Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A METHOD OF PHASED INTEGRATED SEMANTIC SIMILARITY COMPUTATION

MA JUNHONG

Lecturer, Xi'an International University, Private Education Research Center, Shaanxi, China

E-mail: maxiaofei913@163.com

ABSTRACT

This paper proposes a method of phased integrated semantic similarity computation to improve the existing Chinese text algorithms. It computes text similarity in sections from the sentences，paragraphs，to the whole text. Combining with the characteristics of each section，the text semantic factors is also blended in each section，striving to the best accuracy. At last, we have established a similarity computing systems, compared the new method with other common methods in Chinese information processing. The experiments indicate that the improved algorithm has achieved better results.

Keywords: *Texts similarity, Vector Space Model, Semantic similarity, Term weight ， TF-IDF*

1. INTRODUCTION

At present, the network of information resources is growing, many of them is useless, redundant duplication of information ,not only takes up a lot of system space, but also makes the processing efficiency of the system to reduce. Accurate and effective text similarity computation method is one of the methods to solve text processing problems.One fundamental and important work in information processing is text similarity computation , which is the key technology in the textual data mining that related to many important application researches, for example, in the area of document copy text categorization, text clustering, information retrieval, question answering system, and etc. It is worth further research and discussion because of its wide applications.

The existing text similarity computing model has weaknesses such as deficiency in rationale and incompleteness in document properties fitting. Chinese text understanding and processing is more challenging relative to English counterpart[1]. Currently the most widely used method is based on the statistics of the traditional text similarity calculation[2]. Such as VSM used in Chinese text retrieval system, the cosine of the angle of the vector of the Euclidean space used in text classification similarity calculation[3]. However, such methods are complicated, computationally intensive, no semantically and structural relationships. Understanding Chinese language from the view of semantic is more appropriate than from the statistical method. Therefore, the method based on the semantic similarity is growing concern.

The basic idea of the semantic similarity calculation is departing from a simple algebraic statistics, excavations text deeper semantic representation, and makes related applications calculation more accurate[4]. This approach reduces the work of the corpus and the training sets, in particular, is more appropriate for the understanding of the Chinese text.

The aim of this paper is to improve the existing algorithms. For this aim, we compared in detail varieties of text similarity computation method in Chinese information processing and analyze their characteristics and defects. Furthermore, a new method of phased integrated semantic similarity computation, which is an improved method, has been put forward. This method blends in text semantic factors with each section, improves the Chinese text similarity computing efficiency and accuracy.It has certain applicability and feasibility.

This guide provides details to assist authors in preparing a paper for publication in JATIT so that there is a consistency among papers. These instructions give guidance on layout, style, illustrations and references and serve as a model for authors to emulate. Please follow these specifications closely as papers which do not meet the standards laid down, will not be published.

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. CHARACTERISTICS OF THE TEXT SIMILARITY COMPUTING ANALYSIS

2.1 Definition of text similarity in this paper：

The core of the text similarity comparison is comparing the difference between two of the given text (or byte stream, etc.), usually uses a number of [0, 1] as measure [5]. In different fields, the meaning of similarity degree is different, and its intellectual property also have certain effect, so This paper argues that the computation should not be judged solely by the similar words or similar sentences and paragraphs, should also contain semantic understanding. So we make the following definition:

Text similarity is to point to for a given two or more of the text, through the layers of computation of sentence、 paragraph to get the overall similarity between them, and at the same time contains a certain semantic similarity.

2.2 TF-IDF(Term Frequency-Inverted Document Frequency)

In VSM (Vector Space Model), the text is represented on the vectors which are composed of the feature. And the weight of feature is the text Vector Space coordinates[6]. Feature weights for text representation plays a very important role; also affect the text similarity computation results. At present the commonly used method is TF-IDF, IG (Information Gain), MI (Mutual Information), DF (Document Frequency)、χ2(CHI),and etc. Among these methods, IG ,CHI,and MI belong to supervised feature selection algorithm, DF belong to unsupervised algorithm which can be directly used in clustering.

TF-IDF can effectively distinguish the high frequency words and the low differentiate words. In computing, the two aspects are the major considerations:

(1) This feature‘s affects in a text (Usually measured by frequency, namely the TF)

(2) This feature‘s affects in the whole text set (can distinguish between different text, namely IDF)

If L is text total of text set, w represents the weight, then[7]:

$$IDF_t = \log_2 L/DF_{ij} \quad (1)$$

$$w_{ij} = TF_{ij} \times IDF_{ij} = TFij \times \log_2 L/DF_{ij} \quad (2)$$

According to different normalized processing, the use of TF - IDF formula is different.This method is the most commonly used weighting method, can effectively distinguish between high frequency words and low of the degrees of distinction word.

3. METHOD OF PHASED INTEGRATED SEMANTIC SIMILARITY COMPUTATION

3.1 Semantic strengthened weight computation method

TF – IDF is a method based on word frequency. If a word in text preprocessing stage is selected as the feature words, then in the computation of the word frequency the rest of selected words will be indiscriminate treated equally[8]. This method considered is too one-sided, does not take into the semantic factors of context or application fields, etc. Therefore, we propose an improved computation method: In text preprocessing stage it can be combined with the text theme, scope, applications and so on, to give the term weight as f.

At first, using the traditional TF-IDF formula to calculate a term weight wij, then calculated using the formula 3:

$$w'_{ij} = \beta_1 * w_{ij} + \beta_2 * f_i \quad (3)$$

Here introduces two parameters: β1and β2, β1+β2=1. The weights of wij calculated with the traditional formula still holds larger proportion, but need to be adjusted with the semantic factors. The former accounted for by β1, usually be 0.8 or 0.9; the latter accounted for by β2, typically be 0.2 or 0.1refering to β1. And fi will be set according to the specific needs, such as combining with the detection of text classification, subject or keywords, term weights in the field of f can be set to 1, the rest of the non - in terms can be to 0.8, general word can be set to 0.3.

3.2 semantic similarity computations steps

The algorithm based on the semantic understanding often needs to build a new text representation model, through the judgment between text semantic distance, semantic relevance to achieve. At present, Hownet words semantic network has more mature, with the aid of Hownet knowledge structure and the knowledge representation, this paper presents a new fusion Method of phased integrated semantic similarity computation.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The First work is the text level division, namely the text is divided into paragraphs, paragraphs are divided into sentences, sentences into words; secondly , phased calculating the similarity of words, sentences, paragraphs to obtain the text similarity. At each stage the semantic similarity computation has been blended in, completing with the combination of local to global.

The first step: word similarity computing

In Hownet knowledge structure, because the concept is composed of primarily , then the similarity calculation of the words can be seen as the meaning of the primarily similarity calculation. And primarily hyponymy form a tree hierarchy, can calculate by the way of calculating the semantic distance.

General uses the following formula :

$$Sim(p1, p2) = \frac{a}{d + a} \quad (4)$$

Among them, the pl and p2 said two primarily , their semantic distance is d, on behalf of the p1 and p2 path length. a is a adjustable parameters, its value is 0.5 which is on behalf of the similarity of path length.

when computing the primary similarity we consider their relative position on the primary classification tree[9], get the following formula:

$$Sim(p1, p2) = \frac{a \cdot h_1 \cdot h_2}{d + a} \quad (5)$$

That meanings consider the explanatory relations of the two primarily to amend the primary distance, when computing the primary distance.

Second steps: Sentence Similarity Computing

Using the improved TF - IDF feature extraction algorithm, Sentence L1 (including n feature words) and L2 (including m feature words) is represented as a vector form of the following:

$L=(w_1, w_2, \cdots, w_n)$

$R=(w_1, w_2, \cdots, w_m)$

H (L, R) is the two sentence similarity matrix :

$$\begin{bmatrix} S(L_1,R_1), S(L_1,R_2), \ldots, S(L_1,R_m) \\ S(L_2,R_1), S(L_2,R_2), \ldots, S(L_2,R_m) \\ \ldots \quad \ldots \quad \ldots \\ S(L_n,R_1), S(L_n,R_2), \ldots, S(L_n,R_m) \end{bmatrix}$$

Sentence similarity computation formula of the L and R:

$$S(L,R) = \frac{\sum_{i=1}^{n} \max(S(L_i,R_1), S(L_i,R_2), \ldots, S(L_i,R_m))}{n} \quad (6)$$

In order to accurately get the similarity of these two sentences, we calculate Sim (L, R)and Sim (R,L) by the formula 5 , then take the average of the two, ensuring the uniqueness of the calculated results.

The third step: paragraph similarity computing

Suppose there are two paragraphs X and Y, X is divided into t sentences, Y is divided into d sentences, the computation formula is as follows:

$$S(X,Y) = \left(\frac{\sum_{i=1}^{t} \max(S(X_i,Y_1), S(X_i,Y_2), \ldots, S(X_i,Y_d))}{t} \right) * 0.5 + \left(\frac{\sum_{i=1}^{t} \max(S(Y_i,X_1), S(Y_i,X_2), \ldots, S(Y_i,X_d))}{t} \right) * 0.5 \quad (7)$$

In algorithm design, these sentences have been calculated similarity in paragraphs, these words have been calculated similarity in the sentences, are no longer calculate again, directly set the previous value to improve the efficiency.

The fourth step: Text Similarity Computing

The computing method mentioned above means taking the maximums to add up, and then takes an average, which is equal treatment to the words, sentences in the paragraph[10]. Considering that according to the paragraphs position of the text, its importance, namely the influence of the text is also different, should be given different weight. If the key paragraphs are similar, the whole text similarity increases.

Suppose D1, D2 are two texts to be comparing, D1 has m paragraphs; D2 has n paragraphs. In order to convenience, unified use X expressed a paragraph in D1; use Y expressed a paragraph in D2. According to the formula 6 the calculated paragraphs similarity recorded as: S1 (X, Y), S2 (X, Y)... Sm (X, Y), SS (X, Y) said the new paragraph similarity, there is the formula:

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$SS_i(X,Y) = \beta_1 * S_i(X,Y) + \beta_2 * w_i \quad (8)$$

In this formula, $\beta_1+\beta_2=1$, w_i is a paragraph weight whose specific value can be set according to the requirement. This paper takes 0.8 β_1 and 0.2 β_2, the key paragraph weight w_1 is 1.1, the normal paragraph weight w_2 is 1.

According to the formula 8 , D_1 and D_2 similarity computation formula is as follows:

$$Sim(D_1, D_2) = \frac{\sum_{i=1}^{m} SS_i(X,Y)}{m} \quad (9)$$

In addition, in the actual application process, considering the long text paragraphs usually is more and a lot of low similarity is almost zero impact to the whole text, it can set a similarity threshold in advance which paragraphs lower than the threshold value are no longer be joined the weight computation so as to reduce the overhead of the system, further improve the efficiency.

4. COMPOSITION AND IMPLEMENTATION OF SYSTEMS

4.1 modules and flow chart

stem includes a text preprocessing atic w), the text vector repr ure vector selection, and similarity computation, mainly composed of the following modules:

(1) Text library module: for storage the Chinese texts to computing similarity;

(2) Word segmentation dictionary module: management and maintenance the dictionary for Chinese word segmentation;

(3) Chinese word segmentation module: for a given Chinese text word segmentation processing and ambiguity correction[11];

(4) Feature extraction module: according to the word segmentation results and statistics of word frequency analysis, extract delegate to compare text feature vector and each entry contains the corresponding weight;

(5) Similarity computation module: Computation of the text feature vector similarity, then according to the given Simmin threshold to determine the similarity between texts, and obtained the computation results.

The workflow of the system is four steps as follows:

(1) taken out two text from the prepared text set;

(2) Use the Chinese word segmentation tool,IK Analyzer 3.2 for word processing;

(3) Extract text feature vectors and determine the weights,

(4) Depending on the different similarity algorithm steps to calculate the similarity between the texts in set.

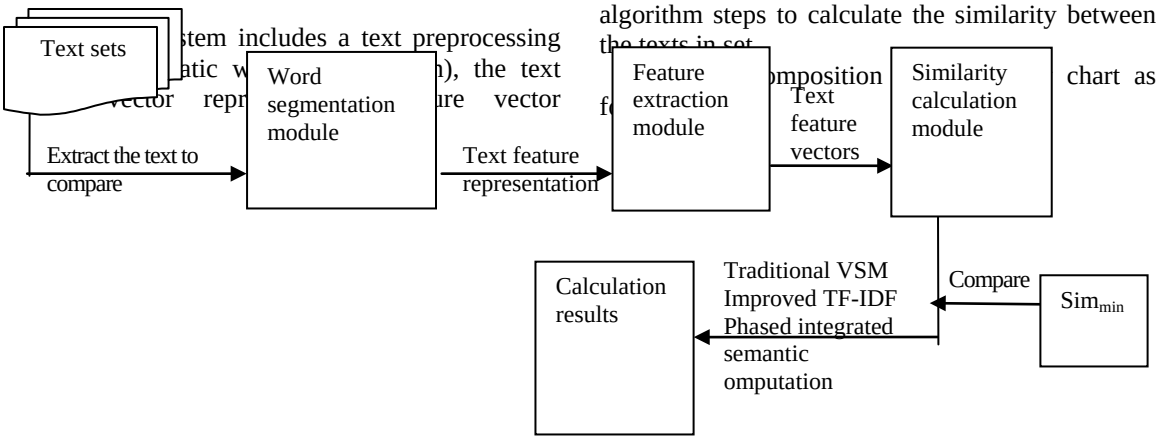

Figure 1: Text similarity computing system and flow chart

4.2 Experiments and results analysis

4.2.1 Data sets of the experiment

This experiment uses the test data set of a research unit which is selected from 40 texts about computer science, Chinese language and literature, economic management fields, and according to the different length divided into 3 groups, respectively

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

using three kinds of similarity computation method to test: the traditional method based on VSM, the semantic strengthened TF-IDF and the phased integrated semantic computation. The Traditional VSM without semantic similarity processing is calculated by the cosine coefficient.

In the traditional method of VSM, We use ($T1_1$, $T1_2$, ..., $T1_n$) and ($T2_1$, $T2_2$, ..., $T2_n$) as the vector of the text D1 and the text D2[12], there is a formula as follows:

$$Sim(D1, D2) = \cos\theta = \frac{\sum_{i=1}^{n} T1_i * T2_i}{\sqrt{\sum_{i=1}^{n} T1_i^2} * \sqrt{\sum_{i=1}^{n} T2_i^2}} \quad (10)$$

The selected text length distribution is as follows:

Table 1: Text length distribution

Grouping Numbers	Text length range	Text counts
1	1000~2000	12
2	2000~3000	16
3	3000~5000	12

4.2.2 The Experimental Results

In view of the preceding discussion, we use precision and recall as evaluation criteria, similarity threshold is set to 0.1, greater than this threshold will be regarded as similar text.

$$precision = \frac{\text{number of correctly detected similar texts}}{\text{number of all the detected similar texts}} \quad (11)$$

$$recall = \frac{\text{number of correctly detected similar texts}}{\text{number of actual existence of similar texts}} \quad (12)$$

The first set of experiments the results are as follows:

Table 2 :Comparison of similarity computation method 1

assessment method	traditional VSM	semantic strengthened TF-IDF	Phased integrated semantic computation
precision	0.7125	0.6831	0.8649
recall	0.5573	0.7597	0.8712

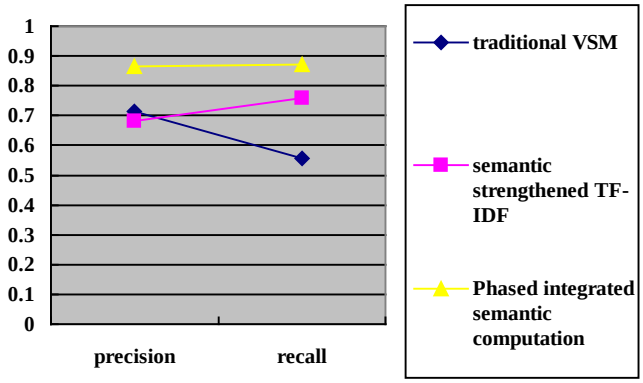

Figure 2.: Comparison Of Similarity Computation Method 1

The second experimental results are as follows:

Table 3:Comparison of similarity computation method 2

assessment method	traditional VSM	semantic strengthened TF-IDF	Phased integrated semantic computation
precision	0.6751	0.7142	0.7964
recall	0.5567	0.7326	0.7783

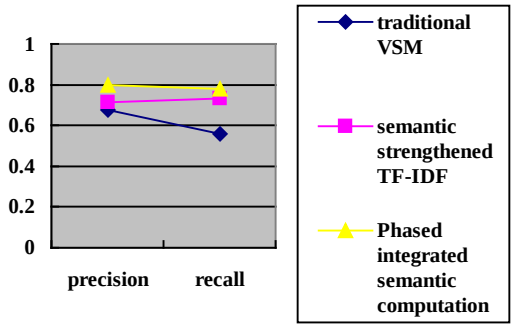

Figure 3.: Comparison Of Similarity Computation Method 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The third groups of experiment results are as follows:

Table 4:Comparison of similarity computation method 3

Assessment method	Traditional VSM	Semantic strengthened TF-IDF	Phased integrated semantic computation
precision	0.7294	0.7359	0.7441
recall	0.7516	0.7462	0.7597

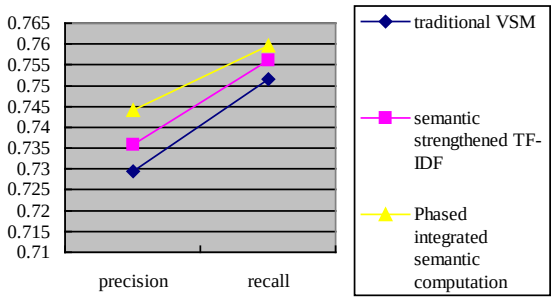

Figure 4.: Comparison of similarity computation method 3

5. CONCLUSION

According to the two formulas of precision and recall rate to analysis,When precision rate is very high but the recall rate is very low, that means the system find similar text percentage is high, the number of miscarriage of justice is less, but not comprehensive find that there are many similar text did not find out; On the contrary, if the recall rate is high and the accuracy is low, that means looked Many non-similarity of the text as similar , which caused a lot of miscarriage of justice, but the actual existence similar texts mostly have been found out. Both cases are not ideal,only both considerable to show that the texts been found out is really similar,and have the actual existing similar text basically all find out.

As it can be seen from the three groups of experimental data, the computation method of this paper whose precision rate and recall rate is fairly basic, while the other two algorithms are slightly some deviation. Compared with the traditional VSM based on cosine algorithm, this method has better effect, and the semantic strengthen TF - IDF algorithm also raised some computational efficiency. But with the increase of text length, the efficiency of our algorithm is reduced somewhat, as can be seen for smaller length of text similarity computing, the algorithm works better.

Chinese text similarity calculation process is very complicated, and there are many uncertainties in the specific application. How to establish a better semantic model, and apply it to more specific fields, such as clustering, automatic abstract, is the next step of our research focus.

REFERENCES:

[1] Salton G,"The state of retrieval system evaluation", *Information Processing and Management*, 441-449, 1992

[2] Yun Peng, Hongxin Wan,"Web Text Clustering and Evaluation Algorithm Based on Fuzzy Set",*JDCTA: International Journal of Digital Content Technology and its Applications*, Vol. 7, No. 1, pp. 11 ~ 18, 2013.

[3] Ong Siou Chin, Narayanan Kulathuramaiyer&Alvin W.Yeo. "Automatic Discovery of Concepts from Text". *In Proceedings of IEEE International Conference on Web Intelligence*, Washington.DC, USA,pp : 1046~1049,2006

[4] Julian Sedding, Dimitar Kazakov. "WordNet-based Text Document Clustering",I*n Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data* , USA,104-113,2010.

[5] Hui-lin Tan, Xian-feng Liu. "The Research of The Selection of Knowledge Reasoning Method Based on Genetic Algorithm ". *Computer Knowledge and Technology*,Vol. 34, No. 12, pp. 55 ~ 59, 2011.

[6] Xi-Qian Jin, "Chinese Text Similarity Algorithm Research Based On Semantic Similarity", Zhejiang University of Technology, 2011.

[7] YAN Wei,Zhang Jiazhong,Wei Daisen,Wang Xiangcheng,Zheng Weibo, "A Novel Self-adaptive Case-based Reasoning Technique to Predict Business Failure",*IJACT: International Journal of Advancements in Computing Technology*, Vol. 4, No. 23, pp. 376 ~ 384, 2012.

[8] ILAMPIRAY.P, "Efficient Resource Utilization of Web Using data clustering and association rule mining", *JATIT:Journal of Theoretical and Applied Information Technology* ,Vol. 37 No.2 ,pp.211~216,2012.

[9] Xin Xu, "Research on Parameters Correlation and Optimization in Text Similarity

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Measurement Specialty" , Central South University,Chang Sha Hunan P.R.C, 2010.

[10] Jing-Fan Wang, "Two-Step Job Information Retrieval based on Document Similarity",Tsinghua University,Beijing, 2008.