Paper Similarity Detection Method Based on Distance Matrix Model with Row-Column Order Penalty Factor

Li Jun 1, 2
1. Hubei University of Technology, Wuhan, China
2. HuaZhong University of Science and Technology, Wuhan, China
Email: 50331615@qq.com

Han Yaqing
Hubei University of Technology, Wuhan, China
Email: 7137100@qq.com

Pan Junshan
HuaZhong University of Science and Technology, Wuhan, China
Email: 58510148@qq.com

Abstract—Paper similarity detection depends on grammatical and semantic analysis, word segmentation, similarity detection, document summarization and other technologies, involving multiple disciplines. However, there are some problems in the existing main detection models, such as incomplete segmentation preprocessing specification, impact of the semantic orders on detection, near-synonym evaluation, difficulties in paper backtrack and etc. Therefore, this paper presents a two-step segmentation model of special identifier and Sharpley value specific to above problems, which can improve segmentation accuracy. In the aspect of similarity comparison, a distance matrix model with row-column order penalty factor is proposed, which recognizes new words through search engine exponent. This model integrates the characteristics of vector detection, hamming distance and the longest common substring and carries out detection specific to near-synonyms, word deletion and changes in word order by redefining distance matrix and adding ordinal measures, making sentence similarity detection in terms of semantics and backbone word segmentation more effective. Compared with the traditional paper similarity retrieval, the present method has advantages in accuracy of word segmentation, low computation, reliability and high efficiency, which is of great academic significance in word segmentation, similarity detection and document summarization.

Index Terms—Paper Detection; Segmentation; Similarity Comparison; Distance Matrix; Model Modular-2 Arithmetic

I. INTRODUCTION

With the development of network and information technology, information sharing and dissemination is becoming more and more convenient, since it has provided an ideal platform for academic researchers. However, it also bred a series of academic misconducts, such as plagiarism, practicing fraud and so on. Besides complete plagiarism, cases like plagiarizing others’ achievements by transposition, paraphrasing and synonym replacement are also very common. These behaviors cause damage to academic quality. Consequently, aiming at eliminating these academic misconducts and improving the quality of paper, paper similarity detection which depends on word segmentation, similarity detection, document summarization and other means are quite significant.

Since word segmentation is the basis and premise of Chinese text similarity calculation, and the adoption of segmentation algorithm of high efficiency can greatly improve the accuracy of text similarity calculation, it is essential to keep on exploring new segmentation algorithm on the basis of existing ones, and improve the integrity and accuracy of word segmentation so as to make the comparison of similarity between texts more accurate, and thus to provide decision support for related business.

Since Chinese has a large volume and variable semantics, copy detection of Chinese paper is more complex and difficult than that of English paper, and the basic resource (such as corpus) for Chinese language processing such as text detection is relatively insufficient, which makes it impossible to apply mature technology and research achievements abroad directly. As a result, information processing of Chinese by computers is more difficult than that of western languages. Meanwhile, there are significant differences between Chinese and English segmentation because Chinese text is a continuous string of large character set, which means no specific separation mark exists among Chinese words, while English text is a fully separated string of small character set by space. Therefore, similarity comparison between Chinese texts
must be inspected based on segmentation systems, and the accuracy of word segmentation determines that of paper similarity calculation. Despite the achievements of word segmentation algorithm, there still exist problems as follows:

1. Disunity in segmentation specification and inconsistency in dictionary design;
2. Incompleteness of segmentation algorithm;
3. Insufficient ambiguity correction mechanism.

Based on similarity calculation, paper similarity detection manages to calculate the similarity between texts automatically. The calculation of text similarity has been widely applied in such fields as retrieval, machine translation, question answering systems and text mining, which is a basic and key problem and has long been a hot topic and difficulty for researchers. At present, many scholars both at home and abroad are studying text similarity calculation problem and have put forward some solutions.

In 1993, Manber [1] from Arizona University proposed the concept of approximate exponential, which was the earliest detection technology and was adopted to measure string similarity among documents. Later, Bao Junpeng [2] and others proposed document copy detection method based on semantic sequence which emphasized position information of words and improved the detection accuracy. Afterwards they proposed corresponding detection model (SSK) [3] which is suitable for copy detection without replacement of words.

In 2005, Jin Bo [4] from Dalian University of Technology firstly extended the scope of text similarity calculation into paragraph based on HowNet semantic similarity. Then he extended paragraph similarity calculation into chapter and proposed text (including words, sentences and paragraphs) similarity calculation formula and algorithms. In addition, he proposed long text similarity copy detection algorithms [5] in 2007 which made it possible to calculate the coverage of similar semantic sequence set by semantic sequence similarity relation and choose each overlap values with the minimum entropy for plagiarism recognition and retrieval.

Hung Chenghui et al. [6, 7] from Zhong Shan University proposed a text similarity calculation method by combining word semantic information with TF-IDF where sentence was regarded as vector space composed of independent entry, and the matching problem of document information was transformed into that of vector in the its space and then the similarity was obtained by dot product method and cosine method. However, TF-IDF which is based on vector space model also has disadvantages. Firstly, it is a statistical-based method and only when the text includes reaches certain length can some words occur repeatedly, which will then reflect its statistical results. Secondly, this method only takes statistical lexical features in context without the semantic information, and thus produces some limitations. Just as Salton form Cornell University says, there is no strict theoretical basis for calculating vector similarity by angle cosine.

To enhance the performance of result merging for distributed information retrieval, a novel merging method was put forward by Wang Xiaohong [8] from Jiang Su University and others, which was based on relevance between retrieved results and query. In 2009, Nie Guihua [9] from Wuhan University of Technology proposed systematic frame model of ontology-based thesis copy detecting system which described framework of paper copy detecting system from three layers: ontology access layer, ontology represent layer and ontology map layer. Semantic and ontology technology was utilized to discuss the build of paper ontology and calculation of paper similarity.

Zhang Huanjiong [10] from Beijing University Of Posts And Telecommunications constructed a new formula to calculate text similarity from Hamming calculation formula based on Hamming distance theory. It has the advantage of simplicity and rapidity but ignores the effects of unequal-length text and word order.

Later in 2011, Chen Yao-Tsung [11] from Tai Wan University proposed a method adopting chi-square statistics to measure similarities of term vector for texts, which reduces the miss rate of similarity detection.

Also, Sánchez-Vega [12] from Mexico proposed a rewriting exponential model based on finite state machine which manages to detect common actions performed by plagiarists such as word deletion, insertion and transposition, but fails to deal with near-synonyms.

Through above analysis, we can still find some problems in main detection methods at present as follows:

1. The statistical property of words in the context is taken into consideration while semantic information is ignored.
2. The effects of unequal-length text and word order are not fully considered.
3. There is limitation in recalling paper source by backtrack.

On the other hand, similarity calculation has different requirements for different applications, such as paper-level, paragraph-level, sentence-level, word-level and morpheme-level, and is often represented by formula or model of similarity calculation. However, with the evolution of plagiarism mechanism and diversification of expressions, problems such as large computation and difficulties in feature extraction and correct understanding of evaluation standards in paper-level and paragraph-level similarity detection still exists. As for word-level similarity, it may deviate greatly from actual value because of its small particle size. In addition, the space and time complexity of detection algorithm are quite large, making it difficult to obtain ideal effect from its application into huge amounts of paper.

A. Contribution

We summarize our contribution as below:

Starting with word segmentation and text similarity research, through integrating several main detection models at present, this paper builds the model specific to plagiarism, replacement of words and transposition based on the analysis of existing detection technologies.
Firstly, a two-step segmentation model is proposed, where we segment words by properties in step one and establish cooperative game model for roots in step two to calculate Sharpley Value, maximizing N-GRAM value.

Secondly, an evaluation score model of new words and popular words in search engine is proposed to make corpus of segmentation preprocessing more complete.

Finally, a distance matrix model with row-column order penalty factor is proposed. This model integrates the characteristics of vector detection, hamming distance and the longest common substring and carries out detection specific to near-synonyms, word deletion and changes in word order by redefining distance matrix and adding ordinal measures, making sentence similarity detection in terms of semantics and backbone word segmentation more effective and the returned results of retrieval better meet the requirements of users. Different from the traditional paper similarity retrieval, the present method adopts modular-2 arithmetic with low computation whose effectiveness and practicability have been confirmed by corresponding experiment.

II. DISTANCE MATRIX MODEL WITH ROW-COLUMN ORDER PENALTY FACTOR

A. Segmentation Based on Special Identifier

In the process of text similarity calculation, segmentation preprocessing must be carried out in the first place, since participial accuracy determines that of paper similarity calculation. There are several common algorithms for Chinese segmentation at present, such as unsupervised segmentation, mechanical word segmentation based on dictionary, segmentation based on linguistic model and segmentation based on character tagging. The first algorithm judges the correlation between two characters by computing their mutual information in corpus, which has good effect on high frequency words but is affected by threshold coefficient [13]; the second one does not have universality because it is affected by specialties of dictionary, and the third algorithm is the most commonly used method which is under research and development. Generally, Chinese Segmentation is confronted with 2 challenges: ambiguity and identification of new words.

According to the recently published Chinese Grammar Questions (Another version of Chinese Grammar) compiled by Professor Xing Fuyi, Chinese words are classified into three categories and eleven small classes, in which nouns, verbs, adjectives, numerals, quantifiers and pronouns are classified as notional words and identification of new words, popular words and common ambiguous segments can be conducted with the help of search engine has unique advantages in processing new words and common ambiguous segments because of its wide applications. Naturally, assessment of new words and popular words can be conducted with the help of search engines’ results.

$$\text{Value(Word X)} = \frac{\text{Number of pages on Baidu}}{100,000,000} \times k_{\text{search}}$$

Within which, N represents the number of pages on Baidu and the denominator 100,000,000 is the maximum number of pages after searching the commonly used words on Baidu.

D. Sentence-level Corpus and Digital Signature

Automatic summarization is divided into key phrase
extraction and text generation, which is used to explain and summarize the content of the original text with a brief description so as to accelerate people understanding information and solving the problem of information overloaded effectively. On the basis of statistics, paper-level and paragraph-level summary is composed of key sentences extracted from the text by Bayesian method and Hidden Markov Model according to word frequency and position information of the text, which is mainly suitable for technical documents with standardized format, and for other types of document, it needs further improvement in plagiarism-detection. Because of its unidirectional mapping mechanism, summary method is of good effect on complete paragraph plagiarism, which is rarely used in actual situation. In addition, it is of great complexity to extract key information with representativeness for paper-level and paragraph-level summary. However, the theory of summary can be used to create a digital signature which is composed of paper title and author information for each sentence and add it to sentence-level corpus for the subsequent backtrack.

A paper is composed of multiple sentences and after the segmentation is conducted, each sentence can be regarded as a set of words and then a sentence-level similarity comparison of the paper to be detected with the compared paper will be conducted. Firstly, the segmentation results of compared paper will be added into database, and secondly each sentence in the paper to be detected with that in the database compared. In order to look up the paper source by backtrack, a digital signature composed of paper title and author information is appended to each sentence. Thus in the similarity comparison process, if two sentences are judged to have a high similarity, their corresponding papers can be retrieved quickly.

E. General Similarity Comparison Model with Sequence Factor

VSM (such as IF-IDF) transforms matching problems of document information to that of vector in the vector space with a comprehensive consideration of term frequency (TF) in all texts and the term’ ability to distinguish catalogs (IDF) [14, 15]. Hamming Distance adopts modular-2 arithmetic, avoids a mass of multiplication during similarity process in Euclid space, and obtains a high calculation speed. All these methods have their advantages, but at the same time, they have such problems as slow paper backtrack and ignorance of word order. For instance, in the following sentences: (a) “lexical collocation methods are various”; (b)” lexical types are various”; (c)” English lexical collocation methods are various”, the similarities which are obtained by VSM, Hamming Distance and the longest common substring, respectively, are approximate if there is no consideration of word order, but obviously, the similarity of sentence a and sentence c is higher than that of sentence a and b. So comparison of sentence-level similarities with consideration of word order will obtain better effects.

Distance matrix model with row-column order penalty factor in this paper integrates the characteristics of VSM, Hamming distance and the longest common substring, and realizes the transformation among these models through different parameter configurations.

In this model, paper to be detected, called A, and compared paper, called B, are divided into two sets of sentences, and each is a segmentation vector. So after the division, paper A is described as: LA={ A_1, A_2, A_3, … A_n }, paper B is described as: LB={ B_1, B_2, B_3, … B_k }, and if the two vectors have a different number of elements, we can add some empty string elements into the one with fewer elements until it has as many elements as the other, and then conduct the following calculations:

$$\text{Sim}(LA, LB) = \frac{LA \otimes LB}{\sqrt{\sum_{i=1}^{n} (LA \otimes LB)_i^2}}$$

$$LA \otimes LB = \left( \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \right) \otimes \left( \begin{array}{c} B_1 \\ B_2 \\ \vdots \\ B_k \end{array} \right)$$

|LA \otimes LB| is the maximum number of words that are identical in two sentences, which is similar to the longest common substring, and the difference is that the effects of multiple identical substrings are taken into consideration in this method. XOR is adopted in the calculation of $LA \otimes LB$, whose result is an N-order sparse matrix. When the element from vector LA is multiplied by the element form vector LB, the result will be 1 if they are identical. Otherwise it will be 0. The first non-zero elements in rows are counted to create a row-sequence table, called TA, saving the row numbers of those elements that have identical elements in LB, and a column-sequence table, called TB, saving the column numbers of those elements that have identical elements in LA. TA and TB represent the word orders of the longest substrings in LA and LB, respectively. Then the similarity of TA and TB is compared by utilizing Euclidean distance and the following function $f(TA, TB)$ is called:

$$f(TA, TB) = \sqrt{\sum_i (TA_i - TB_i)^2}$$

In which, i is the serial number of TA and TB, n is the elements number of TA and TB, and the result of the function will be regarded as the row-column order penalty factor.

III. EXPERIMENTS

A. Segmentation Preprocessing

Experimentally, we select a news page randomly, and conduct individual segmentation by reference to our segmentation model. As shown in Figure 1 (two phases presented), phase 1 refers to the segmentation result according to special identifiers whereas phase 2 indicates the result of Sharpley value segmentation. After these two phases, the accuracy rate reaches 97.2%, and the recall rate 90.2%, satisfactory.
B. The Recognition of New Words and Hot Words

When we input “Gangnam style” and “Gannam style” into Baidu search engine for new words recognition, the former turns out to be: “appears in 17,000,000 relevant pages”, and $k_{search}$ is 18403, while the latter turns out to be 1,790, and $k_{search}$ is 10. According to evaluation score, the former should be a hot word.

\[
\text{Value(Gangnam style)} = \frac{17,000,000 \times 18403}{100,000,000} = 3128.5
\]

\[
\text{Value(Gannam style)} = \frac{10 \times 100,000,000}{1790} = 0.000179
\]

Through repetitive analysis of the result, those words whose values exceed 1000 could be added into vocabulary as hot words.

C. Research on Sentence-Level Corpus and Digital Signature

This paper adopts MD5 algorithm in experiment, which can process messages of any length and produce a message summary with a fixed length, so the message summary obtained through the input of paper title and author information is added into the sentence-level corpus as a digital signature. Figure 2. shows the sentence-level summary.

D. General Similarity Comparison Model with Sequence Factor

The following words are selected to analyze similarities:

we entertained Beijing visitors passionately
(segmentation results of compared sentence in corpus)

Beijing visitors passionately entertained we (Pa: sentence to be detected)

we passionately entertained Beijing visitors (Pb: sentence to be detected)

After segmentation, the sentence vectors are:

\[
s = [101 \ 203 \ 254 \ 357 \ 656];
\]

\[
a = [254 \ 357 \ 656 \ 203 \ 101];
\]

\[
b = [101 \ 656 \ 203 \ 254 \ 357];
\]

"*" means there may be other words among the given ones. Through the vector multiplication and deletion of rows of all zero members, a matrix is obtained as follows:

\[
\begin{bmatrix}
254 \\
357 \\
101 \\
203 \\
656
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 \\
0 \\
1 \\
0 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 \\
0 \\
0 \\
1 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
1
\end{bmatrix}
\]

From the result, TA of sentence \(a \otimes s\) can be expressed as: \((3, 4, 5, 2, 1)\) and TA of sentence \(b \otimes s\) can be expressed as: \((1, 5, 2, 3, 4)\). With the assumption that word order of sentence Ps is: \((1, 2, 3, 4, 5)\), the word order curve graph of three sentences is shown in Figure 3. Figure 4. and Figure 5. show Euclidean distances of Pa and Pb, respectively.

After calculation, the results of \(f(Pa, Ps)\) and \(f(Pb, Ps)\) are: 5.6569 and 3.4641, indicating that penalty factor of Pb is smaller than that of Pa. Namely, similarity between Pb and Ps is higher than that between Pa and Ps, which obviously agrees with manually assessing results, under
the condition that the number of words that are the same between Pa and Ps equals to that between Pb and Ps. If \( f(TA, TB) \) is 1, which means word order is ignored, the model evolves into VSM model with Dice coefficient and if the weight of each word is ignored and multiplication is set as the algorithm, this model evolves into hamming distance model, where \( k \) is a constant related to vector, if \( k \) is set to:

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} ^{ij} \text{and the number of words that are exactly the same in the compared sentence and the detected sentence is defined as the rank, this model evolves into the longest common substring under the condition that word order is ignored. Otherwise it is the longest common substring affected by order penalty factor.}
\]

E. Processing of Replacement of Words and Paraphrasing

For the phenomenon of common replacement of words and paraphrasing, near-synonym processing is adopted, and the corresponding dictionary of near-synonyms, which are mainly from Chinese Synonyms Dictionary and those that are the same word in English, is build. In calculation of (2), the result of multiplying two near-synonyms is 1, the same as that of two identical words. For instance, “increase” and “improve” are of the same meaning, if two sentences include “increase radiation” and “improve radiation”, respectively, the multiplication result of the two words will be 1 after they are judged to be near-synonyms through the dictionary of near-synonyms.

IV. CONCLUSION AND DISCUSSION

This paper proposes a segmentation algorithm based on special identifier and a new word recognition model based on search engine. Sentence-level similarity comparison model can evolve into three main similarity detection models: VSM, hamming distance and the longest common substring model, which obtains good effect on solving the paper plagiarism problem such as word deletion and addition, part copy and replacement of words by add order penalty factor and the process of handling near-synonyms. From the experiment result, it can be seen that the method proposed in this paper can evaluate paper similarity accurately. However, large-scale paper detection needs massive lexicon, data preprocessing and search library. Therefore, it needs more uniform standards in implementation and promotion to get better effects.

ACKNOWLEDGMENT

This work was supported in part by Hubei Ministry of Education. The public number: 20140220092704_1887927089/76.

REFERENCES


