OPEN ACCESS Freely available online

Journal of Research and Development

Review Article

Research on Computational Logic of Academic Detection System

Hao Jiangfeng*

Department of Civil and Commercial Law, Macau University of Science and Technology, Cotai, China

ABSTRACT

The problem of academic misconduct has a long history. Today, with the sufficient information exchange and the rapid development of science and technology, the technical difficulty of "plagiarism" or "plagiarism" is decreasing day by day, and the corresponding governance ability of academic misconduct has also been significantly improved. Relying on the progress of computer science and artificial intelligence, the judgment of the similarity of works can be solved by many non-artificial methods. Correspondingly, for the identification of the similarity of works, the range and proficiency of people's grasp of relevant information become the rationality of computer code. Since the "Zhai Tianlin incident", the ministry of education has increased the governance of academic norms, which is reflected in the operational level of algorithm improvements through the academic misconduct inspection system, regular sampling of dissertations, and changes to the similarity requirements for dissertations. The latter two requirements are based on the completeness of the check system, so there is no room for the audience to discuss whether the black box of the algorithm of the system is reasonable. This article intends to start from the perspective of law, combining the computer logic of the duplicate checking system and the reverse engineering of the results to analyze the rationality of its calculation logic.

Keywords: Duplicate checking system; Computational logic; Copyright law; Works; Academic misconduct

INTRODUCTION

The drivers of human progress

Life is the great work of the earth and consciousness is the evolutionary result of natural selection. Consciousness, as an active reflection of material conditions, has a reason for the generation and existence of any consciousness. Some people joked that "laziness" is the driving force for the development of human society. In order to be lazy, human beings have invented all kinds of tools to assist survival. In fact, the results of natural selection show that activities conducive to gene inheritance are more advantageous than gene preservation. Whether it is human aesthetic tendencies that imply reproduction and reproduction, or the habits and likes and dislikes of daily life, they are the result of genetic selection. Therefore, the real driving force of human development is the inheritance of genes. The so-called "lazy" biological explanation is the preservation of energy. Under the harsh living conditions in the ancient collection period, the scarcity of food will inevitably lead to the lack of calories. Even after entering into agricultural civilization, the backwardness of means of production and the lack of production still mean starvation and death [1].

With the passage of time, the progress of productivity and the improvement of efficiency have brought about "division of labor", and the division of labor has further promoted the development of productivity. In modern times, mankind has realized that science and technology are the primary productive forces. After the completion of the first industrial revolution, "the productive forces created by the bourgeoisie during its less than one hundred years of class rule are more than all the productive forces created by all previous generations more. Therefore, the progress of science and technology has become the top priority pursued by various countries and regions, and the number of papers and research quality active in the frontier of science and technology has become a direct manifestation of scientific and technological progress. Before the information age, most of the review of papers by scientific research institutes was

Correspondence to: Hao Jiangfeng, Department of Civil and Commercial Law, Macau University of Science and Technology, Cotai, China; E-mail: ttlxhjf@qq.com

Received: 27-Dec-2022, Manuscript No. JRD-23-21177; **Editor assigned:** 29-Dec-2022, PreQC No. JRD-23-21177 (PQ); **Reviewed:** 12-Jan-2023, QC No. JRD-23-21177; **Revised:** 03-Mar-2023, Manuscript No. JRD-23-21177 (R); **Published:** 10-Mar-2023, DOI: 10.35248/2311-3278.23.11.214

Citation: Jiangfeng H (2023) Research on Computational Logic of Academic Detection System. J Res Dev. 11:214.

through peer review and anonymous review. Humans are animals of information, and the premise of making judgments on things is the input of information. That is, people who make decisions about new things are constrained by prior input information [2]. However, it is difficult to exhaust the experience. As an incomplete sample, the reviewers are bound to be limited by their own knowledge. Scholars with a little lack of knowledge can pick up some words or scientific research results in the middle of nowhere. After a bit of patchwork, they can form a "thesis" that has taken shape. The gold and jade are outside, and the corruption is in it. The thesis is an article that explains the reality and solves the problem. Its core is to carry out research in various academic fields, or to describe the results of academic research, and the "thesis" that is mass-produced by similar industrial methods is really difficult to discuss academic achievements. Therefore, in the contemporary era of highly developed computer technology, an academic norm detection system with the help of artificial intelligence algorithms has emerged as the times require. Since the "fixed" attribute of text can be mechanically detected by the system, its core function is to use programming tools to compare the characters of the article to be checked with the existing articles, and to detect duplicate fields with the same detection mechanism as the standard. The ratio of the total article to determine whether the article is suspected of plagiarism or plagiarism, and then determine whether the author has academic misconduct, and use this tool to regulate the academic atmosphere and scientific research spirit to promote the development of humanities or science and technology [3].

LITERATURE REVIEW

The origin of plagiarism

The historical origin of plagiarism: According to the definition of "modern Chinese dictionary", the word plagiarism has three meanings, which are:

- Detour and attack.
- Copy others' articles or works without permission and publish them in their own name, and achieve substantial similarity.
- Copy others regardless of reality here we focus on discussing the legal implications of the latter two. In addition, the related concept of "plagiarism" is "plagiarism", "plagiarism" is defined as robbery, plundering, and "plagiarism" is defined as another expression of "stealing".

Therefore, "plagiarism" refers to plagiarism and stealing, including but not limited to other people's writings, graphics and other works. At present, there is no clear definition of related concepts such as "plagiarism" and "homage" in my country's copyright law. In practice, "plagiarism" and "plagiarism" are usually used in the same term. According to the "reply of the copyright administration and department of the national copyright administration on how to identify plagiarism to the Qingdao municipal copyright administration", in the dimension of copyright law, plagiarism and plagiarism is an equivalent concept. In academia, researchers have relatively established that the term plagiarism is more precise and strict in expression than plagiarism by comparing the literary meaning, directionality and focus of the two words [4]. In the revision of copyright law in 2001, the word plagiarism was deleted from the expression, which further explained the fact that plagiarism has the same but more accurate meaning than plagiarism. However, the author believes that the focus of the word "plagiarism" is "plagiarism", which itself has the meaning of illegality and negation, and such wording objectively undertakes the work of the judge's conviction, and it is inevitable to have preconceived suspicions. "Plagiarism" focuses on the behavior of "plagiarism". The empirical concept holds that the behavior itself does not have the attribute of being correct or not, so there is no concept of illegality. Adhering to the objective and neutral spirit of the rule of law, the author believes that the term "plagiarism" is more in line with the realistic concept of the rule of law than "plagiarism" [5].

Plagiarism, as a typical behavior of borrowing, has a typical economic explanation. Compared to spontaneous growth, plunder. It's the quickest way to make money. According to the assumptions and cost-benefit analysis of rational people, once the illegal cost of plundering is less than the gain of plundering, when the income brought by plundering is greater than the spontaneous growth, plundering will occur; when the marginal benefit of plundering is less than the spontaneous growth or plundering. The looting stops spontaneously when the marginal cost of the law is greater than the cost of illegality or spontaneous growth. From a normative point of view, the proof of plagiarism needs to meet the three parts of "contact", "substantially similar" and "defense reasons". Once there is a negation of one factor, the proof of plagiarism is invalid. In reality, it is very difficult to prove the "contact" of behavioral elements alone. Article 64 of my country's civil procedure law stipulates that the parties have the responsibility to provide evidence for their claims. That is, the usual proof rule "whoever claims, whoever gives evidence". The academic interpretation is that "the party who asserts the positive facts bears the burden of proof". Another meaning of the burden of proof is that "the party who cannot fulfill the burden of proof will bear the risk of losing the case", so from a practical point of view, there is a huge risk of losing the case only in the operation of the "contact" link: The publication of the work is to show the public to the public. Show your creative output, the work is specific, and the public is endless [6]. Who, when, where, and in what way achieved "contact", any of the above means a huge cost of proof. If the party is unable to complete the acquisition of evidence, he can also apply to the court or other agencies for evidence collection, document writing, transportation, etc. Administrative efficiency and permission are uncertain factors. In the next step of "substantially similar" identification, there is a huge uncertainty. Substantial similarity is a description of the degree of similarity, and the infringing work reproduces or partially reproduces parts of the original work, thus causing the allegedly infringing work to be similar to the prior work. As far as the substantial similarity is concerned, the subject of its identification is the court, specifically the presiding judge of the case at that time. Different subjects lead to uncontrollable facts. First of all, there is a difference in the identification of the parties at the professional level with the presiding judge, that is, judges who are both judicial personnel, and there are also huge differences

in the identification of different courts or trial levels [7]. See Shaanxi "Monkey-shou case". In this case, the court of first instance and the court of second instance contradicted the determination of substantial similarity, which led to the opposite of the trial result. In the "reasons for defense", in accordance with articles 24 and 25 of my country's copyright law, 14 situations of restriction on copyright are listed successively. That is to say, when the use of the work involved is as described above, the copyright owner can already give up the so-called right relief lawsuit idea.

Explicit and regulation of rights: "Today a rabbit walks away, and a hundred people chase it. It's not that one rabbit is enough for a hundred people. It's undecided. If it's undecided, Yao will bend his strength, but what about all the people? The market is full of rabbits, and those who walk don't take them. It's not because they don't want rabbits. The division has been set, the division has been set, and although people are contemptible, there is no contention. Therefore, to govern the world and the country, it is all about setting the division." The rights need to be clearly stated, and the rights of "the divisions have been set" are "Though people are contemptible, there is no dispute", so the declaration of rights is as important as making a five-colored stick [8].

Title page statement: The earliest copyright declaration in Chinese history was during the Shaoxi period of the Southern Song Dynasty, when Cheng Sheren, a native of Meizhou, Sichuan, engraved Wang Cheng's "a brief history of the Eastern capital". No cladding" statement and the subsequent copyright statement are roughly similar. Limited by the technical means of farming civilization, the circulation of books is nothing more than movable type printing, engraving printing and manual transcription. There are few printing methods. Even for ordinary families with ordinary economic conditions, manual transcription is the main means. With reference to the reasonable provisions for personal use in article 24 of the current copyright law, ancient China has always lacked or even failed to achieve effective regulation at the regulatory level; Paiji statement. The so-called Paiji, also known as Muji, Mowei, tablet or book, also appeared in the Song Dynasty in ancient China, which had a highly developed civilization compared to the border countries. A card record, as the name implies, is to record the author's surname, church number, font size, publication time and related matters in a fixed mark. Common marks are rectangles, tripods, and bells. As early as the Southern Song Dynasty Cheng Sheren's "Dongdu Shilu", later as Zheng Neng's block-printed "twelve poems of the former Tang Dynasty" in the Wanli period of the Ming Dynasty, in which it was a copy of Cen Shen's "Cen Jiazhou collection" in the Tang Dynasty. Yuzhai version, it is not allowed to re-engrave." In Hangzhou in the Ming Dynasty, Hengqiu Pavilion inscribed the words "Guiguzi" card: "Hulin Jiashuli Zhang Ya is issued, and it will be investigated for thousands of miles." In addition to the statement of the card, it is still on the title page. Engraved with the words "no reprinting, violations of thousands of miles will be dealt with" for double insurance, but unfortunately it is not possible to obtain whether this volume has been reprinted or other matters of infringement and investigation; advertising statement [9]. In the Ming Dynasty, Wanli Xin'an Wu Jishi Xichunlou engraved version of "the six classics", with an advertisement engraved on the title page: "Subu is a book, such as obtaining a jade, can't bear to keep it privately, today in the high seas. The image is exquisite, the writing and paper are beautiful, if you look at the Songban, the correction and engraving will be correct. If you look at your husband's rash intentions to add or change it, there is a huge disparity, and a liberal gentleman should learn from it. Huang's jade qingzhai block-printed edition of "compilation of rites and music": "Benya Tibetan plate. It must be studied for thousands of miles." At the end of the Ming Dynasty, Yunjian Pinglutang block printed edition of "Huangming Jingshiwen compilation": "Benya" collected editions, reprints and engravings must be investigated for thousands of miles." In the ninth year of Chongzhen, the lotus nunnery inscribed the red and ink overprint "Guangjin Shiyunfu" in the card: "The Zhuwen on cotton paper, the price is one tael. This ya Tibetan plate. Different from the Song and Yuan Dynasties, the Ming Dynasty' s declaration of rights contained the word “ya” with an official color instead of the “applied to the superior” within the recording organization. In addition, it shows that the idea of copyright protection has emerged from the folk and has been accepted by the official public, so that the official statement is used to intimidate potential infringers and prevent the occurrence of infringement [10].

In the period of farming civilization, ancient China had splendid artistic achievements, among which all dynasties and dynasties had typical literary and artistic achievements of ancient China. Tang poetry and Song poetry, Han Fu Yuanqu, were limited by the printing technology and rights protection conditions at that time, as well as the "Zihan" In the cultural background of “Yan Li”, the signatures at that time were mostly written literally, and rights protection was mostly through the proof of the memory of the villagers or the drive of the conscience of the parties; at the official level, or the creation of court works under the decree and the local government’s graphic records, or related works are included in official historical writings, that is to say, in an atomized society, the rights publication of works relies heavily on the protection of public power, and even the protection of works by large families before officials became famous. It is also difficult to rely on private resources for relief and protection. There is an anecdote in the Tang Dezong period: Li Bo, who was then the prefect of Qizhou, received a poem from a scholar named Li. After reading it, Li Bo found that the paper, the content of the poem, and even the handwriting were very familiar. Only then did he realize that the poem was written by himself. The handwritten book ten years ago, after negotiating, the show admitted that it was an old work he bought at the book market, and tried to change Li Bo's signature and lied about what he had done and spoofed. Broadcast was rogue". This incident reflects the convenience of copyright infringement in ancient China and the difficulty of relief, coupled with the backwardness of the official evidence system and the low level of evidence collection, in reality, the copyright owner can only promote it on a large scale to obtain the time and scope of copyright recognition.

Due to the technical characteristics of intellectual property, it is universal, and the differences are relatively small in the east and

the west where technology is relatively synchronized. Before printing technology was introduced to Europe, the local cultural transmission also relied on manual copying as in ancient China. After Gutenberg invented and improved the printing press, a large number of printings can be carried out with only one purchase of the machine and several typesetting adjustments. This technological advancement will inevitably bring about adjustments to business models and regulations. Although there was no modern concept of copyright at that time, the parties concerned already felt the necessity of rights protection. The pamphlet "warning to printers," published in 1525 by the reformation leader Martin Luther, sharply accused the printers of such behavior as being tantamount to road blocking robbers. Correspondingly, the consequences of vicious competition led to a series of rights demonstrations by some booksellers in the British parliament, which eventually led to the passage of the world's first copyright law by the British parliament in 1709. The law on the rights of books for a certain period of time (referred to as the Queen Anne Decree). Since then, copyright has been recognized as an intangible property right. Within a reasonable range, the official should give adequate protection to copyright and coordinate the interests of copyright owners and printers. This has become a basic civil and commercial legal principle. After the specific printing right (copyright, which is called copyright or copyright in modern times) is infringed, the right holder can apply to the local court for an injunction and require the infringer to bear the infringement liability. The above behavior largely has the prototype of modern intellectual property rights.

Contemporary plagiarism and norms: As far as daily life is concerned, "plagiarism", as a general term in the folk perspective, is actually a function of definition and characterization. The denial of "copying" behavior in the literal sense completes the evaluation of violation of public morality and even the law in nature. In the field of judicial practice and theory, "plagiarism" and "plagiarism" are not appropriate terms. In the current legal provisions, article 1185 of the civil code: If the intellectual property rights of others are intentionally infringed, and the circumstances are serious, the infringed party has the right to request corresponding punitive damages. The verb in the civil code is the normative word "infringement", and there is only a judgment on the value level, but there is no qualitative before the judgment. Since then, more detailed regulations on intellectual property infringement have been regulated by other lower laws. Article 52 to 57 of the copyright law stipulates that the verbs of action are "infringement", and the verbs of article 58 are "infringement", which means that there is no so-called "plagiarism" or "plagiarism" in the provisions of the copyright law. In a word, such qualitative terms do not exist in jurisprudence and judicial practice. But the legal activities themselves are the adjustment of daily life, the guidance and use for the general public. In Article 37 of the regulations for the implementation of the copyright law, "there is an act of 'infringement' listed in article 48 of the copyright law, which at the same time damages the public interests, in which the word “infringement” still, applies. The above, by sorting out the terms of copyright involving "plagiarism" or "plagiarism", shows that the core concept in the contemporary norm system-rights and obligations are the center of norm extension. For other peripheral concepts, it is not commonly used by the official specification.

In the contemporary norms of copyright infringement, there is not much difference in essence from the ancient times, and it also requires the copyright owner to "discover", "obtain evidence", and litigate. Works with text as a carrier have a long history of development, and the content is fixed, and the format is relatively easy to transform. Today, with the rapid development of computers, computer programs can be used to identify and judge. For other works with higher complexity, greater difficulty in production, and multiple communication channels, such as audiovisual works, new media works, and even relatively simple musical works, the identification of pre-infringement and the determination of substantial similarity are still relatively backward regulatory areas. For example, in Shaanxi's "monkey longevity case", Shaanxi Ren Xinchang published his own "monkey longevity picture" in 1998. In 2007, Shaanxi Ren Xinchang discovered Li Zhongyuan's "Zhongyuan calligraphy and painting" webpage and his comments on "Tai Chi" on the website page. The introduction of "monkey shou" is also composed of a cursive "shou" character and a monkey-shaped pattern. Ren Xinchang believes that Li Zhongyuan's "monkey shou" works are different from the "monkey" in his "monkey shou tu" only in the tail part. Except for the size of the picture and the position of the composition, the rest are highly similar or even the same. As for the protection of copyright, in April of the same year, Ren Xinchang sued Li Zhongyuan in court for allegedly infringing his copyright and demanding compensation. Similarly, in "Lu Xun's printmaking case", "Qiong Yao Yu Zheng case" and "meeting Aobao on the moon", the original copyright owner discovered the suspected infringement, and then collected evidence in detail and proceeded with litigation. In the field of writing, especially the "duplication check" of papers in the graduation season of colleges and universities has become an important way to "discover" rights, and it is also an opportunity for students to experience the high similarity rate of written expressions. Colleges and universities only mechanically apply the academic misconduct detection system (hereinafter referred to as the weight check system), but cannot examine the rationality of the system at the academic level.

The duplicate checking system uses the high-efficiency computing features of computers to convert written works into computer language, identify duplicate content by comparing the similarity between characters, and present the occupancy ratio of similar content in the form of numbers to achieve duplicate articles rate judgment tool. The surface of this system is computer software code, but the application of mathematical logic is still behind it. In the development of human history, mathematics plays an irreplaceable role. Mathematics is a universal means for human beings to strictly describe the abstract structures and patterns of things, and can be applied to any problem in the real world. Therefore, the same applies to legal issues in the social sciences. In intellectual property law, where law and science and technology are relatively closely integrated, mathematical logic can of course be used for legal analysis and research. In the contemporary text recognition of

intellectual property infringement, it is necessary to explore the rationality of the academic detection system in particular.

DISCUSSION

Mathematical logic deduction of mainstream algorithms

The materialist dialectics of marxism summarizes the methods of researching science and technology, which are usually induction and deduction. In reality, human cognitive activities can only first come into contact with individual things, and then organize them into general ones. After they are theorized and generalized, they can be extended from the general to the individual, and so on and so forth, promoting the deepening of cognition. Induction is from the particular to the general, and deduction is from the general to the particular. For the study of academic normative systems, induction to deduction can also be used for analysis. Since CNKI's academic misconduct detection system and PaperPass detection algorithms are not public, the author conducts a deductive analysis of the current general methods, and reverses the results of the above algorithms at the end of the article to obtain their mathematical logic and conduct legal studies.

Hamming distance algorithm: The hamming distance algorithm, or hamming distance, was originally an algorithm used in data transmission error control coding. Because of its error control characteristics, it can be understood as the function of controlling and identifying different characters, and then identifying the distinguishing parts. Stripped from the original, and then compared. Specifically, the hamming distance is a concept that represents the number of characters in the corresponding positions of two (same distance) strings. Usually, d(x, y) represents the hamming distance between two words X, Y. By comparing or operating two strings, and counting the number of 1s, this number is the hamming distance. Divide this number by the amount of zongzi characters to convert to a percentage.

The basic idea is to identify->repeated part extraction->calculation.

- Identify, enter the article to be checked and record the key value.
- The repeated part is extracted, and the continuously repeated part is extracted. What needs to be added here is the limit of "continuous", which is adjusted according to the requirements of the reviewer. It is said that the continuous part of CNKI is identified as 13 Chinese characters.
- Calculate, divide the number of consecutive parts by the total number to get the overall similarity rate, where the dividend is the article to be checked.

Specifically, the python language expression of the hammering distance algorithm is:

```
def checkfun(namestr):

subject={} # record the duplicate check result, key value pair, original sentence+repetition rate

summary={}

# 1 Find the historical data of the comparison library

checkpath = "../CheckRepeat/database/OrigCorpus/cutdatas.txt" # Compare project corpus in database

with open(checkpath,"r",encoding="utf-8") as f:

checklist=[line[:] for line in f.readlines()]

subjectname=[sub for sub in checklist if "subject" in sub] # item name

summaryname=[summ for summ in checklist if "summary" in summ] # Project introduction

if "subject" in namestr:

# 2 Perform article name verification operation

for rline in subjectname:

line = ''.join(str(rline).split(' ')[2:])

subp = difflib.SequenceMatcher(None,namestr.split('\n')[0].replace('subject',''),line).ratio()

subject[line]=float('%.4f'%(subp))

if "summary" in namestr:

# 3 Perform article content verification operation

for rline in summaryname:

line = ''.join(str(rline).split(' ')[2:])

sump = difflib.SequenceMatcher(None,namestr.split('\n')[1].replace('summary',''),line).ratio()

summary[line]=float('%.4f'%(sump))

# 4 Print the test results

outreslut=""

sort1=sorted(subject.items(),key=lambda e:e[1],reverse=True) #sort

outreslut +="Article name: "+"*"*5+"["+namestr.split('\n')[0].replace('subject','') + "]"+"*"* 5+" duplicate check results are as follows:\n\n"

for item in sort1[:1]:

if item[1] >= 0.5:

outreslut += "with \t[<span style=\"color:red\">"+item[0].replace("\n","")+"</span>]\t in the database The highest similarity rate: <span style=\"color:red\">"+str(item[1])+"</span>\n"

else:

outreslut += "<span style=\"color:green\">No duplicate item profiles found</span>\n"

sort2=sorted(summary.items(),key=lambda e:e[1],reverse=True) #sort
```

```
outreslut+="\n\nProject introduction:"+"*"*5
+"["+namestr.split('\n')[1].replace('summary','')+"]"+The duplicate check result of "*"*5+" is as follows:\n\n"

for item in sort2[:1]:

if item[1] >= 0.5:

outreslut += "with \t[<span style=\"color:red \">"+item[0].replace("\n","")+"</span>]\t in the project library The highest similarity rate: <span style=\"color:red \">"+str(item[1]) +"</span>\n"

else:

outreslut += "<span style=\"color:green\">No duplicate item profiles found</span>\n"

# 5 Write to file

with open("../CheckRepeat/database/DealCorpus/checkout.txt",'w',encoding='utf-8') as f:

f.write(outreslut)

print(outreslut)
```

By observing, we can find that the core content of the above code is:

```
for rline in summaryname:

line = ".join(str(rline).split(' ')[2:])

sump= difflib.SequenceMatcher(None,namestr.split('\n')[1].replace('summary',''),line).ratio()

summary[line]=float('%.4f'%(sump))
```

That is to say, the recognition logic of the content of the article is the key point, and the core of hammering distance is to recognize and replace the repeated text and check after replacement, and its recognition range is the set database. Then, for authors, the object of determination of similarity is not limited to one or some authors, but all articles in the database.

Hash algorithm: Hash algorithm is a general term for an encryption algorithm. This algorithm is a "cryptographic" algorithm that can only be encrypted but not decrypted. Since the original design of the computer is a binary algorithm of closed circuit 0 and channel 1, it is necessary to convert human language and digital language into mechanical language that can be recognized by the computer, that is, characters. This algorithm can convert any length of information into a fixed-length string. Due to the characteristics of the base, the output string has two characteristics: the local change of the input value will lead to a huge difference in the output hash algorithm value. Only the exact same input value can get the exact same output value. There is no law between the input value and the output value, so the input value cannot be calculated from the output value. To find the specified output value, we can only use the heuristic enumeration method, which also makes it possible to identify and determine the text that can be converted into the same indicator.

For easy understanding, examples are used to explain the generation rules of the improved hash algorithm in detail. The generation of simhash is divided into five steps: Word segmentation->hash->weighting->merging->dimensionality reduction.

Participle: First, the text segmentation is judged to form the characteristic words of this article. Then, a word sequence with noise words removed is formed. Finally, add weights to each participle. We assume that the weight is divided into 5 levels, the parentheses represent the importance of the word in the entire sentence, the larger the number, the more important.

Hash: Use the hash algorithm to turn each word into a hash value, and turn the string into a string of numbers. Remember what I said at the beginning of the article? To improve the performance of similarity calculation, it is necessary to convert the article into digital calculation. Now is the time for the dimensionality reduction process.

Weighted: After hashing the result in the second step, it is necessary to form a weighted number string according to the weight of the word. For example, the hash value of "higher replaceability" is "100101", which is calculated as "4 -4 -4 4 -4 4" by weighting; the hash value of proper nouns and new words is "101011", which is calculated as "5 -5 5 -5 5 5" by weighting.

Merge: Accumulate the sequence values calculated by the above words to become only one sequence string. For example, "4 -4 -4 4 -4 4" for the subject above, "5 -5 5 -5 5 5" for proper nouns, accumulate each bit, "4+5 -4+-5 -4 +5 4+-5 -4+5 4+5" ==> "9 -9 1 -1 1 9". As an example, only two words are counted here. The real calculation needs to accumulate the sequence strings of all words.

Dimensionality reduction: Turn the "9 -9 1 -1 1 9" calculated in step 4 into a string of 0 1 to form our final simhash signature. If each bit is greater than 0, it is recorded as 1, and if it is less than or equal to 0, it is recorded as 0. The final result is: "1 0 1 0 1 1".

Specifically, the python language expression of the hammering distance algorithm is:

```
private BigInteger hash(String source) {

if (source == null || source.length() == 0) {

return new BigInteger("0");

} else {

char[] sourceArray = source.toCharArray();

BigInteger x = BigInteger.valueOf(((long) sourceArray[0]) << 7);

BigInteger m = new BigInteger("1000003");

BigInteger mask = new BigInteger("2").pow(this.hashBits).subtract(

new BigInteger("1"));

for (char item : sourceArray) {

BigInteger temp = BigInteger.valueOf((long) item);

x = x.multiply(m).xor(temp).and(mask);

}

x = x.xor(new BigInteger(String.valueOf(source.length())));
```

```
if (x.equals(new BigInteger("-1"))) {
x = new BigInteger("-2");
}
return x;
}
}
public BigInteger simHash() {
int[] v = new int[this.hashBits];
List<String> words = cutSentenceToWords(sentence);
for (String word : words) {
BigInteger t = this.hash(word);
for (int i = 0; i < this.hashBits; i++) {
BigInteger bitmask = new BigInteger("1").shiftLeft(i);
if (t.and(bitmask).signum() != 0) {
v[i] += 1;
} else {
v[i] -= 1;
}
}
}
BigInteger fingerprint = new BigInteger("0");
for (int i = 0; i < this.hashBits; i++) {
if (v[i] >= 0) {
fingerprint = fingerprint.add(new BigInteger("1").shiftLeft(i));
}
}
return fingerprint;
```

The core part of it is:

```
int[] v = new int[this.hashBits];
List<String> words = cutSentenceToWords(sentence); for (String word : words) {
BigInteger t = this.hash(word);
for (int i = 0; i < this.hashBits; i++) {
BigInteger bitmask = new BigInteger("1").shiftLeft(i);
if (t.and(bitmask).signum() != 0) {
```

That is to say, when the strings of similar parts are recorded, and the combined weight after combined calculation is greater than the value designed by the examiner, the computer will mark the part of the characters in red to realize the identification of the similar parts. The calculation is a binary mathematical operation, and the result can be obtained by dividing by K and taking the remainder.

In addition, in the process of character conversion, there is a problem in the bitmask naming of the hash algorithm, that is, for the modification of a single character, the recognition sensitivity of the hash algorithm is low, even in large-scale plagiarism and copying, once the individual Chinese characters are changed, its recognition effect will be greatly reduced.

Correspondingly, there are also the Euclidean algorithm, the improved simhash algorithm, the improved minhash algorithm and the cosine theorem algorithm. Although the methods are different, the logic behind them is the same that is, identifying and extracting similar parts through the tool features of the computer, and obtaining the similarity value through the overall proportion of similar parts for the reviewer to judge. So the question here is to judge the rationality behind the algorithm.

The inspection of the rationality of the algorithm

Through the deduction of the above-mentioned main algorithms, we know that the computer algorithm is only an appearance, and the problems involving the legal level are still being adjusted by people as social subjects. The premise of negotiation is equality, and the premise of equality is professionalism. However, in the face of the identification of CNKI or other duplicate checking software, non-computer disciplines are unable to discuss, which provides the premise of this article.

CNKI's duplicate checking rule is that the repetition of 13 consecutive Chinese characters is considered a repetition. By BigIntegerization of the identification of 13 consecutive Chinese characters and their paragraphs, they are compared with the CNKI database. The CNKI database includes academic journals, journal literature, the number of journal issues and literature, including academic journals, newspapers, dissertations, yearbooks, reference books, conference proceedings, foreign literature, and other major research and study-related documents. Resource In terms of the quantity and quality of various resources, it is far superior to other similar products, and in terms of years, the collection time is also relatively long.

Substantially similar logic of algorithm rationality: Substantially similar. The so-called "substantial similarity" means that the infringing work reproduces or partially reproduces the original part of the original work, thus causing the allegedly infringing work to be identical to the prior work. This provision is the prerequisite for the establishment of intellectual property rights for the determination of "infringement of civil rights and interests". So the premise here is that the scope of "prior works" is all historical prior works? Or some or several prior works? In terms of the scope of the comparison, it seems that the how net system, which represents the official academic direction, uses the former concept. So what is the rationale for this provision?

Copyright law is a law that coordinates the interests of authors, the surrounding of works, and the interests of the public. It not only protects the interests of creators in works, but also protects the interests of civil and commercial subjects around works, and at the same time, it should not limit the creative space of the public. Therefore, the protection of creators on the one hand

reflects the protection of the past and at the same time promotes the birth of future intellectual achievements. In the field of written works, the specific description of a specific field is fixed. For example, in the field of law, the names of legal norms, government regulations, and public documents are typical. In addition, the name of the country is in the front, and it is easy to cause more than 13 consecutive Chinese characters. Repeat if a certain fact and norm are defined in the literature review, the creator's room for play will be squeezed to the extreme. What is good at the top is necessary at the bottom. Under the leadership of the CNKI detection system, a number of commercial detection systems have emerged as the times require. The author personally tested a certain detection system. The detection algorithm not only included the name and definition of legal norms into the detection, but even included the author's unit, region, zip code, and date of birth into the detection scope. Since the company protects the algorithm in the form of trade secrets, the author cannot obtain its algorithm code and cannot analyze it from the perspective of mathematical logic and computer, but from a practical point of view, this approach has obvious problems.

Tolerance for "coupling". The colorfulness of the world lies in the diversity of ideas, and the descriptions of the works also have a myriad of expressions. However, for a specific thing, the types and forms of its description and expression are limited in number. For example, in the field of philosophy, where language is highly concise, the development or decline of things is collectively referred to as "movement"; in the field of law, which is also highly generalized, The name of the clockwork specification and the definition of things have been polished and tested by countless scholars, and it can even be used to describe the precision of its definition by "rewarding a thousand pieces of gold". The diversity of Chinese language brings us the image experience of hieroglyphics, but there are still its own laws behind the language. The arrangement of subject-verb-object can be reversed as subject-verb-host, but after a long practice of several points and the writing of thousands of scholars, many expressions have actually emerged one after another. Machinery has no subjective initiative, and the mechanical division of the detection system is really strict on the creative concept of later creators. The law still allows “coincidental” situations, that is, other requirements for infringement are met and no punishment is imposed. Why is the detection system in the major link of degree conferring instead of the academic committee's responsibility to check the paper? People are not tools, people are goals. And after the "Zhai Tianlin Incident", every graduation season online community is bound to be full of vulgar words. The helpless act of lack of manpower is a temporary compromise, and it is quite suspicious of replacing the main position of people with this. Therefore, whether it is to comply with philosophical guidance or to implement humanism, the review and revision of mechanical algorithms has now been proposed, and the need for manual intervention is also calling.

Logical remodeling of algorithmic decisions: Substantial similarity is the core element of infringement judgment. The inspection of substantive similarity focuses on the fixed "expression", and does not delve into the "content" part. But from the point of view of expression, the fixed expressions of "plum blossom brand" and "Gong Suo Liancheng" are different. Therefore, according to the mechanistic algorithm, of course, the two cannot be regarded as substantially similar, and thus do not constitute infringement. But people have subjective initiative, and people's consciousness will actively mediate the expectations of things, and this process is difficult for even the subjects. The judgment idea in "Qiong Yao Yu Zheng case" is to use the "content" part to complete the judgment on the substantial similarity of the two works. Three years after the judgment, Yu Zheng apologized to Ms. Qiong Yao at the end of 2020, which further confirmed the existence of the original "plagiarism". There are countless such cases, such as "Shaanxi Monkeyshou case", "Zhuang Yu Guo Jingming Case" and "Han Han plagiarism case". Therefore, by analogy to the article, we can identify and judge the similarity objectively, but the inspection of the content should not be ignored. After the mechanical inspection, manual inspection can be introduced to the articles that are more controversial, and finally artificial inspection can be used. Verification shall prevail. After all, artificial intelligence is still in its infancy. Machine intelligence in the era of weak artificial intelligence does some relatively simple repetitive tasks, and matters involving fact judgment and value selection should still be people-oriented.

Things are changing. Things are generally moving, and when we allow the duplicate checking algorithm to detect articles, the logic behind it will change. In the process of writing the article, due to the interference to the detection algorithm, the language will change to a certain extent, and this movement is not developed from the perspective of truth, goodness and beauty, but to avoid the algorithm's duplication check. Similarly, in the writing of papers at the end of scientific research, the focus of researchers is no longer on scientific and technological progress or content innovation, but on the publication of articles. After all, researchers also need to complete projects to support their families. The world is full of interests, and the "weight reduction" of articles has also become a business. By developing and commercializing algorithms that avoid the weight checking algorithm, both economic losses and social efficiency can be regarded as major losses. Such losses are often invisible and inconspicuous, making it difficult for officials to detect and regulate. In addition to the submission in the Chinese context, the author also contributes a little in the foreign language context. The review of domestic editors focuses on repetition rate and writing. There seems to be a certain degree of vigilance against conceptual innovation and technical improvement, perhaps due to the internal repetition rate requirements of editors and the benefit measurement of merit and demerit dialectics. The technically advanced undergraduates, masters and doctorates are not particularly friendly, and there are often tens of thousands of words for thesis requirements and discrimination against associate professors and below. In the foreign language environment, foreign editors seem to pay more attention to the innovation of ideas and the development of technology, and everything is subject to content and innovation. My country has now become the second largest country in terms of economic volume. It is undeniable that there is still a certain gap with developed countries in some fields. The reality does

not represent the future. After correcting improper concepts and rules, it may catch up with other countries or even complete the catch-up. Overtake the gap still exists, and comrades still need to work hard.

CONCLUSION

The difference between humans and animals lies in whether they can make and use tools. The development of tools represents the progress of humans and society. When mathematics was invented and developed, it has become a powerful tool for human beings to understand the world and transform the world. The computer text check system after the transformation of mathematical ideas has greatly reduced the workload of human beings, and mechanical plagiarism and plagiarism can be good regulation, and then to a certain extent, promote the progress of society.

In addition, there are many problems that should not be ignored in the duplication checking algorithm. The disadvantage of mechanism is that it denies the view of the connection and development of things. In the absence of the initiative of the main hexagram, the requirements for the choice of words and sentences will appear higher. As the result of scientific research-the core essence of the paper is innovation, not word innovation. The times must progress, and human beings must develop. Under the current changing international situation, only self-improvement can stand in the forest of the times. Contradiction is the driving force for the development of things, and the solution to the contradiction becomes the driving force of social development. Innovation still has a long way to go, and we still need to work hard to make progress.

People are ends, not means. The development of tools is to serve people better, not to limit the development of people, or even enslave people as subjects. People with free will are the main body of society, and the advancement of any tool is to make people a better person. The mechanical application of the duplicate checking algorithm is somewhat contrary to "rationality" and the historical fact of the pursuit of "perfect goodness". When the direction is wrong, stopping is progress. When we, as the subject, are lost in the jungle of mechanical tools, we should stop and reflect on ourselves. The objective existence of material reality does not disappear because it is not reflected by consciousness. From the point of view of human progress and development, if it cannot be recognized by the world of consciousness, there will be no chance to be driven by the spirit of Nus, through the operation of laws. Achieving a higher, better, more beautiful purpose, and the world will become a state of "stagnation". Whether from a national level, a social perspective, or human civilization, there is room for debate on the norms of articles that crystallize wisdom. However, as mentioned above, associate professors are still discriminated against, so what about students? I don't want to write this article until some discipline prosecutors who may see people's will.

REFERENCES

1. Liao HJ, Lin CH, Lin YC, Tung KY. Intrusion detection system: A comprehensive review. J Netw Comput Appl. 2013;36(1):16-24.
2. Abdallah A, Maarof MA, Zainal A. Fraud detection system: A survey. J Netw Comput Appl. 2016;68:90-113.
3. Smaha SE. Haystack: An intrusion detection system. In Fourth Aerospace Computer Security Applications Conference. Orlando, FL, USA, 1998.
4. Sun Z, Miller R, Bebis G, DiMeo D (2022) A real-time precrash vehicle detection system. InSixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings. Orlando, FL, USA, 2002;171-176.
5. Papageorgiou C, Evgeniou T, Poggio T. A trainable pedestrian detection system. InProc. of Intelligent Vehicles 1998;241-246.
6. Wang RJ, Li X, Ling CX. Pelee: A real-time object detection system on mobile devices. Advances in neural information processing systems. 2018;31.
7. Vigna G, Kemmerer RA. NetSTAT: A network-based intrusion detection system. Comput Secur J. 1999;7(1):37-71.
8. Hoque MS, Mukit M, Bikas M, Naser A. An implementation of intrusion detection system using genetic algorithm. arXiv preprint arXiv:1204. 2012;1336.
9. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity. 2019;2(1):1-22.
10. Javaid A, Niyaz Q, Sun W, Alam M. A deep learning approach for network intrusion detection system. In proceedings of the 9th EAI international conference on bio-inspired information and communications technologies (formerly BIONETICS). 2016;21-26.