Duplicate checking Strategies for Selecting Topics for Graduation Thesis Based on Maximum Public Sequence

Duplicate checking Strategies for Selecting Topics for Graduation Thesis Based on Maximum Public Sequence

Guorui Yu[1, 2], Li Huang[1, 2], Jiewen Sun[1, 2], Shangchun Liao[3] *

[1]School of Computer Science and Technology, Wuhan University of Science and Technology.

[2]Intelligent Information Processing and Hubei Real-time Industrial Systems Laboratory.

[3]Key Laboratory of Metallurgical Equipment and Control Technology, Wuhan University of Science and Technology, Ministry of Education.

867209633@qq.com, Lihuang82@wust.edu.cn, 30970204@qq.com, 2543861131@qq.com.

*Corresponding Author: Shangchun Liao; email: 2543861131@qq.com; phone:15071193491.

Abstract: The system is based on the undergraduate thesis management system as the practice platform, focusing on the realization of the thesis topic check function. The current check-up detection system is based on the entire contents of the query, and does not meet the requirements for an undergraduate thesis query. According to the characteristics of the thesis, the system first divides the keywords into the topic, selects the irrelevant parts, then removes the irrelevant words, and finally checks the keywords in the topic, and realizes the efficient and accurate check for inspection.

1. Introduction

Graduation design and thesis of undergraduate course are very important links in personnel training and teaching. Graduation design and thesis management of undergraduate students have become the most important work of the school's corresponding educational administrators, among which the most complex and troublesome step is the re-examination of the title of an undergraduate thesis. With the gradual deepening of the enrollment expansion and the continuous expansion of the scale of running colleges and universities, the number of graduates is increasing, and the topics of undergraduate thesis are complex, inevitably duplicate or partly duplicate, which adds trouble to the duplicate examination of undergraduate thesis topics. In order to reduce the workload of educational administrators and improve their work efficiency, we designed the system.

2. Current situation and development trend at home and abroad

In the checking and testing of undergraduate thesis titles by computer, due to the different educational mechanisms at home and abroad, the corresponding management system cannot meet our requirements. Because of the different school management systems at different levels in China, there is also a lack of a system to meet the different school management.

At present, there are three major graduation thesis duplications checking systems in China: HowNet, Wipe and Wanfang. Among them, the resources are constantly updated. Every year graduates papers except for confidentiality requirements are basically included in the three systems as a comparative resource bank. HowNet, Wipe and Wanfang are not open to individuals. Wipe and Wanfang are open to people. Wanfang does not test the Internet and English. HowNet and Wipe both test the Internet and English.

At present, there are still some problems in the duplication checking system, as follows:

1. Many books have a long tradition, and these materials are not included in the testing system. Most of the databases are articles of previous student papers and periodicals. Books and government work reports are not in the library yet, so the similarity of the papers cannot be found.

2. Chinese HowNet duplicate checking is based on sentences. That is, the article is divided into sentences, and then compared with the articles in HowNet database sentence by sentence. If there are the same principal contents (i.e. notional words, such as nouns, verbs, professional vocabulary, etc.), it will be marked red. If a large number of red sentences appear in a paragraph, the paper repetition rate is calculated. Changing the order of sentences or changing the structure of the subject, predicate and object will reduce the repetition rate of papers.

3. Some scholarly documents on the network have not been included in the paper duplication system. If students cite the content on the web page, it is likely that the system cannot accurately identify the duplication rate of papers.

4. The sensitivity of China HowNet to system detection has been set a threshold, and it cannot be detected if the sensitivity is less than 3%. That is to say, if we extract some sentences from several documents at the same time, although they constitute plagiarism, they cannot be detected systematically.

Therefore, it is particularly important to develop a duplication checking system for undergraduate thesis titles, which can simplify the tedious work in duplication checking of undergraduate thesis titles, assist educational administrators to screen the topics of undergraduate thesis quickly and conveniently, and improve work efficiency. This is not just an important means to realize the modernization and networking of education management, but also a way to promote the reform of traditional teaching mode. It is of great importance and practical value to improve teaching management and teaching quality.

3. Word segmentation processing

The first paragraph after a heading is not indented (Bodytext In the process of researching the duplication detection system of graduation thesis titles, we found that many irrelevant words would affect the whole process of duplication detection and the final result of duplication rate by studying the data of our college of Computer Science in the past three years. Therefore, it is necessary in order to remove irrelevant words and check the duplicate of those variables directly. This will enhance the efficiency and accuracy of duplicate checking. The first step in removing extraneous words is word segmentation [1].

When dealing with text analysis, word segmentation is often a problem. Especially in the process of duplication detection of graduation thesis titles, most text data are in Chinese. There are 3 kinds of commonly used methods for word segmentation, one is word matching [2], the other is word semantics distribution [3], and the other is dictionary matching [4]. At present, word segmentation method based on words is effective. The algorithm based on dictionary matching is too complex, and the algorithm based on probability analysis is relatively simple. After matching word segmentation with network thesaurus, nearly three years' graduation thesis title of Computer College is processed.

After comparing with the network thesaurus, the paper title data of the Computer Academy in the past three years are processed by word segmentation, and the results are shown in Figure 1.

As can be seen from the figure above, in the process of word segmentation, each topic is divided into several keywords. A number of keywords form a topic, which means that different key words in

each topic are likely to repeat. The repetition of these key words greatly affects the results of repetition checking, so it is necessary to remove irrelevant words [5].

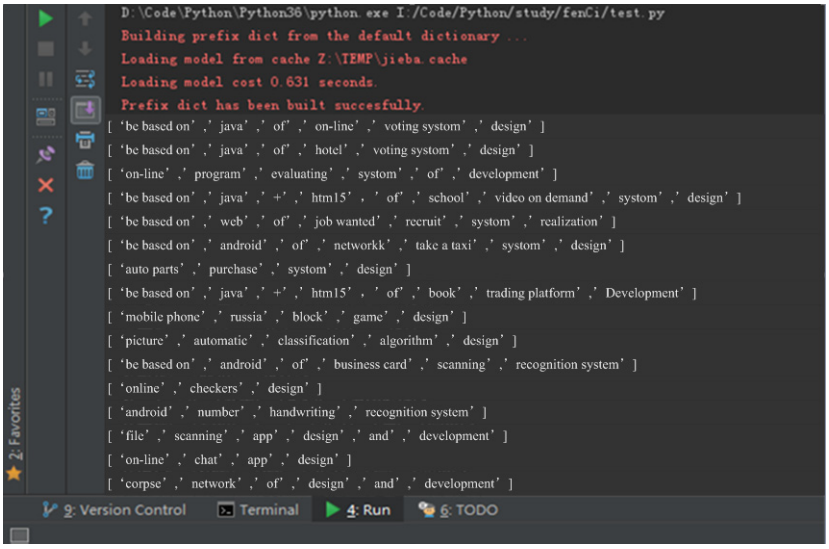

Fig 1. result of word segmentation.

4. Removal of Irrelevant Words

The first paragraph after a heading is not indented (Bodytext The keywords "design", "realization" and "development" are obtained by splitting the topic data of the paper, totaling 147 words and 892 words. The experimental results are sorted according to the frequency of keyword occurrence, and the following results are obtained, as shown in Figure 2.

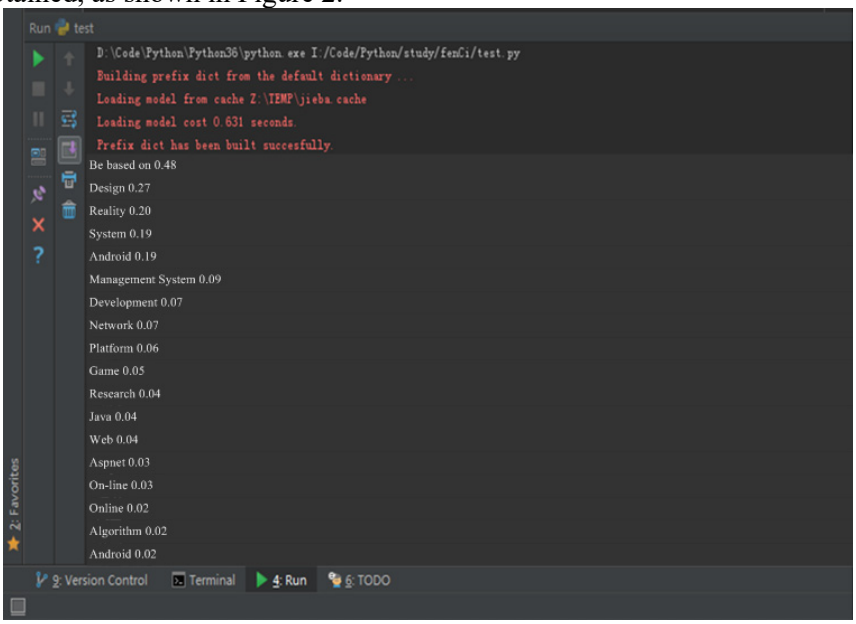

Fig 2. duplicate data.

According to word segmentation, only the first 20 keywords appear more frequently. The frequency of keywords after the twelfth data "on-line" is less than 2%, which has little influence on the results in the process of checking duplication. Therefore, the key after the twentieth is ignored, not included in the statistical scope, and only duplicated. The top 20 keywords were counted and tabulated, as shown in Figure 3.

20.	Design	Reality	System	Android	Management system	Development	Network	Platform	Game
0.47	0.26	0.19	0.19	0.09	0.09	0.07	0.06	0.06	0.05
researth	On-line	Online	Algorithm	Android	Software	C #	Mobile phone	Information	Application
0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Fig 3. duplicate data.

According to the content shown in the figure, most of the irrelevant words are embodied as "based", "design", "realization". They basically conform to the basic facts of Computer College. On the basis of factual demonstration, it can be concluded that these irrelevant data are split correctly and can be checked in the next step.

According to the statistical results of the data, we can clearly see the distribution of the initial 20 keywords. Most of the data sets are based on the words "based on" and "design". In order to express the distribution of the data more intuitively, a column is made according to the first 20 repeatability data as shown in Figure 4.

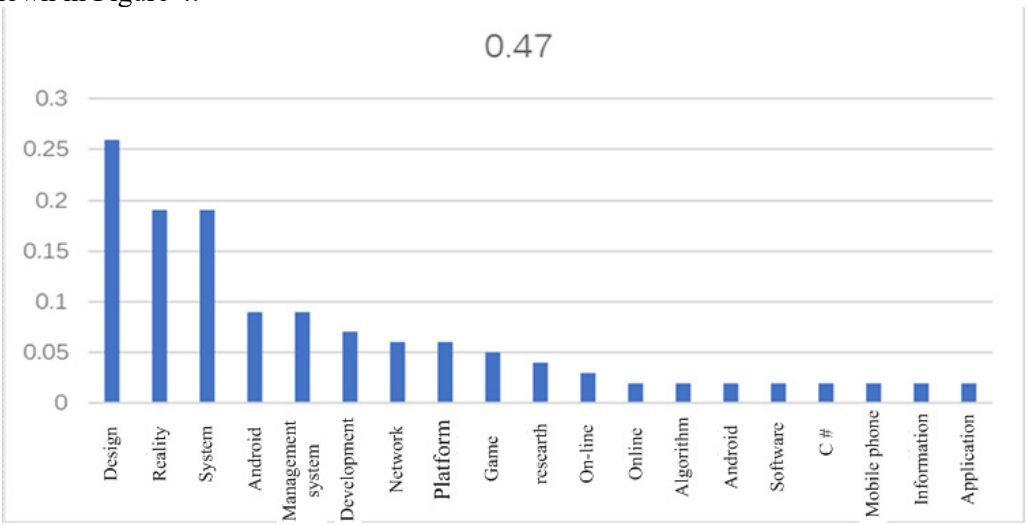

Fig 4. histogram.

Through the analysis of the histogram, we know that the extraneous words are not evenly distributed, but are ladder-shaped, and their probability is decreasing. Furthermore, irrelevant words in different paper titles are repeated in several words, which confirm the practical significance of removing irrelevant words. After removing irrelevant words and statistics before duplication checking, it will greatly improve the efficiency and accuracy of duplication checking, and make the system more intelligent and humane.

5. Implementation and Testing of Duplicate Checking Algorithms

After eliminating the unconnected words, we will check the variable vocabulary. In the process of checking the variable vocabulary, the system uses the LCS algorithm calculation process [6], optimizes and enriches according to the actual situation. Now the introduction and use of the system using the algorithm are described.

When using. NET, MVC, JAVA and other technologies to build a separate paper management system, the longest repetitive substring is utilized to check the title of the paper.

(1) LCS: In a specific string str, repeated substrings are called repetitive substrings. If there are multiple repetitive substrings in the string str, the longest of them is called the longest repetitive substring. For example, str='abcdacdac', the substring'cdac'is the longest repetitive substring of str.

LCS problem: longest common subsequence

For example:

A sequence S is called the longest common subsequence of a known sequence if it is the subsequence of two or more known sequences and is the longest of all sequences that meet this condition.

The Branch of LCS Problem: Longest Common Substring and Longest Common Subsequence[7]

Substring is a continuous part of a string. Subsequence is a new sequence obtained by removing any element from the sequence without changing the sequence. To be more concise, the position of the characters of the former must be continuous, while that of the latter (subsequence LCS) must not. For example, the longest common substring of the string acdfg and akdfc is df, and their longest common substring is adf.

LCS strategies for solving problems:

One: exhaustive method, high complexity, difficult to achieve;

Two: Matrix, that is, the LCS problem of dynamic programming section, to reflect its advantages;

Question description of the twith idea, see Formula 1:

$$c[i,j]=\begin{cases}0 & \frac{if\ i=0\ or\ j=0.}{if\ i,j>0\ and\ x_i=y_i.}\\ c[i-1,j-1]+1 & \\ max(c[i,j-1],c[i-1,j]) & if\ i,j>0\ and\ x_i\neq y_i.\end{cases} \quad \text{Formula (1)}$$

An example of the algorithm is shown in Figure 5:

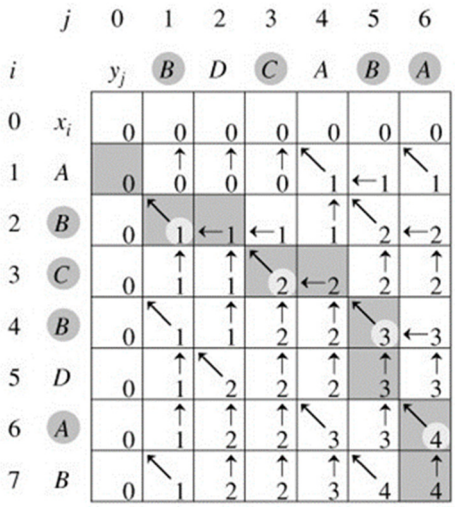

Fig 5. algorithm example diagram.

If XM = yn, then ZK = XM = yn and Zk-1 is the longest common subsequence of Xm-1 and Yn-1.

If xm_yn and zk_xm, Z is the longest common subsequence of Xm-1 and Y.

If xm_yn and zk_yn, Z is the longest common subsequence of X and Yn-1.

The time complexity and space complexity of the algorithm are n*m and n*m respectively.

In addition, it is more complicated to record paths.

LCS solves lis problems:

It is necessary to sort first, and then find the longest common subsequence with the original array.

The following conclusions are drawn from the test of "Flash-based Aircraft Shooting Game Design" based on the duplication algorithm, as shown in Figure 6.

Title	Type	description	Adding time	examination	File	Applicant's name	Applicant number	Title ▼ Design of Aircra search Flight Shooting Game
Flight Shooting Game Based on Flash	paper	Nothing	2017/3/14	Have passed	old.docx	student	000000000000	Similarity degree:100%
Flying shooter	paper	Nothing	2017/3/14	Have passed	old.docx	student	000000000000	Similarity degree:75.00%
Design and Implementation of Flight Shooting Game Based on Android	paper	Nothing	2017/3/14	Have passed	old.docx	student	000000000000	Similarity degree:44.44%
Longitudinal Design Game Based on Coos 2D-x	paper	Nothing	2017/3/14	Have passed	old.docx	student	000000000000	Similarity degree:40.0%
Design of Airplane Fighting Game Based on JAVA	paper	Nothing	2017/3/14	Have passed	old.docx	student	000000000000	Similarity degree:37.50%

Fig 6. test results

By using this algorithm, we can check the duplication of undergraduate graduation thesis. By calculating the similarity, we can know whether there are similar topics in the last three years' graduation thesis topics. After the similarity is found, it can be confirmed by the administrator. After checking the data of this year and the data of the past three years, the results are shown in figure 7.

Design of Online Voting System Based on JAVA	Game Character Voting System Based on Nodejs	66.67%
Design of hotel management system based on Java	Hotel Management System Based on JAVA	92.86%
Development of Online Program Evaluation System	Design and Implementation of Online Ordering System	62.50%
Design of Campus VOD System Based on java+htm15	Video Broadcasting website based on asp.net	50.0%
Realization of Job Search and Recruitment System Based on Web	Design and Implementation of Online Job Search and Recruitment System	66.67%
Development of Network Taxi System Based on Android	Design and Implementation of English Learning System Based on Android	55.56%
Design of Purchasing System for Automobile Parts	Computer Parts Inventory Management System Based on Web	50.0%
Development of book trading platform based on java + htm15	Information push system of steel trading platform based on Java	57.89%
Mobile Tetris Game Design	Russian Tetris Based on JAVA	77.78%
Design of Automatic Image Classification Algorithms	Design and Implementation of Automatic Classification Algorithms for Patterned Fabrics	75.00%

Fig 7. results of checking results

6. Conclusion

In the process of duplicate checking of undergraduate papers, the first step is to segment the Papers'titles. After obtaining the results of word segmentation, the keywords with the top ten repetition rates are treated as irrelevant words. After deleting these irrelevant words, the results of duplicate checking for each topic will be more accurate.

Acknowledgements

This research is financially supported by the training program of innovation and entrepreneurship for college students in hubei province in 2018 (No:201810488082).

References

[1] Ge, Y.L., Li S.R., Chang P., Lu S.L. (2016) Optimization of ASP flooding based on dynamic scale IDP with mixed-integer.New York. Applied Mathematical Modelling., 145~189.

[2] Wan, A.P., Gu, F., Jin, J.M.,Gu X.J., Ji Y.J. (2016) Modeling and optimization of shutdown process of combined cycle gas turbine under limited residual natural gas. Applied Thermal Engineering.,.19~26

[3] Xu, H.T. (2015) Research on Chinese Word Segmentation Domain Adaptive Method Based on Active Learning [D]. Beijing Jiaotong University.

[4] Ge Y.L. (2017) Applied Mathematical Modelling Optimization of ASP flooding based on dynamic scale IDP with mixed-integer. Built Environment Project and Asset Management. 33~34.
[5] Li Z.Q. (2017) Development of a web-based system for managing suppliers' performance and knowledge sharing in construction projectEmerald. 1~9.
[6] Yu, H.Y. (2014) Comparisons of LCS and GST Algorithms in String Similarity Measurement. Chengdu. Electronic technology. 21~27.
[7] Guo, Z.K., OuYang, L.Y., Li, Q. (2015) Research on Duplicate Checking and Editing Distance Algorithms. Beijing. Information communication. 125~178.