# Automatic Chinese Summarization Method Based on the HowNet and Clustering Algorithm

**Gang Bai Dongmei Wang Zongyao Ding Yi Zhu**

College of Information Technical Science, Nankai University, Tianjin 300071, P. R. China

## Abstract

To solve the problems in traditional automatic Chinese summarization, a new method based on the word concept and clustering is presented in this paper. Different from the normal statistical method, concept is used as feature instead of word. Also, instead of word frequency statistics, word concept frequency statistics (WCFS) is used in our approach. For each paragraph, a conceptual vector space model is established, and then the clustering algorithm is used for multiple topic partition. The evaluation results show that the method proposed in this paper is more efficient and robust than the traditional one.

**Keywords**: Automatic summarization, Clustering, HowNet, Conceptual vector space model, Topic partition

## 1. Introduction

Along with the development of Internet, the electronic text emerges massively, and automatic summarization is quick, effective, objective. Such superiorities with which manual digest is incomparable have fully manifested its good practical value.

In general, automatic summarization is defined as the process that the abstract of a document is generated automatically by computer [1]. It is viewed as one of the most important applications of Natural Language Understanding.

The automatic summarization system is divided into two kinds: mechanical summarization based on statistics and understanding summarization based on meaning.

The mechanical summarization extracts some sentences of the original text to compose the abstract according to the article external characteristics, such as word frequency, title, position, syntax structure, clue word, instruction phrase and so on.

The understanding summarization based on meaning uses linguistics knowledge to gain language structure; moreover, it uses the domain knowledge to judge and reason, then obtains the meaning expression of digest, and finally produces the abstract from the meaning expression [2]. Understanding summarization not only requires the computer to have the ability of natural language understanding and production, but also requires it can express and organize each kind of background and domain knowledge.

Both of the two methods are not easy to implement, but the mechanical summarization is simpler and easier. This paper uses the mechanical summarization as the method.

The background knowledge about HowNet and how to establish the concept vector space model based on HowNet will be introduced in the second part. And we explain how to carry on topic partition based on clustering algorithm in the third part and introduce how to select summarization sentences after clustering in the fourth part. At last, the result of experiment will be shown to evaluate our method.

## 2. The word concept and the conceptual vector space model based on HowNet

It is generally supposed that each pair of feature vector is orthogonal in the vector space model (VSM). However, when VSM is used in the natural language processing domain, it often uses small language unit like character, word or phrase, as feature, thus the phenomenon that a word has several meanings and several words have the same meaning is inevitable. In order to solve this problem, this paper takes HowNet as the word's semantic knowledge library, and then uses it to establish the conceptual vector space model.

### 2.1. Introduction of HowNet

HowNet is a knowledge library that takes the concept represented by Chinese and English words as the description object [3]. In this paper, we make use of the important resource inside HowNet to evaluate the meaning for each word. Thus, we may obtain some

useful information from HowNet through processing. It can be described as follow:

WX：Word

GX：The lexical category of the word

DEF：The definition of the word

For example:

WX = 工作 (means 'work')

GX = N (means 'nouns')

DEF = affairs|事务， $undertake|担任

## 2.2. Concept obtaining from the HowNet

After the word segmentation on the text, each word is labeled by its lexical category. For a general language, when we express the subject of an article, the nouns play the main role, and the number of verbs must be far less than the noun's. Moreover, the meaning of noun is stable than the verb [4], so only the nouns are withdrawn in this paper. In this way, it may save the running time of program on one hand; on the other hand, it can enhance the precision when we extract the key concepts from the text.

In our approach, we obtain the concept of each noun based on the HowNet knowledge library and extract DEF item from useful information. The synonyms have the same concept, while a polysemant can have several different ones. If a word has only one meaning and exists in the HowNet dictionary, we may obtain its meaning directly. For the polysemant, we count the appearance frequency of each concept of this polysemant separately in the text, and then select the concept that has the highest appearance frequency as the concept of the word [5]. If a word does not exist in the HowNet dictionary, we use the word prototype as its concept. After the processing, the concept of each noun is obtained. This process eliminates the phenomenon that a word has several meanings and several words have the same meaning.

## 2.3. Concept weight computation

After we obtain and merge all concepts of words in the text based on HowNet, a concept set $\{C_1,C_2,\ldots,C_k\}$ is obtained. $C_i$ is the word set $\{W_1,W_2,\ldots,W_h\}$ in which the words have a same concept (DEF item). Then the frequency $F(C_i)$ of the concept $C_i$ is: [5]

$$F(C_i)=\sum_{j=1}^{h}F(W_j) \qquad (1)$$

$F(W_i)$ is the appearance frequency of the word $W_i$ in the text. Then the concept weight $w(C_i)$ is calculated as follow:

$$w(C_i)=\lambda\times F(C_i)\times(\log_2\frac{n_i}{N}+1) \qquad (2)$$

$\lambda$ is the position weighting coefficient. If concept $C_i$ appears in the title, $\lambda$ is 1.5, otherwise $\lambda$ is 1. $F(C_i)$ is the appearance frequency of the concept $C_i$ in the text. $n_i$ is the number of paragraphs in which the concept $C_i$ appears and $N$ is the total number of paragraphs in the text.

By formula (2), we can calculate the weight of each concept, and then sort, choose several concepts on top of the weight list as the key concepts of the text to establish the concept vector model for each paragraph and compute sentence weight.

## 2.4. Establishment of conceptual vector space model for each paragraph

For the paragraph i, we establish the conceptual vector $P_i = \{C_1, W_{i1}; \ldots; C_j, W_{ij}; \ldots; C_n, W_{in}\}$. The element $W_{ij}$ is the weight of $C_j$ in the paragraph; it can be calculated based on the formula below:

$$W_{ij}=\frac{\lambda\times tf_{ij}\times\log_2(n_j/N+1)}{\sqrt{\sum_{j=1}^{n}(\lambda\times tf_{ij}\times\log_2(n_j/N+1))^2}} \qquad (3)$$

$\lambda$ is the position weighting coefficient; $tf_{ij}$ is the appearance frequency of $C_i$ in paragraph $P_i$; $n_j$ is the number of paragraphs in which the concept $C_i$ appears; N is the total number of paragraphs in the text.

The similarity of two paragraphs should be calculated in the process of topic partition. Because the paragraphs are already projected to n-dimension vector space, we can take the cosine of angle between the two paragraph conceptual vectors as the similarity of the two paragraphs.

If there are two conceptual vectors $x=(x_1,x_2,\ldots,x_n)$, and $y=(y_1,y_2,\ldots,y_n)$, we can obtain the similarity of two paragraphs based on the formula below [6]:

$$Sim(x,y)=\frac{\sum_{k=1}^{n}x_k y_k}{\sqrt{(\sum_{k=1}^{n}x_k^2)(\sum_{k=1}^{n}y_k^2)}} \qquad (4)$$

## 3. Topic partition based on clustering algorithm

The traditional automatic summarization methods extract sentences only according to the weight of the sentences. These methods only cover the main topics, while neglecting the less important ones, so the integrity is always not very good. Therefore, we use the clustering algorithm to carry on the topic partition in the text in our approach. Not only can this approach guarantee the summarization cover the topic comprehensively, but also enable the summarization to cover the most information of the text.

We use K-means algorithm as the clustering algorithm. Assume that conceptual vector of each paragraph is a sample point in n-dimension feature space, and the paragraph clustering problem can be changed into a normal clustering problem of M sample points in n-dimension feature space, while n is the number of key concepts in the text and M is the total number of paragraphs in the text.

### 3.1. K-means clustering algorithm

The K-means algorithm is described as follow:

(1) Choose K paragraph conceptual vectors randomly as initializing cluster centers.
(2) Compute the similarity between each paragraph conceptual vector and each cluster center, and then reassign each paragraph conceptual vector to the cluster with the highest similarity.
(3) Compute the mean of the samples as new cluster center in each cluster.
(4) Repeat (2) and (3), until the mean of each cluster center does not change any more or less than some predefined threshold.

### 3.2. The automatic selection of parameter K

In order to solve the problem that the traditional K-means cluster method can not decide the value of K automatically, we propose a new method to determine parameter K. We continue merging until only a single cluster remains and at each step plot the similarity between the two clusters selected for merging. Fig. 1 shows the plot of these similarity values for a text example. We observe a sudden drop of the similarity when the algorithm tries to merge two real clusters. In the experiments on other data sets, we also observed similar trends from the plots. This suggests that we can use the occurrence of a sudden similarity drop as a heuristic to determine k [7].

In this way, we can determine the value of K (the number of topics in the text).

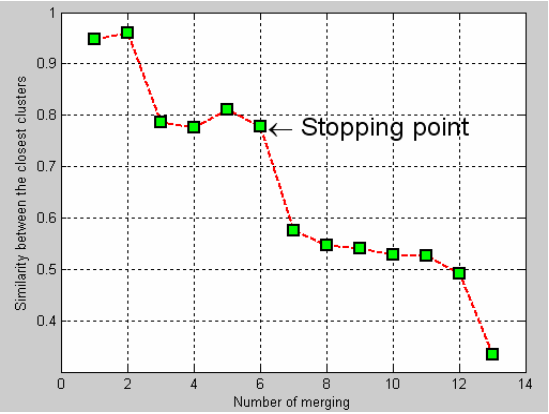


Fig.1: The similarity between the closest clusters.

## 4. The selection of summarization sentence after clustering

### 4.1. The weight of sentence computation

After the analysis of massive test results, the sentence weight is mainly related to these factors: (1) the importance of the key concepts that appear in the sentence, (2) the position of the sentence in the text, and (3) whether there is special mark in the sentence. So the weight of sentence is:

$$w(S_j) = \lambda_1 \lambda_2 \frac{\sum_{i=1}^{n} F_{ij} w(C_i)}{L(S_j)} \qquad (5)$$

$w(C_i)$ is the weight of the concept $C_i$ in the text; $F_{ij}$ is the appearance frequency of concept $C_i$ in the sentence $S_j$; $L(S_j)$ is the length of sentence $S_j$; $\lambda_1$ is the position weighting coefficient of sentence $S_j$, if the sentence appears in the first paragraph of the text, $\lambda_1$ is 1.5, if the sentence is the first sentence of paragraph except the first paragraph, $\lambda_1$ is 1.3, otherwise $\lambda_1$ is 1; $\lambda_2$ is the special mark weight proportionality factor, if there is a hint(e.g. “this article elaborated”, “this article proposed”, “in brief”, “in summary” and so on), $\lambda_2$ is 1.5, if there is a hint that means to give an example(e.g. “for example”), $\lambda_2$ is 0.5, otherwise $\lambda_2$ is 1 [8].

## 4.2. Proportion computation

After topic partition, we assign the number of abstract sentences to each topic according to the importance of the topics, then select several summarization sentences from each topic according to the weight of the sentences, and at last, output the summarization sentences according to the order of appearance in the text.

Assume the number of summarization sentences that we want withdraw from the text is L, the number of the topic in the text is n, and the weight of the topic i (i=1,2,…n) is $wt_i$, which is the weight sum of the sentences contained in the topic. The number of sentence in the topic i is $P_i$ [9].

$$P_i = \frac{L \times wt_i}{\sum_{i=1}^{n} wt_i} \tag{6}$$

# 5. System Evaluation

## 5.1. Exterior evaluation

In order to evaluate the system objectively, it is better to have many texts that have summary which experts have written already. Because these language materials are deficient at present, we use exterior evaluation at first.

We have 2816 texts that belong to different types (politics, military, art, environment, education, computer, medicine, sports, economy, and transportation). Then, we selected 300 texts that have more than 400 characters as testing language materials. We do summarizations respectively on these texts by our approach and the traditional summarization method under different compression ratio. Then we take these summarizations as the test samples and use a text classifier to classify, the results of classification show in Fig. 2.

The result proves that our method is more efficient than traditional summarization method. When the compression ratio is smaller than 10%, the precision of text classification on the summarizations abstracted by the two methods are almost the same, and the precision is very low because it loses a lot of information when compression ratio is too low. When the compression ratio is more than 10%, the precision of text classification by our method is better than the one by traditional summarization method. When the compression ratio is between 20% and 60%, the distance between two precisions is longer and longer. When the compression ratio is more than 60%, the distance begins to reduce. When the compression ratio achieves 70%-80%, the precision achieves tiptop, because under this compression ratio, our method not only maintains the main information, but also clears up the noise. When the compression ratio is 30%, the precision by our method is almost the same with the precision that use the original text as test sample to classify directly. So, 30% is an ideal compression ratio when we do automatic summarization. Not only does our method maintain the original information in the text, but also compresses the text.

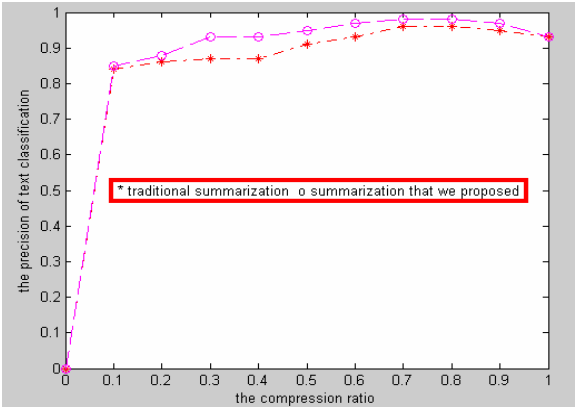


Fig. 2: The result of classification.

## 5.2. The precision of topic partition

Because the topic of a text is a subjective concept, there is not an objective standard for judgment. In order to test the precision of topic partition based on our method, we use a compromised method to test. We extract some text with short length from the language material corpus. Because the length of each text is short, we can suppose that they are in the certain semantic level and each short text belongs to a topic. We use these texts to compose a long text. So it will have an objective standard for these long texts to carry on topic partition.

If each paragraph of the text is assigned to the topic which it belongs to originally, then it is a correct partition, otherwise it is a wrong partition. The precision of topic partition are:

$$p = \frac{n}{N} \times 100\% \tag{7}$$

p is the precision of topic partition, n is the number of texts that have a correct partition; N is the number of long texts that we have made [10].

2600 texts with short length are chosen randomly from language material corpus according to the principle above, and these texts are used to compose 250 long texts. Each long text contains 2-7 short texts that belong to different types and the number of paragraphs in each long text is more than 10. Then another 250 long texts are composed, these long texts contain 2-7 texts that belong to the same type (they have the similar content). The result is shown in Table 1.

| Testing texts | correct partition | Number | precision |
|---|---|---|---|
| Long texts composed of texts that belong to same type | 146 | 250 | 58.4% |
| Long texts composed of texts that belong to different types | 219 | 250 | 87.6% |
| Total | 365 | 500 | 73% |

Table 1: The result of test.

From the Table 1, we can find that the precision of topic partition based on the cluster algorithm on the multi-subject article is 73%, and the precision of topic partition on the long texts composed of the short texts that belong to the different types is higher than the precision of topic partition on the long texts composed of the short texts that belong to the same types. The result conforms to the fact.

# 6. Conclusions and future work

This paper proposes a new summarization method based on the conceptual vector space model and clustering algorithm. This method can eliminate the situations that a word has several meanings and several words have the same meaning. It can also solve the problem that the assignment of summarization sentence is not balanced in multi-subject text. It has a certain coverage and completeness. The experimental results indicate that the method we proposed is more efficient than traditional ones, but the problem of summarization continuity has not solved yet, which will be done in our future work.

# Acknowledgement

We should thank http://www.keenage.com for the HowNet tools that it offers.

# References

[1] T. Liu and K.Z. Wang, Four Kinds of Main Methods of Automatic Abstracting, *Journal of Information*, 18:12-19, 1999.

[2] B. Jin, Y.J. Shi and H.F. Teng, Automatic Abstracting Technology and Its Application, *The Research and Computer Application*, 12:13-15, 2004.

[3] C.K. Sun, L. Li and X.L. Yang, Research and Implementation of Knowledge based Text Summarization Systems, *The Research and Development of Computer*, 37:874-881, 2000.

[4] T.S. Yao and Q.B. Zhu, *Nature Language Understanding: The Research of Human Language that Enables the Machine to Understand*, Published of Tsinghua, 2002.

[5] M. Wang, T.T. He and D.H. Ji, Automatic Chinese Text Summarization System Based on Conceptual Vector Space Model, *Journal of Chinese Information Processing*, 19:87-93, 2004.

[6] H. P. Luhn, The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 2:159-165, 1958.

[7] X.Z. Fern and C.E. Brodley, Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pp.186-193, 2003.

[8] P. Hu, *A Study of Chinese Text Summarization Based on Adaptive Clustering Algorithm*, M. S. Degree Thesis, Wuhan, 2005.

[9] H. Li, *An Abstract Method Based on Semantic*, M. S. Degree Thesis, Nanjin, 2004.

[10] J.L. Fu and Q.X. Chen, Study on Topic Partition in Automatic Abstracting System, *Journal of Chinese Information Processing*, 19:28-35, 2005.