A Method on Chinese Thesauri

Fu Chen[1(✉)], Xi Liu[1], Yuemei Xu[1], Miaohua Xu[2], and Guangjun Shi[3]

[1] Department of Computer Science and Technology, Beijing Foreign Studies University, Beijing, China
chenfu@servst.com
[2] Hebei Surveying and Mapping Institute, Heibei, China
[3] Computer Network Information Center, Chinese Academy of Science, Beijing, China

Abstract. In recent years, text analysis has become increasingly heated in many fields. And now, majority methods of text analysis are using Word2vec, Naïve Bayes or so on to classify the large number of texts. But for the text itself, not all samples are useful for some high-requirement researches and only use one keywords to get the related sample is definitely not enough. In this paper, we provide a novel model of second text filtering with Chinese Thesauri. It includes roughly 5 steps: sample collecting, thesauri establishment, word-segment algorithm, word-frequency statistics and the calculation of text relevance. Its main purpose is making the sample texts more accurate with the keywords which are input by the user and avoiding the needless time and space waste.

Keywords: Chinese thesauri · Text analysis · Semantic distance

1 Introduction

In every kind of language, the relation of word to word is obviously not totally isolated. And this is also the basic of all those algorithms of calculating semantic distance in order to classify a text into a kind of assay. In this report, we will build a tree of all Chinese thesauri. And we will use this tree to find the thesauri of the keyword which is given by the user. In the meantime, every word will be attached with a weight to show its similarity or correlation with the user's key words, and the value will be used in the later filter procedure.

After the calculating of the weight of every thesaurus, the next step is to statistics the number of the keyword and thesauri in each assay. And compared with the text analysis of English assays, Chinese assays seems more difficult because the word in Chinese is not segmented by blank space and only have a punctuation after an integrated sentence or sense-group. So, before all the text analysis of Chinese assays, they should be word segmented. Only after the word segmented can Chinese assay be used to statistics the word that you want to count. Then, use the number and weight to calculate the correlation value with the keyword that user input earlier. And considering the sum of words if different from assay to assay, the correlation will be divided by the sum of words. And in order to make the comparison more obviously, we choose to use the thousandth to express the assay's correlation. At last, the correlation value can be a

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 601–608, 2017.
DOI: 10.1007/978-3-319-59288-6_60

criterion of the second text filter. It can be very useful to get rid of the texts those are not up to the user's standard.

We choose a specific word Confucianism as the keywords which are supposed to be input by the users. First, we set a levels –tree based on the keywords' thesauri. Second, we calculate the thesauri' correlation value preparing for the latter calculation. This procedure is based on the shortest path of two nodes in a tree. Then we collect ten texts from Internet as our sample. And deal them with word-segment algorithm and word-frequency statistics. Finally, we use the value of words' correlation value, word frequency and the total number of word related with every sample text to gain the ultimate degree of correlation between each sample and the keywords. Figure 1. Shows the process diagram of our study.

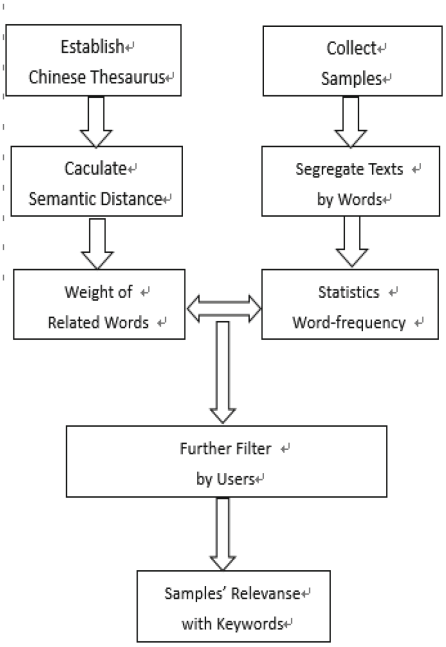

Fig. 1. The process diagram of our study

The remainder of this paper is structured as follow. In Sect. 2, we review and discuss the related work. Section 3 describes the establishment of the level-tree of Chinese thesauri and the algorithm of correlation value calculating between two words which is based on that level-tree. Section 4 shows the process of word-segment algorithm and word-frequency statistics in Chinese text. Section 5 is about the calculation of the relevance between the text and keywords. At last, we will talk about conclusion and future work in Sect. 6.

2 Paper Preparation

This section discusses related work on two key aspects of the article. Firstly, Subsect. 2.1 presents the origin and construction of Chinese Thesauri. Then, Subsect. 2.2 introduces the algorithm of text similarity computation.

2.1 Origin and Construction of Chinese Thesauri

A thesauri is a set of items (phrases or words) plus a set of relations between these items. And it can be divided to two types, manual and automatic. And in this paper, we will use the manual thesauri. There are two kinds of manual thesauri. The first are general-purpose and word-based thesauri like Roget's and WordNet. Those thesauri contain sense relations like antonym and synonym but are rarely used in IR systems. The second are IR-oriented and phrase-based thesauri like INSPEC, LCSH(Library of Congress Subject Headings), and MeSH(Medical Subject Headings). Those manual thesauri usually contain relations between thesaurus items such as BT(Broader Term), NT(Narrow Term), UF(Used For), and RT(Related To), and can be either general or specific, depending on the needs of thesaurus builders. This type of manual thesauri is widely used in commercial systems [1]. Thus, we chose the second kind of manual thesauri. But because of the expense, we only construct a part of the whole thesauri and just used for our experiment. Whenever the reader wants to use the algorithm in this paper, you can just construct your own thesauri as long as it can satisfy the requirement.

2.2 Algorithm of Text Similarity Computation

Displayed equations or formulas are centered and set on a separate line (with an extra line or halfline space above and below). Displayed expressions should be numbered for reference. The numbers should be consecutive within each section or within the contribution, with numbers enclosed in parentheses and set on the right margin.

For text-based semantic similarity, perhaps the most widely used approaches are the approximations obtained through query expansion, as performed in information retrieval (Voorhees 1993), or the latent semantic analysis method (Landauer, Foltz, & Laham 1998) that measures the similarity of texts by exploiting second-order word relations automatically acquired from large text collections [1]. In this paper, we choose to use the Chinese Thesauri to evaluate the text relevance with the keywords which is input by users. And it can be regarded as the second step after the calculating of word-frequency of the related words in texts and this step will be very useful to filter the samples more accurate.

3 Establishment of Chinese Thesauri

Thesauri in Chinese means the related words and alternative words of a keywords. And in the structure of level-tree, the parents node and brothers node can also shows special

relations of two words. So, we try to classify all the Chinese words into a huge level-tree. The principle of the structure is the inclusion relation and the coordinative relation. And the shortest path between two nodes is obviously an important character of the tree. In the field of text analysis, the shortest path can be used to calculate the correlation of two words, which is also used in this paper.

3.1 Establish the Level-Tree

In order to make the theory more easy-to-understand, we choose a keywords to start our experiment as an example. The keywords we chose is "Confucianism" and it is of course that the user can chose any else words as long as he need. And in order to establish the thesauri level-tree, we add some related words like "Confucius", "Analects"(The Analects of Confucius) and so on. Figure 2. Shows the level-tree, which is a part of the integrated Chinese thesauri, used in the example experience in this paper.

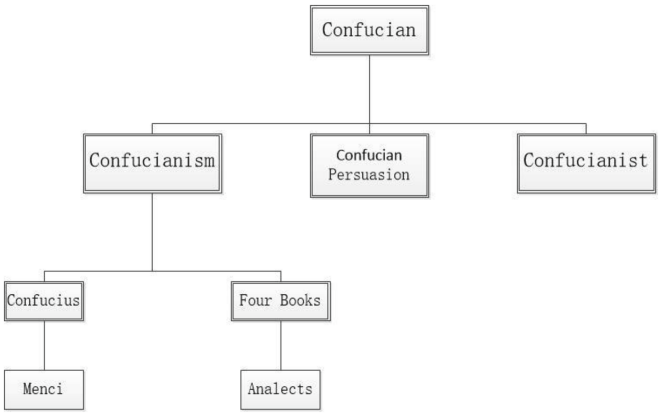

Fig. 2. The thesauri level-tree of our experience

After the establishment of the thesauri tree, we will use it to gain the correlation value of each related word. And based on the structure of level-tree, we choose to apply the Dijkstra algorithm to realize the process of traversing all the nodes, and get the shortest path between every related words and the keywords. Considering the principle of easy-to-understand and practical, we choose a simple formula as following show to calculate the semantic distance of the related words and the keywords:

$$\mathrm{Dist(Ci)} = 1/(2\text{^}\mathrm{Li}) \tag{1}$$

The Ci represent the related words. Li means the length of the shortest path of the related words and the keywords. Dist(Ci) shows the degree of the relation of word Ci and the keywords. We definite the weight of the keywords "Confucianism" is 1, and any other words' weight will be based on the semantic distance between it and the keywords. Using this formula, we arrive at the conclusion that the correlation of "Confucian", "Confucius", "Four Books" is 0.5 and for "Confucian Persuasion", "Confucianist",

"Menci", "Analects", the value is 0.25. And the weights of those related words are showed in Fig. 3.

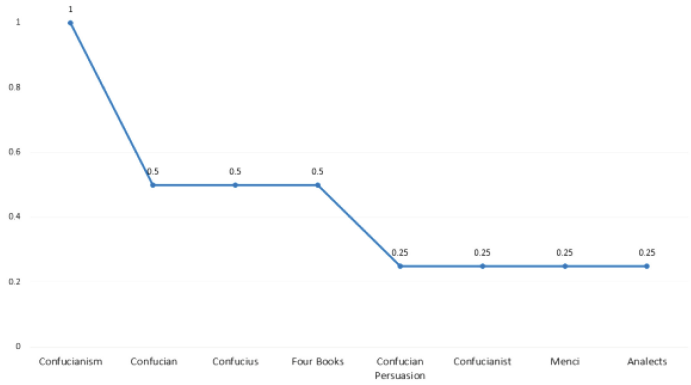

Fig. 3. The weights of related words

4 Word-Segment Algorithm and Word-Frequency Statistics

Differ from English text, the words in Chinese text don't have the blanks to divided from each other. It even ever made some trouble in the development of the text analysis in China. But not very long after the emerge of that problem, Jie-Ba algorithm was came up with. And in our experiment, we also choose to use it in order to divide the words in Chinese text. The Jie-Ba algorithm is based on the trie tree, which is a famous prefix tree. It includes more than 20,000 words, approximately cover the all common words that may be used in our ordinary life. And all these words are collected in a txt file. It is similar with the verb collocations in English, presenting the correlation between words and words like Fig. 4 shows.

And according to the trie tree, we are able to distinguish the divide method of the sample. Once the word matched and have no word longer than it can match, a blank will be added. And text will be segregated word by word. Jie-Ba algorithm provides three modes named full mode, default and search mode. Full mode can scan all the letters that can become a word, its speed is quick but can't solve the problem of ambiguity. The default mode is trying to segregate the text most accurately, so it is suitable for text analysis and we choose this mode for those reasons. As you can see by its name, the search mode fits the search engine and under this mode, text will be default segregated and long words will be cut again for a second time.

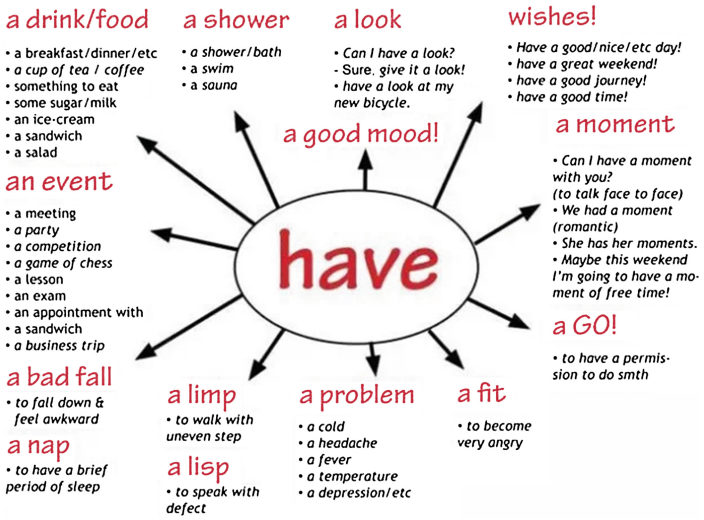

Fig. 4. The verb collocations in English

After word-segment, the text is divided word by word and segregated by blanks. So, it offers the condition of word-frequency and we don't need to worry about take the wrong match of words. And we use the same dictionary which is used by Jie-Ba algorithm to calculate the word-frequency in each sample text. It is easy to realize because it just need some cycles to match the words, gain the final data and output the result.

5 Calculation of Text Relevance

Provided the related words' weight and word frequency, the last step is to count the text relevance with the keywords. And we choose to use the following formula:

$$\mathrm{R(Assayk)} = \left\{ \sum_{i=1}^{n} [Dist(Ci) * Nk, i] \right\} / MK \tag{2}$$

In the formula, Assayk represent the text. 'n' means the number of keywords and related words. Ci is the NO.i word, N(k,i) means the sum of word Ci in Assayk. And Mk is the sum number of word in Assayk. After calculation, we arrive at the conclusion of following data, which is showed in Figs. 5 and 6:

R(text1)=21.09‰

R(text2)=0.89‰

R(text3)=14.97‰

R(text4)=16.59‰

R(text5)=36.52‰

R(text6)=17.94‰

R(text7)=4.56‰

R(text8)=22.94‰

R(text9)=10.30‰

R(text10)=17.16‰

Fig. 5. Text relevance of samples

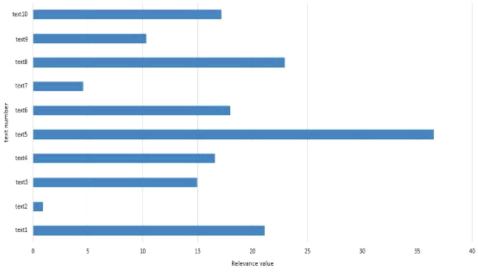

Fig. 6. Histogram of text relevance

And as the figure shows, every text is attached with its own value of the relevance between it and user's keyword. Using those data, we can offer the users a second select chance, limiting the relevance value of those texts that he gain from input the keywords in search engine only. And this step can further filter the samples to attain different users demand and class the texts, which contain the keyword, by the text relevance. It will be used in many high requirements research, because the texts is no more just texts but hold a relevance with the keywords in different degree.

6 Conclusion and Future Work

This paper presents a novel model to secondly filter the texts those are collected by input the keywords on the Internet. It is based on the Chinese Thesauri and related to the establishment of thesauri, word-segment algorithm, word-frequency statistics and the calculation of text relevance. Through the experiment, we validate the effectiveness and accuracy of our method. Using the Chinese Thesauri, we limit the relevance of all the samples, which can avoid the waste of analyzing the useless texts and make the samples attain the higher quality after the second filter.

Future word will include a number of aspects. Firstly, the existing thesaurus needs an more accurate standard and cover larger scale. It requires a more professional knowledge of linguistics and graph theory. Secondly, the language is always in the proceeding of change. At the same time, the Chinese Thesauri will also be changing. So, we need to continually update the structure of our thesauri into a more scientific framework. Finally, in order to make this model be used by not only the technical scholar but also the layman, we are supposed to get the entire algorithm into software. After this step, it will only require the user input the keywords and smallest relevance of the texts and keywords. Then, sample collecting, thesauri establishment, word-segment algorithm, word-frequency statistics and the calculation of text relevance will all be done automatically. It will largely enhance the features of using-friendly and efficient.

Acknowledgements. The research was supported in part by the National Science Foundation of China under No.61672104, 61170209, 61502038,U1509214;Program for New Century Excellent Talents in University No.NCET-13-0676. Key Program of BFSU 2011 Collaborative Innovation Center No.BFSU2011-ZD04.

References

1. Jing, Y., Crof, W.B.: An Association Thesauri for Information Retrieval (1994)
2. Mihalcea, R., Corley, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity (2006)
3. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods (2010)
4. Scott, S., Matwin, S.: Text Classification Using WordNet Hypernyms (1998)
5. Roberts, C.W.: Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcript. Lawrence Erlbaum Associates, Mahwah (1997)
6. Lacity, M.C., Janson, M.A.: Understanding qualitative data: a framework of text analysis methods. J. Manage. Inf. Syst. **11**(2), 137–155 (1994)
7. Stone, P.J.: Thematic text analysis: new agendas for analyzing text content. In: Roberts, C. (ed.) Text Analysis for the Social Sciences. Lawrence Erlbaum Associates, Mahwah (1997)
8. Lehnert, W., Sundheim, B.: A Performance Evaluation of Text-Analysis Technologies. www.aaai.org
9. Soergel, D.: Indexing languages and thesauri: construction and maintenance (1974). www.dsoergel.com
10. Wang, Y.-C., Vandendorpe, J., Evens, M.: Relational thesauri in information retrieval. J. Am. Soc. Inf. Sci. **36**(1), 15–27 (1985). America
11. Larsen, H.L., Yager, R.R.: The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. IEEE Trans. Syst. Man Cybern. **23**(1), 31–41 (2002)
12. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures (2001)