Analysis and Implementation of Job Repeatability Detection Algorithm

Qing Tian[1,a], Jun-ling Zhu[1,b*], Qun-le Fang[2,c], Qiao-wen Long[1,d]

[1] School of Computer Science, Zhaoqing University, Guang Dong, 526061,China

[2] The 3rd School, Airforce Early Warning Academy, Wuhan, 430019, China

[a]email: zdtina@qq.com, [b]email:zhuzhu000011@163.com, [c]email: phylqfang@163.com, [d]email: 1275928600@qq.com, *Corresponding Author

Abstract: With the office automation and development of paperless needs, the electronic homework is more and more widely used in colleges and universities, but the characteristics of being easier to copy makes copying phenomenon more and more serious, the cost of copying homework is lower and lower. Therefore this paper specializes in the repeatability of detection algorithm targeting the features of Chinese homework, which puts its focus on the study of space vector algorithm and its implementation process, besides comparing JaccardSimilarity method, ShingLing algorithm and Lenvenshtein distance algorithm, the result can show us that vector space model algorithm is more applicable to homework repetition detection.

Keywords: Vector space model, Job repeatability check, JaccardSimilarity, Lenvenshtein distance, ShingLing algorithm

Introduction

In the twenty-first century, with the rapid development of Internet and intelligent devices, electronic office has began to develop into a trend, and electronic work is gradually replacing the paper work of schools and institutions. But as for electronic homework, there is a congenital deficiency, namely the plagiarism situation is very serious, the electronic version of text is easy to copy and paste, thus part of students may summit homework by copying because of being lazy, which can result in the monotony of electronic homework, thus homework itself will lose the meaning. Thus this paper is to alleviate this problem, by learning from the wisdom of predecessors, making research on the operation repeatability detection algorithm, as well as the implementation of the algorithm. However, due to the complexity of the type of the job, assignments that have exact answers such as mathematics, chemistry, geography, biology and so on can not included. While composition, membership application, practice, summary, reading review, social experience, and a series of experimental report in the form of text display, which can be regarded as the main object of repeated detection. At present, there is a specialized software for university plagiarism in the foreign countries such as: Ferret[1], Wcopyfind[2], these software can design a set of related algorithm based on the operation characteristics, which is mainly used for detecting English operation, but the amount of special software for domestic plagiarism detection is rare, therefore the related research is relatively small, but there are more and more studies on natural language text copy detection algorithm.

As for the research of text replication technology, it mainly focuses on text similarity algorithm in China. In essence, text similarity algorithm is the core of job copy algorithm. Therefore, in order to study text similarity algorithm, the operation plagiarism algorithm must be studied. At present, there are many kinds of methods to study text similarity algorithm, such as similarity comparison based on paragraph, similarity comparison based on sentence, similarity comparison based on semantic meaning, similarity comparison based on text similarity comparison, similarity comparison based on comparison of various methods and so on[3]. There are many kinds of algorithms, such as VSM vector space model algorithm[4], Simhash algorithm[5], Hamming distance algorithm[6] and shingling algorithm, Levenshtein distance algorithm[7], JaccardSimilarity algorithm[8] and so on. Each has its own advantages and disadvantages in text similarity detection. In this paper, for the convenience of study, it can select VSM vector space model algorithm, Shingling algorithm, JaccardSimilarity algorithm and Levenshtein Distance algorithm these four methods to carry out comparative study, so as to determine the advantages and disadvantages of various algorithms in the plagiarism algorithm.

Introduction of Text Similarity Algorithm

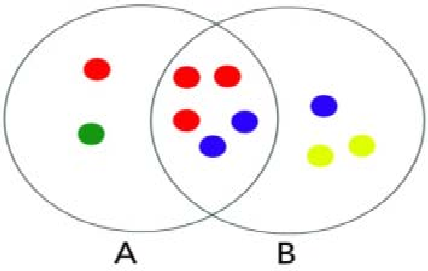
Cosine Algorithm Based on Space Vector

The implementation principle of vector space model is to convert the document to a set of features (keywords), then it can regard document as a n-dimensional vector represented by weights of *n* feature items.

To use this model, we must assume that there is no relationship between the words, so vector space model itself can make semantic similarity not introduced into this model, therefore it is impossible for combing the similar semantic similarity detection methods based on vector space model. But vector space itself can have a great advantage, it can reduce the complexity of the text comparison, document can convert from text to geometric vector, so that the two texts can be studied to calculate the similarity by using the method of comparing vector.

JaccardSimilarity Algorithm

JaccardSimilarity Algorithm, it is not too hard to understand the principle of calculating the similarity, in fact its calculating method is easier to understood by people. namely, the number of the shared part of the characteristic sets of these two texts can be divided by the total number of them, what they can get is the similarity of JaccardSimilarity, the realization of the principle can be shown in Fig. 1.



The similarity degree of set A and set B is: 5/10=.5

Fig. 1 Schematic Graph of Jaccard and Similarity

Shingling Algorithm

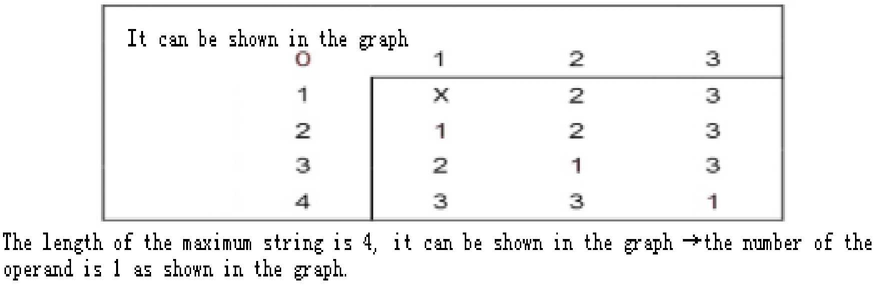
Shingling algorithm treats document as a sequence of text, ignoring some of the tiny details in the document, such as text formatting, letter case, and so on. A collection of substrings followed by text combinations *S* (*S1*, *S2*) can be used to represent document. These substrings of *S1* and *S2* can be called as shingle. Supposing that there is a document *D*, let's set a fixed value to a-shingling as a, then we can use this value a to make the adjacent a word into the shingle. For example, a text segmentation is for (*I, am, Zhaoqing University, 's , a, student*), the text of the 3-shingling is the set: {(*I, am, Zhaoqing University*), (*am, Zhaoqing University, 's*), (*Zhaoqing University, 's, a*), (*'s, a, student*)}, through this processing, the similarity r of these two documents *D1* and *D2* can use JaccardSimilarity method to calculate.

Levenshtein Distance Algorithm

Levenshtein Distance can be shorted for LD, which represents a given string into another designated string, which increase or deletes the number of operations.

The so-called addition and deletion operation refers to the deletion of a specified character from the string, the conversion of a specified character in the string to another target character, as well as the insertion of a specified character in the string.

By using this method, the steps of calculating the similarity are as follows: first of all, we should compare the length of two texts (strings), taking the largest one, denote as *m*, and then we can calculate out the number of operations *M*, then we can get the ratio of *M* to m, using 1 minus the ratio, so as to get similarity, the realization of the principle can be shown in Fig. 2.



It can be shown in the graph

0	1	2	3
1	X	2	3
2	1	2	3
3	2	1	3
4	3	3	1

The length of the maximum string is 4, it can be shown in the graph →the number of the operand is 1 as shown in the graph.

Fig.2 Schematic Graph of LD Algorithm

Implementation of Vector Space Model Algorithm

Vector space model algorithm is the most common and widely used algorithm in text similarity detection algorithm. Therefore, in this paper, it analyzes the implementation process of this method and gives the key program code. The vector space model algorithm can give the similarity of text through text reading, denoising, calculating word frequency as well as the calculation of cosine angle.

Read Word Document

The text is generally word document, because the format of Word as well as the way of encoding is special, Word document can not be read and written with the inputting and outputting stream of Java, while the development tools of java in Eclipse are not integrated with this function, so it requires us to implement by ourselves, java can offer more technical supports for word documents to read and write. Generally used are as follows: Jacob, java2word, poi, rtf, itext, worddemo, WordExtractor and so on.

It is worth noting that the use of java2word can not enter the the word document in 07 version, the interface provided by poi can not support the word document to read, the word document through WPS document processing, WordExtractor can not read and write properly.

The corresponding code is as follows:

```
public static String readDoc(String doc) throws Exception {
    // Establishing an input stream to read doc document
    FileInputStream in = new FileInputStream(new File(doc));
    WordExtractor extractor = null;
    String text = null;
    // Establish WordExtractor
    extractor = new WordExtractor();
    // Extracting doc document
    text = extractor.extractText(in);
    return text;
    }
```

Implementation of Word Segmentation Technology and Removing Noise Words

Having Chinese segmentation for the text needs to use word segmentation technology [i], which refers to use certain standard to cut the word string into single words or phrases one by one. Standard English will take space as the basis of segmentation, but when we write Chinese, it does not need to be separated by space, so it can not take space as the basis for segmentation like English, at the same time, sometimes the phrase can be composed by two, three, four or more words, so as a result, there is no way to cut through the words string according to the number of words. On the level of the phrase, the segmentation in the middle part is very complicated, which can not be generalized to talk about, so we need a reliable Chinese dictionary as the basis for word segmentation, and the words in the dictionary can determine the result of the word segmentation. Chinese segmentation technology has been for a long time, thus there is no need for us to do a detailed understanding, while the commonly used word segmentation techniques including: Paoding Analysis, LingPipe, JE-Analysis, IKAnalyzer Chinese Segmentation, word Analyzer, Jieba Analyzer, Ansj Analyzer, HanLP Analyzer, in this paper, the technology used in the experiment is word Analyzer and Lucene text retrieval tool.

Calculate Word Frequency

Word frequency refers to the ratio of the number of a word that appears in the text to the total number of word in the text (the total vocabulary used here refers to the number of words that are removed from noisy as well as the number of non Chinese characters). The method for calculating the frequency is not difficult, we can assume that we can get a keyword text 1, text 1 can save all keywords, we need to get the statistic number from all keywords of of text 1, then saving the results into the map, outputting it to a new keyword of text number 2, then it can obtain the word frequency of key word. It can save them to map 1, then output them to the new word frequency text 3.

The statistics code of keyword frequency is as follows:

```
while ((lineTxt = br.readLine()) != null) {
              String[] names = lineTxt.split(" ");
              a = Integer.parseInt(names[1]);
              map.put(names[0], a);
              count += a;
              System.out.println(names[0]+" "+a);
              }
```

Calculate Keyword Weight

When the weight of the keyword is calculated, we have to mention the reverse text frequency IDF, which is the frequency of the reverse file, it can indicate the number of text that the keyword appears in the text library. IDF of keywords, refers to the ratio of the total number to the number of the text with text keywords corpus, using this ratio to acquire the logarithmic number, which can get the keywords inverse document frequency in the corpus, the calculation method can be seen as Formula 1.

$$\mathrm{IDF}(ti) = \log(n / N) \tag{1}$$

Among them, *N* is the number of text that can appear keyword *ti*, and *n* is the number of total text.

In the upper forum, when *N* is zero, there is obviously an error, therefore it needs to be judged. When *N*=0 is used, the formula for the weight should be changed, which can be shown in Formula 2.

$$\mathrm{IDF}(ti) = \log(n / 1) \tag{2}$$

Calculation of Angle Cosine

As for the text comparison algorithm of vector space model, it can extract keyword after removing the noise word from the text, as well as the operation of decreasing noise dimension, the words will not be the same between two different texts, so in the process of calculating the cosine angle of these two texts must complement each other for the lack of weight of words, adding zero in the place that there is absence of vector in the corresponding vocabulary.

The method of calculating the cosine of weight vector is the same as that of the geometric vector method. The formula can be shown in Formula 3.

$$\mathrm{Sim}(D1, D2) = \cos\theta = \sum_{k=1}^{n}(W_{1k} * W_{2k}) / \sqrt{(\sum_{k=1}^{n} \mathrm{W}_{1k}{}^{2})} * \sqrt{(\sum_{k=1}^{n} W_{2k}{}^{2})} \qquad (3)$$

Among them, *D* is the compared text, *p* represents its keyword, a collection can be composed by the keyword, which can be used as text *D* (p_1,... p_k,... p_n), among them: *1<=k<=N*. Generally speaking, we also can calculate the weight *W* for its each feature item, so as to replace the original feature item with the weight, then change the article into geometric vector, namely, *D1*（W_1, W_k, …, W_n), thus *D1* can be called as weight vector. Bring the weight of *D1* and *D2* into Formula 3, the cosine value of these two texts can be calculated out, the bigger the value is, the greater the similarity of these two is, and vice versa.

Algorithm Test Analysis

After program realizing these algorithms, we can use these algorithms to test several electronic jobs. In view of the characteristics of the job, in the test, we do not require large text as test data, because the job can have the characteristics of having the same goal and purpose, when the repetition rate is below 40% in the work, we can regard it as normal, namely there is no plagiarism. Thus, we select twenty five the experimental reports of e-commerce as testing cases, having artificial prediction on these twenty five documents, and the result can be shown in Table 1.

Table 1 Artificial Judgment Result

Repetition situation	<5%	20%~40%	40%~80%	>80%
Document ID	1~5	6~10	11~20	21~30

With the above analysis algorithm, combined with the characteristics of the experimental report of electronic commerce, we regard it as no plagiarism with the repetition rate below 40%, while the repetition rate is from 40% to 80%, which can be regarded as generally plagiarism, when the repetition rate is more than 80%, we regarded it as serious plagiarism. The test result can be shown in Table 2.

Table 2 Test Result of Each Algorithm

Plagiarism judgment	**Levenshtein method**		**Shingling method**		**Jaccard method**		**Vector space model method**	
	accuracy rate	recall rate	accuracy rate	recall rate	accuracy rate	recall rate	accuracy rate	recall rate
No plagiarism	100%	60%	100%	70%	100%	80%	100%	90%
General plagiarism	67%	80%	75%	90%	91%	100%	91%	100%
Serious plagiarism	83%	100%	91%	100%	100%	100%	100%	100%

From the above results we can see, from the vertical direction of the table, when the repetition rate of these four kinds of methods is small, it can have an accurate judgment, which can find out the lower probability of all plagiarism documents with high repetition rate, that is to say, the higher the degree of the repetition is, the higher the accuracy of the judgment algorithm is. In lateral view, the preparation rate of Levenshtein method is rather low, while the shingling method is in the second place, while Jaccard method is relatively better, the determining result of the vector space model method is the best, therefore, the vector space model method can be widely used in a variety of check and document retrieval.

Conclusion

This paper analyzed several commonly used text repetition algorithm, and put its focus on the vector space model programming process, and finally tested the comparison of several algorithms through experiment. The result showed us that vector space model was the closest one to the real situation, which can be an ideal algorithm for detecting the repeatability of the job. However, the test still existed one shortcoming in this paper, that is, the amount of sample is rather less and the accuracy of the test results is not too high. In the late stage, a large number of jobs should be accumulated to form a sample library, which can provide the basis for the accuracy of the test.

Reference

[1] Junpeng Bao,Caroline Lyon. Copy detection in Chinse documents using Ferret[J]. Language Resources and Evaluation, 2006(40): 357-365. http://www.springerlink.com/content/r54v63q53827vtn5/fulltext.pdf.

[2] Wcopyfind.The Plagiarism Resource Site Charlottesville, Virginia Wcopyfind 2.6 Instructions.http://plagiarism.phys.virginia.edu/Wsoftware.html[2007-4-22].

[3] Shi Kailun. Research and Implementation of Semantic Dimilarity Based on the Combination of Knowledge Base and Corpus[D]. Beijing Jiaotong University, 2016:7-24.

[4] Yan Chunmei. Research on Similarity Algorithm of Paper with Vector Space Model and Semantic Understanding[D]. Chengdu: Xi'an Jiao Tong University, 2015: 26-27.

[5] Xu Jihui. Research on Massive Document of Anti-cheating Technology Based on Simhash Algorithm[J]. *Computer Technology and Development*, 2014, (09): 103-107.

[6]Zheng Kai, Ouyang Linyan, Lin Qiang, Liu Fangbing. Research on LCS Algorithm and Editing Distance algorithm[J]. *Information and Communication*, 2015, (05): 22-23.

[8] Yu Yongyan Multi Model Estimation Based on Jaccard Distance and Conceptual Clustering[J]. *Computer Engineering*, 2012,(10): 20-24.