International Journal of Knowledge and Language Processing
Volume 8, Number 1, **2017**

www.ijklp.org
KLP International ©2017 ISSN 2191-2734
pp.13—24

Literature Review of Duplicate Checking System

Jin Zhang[1]
[1]Institute of Scientific and Technical Information of China
Beijing, China
Zj729929844@163.com

Yao Liu[2]
[1]Institute of Scientific and Technical Information of China
Beijing, China
liuy@istic.ac.cn

Received March 2017; revised March 2017

ABSTRACT . *This present literature review focuses on the research status and new progress of the duplicate checking system about Chinese and English papers, which introduces the concept, characteristics and application of the duplicate checking system. This paper puts forward the problems existing in the system and summarizes the latest progress of the duplicate checking system. Three models are introduced about the checking system of English paper, including the vector space model, the generalized vector space model and the invisible semantic indexing model. For the Chinese duplicate checking system, we list four similarity analysis methods based on the attribute theory, the Hamming distance, the semantic understanding and the string matching algorithm. This paper also discusses the related problems and future directions about the research and development of the duplicate checking system. By comparing with the latest research progress of Chinese and English papers, the author finds it difficult to implement the Chinese natural language understanding via computer.*

Keywords: duplicate checking system, models, similarity analysis methods

1. **Introduction.** The duplicate checking system is a kind of technology of copy detection, which is the method of academic misconduct detection system for the identification of the main body of the thesis. At present, most of academic journals in editorial departments

have been used this system for quality testing of academic papers, and higher institutions learning have also taken it as an effective way to ensure the quality of graduate dissertations.

This paper studies the research status and new progress of the duplicate checking system for Chinese and English papers. The concept, characteristics and application of the duplicate checking system are introduced. The paper puts forward the problems about five parts: the flaws in detection system, the recessive academic misconduct behavior, the utilitarian tendency of education, the influence of the quality and innovation of dissertations, the specifications of reference forms. We also summarizes the latest progress of the checking system including three models and four methods. These models are introduced about the checking system of English paper, including the vector space model, generalized vector space model and invisible semantic indexing model about the English Checking system. We list four similarity analysis methods based on the attribute theory, the Hamming distance, the semantic understanding and the string matching algorithm with regard to the Chinese checking system. The related problems and future directions about the research and development of the checking system are discussed in this paper, we find the computer Chinese natural language understanding are more difficult than English based by comparing the latest research progress of Chinese and English papers.

2. **The Review of the Concept, Application and Characteristics of Duplicate Checking System.**

The duplicate checking system, namely the similarity detection, can be as a way to academic misconduct detection for the identification of the main body of the thesis. After the calculation and operation of system, users can get "Copy Rate", "Repetition Rate" and other forms of visual detection results.

2.1. **The Concept of the Study on Duplicate Checking System.**

Duplicate checking system is a kind of technology of copy detection, whose idea is seen each paper as a series of Token set. These Tokens can be characters, words, sentences, paragraphs and chapters. In their mathematical models, the collections of these Tokens can be computed. Suppose getting respective Token sets A and B from some algorithm that extracted from thesis a and b:

① If $A = B$, a is a duplicate or all copies of b,

② If $A \subset B$, $A \neq B$, a is a subset or parts of plagiarism of b.

According to the above relationships, Broder discussed the resemblance $r(A, B)$ and containment $c(A, B)$ between two documents a and b, and attributed two relations to set the intersection problem.

If w successive Token sequence (i.e. w-shingling) is extracted from A and B, the resemblance between a and b is [1]

$$r_w(A, B) = \frac{|S(A, w) \cap S(B, w)|}{|S(A, w) \cup S(B, w)|}$$

where $S(A, w)$ and $S(B, w)$ is respectively the shingling subset of all the length of the w,

$r_w(A, B)$ is also called document similarity. Thus the containment can be expressed as:

$$c_w(A, B) = \frac{|S(A, w) \cap S(B, w)|}{S(A, w)}$$

Document similarity $r_w(A,B) \in [0,1]$, and $r_w(A,A) = 1$. Thus, the resemblance distance for two documents is defined as $d_w(A,B) = 1 - r_w(A,B)$.

Although Broder use the mathematical definition of $r_w(A, B)$ 、 $c_w(A, B)$ and $d_w(A, B)$ is to identify “roughly the same” and “roughly contained”, but they can also constitute the theoretical basis of the check method for the duplicate checking [2].

2.2. **The Application of Duplicate Checking System.** At present, most of academic journals in editorial departments of China have been used this system for quality testing of academic papers, and higher institutions learning have also taken it as an effective way to ensure the quality of graduate dissertations. The publication of academic papers or degree applications is possibly only after the examination of academic misconduct detection systems. It can be said that thesis testing has standardized academic behavior to a certain extent and become a “yardstick” for detecting academic behavior" [3].

At present, there are many domestic academic similarity detection systems, among them, CNKI, Wanfang Data, VIP detection system is widely used. Different testing systems contain different databases, and most of them contain the literature resources of the whole subject.

Periodicals, dissertations, newspapers, the Internet and other resources are contained, but the length and scope of each test system is different, which is an important factor affecting the results of the test. Another measure that affects the result of the test is to set the test target. Take CNKI,Wan Fang and CQVIP as an example； CNKI have detection index what can remove the copy ratio of citation or document, and the number of suspected coincidence words or paragraphs; Wanfang have reference literature similarity ratio index, residual similarity ratio index; CQVIP have "reference rate" and "carbon rate" and "self writing rate" index [4].

At present there is strict requirements on the similarity detection results for academic papers which written by domestic university graduate, generally within 10% to 30%, unqualified students revise it according to the detection ratio, the less is required modified again, the serious may be postponed refused to award PhD.

2.3. **The Characteristics of Duplicate Checking System.** Typically, text similarity computation formulas or models that need to satisfy certain conditions, thus the designed text similarity algorithm also needs to have the following properties [5]:

1. Reflexivity: The similarity of text, sentences, paragraphs and other text fragments is 1 comparing to themselves.

2. Monotonicity: The similarity of texts is monotonically increasing or monotonically decreasing within a certain range.

3. symmetry: If the text W1 is similar to the text W2, the similarity between W2 and W1 is determined.

4. transitivity: If there are three text W1, W2 and W3, W1 and W2 are similar, W2 and W3 are similar, then W1 and W3 are similar.

3. **The Research Status of Duplicate Checking System.**

3.1. **English Duplicate Checking System.** Since the application of WordCheck software for querying duplicate fund applications in 1991, the technology of plagiarism identification in natural language has developed greatly in foreign countries, and there have been many plagiarism identification systems.

In 1994, Mander developed a similar query file system named SIF(later renamed SIFF)[6] which used in large-scale file system. It can query binary and text files, although there is no emphasis on the natural language text of the query, but first applicate digital fingerprinting technology to calculate the similarity between documents, which provides a new idea for plagiarism identification technology.

In 1995, in the program named "Digital Library" studied by Stanford University, Brin used digital fingerprint technology to develop copy detection system for digital documents, which is used for a complete copy of identification document and partial replication. COPS adopts document registration mechanism, and its system architecture is adopted by most subsequent plagiarism identification systems.

In the same year, the team led by Shivakumar used relative frequency model to develop a new replication detection system SCAM, and improved the COPS system. From the experimental results, the SCAM system is superior to the COPS system. SCAM learned from the vector space model in information retrieval technology, and uses improved cosine method to calculate document similarity. Later, Garci, a-Molina continued to study the document replication detection system dSCAM in the distributed database environment, and discussed near-replicas document query that based on the Web.

In 1996, Heintze developed plagiarism and recognition prototype KOALA based on the web with the method of digital fingerprinting technology and released it online for free testing. In 1997, the team led by Broder used the method "shingling" to classify the 150 GB Web document set, and got better experimental results.

The same year, the plagiarism identification prototype system CHECK is formed for Latex format documents, CHECK decomposes a Latex document into a tree structure first, then use vector dot product method to compare the similarity of documents, and it is the first time to use information documents to reduce unnecessary calculation.

In 2000, Monostori has high recognition accuracy by using MDR (match detect reveal) method to determine the overlap degree of the document, MDR (suffix tree) search the maximum substring between strings with the suffix tree, then use the suffix vector to store the suffix tree so as to improve the efficiency of recognition.

In 2001, Finkel proposed the method SE (signature extraction) to detect the overlap of documents, SE methods include chunking, culling, digesting, sorting, comparing and other steps, the recognition accuracy is higher.

In 2002, the team led by Chowdhury studied the rapid detection method of duplicate documents in large scale document collections, using I-Match systems that the same like

SIFF system. I-Match preprocess data intelligently, removal of high-frequency words and low-frequency words, the advantage is to improve system efficiency. I-Match system works well when identifying documents with a high similarity, but each document has only one Hash value, which can't deal with partial plagiarism.

In the same year, Hoad and Zobel integrated the word frequency statistics and digital fingerprinting methods to solve the problem of derivative document recognition. By searching a large number of XML data and Linux files to find a better algorithm for plagiarism recognition.

In 2003, Schleimer proposed a Winnowing algorithm based on digital fingerprinting to accurately identify replication problems of documents, and was applied to MOSS, an online service for plagiarism recognition. In addition, there are several websites that provide online plagiarism identification services for Internet, such as Plagiarism, Integriguard and EV, E2 and so on. In China, Qinbao Song proposed detection algorithm relate to illegally copied digital goods, through a multi-level and multi granularity text representation of digital frequency statistics to construct the overlap measure based on the algorithm, and achieved good results.

3.2. **Chinese Duplicate Checking Systems.** At present, there are mainly three duplicate checking systems in Chinese Dissertations: the academic misconduct detection system of CNKI, the similarity check system of Wanfang Data, and VIP duplicate checking systems.

China National Knowledge Infrastructure, launched in 1996 by Tsinghua University and Tsinghua Holding Group dedicates to mass digitization of China Knowledge resources, as well as creating the platform for global dissemination and value-added services. CNKI has built the most comprehensive system of China academic knowledge resources—China Integrated Knowledge Resources Database, which collected over 90% of China knowledge resources, comprehensive coverage of journals, dissertations, newspapers, proceedings, yearbooks, reference works, encyclopedia, patents, standards and laws & regulations.

Based on the vast amount of academic literature resources of Wanfang Data, Wanfang is mainly check the achievement of scholarship, and the objective and informative test reports are provided to support academic publishing, scientific research management and dissertations management.

VTIMS is a company with access to VIP Hengyuan (Beijing) Information Technology Co., Ltd. and Beijing many major colleges and universities jointly developed, combining the data resources and access to VIP information of the data mining technology, and successfully applied in large-scale text comparison in the field of innovation the product, is a one-stop platform for thesis writing guidance and management [6].

The duplicate checking system mainly use CNKI system, the system based on the massive journal articles, and conduct a comprehensive comparison of the thesis that put it in [7].

3.3. **Problems in The Research of Duplicate Checking System in China.**

3.3.1. **The Flaws in Detection System.** The types of checking databases are not enough now. References to books and foreign materials are of vital when writing dissertations. However, this part is the weakness of the database. The current comparison resource library is mainly periodicals, dissertations, newspapers, meetings, etc, and books are not included.

Although some dictionaries, encyclopedias, handbooks, catalogs, table spectrum, directory are included in CNKI, but it is far from enough. Moreover, as an important part of the thesis writing, there is a great gap in the total amount of foreign language documents.

Detection technology needs to be improved, and the main idea and structure of this paper are difficult to predict. As is known to all, academic texts are not the author's innovation, innovation only exist in the part viewpoint of it in the paper, the core point of view and the structures is the soul of the thesis.

The core point of view needs to be demonstrated by citing existing academic achievements. However, the detection process is realized by the computer program, the computer program using a variety of analysis techniques and the semantic analysis of sentences, but it is difficult to analyze the core viewpoint and the structure of the article, thus it is difficult to achieve the desired effect.

3.3.2. **Easy to Induce Recessive Academic Misconduct Behavior.** Under the impetus of the students' demand, "pre-check" and "counter-check" industry appeared. According to media reports, in the graduation season, some online stores that only provide papers "pre-checking", a paid service for graduates, is monthly more than 1 million yuan per month. Many graduate students purchase the online software or detect in advance on the Internet, and then change paragraphs and sentences to identify plagiarism of sentences or expressions, even some in full view of almost all plagiarism case, can still be detected by. The implicit academic misconducts come along whose patterns changing constantly later, bad academic atmosphere prevails, and academic norms completely discredited.

3.3.3. **Bring About The Utilitarian Tendency of Education.** The use of the detection system makes the paper evaluation system simple, teachers and students emphasize on the "Repetition Rate" and other indicators, and the computer is used for replacing the human brain, judge the quality of academic papers by computer systems. Graduate education and management are gradually influenced by instrumental rationality and focus too much on explicit indicators.

In the "requirements and guidelines" of these indices, the people in University only work out all kinds of data blindly, attach importance to the ranking of academic achievement; teachers are fond of the teaching of scientific research; while students are fond of grading research. Education is increasingly utilitarian, which is one of the reason that graduate students’ basic skills are too poor, scientific research abilities are low, and lack of interest in learning, only can use various means to muddle through.

3.3.4. **Influence The Quality And Innovation Of Dissertations.** The main process to duplicate checking of papers is the copy ratio of the text in the paper. In fact, it is difficult to judge the quality of the paper. Under the pressure some students are afraid of using the existing research content in order to reduce the repetition rate. In essence, it is also contrary to the law of academic research.

Second, scientific research has a high degree of dependence on data, that is, how much data to do." It can be said that a large amount of data access is the basis of writing, reasonable references and reference materials are the necessary conditions for writing.

References to documents should be limited to the actual needs of the discussion, not to the provisions of "duplicate checking ".

This kind of problem is more prominent especially in social science research. The excessive caution in quoting literature is likely to make the research atmosphere divorced from the literature’s. In order to avoid repetition, the post is full of "own words" after adaptation, thus the language of the thesis is difficult to show. In order to reduce "repetition rate", some students often change their papers beyond recognition. The readability, internal logic and scientific results of the paper are greatly reduced, which seriously affect the quality of academic papers [8].

3.3.5. **Specifications of Reference Forms.** It has a direct impact on the degree of integrity of dissertations that the description format specification references whether the reference document format is standard or not. The similarity comparison results showed that the thesis is a reference of others research results, if the reference format is not standardized, the similarity comparison results showed the thesis is plagiarism of scientific research, which would belongs to the academic moral problems. Part of the student's graduation thesis is not standardized because of the description format during the paper examination, although it is quoted literature, but the test results show that the cited documents are red [4]. As qualified doctoral graduates, we must have the correct academic morality. When quoting other people's research results, the format must be standardized. The state's latest Bibliographic Description Rules: GB/T7714-2015 "Information and documentation reference rules" has been formally implemented in December 1, 2015, which provides a theoretical basis for doctoral dissertations.

4. **Technology and Research Plan of the Duplicate Checking System.**

4.1. **Technology and Research Scheme of English Duplicate Checking System.**

4.1.1. **Space Vector Model (VSM).** The core idea of the famous vector space model (VSM) is based on the knowledge of statistics. First, the word text need to preprocess, drawed the feature extraction based on word frequency, established a feature vector of the text. This text can be expressed through these feature vector to assign weights reflect the important degree of key words. The whole document has the ability to identify and describe. The similarity of the text is calculated by mapping the text into n-dimensional space vectors based on the methods of word frequency statistics and vector dimensionality reduction. The most commonly used method is to compute the angle cosine between vectors. At present, VSM algorithm has become the basis of many text similarity algorithms, and has also been widely applied to other fields.

4.1.2. **Generalized Vector Space Model (GVSM).** S.K.M. Wong et al established the General Vector Space Model (GVSM) by extending the vector space dimension, the new model improves the traditional VSM model. based on the correlation between words, its basic idea includes improving the spatial dimension of the dissipation condition in text feature space orthogonal, using Boolean algebra the minimum N dimensional space vector representation as space vector N2 dimension, and the vector angle similarity representation.

4.1.3 **Invisible Semantic Indexing Model (LSI).** The basic algorithm idea of the implicit semantic indexing algorithm (LSI) is to transform the text feature matrix into singular matrix by using singular value decomposition of matrix. The singular value decomposition of Singular Value Decomposition (SVD) is the use of formal analysis in the unitary matrix matrix diagonalization method for matrix decomposition, it has important applications in information processing, pattern recognition, statistics and other fields.

In the text to be compared, if there are a large number of the same words, the text has potential similarity:

The first step: the text library creates a text feature matrix, the numerical matrix each is integer; a vector dimension represents the text library all the features, another dimension represents the text library text, numerical meaning represents the number of a feature in a text.

The second step: the special matrix singular value is resloved the small singular value matrix is removed according to the specific rules, singular value matrix representation of text database query vector is obtained, and two vectors are mapped to the same subspace.

The third step: calculate the cosine angle value of the two text subspace vectors, and use the calculated results to indicate the size of text similarity.

The characteristics of LSI algorithm is to remove the small singular values, because the smaller singular values represent the characteristics are not obvious .According to the size of the singular values, the feature items are selected, The characteristics of retained features has obvious effect in the text of the vector. The main idea of the LSI algorithm model is based on the singular value decomposition of the text feature space into text concept space, calculating method of space instead of the original vector cosine vector text concept, improve the algorithm accuracy.

The shortcomings of the LSI algorithm model depend too much on the text information of the corpus. When the number of texts is small, the latent semantics can not be expressed well. Using singular value decomposition (SVD), we can solve the problem of semantic multiple words by using the information of text up and down, which is more reliable than the spatial vector model.

4.2. **Technology and Research Program of the Duplicate Checking System.**

4.2.1. **Text Similarity Analysis Method Based on Attribute Theory.** In 1999, Pan et al used the attribute theory to calculate the similarity of text extraction, created the model with text basic attributes, according to the text attribute barycenter created model affecting factors of text similarity, established the text similarity calculation method based on these factors. According to the experimental results of neural science and basic principle of philosophy and logic, the attribute theory takes the psychological process of human cognition as the main line, applying the advanced mathematical theory of category theory as the tool, and making full use of the results of artificial intelligence, thinking science, cognitive science and computer science to build a mathematical model, which can adapt to the numerical value of human thinking the needs and can adapt to the non numerical needs. The text attribute barycenter coordinate model can calculate the correlation between keywords and keywords by calculating the distance between coordinate points and coordinate points, the correlation of keywords and text through the relationship between the coordinates and simplex calculation, the similarity of the text and the text through the relationship between simplex and simplex. Therefore, the text attribute barycenter coordinate model can express more semantic information, which provides another possibility for the processing method of similarity [9].

4.2.2. **Text Similarity Analysis Method Based on Hamming Distance.** In 2001, Zhang et al proposed that the Hamming distance in information theory can be used for the calculation of text similarity [12], feature extraction of text can be express using Hamming code, and so the distance between the texts can be replaced by the Hamming distance. The biggest feature of this method is that the efficiency of text similarity calculation is improved due to the simple modulo 2 and adding equal operation greatly compared to traditional Euclidean space complex multiplication. The method of computing text similarity based on Hamming distance uses the concept of Hamming distance in coding theory to calculate the similarity of text by calculating the Hamming distance between text and query. Compared with other methods, it has the advantages of simple calculation and so on.

Hamming distance is a basic concept that describes the distance between two N long code words. It can reflect the difference between two code words, and then provide an objective basis for the similarity between code words. For the text, firstly, according to the relevant information, such as the information of keywords and abstract, arranged in a n sequence of code words. By using of these codes, text information make the 1- 1 correspondence of text and code. Similarly, query expressions are also represented by code words. For the original text set, it can be 1- 1 corresponding to the set of code words, study the text similarity relations in text sets, use the Hamming distance between characters to represent, and better reflect the relationship between texts. For D (M1, M2), their distance is between 0 and N, and when the text query by using the N code said completely at the same time, distance value is n, when the text and the query code are completely the same, then the distance is 0, which quantitatively describes the difference between the text the different degree of the similarity function, the quantitative description of the differences between the texts [10].

4.2.3. **Similarity Computation Method Based on Semantic Understanding.** Compared with the text similarity calculation method based on statistics, the method of text similarity calculation based on semantic understanding does not need the support of large-scale corpus, nor does it require long training, and the accuracy is higher. The typical method is the method of similarity calculation by using the knowledge structure of HowNet and the syntax of its knowledge description language, proposed by Jin Bo, Shi Yanjun, et al. (2005). In the calculation of the similarity of words, using HowNet sememes tree structure and HowNet knowledge network knowledge, comprehensive and reliable calculation; through the similarity of the notional set calculation to efficiently compute the sentence similarity; then the similarity calculation based on HowNet semantic understanding is extended to the paragraph and text, the similarity calculation of more practical value. At present, the computation of text similarity based on semantic understanding is mostly limited to the range of words or sentences. The computational efficiency is not high, and more large-scale applications need to be further studied [9].

4.2.4. **String Based Matching Algorithm.** String matching algorithm is the basic unit of string representation of text, the similarity is compared through the matching between two strings, and the algorithm is different from the statistical method of feature items. At present, string based matching algorithms have the following two kinds:

Edit distance based LD algorithm

The edit distance refers to the number of operations between two strings in the transformation, the specific operation has three cases including insert, delete and replace. The LD algorithm is mainly based on the edit distance between two strings to express the size of the two string size difference. For example, the string A: "he's the star of the troupe" and the string B: "he's a star in the theater." the edit distance.

Step 1: replace the troupe with the theater in A";

Step 2: delete "Ming" in A;

Step 3: add "one" in A";

Step 4: The resulting string A and B is the edit distance of 4.

The basic principle of LD algorithm is simple, but the time complexity and space complexity of the algorithm can reach O (M*N) when calculating the matching path in the LD matrix. Therefore, the LD algorithm is not suitable for long text similarity calculation. In the case of short strings, it can obtain good performance, and is suitable for information retrieval, Machine Translation and so on.

LCS Algorithm Based on Longest Common String

The principle of the longest public string method is to match the longest part of the string between the two strings, and calculate the size of the similarity according to the proportion of the same length of the same string. Subdivision, LCS algorithm is also divided into many kinds, the most common is the Needleman/Wunsch algorithm. Needleman/Wunsch algorithm is similar with LD algorithm. get the longest public string through the insert, delete and modify three kinds of operations, is also similar with the LD algorithm in time and space complexity.

The text similarity calculation process based on string matching mainly adopts the idea of dynamic programming, compared between the methods of statistical segmentation based on a breakthrough, while it is not suitable for the similarity comparison between long texts because of the complexity of the algorithm, but it is still worth us to learn and think [10].

4.3. **Problems and Development Directions of System Technology and Research Scheme.** To implement the natural language understanding by using computers, Parataxis of Chinese has more difficulties than Hypotaxis of English. Characteristics of the Hypotaxis language are the changes of words conforming to the morphological rules and paying attention to the syntactic plane, while Chinese is a Parataxis language with reasonable collocation of words and the semantic plane. Therefore, the similarity computation of Chinese text is more challenging, and there are some problems about the present similarity computation methods of Chinese text [11].

According to the statistical methods, there are some methods to achieve the similarity computation, for example, the TF-IDF method based on the vector space model, the approach with the latent semantic index model (LSI), the calculation method based on the hamming distance algorithm and the method based on the attribute theory. These training process have certain limitations such the support of large-scale corpus and time consuming, because the construction of large-scale corpus requires lots of human and time work, and there is also the problem of data sparseness [12]. The TF-IDF method based on vector space model only considers the statistical properties of words in context, discarding structural information and semantic information of the text, so it is only valid if the text contains enough words, and the drawback of LSI is the corpus is too sparse to reflect its potential meaning; the calculation method based on Hamming distance is usually used for the field of sentence fast fuzzy matching, however, the operations of its edit are not flexible, the method is based on the word as the basic unit of measure, but a single word often means insignificant in Chinese; the establishment and calculation model of attribute method the theory are too complicated to suitable for large-scale application.

5. **Conclusions.** In the future text similarity calculation should be transferred from the statistical method to the semantic based method, which is more in line with Chinese language features and language habits. The computer field and Chinese language disciplines need to cooperate more fully mining the original meaning structure of Chinese grammar, to explore a more practical and effective method for the analysis of syntax, to expanded to paragraphs and the whole text from the scope of words and sentences gradually. The text similarity computational efficiency based on semantic understanding remains to be further improved, this study not only requires more efficient algorithm, but also to further optimize the use of the data structure and database technology. In short, the calculation of similarity of Chinese text is quite complex, and it is difficult to find a universal and efficient computing method. We need to continue to explore, find and perfect in practice.

Acknowledgment. This work is partially supported by Institute of Scientific and Technical Information of China (ISTIC) and Peking University (PKU). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Shi Yanjun, Teng Hongfei, Jin Bo. Research and progress of plagiarism identification. Journal of Dalian University of Technology, 2005, 45 (1): 50-57.

[2] Manber U. Finding Similar Files in a Large File System// Usenix Winter Technical Conference. 1994:1- 10.

[3] Baker B S, Manber U. Deducing Similarities in Java Sources from Bytecodes. 1998:15- 15.

[4] Wu Jinqiong. Similarity detection and analysis of doctoral dissertations. Taking Guangxi University as an example. agricultural network information, 2016 (6): 120- 125.

[5] Liu Rongjie. Reflection on the similarity test of graduate degree thesis. Journal of Anqing Teachers College (SOCIAL SCIENCE EDITION), 2015 (4): 146- 148.

[6] Tang Xinghua, Hu Ling, Liu Yong. Discussion on setting up the duplicate database of academic journals on Internet. Journal of Southwestern University (SOCIAL SCIENCE EDITION), 2005, 31 (3): 190 - 192.

[7] Li Zhiming, Wan Fang, VIP. HowNet paper similarity detection system. comparative study of University Library and information science, 2015, 33 (1): 61-64.

[8] Li Ji. Said "checking weight". On the style of study [J]. China graduate student, 2015 (9) .

[9] Shen Bin. Research on Chinese text similarity calculation based on word segmentation [D]. Tianjin University of Finance Economics, 2006.

[10] Sun Runzhi. Research and implementation of text similarity computation based on semantic understanding [D]. Graduate University of Chinese Academy of Sciences (Shenyang Institute of computing technology), 2015.

[11] Jin Yaohong. Computer engineering and applications, [J]. text similarity computing based on context framework 2004, 40 (16): 36-39.

[12] Kong Yuanyuan, Deng Yan. Wan Fang, CNKI, VIP and adaiah paper similarity detection system comparative study of [J]. industry and Technology Forum, 2015, 20 (12): 82-83.