The Design of the Majority Voting Algorithm Based on Search Engine for the Text Copyright Detection Crawler

Yu Liu
Zhejiang University of Software College, Zhejiang Ningbo 315000, China
service@52exe.cn

Abstract. With the rapid development of the Internet era, it is very common for the websites to illegally reproduce the written works, which is a special form of copyright infringement with the characteristic of high covertness, fast speed and extensive influence. This behavior seriously violates the information website dissemination copyright of the authors and their authorized publishing websites, which has extremely serious damage to the healthy development of the text industry! To this end, the four state departments of State Copyright Administration, State Internet Information Office, Ministry of Industry and Information Technology, and Ministry of Public Security jointly carry out several times of "Jian Wang" actions to conduct the special harnessing and combating of network piracy. Against the phenomena of online illegally reproduction written works, and through researching the Internet crawler and the websites with illegally reproduction written works, this paper puts forward the text copyright detection crawler system based on search engines and majority voting algorithm.

1. Foreword
The browsing, downloading, using and modifying of a written text is out of control when it is transmitted over the Internet. It is because of the technical, virtual, widespread and complex characteristics of network communication that the copyright of written works has become one of the bottlenecks of the sustainable development of the digital publishing industry [1]. Especially in recent years, with the rapid development of network crawlers and various text acquisition software, the problem of illegal reprint of text is even worse than before. Usually, when we search a text on a search engine, we would find that a large number of websites have the same work. It is difficult to determine that which one authorized and which one illegally reproduced. In recent years, relevant departments and organizations have combated online piracy of written works for many times. At the same time, with the improvement of public consciousness of copyright, compensation for the economic case become common for the network text in the event legal action illegally reprinted. It is more typical of the Baidu Wenku infringement scandal that hundreds of the media people signed complaint information violations, as well as "news porter" copyright disputes (the case of tencent accused Jinri Toutiao) [2] [3] [4].

Illegal reproduction of text works on the Internet are characterized by the rapidity, concealment and extensive influence. At the same time, the administrators of infringing websites are often unable to be contacted while safeguarding their rights. The cost of copyright detection and safeguarding rights is too high. Thus, it is particularly important to put forward a kind of test system with high efficiency, more
convenient, less investment, and accurate result. The traditional test form is based on manual testing and setting up of the copyright database. And the infringement of manual testing requires a lot of artificial, time consuming, also the possibility of artificial laxity is extremely high, as well as the high cost of setting up database copyright. It is hard to take a lot of small and medium-sized enterprises or individuals, so this paper points out a search engine and majority voting algorithm based on the text copyright testing crawler system and the major search engines on the Internet. By means of network crawlers, websites with possible text infringement can be grabbed. And by means of majority voting algorithm, the illegal reproduction of the text works detected is the final output. Compared with the traditional method, this method reduces the cost of labor input and hardware construction with the features of less input, less procedure, and more accurate test results.

2. Traditional methods for detecting the copyright of written works

2.1. The artificial detection method
This method usually refers to the artificial network retrieval of the text works that already known. For example, do web retrieval through BBS, search systems of the communities, or search engine systems. This method is characterized by relatively low technical difficulty in execution. It only needs to carry out data retrieval on the Internet at regular or irregular intervals, and then determine whether it is illegally reproduced or not manually. However, the disadvantage of this method is that it requires a lot of human resources and a large amount of time. At the same time, it is likely that human factors will lead to the misjudgment or omission of some data. This method is only applicable to detect the copyright of few works, not the copyright of a large number of written works.

2.2. The detection of the matching of text copyright database
Usually it refers to the method that text copyright database is established, and constantly does data acquisition on the Internet through the web crawler technology, then stored in the database. When there is a need for copyright protection, with matching data automatically in the data, and it will continuously get new data, according to the crawler for updates the match, finally get a more complete copyright test results.

This method can be used to detect the copyright of written works on a large scale. However, this method has relatively high technical requirements, and at the same time, a great deal of investment is required in the early stage for the construction of basic hardware. The efficiency of this method has high requirements on the data storage capacity in the database, the efficiency and the quality of crawler. At the same time, this method is not suitable for SMEs or individuals to test the copyright of written works [5].

3. Design of text copyright detection crawler

3.1. The overall system design
The development of the writing on the network needs to create a fair and reasonable order, for illegally reproduced stripped the reasonable return of creative labor and investment of authors and publishers, and reduced the national tax revenue and public employment opportunities, also made the network users lose the opportunity of enjoying better products. But for small and medium-sized enterprises or personal websites, it is difficult to have a lot of energy or money for each writing test copyright or copyright monitor, therefore it is particular important to use a copyright identification system which is simple, cheap, accurate and efficient by means of technologies. Based on the study of traditional text copyright detection methods and their advantages, disadvantages and characteristics, a majority voting algorithm based on search engine is designed to detect text copyright crawler.
The system mainly reads the text works prepared for detection, and obtains several paragraphs at random, then automatically searches for these paragraphs through the crawler module to obtain some top-ranking items, finally makes the judgment of illegal reproduction through the majority voting algorithm.

3.2. The text splitting module design
Text split module is mainly for the detection of text works split processing. Take article A for example, the word number of the article is \(a\_\text{count}\), and the number of target random paragraphs is \(d (d_1, d_2, d_3 ...d_d)\), also the number of each text is \(c (c \ll a\_\text{count})\), then \(d\) random integers are obtained in the range \([0, a\_\text{count}]\), and the result of sorting from small to large is \(e_1, e_2, e_3 ...e_d\). At this point, the text range of each random paragraph is (\(S\) denotes the random number corresponding to this paragraph):

\[
\begin{cases} 
[s, s + c] & (k \geq c) \\
[s - (s + c - a\_\text{count}), a\_\text{count}] & (k < c) 
\end{cases}
\]

\(k\): The number of intermediate words between the beginning random number and the end of a written work.

The number of intermediate words between the beginning random no. and the end of a written work

3.3. The Crawler module design
According to the text split module, the random paragraph number of the split is \(d (d_1, d_2, d_3 ...d_d)\), and the module will send the paragraphs to \(b\) search engines separately. At this time, the search engines record as \(e_{r_1}, e_{r_2}, e_{r_3} ...e_{r_b}\), and takes the \(m\) (\(re_1, re_2, re_3 ...re_m\)) former results of each search engine [6] [7] [8]:

**Figure 1.** Overall design scheme of the system

**Figure 2.** Data diagram of a test result of the text to be tested
3.4. The design of voting module
According to the results obtained by the crawler module, data processing is carried out to integrate all search engine results of each paragraph. If the target pages are consistent, then the results are merged as \( RES(d_{1res1}, d_{1res2}, d_{1res3}, \ldots, d_{1resn}) \). The combined results are denoted as the sum of the results of the \( d_1 \) paragraph in the \( n \) preceding items on different search engines \((n \leq m)\).

Then the results of all paragraphs are combined as \( tre \), which represents the sum of the results of \( m \) preceding paragraphs in the search engine of \( d \) random paragraphs of the article. The Majority Vote Algorithm was finally adopted.

The result is \( tre \) (it stands for one of the results of \( tre \)), if the number of times of \( RES \) appears in different paragraphs is more than \( d/2 \) times, it is considered that the result has the feature of illegal reproduction, otherwise it does not have the feature of illegal reproduction [9] [10].

3.5. The impact of variables to the system
In this algorithm, there are four core variables as \( d, c, b, m \) and several derivative variables. Letter \( d \) refers to the number of paragraphs randomly divided into the text works to be tested. The larger the number, the higher the matching degree of the result, but some results may be omitted. Letter \( c \) \((1 \leq c)\) means the number of words in every random paragraph. The smaller the number, the more comprehensive the result will be. But it can generate a lot of junk data. Letter \( b \) represents the number of search engines. The larger the number is, the higher the result range and accuracy it will be. Letter \( m \) is to take the results of the before items of each search engine. The larger the number is, the wider coverage of the results, and the more complete of the final data.

3.6. The implementation of copyright monitoring function
Generally speaking, illegally reproduced for a text detection or copyright, did not once get the final result, but it will need to conduct a test every once in a fixed time. On one hand it can increase the emergence of new illegally reproduced as a result, on the other hand it can also improve the result data to achieve the purpose of monitoring. For the purpose of monitoring, monitoring modules can be added in the above system.

![Figure 3. Improved flow chart of text copyright crawler (text copyright monitoring crawler)](image)

Text copyright monitoring function can be realized by improving the majority voting algorithm based on search engine to detect the crawler of text copyright. That is to say, the data is obtained from the search engine continuously and it is judged whether there is illegal feature of the reproduction. If so, the data is recorded and fed back to the user; otherwise the data is not recorded and not fed back [11].
4. Experiment and simulation

4.1. Detection of copyright of a single text

The text copyright of text A is tested. The basic situation of text A is as follows:
- Chapter total words: a_count = 914
- Random paragraphs: d = 5
- Number of words per paragraph: c = 30
- Number of search engines: b = 1 (baidu search for example)
- The m items before searching engine: m = 150

The above algorithm is implemented through the code. It can be seen from the detected titles that the system's automatic detection results are very accurate. Each title basically includes the original title (I would like to be a red dust, mud fragrance floating ten thousand miles). The results of manual test and system monitoring are as follows:

Table 1. The results of manual test and system test

| The original title | I would like to be a red dust, mud fragrance floating ten thousand miles |
|--------------------|------------------------------------------------------------------------|
| Manual check quantity | 16                                                                      |
| Manual detection time | It took 3 people 130 minutes                                           |
| The number of system detection in this paper | 18                                                                     |
| Detection time of the monitor of this paper | 432 seconds                                                           |
| Compare the system omission in this paper with manual detection | -1                                                                    |
| Manually review error number in this system | 0                                                                     |

According to the table, it can be seen that the artificial detection results are time-consuming, and some results have been misjudged, which is a great investment of manpower. However, most voting algorithms based on search engine are much faster and more accurate in detecting crawler system.

4.2. Copyright detection of many text works

Preliminary statistics of copyright examination of text works:

Table 2. Copyright test and preparation data of many text works

| Number | Article word no | Random paragraph | Each passage word number | Search engine number | The m items results before searching engine: |
|--------|----------------|------------------|--------------------------|----------------------|------------------------------------------|
| 1      | 1504           | 5                | 50                       | 2                    | 120                                      |
| 2      | 964            | 5                | 50                       | 2                    | 120                                      |
| 3      | 1532           | 5                | 50                       | 2                    | 120                                      |
| 4      | 2130           | 5                | 50                       | 2                    | 120                                      |
| 5      | 743            | 5                | 50                       | 2                    | 120                                      |

Table 3. Copyright test results of many text works

| Number search engine 2 time (s) results manual results |
|-------------------------------------------------------|
| No | Search engine 1 time (s) | Search engine 2 time (s) | Results | Manual results |
|----|--------------------------|--------------------------|---------|----------------|
| 1  | 435                      | 443                      | 20      | 17             |
| 2  | 445                      | 461                      | 19      | 18             |
| 3  | 443                      | 437                      | 21      | 22             |
| 4  | 442                      | 446                      | 17      | 17             |
| 5  | 450                      | 453                      | 43      | 40             |
Based on the results of the copyright detection experiment and simulation of multiple text works, it can be seen that the majority voting algorithm based on search engine designed in this paper has the features of higher detection efficiency, higher accuracy and less input compared with manual detection.

5. Summarize
Illegal reproduction has seriously hindered the development of written works on the Internet. To build a healthy environment for the development of written works on the Internet, we must resolutely fight against the increasingly popular illegal reproduction. The traditional infringement detection and monitoring of text works are mainly through artificial detection or the establishment of text copyright database. The former requires a lot of effort, what extremely easy to create misunderstandings at the same time, and the latter requires a lot of money. On the basis of the construction of hardware facilities, the two kinds of detection mechanisms, for small and medium-sized enterprises and even individuals are not very appropriate, while this paper use search engines to search results, obtained from the crawler technology batch search results, after using the random section form, and through the majority voting algorithm on the basis of to judgment, it could be the result of illegal reprint. Finally by increasing the monitoring module, make this system to monitor the illegal reprint written works with good effect. In order to verify the feasibility of the system, experiments and simulations were carried out by writing relevant programs. According to the results, the system was proved to be more efficient, more accurate and less input than manual detection.

Acknowledgments
Liu Yu (1993- ), as a male, was born in Jilin City, Jilin province, graduated from school of software, Zhejiang University, who is also the CEO of Jilin Qunlong Technology Co., LTD. And He is the developer of Anycodes online programming platform (2015SR164783), also the copyright owner of Anycodes online programming platform (2017sr164783) and online programming software (2017SR105637). He is the owner of the online programming patent Anycodes also.

References
[1] Wang Ying, Xang Xiaoyan. Management of illegal reprint in online literature [J]. Publishing science, 2016, 24 (3): 69-72.
[2] Zhang Hongbo. Online reprint refused free lunch [J]. China copyright, 2009 (4): 60-61.
[3] Zhang Hongbo. Cracking down on illegal media reprinting requires a hammer [J]. News world, 2015 (5): 1-1.