 2001. A new approach to unsupervised text summarization. *In Proceedings of ACM SIGIR'01*, pages 26-34. ACM, New York.

Patrick Pantel and Dekang Lin. 2002. Document clustering with committees. *In Proceedings of ACM SIGIR'02*, pages 199-206. ACM, New York.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *In ANLP/NAACL Workshop on Summarization*.

Gerard Salton, Amit Singhal, Chris Buckley and Mandar Mitra. 1996. Automatic text decomposition using text segments and text themes. *Hypertext 1996*: 53-65.

Gerard Salton, Amit Singhal, Mandar Mitra and Chris Buckley. 1997. Automatic text structuring and summarization. *In Information Processing and Management*, 33(2):193-208.

Ji-Cheng Wang, Gang-Shan Wu, Yuan-Yuan Zhou, Fu-Yan Zhang. 2003. Research on automatic summarization of web document guided by discourse. *Journal of Computer Research and Development*, 40(3):398-405.

Xiao-Lan Yang and Yi-Xin Zhong. 1998. Study and realization for text interpretation and automatic abstracting. *Acta Electronica Sinica*, 26(7):155-158.