A Study of Natural Language Plagiarism Detection

Byung-Ryul Ahn *, Heon Kim **, Moon-Hyun Kim **
* anbr0305@skku.edu
** heonyong@hanmail.net
** mhkim@ece.skku.ac.kr

Abstract. Vast amount of information is generated and shared in this active digital As the digital informatization is vividly going on now, most of documents are in digitalized forms, and this kind of information is on the increase. It is no exaggeration to say that this kind of newly created information and knowledge would affect the competitiveness and the future of our nation. In addition to that, a lot of investment is being made in information and knowledge based industries at national level and in reality, a lot of efforts are intensively made for research and development of human resources. It becomes easier in digital era to create and share the information as there are various tools that have been developed to create documents along with the internet, and as a result, the share of dual information is increasing day in and day out. At present, a lot of information that is provided online is actually being plagiarized or illegally copied. Specifically, it is very tricky to identify some plagiarism from tremendous amount of information because the original sentences can be simply restructured or replaced with similar words, which would make them look different from original sentences. This means that managing and protecting the knowledge start to be regarded as important, though it is important to create the knowledge through the investment and efforts. This dissertation tries to suggest new method and theory that would be instrumental in effectively detecting any infringement on and plagiarism of intellectual property of others. DICOM(Dynamic Incremental Comparison Method), a method which was developed by this research to detect plagiarism of document, focuses on realizing a system that can detect plagiarized documents and parts efficiently, accurately and immediately by creating positive and various detectors

Keywords: Multimedia Contents Protection, documents piracy, Plagiarism Detection System

1 Introduction

As the development of technology and the importance of information become more crucial, the cases of plagiarism and intellectual rights violation is increasing. Illegal plagiarism is thriving but research and solutions on this issue is needed domestically and abroad. Discriminating plagiarism and personal feelings the issue requires much time and sources for them to passed on from person to person. A more efficient theory and an objective system are needed.

Selecting copied documents among numerous documents takes a great deal of time and labor, and it is also a complex job to measure resemblance between similar documents.

The research has focused on this point, on how to compare many documents and how to search for reliable documents that are quickly compared with other documents.

This was the beginning point of research and establishment. Of the many methods to detect plagiarized documents, the most common method is to analyze documents with matching sentences or words.

By extracting key-word centered Core Detectors from input duplicates as the main detector, these detectors are dynamically produced and managed to obviate the unnecessary complexities.

The Core Detector is an aggregate including the elements of sentences often used when plagiarized, which detects the possibility of duplication. When actually distinct as a highly possible duplicate, an accurate comparison is done on the falling document and the similarity percentage is output.

The methods on the production process these Core Detectors and production and management of detectors with various lengths will be introduced in the main context.

2 Methodology and software of natural language piracy detection

There are mainly three ways to detect and judge pirated digital documents: the fingerprint method, as a kind of statistical method, examines similarities of used words in the original and copied documents or the frequency of used words; the clustering method measures the coincidences of the original and copied documents; the hybrid method combines the said two methods[1-3]. Other than these, the structure related method focuses on the overall flows of sentences. However, as it is impossible to understand such structures in natural language documents, it is not used widely but for analysis of program source codes.

The natural language detecting software usually uses the fingerprint method, which extracts statistical features or the hybrid method rather than examine structural features of a document.

Currently, foreign natural language piracy detecting softwares are Findsome[4] of Digital Integrity, EVE2[5] of CaNexus, Turnitin of iParadigms[6], CopyCatch[7] of CFL Software Developments, WordCHECK of WordCHECK[8] Systems, etc. As domestic softwares, there are Professors' club[9], Clonechecker[10], and LOFC (Linear Order Function Call)[11], but only Professors' club has proper functions.

COPS[12] converts letters of various formats to ASCII letters and group them by sentence to save in databases. When a document comes in, it examines it in terms of overlaps with existing sentences saved in the databases.

SCAM[13], which is an improvement of COPS, groups letters not by sentence but by word and save them in the database. When a document comes in, it shows the frequency of used words in vectors and detects similarities through the dot-product between these vectors[14].

3 DICOM(Dynamic Incremental Comparison Method) system architecture

Despite the research on algorithms to prevent natural language and measure the
plagiarism rate has been continued and applied, it is not successful.

For example, in natural language, it is difficult to maintain a special structure as the program code, and with a small alter though the meaning stay be the same the code will be recognized as a totally different sentence, which makes it hard to detect.

If it is a small copied file with a few partially-revised sentences, it will be possible to detect piracy through line-by-line comparison. However, it will require substantial time and computer resources to compare numerous documents large in size one by one through line-by-line comparison. In order to overcome the shortcoming, ways to detect piracy quickly and accurately without line-by-line or word-by-word comparison have been sought.

In this chapter, the new system DICOM (Dynamic Incremental Comparison Method) will be introduced, which can overcome the weak points of existing systems and reduce cost and complexity effectively.

3.1 Detector creator

The self document is an original document that should be protected, whereas the typical sample document is a document highly likely to be copied. Among various copied documents, documents high in keyword frequency and similarity are utilized as a detector-creating sample document. As detector-creating sample documents exert a great influence on detector creation and detection effects, documents of high confidence should be selected. In the typical sample document which is extracted from copied documents for detector creation, sentences concerned with the keywords of the original are filtered. Sample document sentences including the keywords and original document sentences including the keywords are compared in the right and left sides by token according to keywords. When both of them conform, the token is lengthened, whereas when they do not, it suspends lengthening.

Various detector cells that went through the keyword comparison in copies and originals are generated. As a detector is made by comparing copy types in actual copies, the detection feasibility has been enhanced. Detectors of various sizes such as phrase, clause, sentence, document, etc. are created. These detectors are collected by a detector collector to act as a detector.

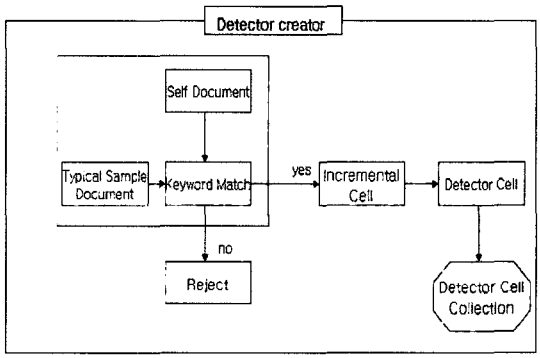

Fig. 1. Detector creator structure

3.1.1 Detection service creation algorism

A: original sentence B: copied sentence

A is composed of n tokens A: 1.....i .....n

B is composed of m tokens B: 1.....j.....m

i is the position of A's start token, j is the position of A's start token

D(A,a): data of A's No. a token, D(B,b): data of B's No. b token

If D(A,a) = D(B,b) : a= a-- , b= b-- (until-> a,b >= 1)

Else stop

$\Delta\alpha$= i-a

If D(A,a) = D(B,b) : a= a++ , b= b++ (until-> a<=n, b <= m)

Else stop

$\Delta\beta$= a-i

C: Core detector C = concatenate { D(A, i- $\Delta\alpha$),....,D(A, i+ $\Delta\beta$)}

3.2 Construction of Core Detector

There are three main factors that decide the generation of the Core Detector.

Core Detector = K(Ct) ∪ R(Ct) ∪ F(Ct)

There are three main factors that decide the generation of the Core Detector.

- K(Ct) is the elements of the Core Detector that are generated from the analysis between a typical sample extracted from a copy and keywords. The numerous keywords used by the user become the main keys to generate the Core Detector.
- R(Ct) are the elements of the Core Detector that are generated from the analysis between a typical sample extracted from a Typical Sample and Related words. The Relation word is a group of related words that come from the experience of being saved in the DB. Because the Related words are grouped from experience, the group continuously grows .
- F(Ct) are the elements of the Core Detector that are generated from the comparison between a Typical Sample and frequently used words, after checking the frequency of the words used in the input document.

3.3 Dynamic Controller

The Dynamic Controller is a manager which manages created detectors and has active processes to enhance the efficiency of detection. The detector collector collects and save various created detectors by using keyword-centered matching algorism, world frequency information, relation information, etc. The detectors in the detector collector are compared to see how much they match with documents suspicious of copy. The number of detection of each detector is recorded and saved in log files.

It is possible to distinguish efficient detectors from those that are not by compiling statistics of log files. Matching detectors are reported to the Dynamic Controller and detectors of low efficiency are deleted through relative comparison with others.

As detection is performed mainly by detectors of high efficiency among numerous detectors, time and resources can be saved.

The Dynamic Controller actively manages detectors and plays the roll of filtering efficient detectors to maintain the optimum number of detectors.

3.3.1 Maintenance Core Detector Validity

A generated Core Detector Set contains numerous elements.

Each element is compared to the sample, which decides whether it was plagiarized which puts it in the core position.

Thus maintaining the effectiveness of the Core Detector Set is the key point of discriminating plagiarism.

- Each element of the Core Detector Set contains an index of the used frequency.
- The elements with low frequency are discharged from the Core Detector Set. When the number of the elements fall under the limit, the algorithm is performed to supplement elements of the Core Detector.

Relative matching rate of core detector

- Core detector A's matching rate = the sum of core detector A's matching rate for all copies / the sum of all core detectors' matching rate for all copies * 100

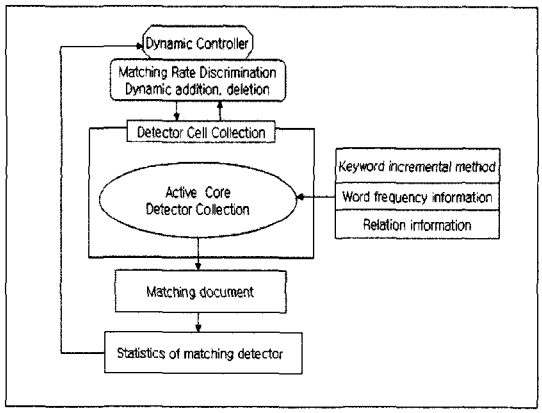

Fig. 2. Dynamic controller structure

4 System Implementation and Experiment

Tests were carried out in terms of word access pattern, sentence access pattern and core detector access pattern.

Table 1 shows the test results of 100 sample documents. As shown in Table 1, the sentence access pattern is the most excellent in terms of speed and complexity.

However, it shows the lowest performance in terms of matching rate and copied document detection rate, which represent the efficiency and suitability of document detection. The word access pattern has a good performance in the matching rate and document detection rate but is the highest in complexity.

The core detector access pattern suggested in the study shows above-the-average performances in speed, complexity and matching rate. In particular, as it has the best document detection rate, it is considered the most suitable for pirated document detection

Table 1. Performance evaluation by pattern (The averaged test results of 100 samples)

	Plagiarism detection rate	Speed	Complexity	Matching rate	Reliability
Word access pattern	26.91%	120	82.55	79.45%	42
Sentence access pattern	35.83%	249	56.6	32.53%	51
Core detector access pattern	43.78%	253	69.76	62.58%	76

Fig. 3 shows the contribution trend of each pattern's copy detection rates according to time with a cumulative broken line, while Fig. 4 & 5, the trend of positive and negative defects according to time. In Fig. 3, the core detector access pattern has better performance in the copy detection rate than other patterns as time passes, whereas the sentence access pattern, the worse performance.

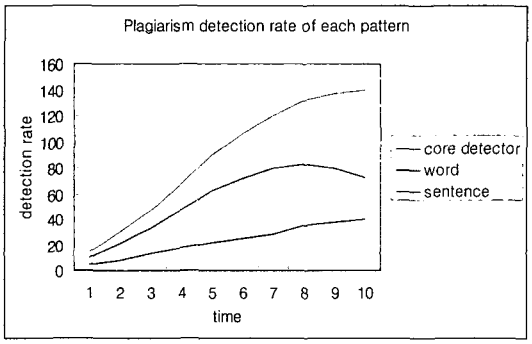

Fig. 3. Plagiarism detection rate of three main patterns

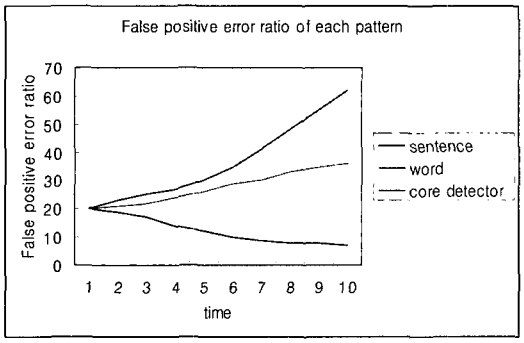

Fig. 4. Comparison of False positive

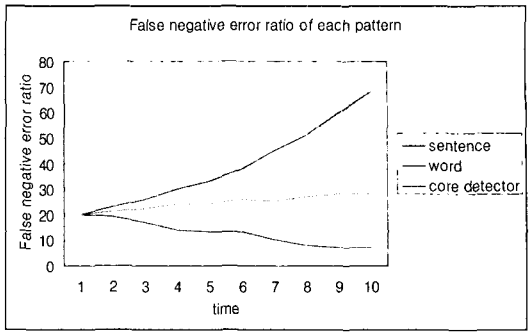

Fig. 5. Comparison of False negative

According to Fig. 4 & 5, the positive defect is highest in the word access pattern, whereas the negative defect is the lowest in the sentence access pattern, which lead to a conclusion that the word access pattern is weak in the positive defect and the sentence access pattern in the negative defect.

It was confirmed that the core detector access pattern is an alternative to overcome the positive defect of the word access pattern and the negative defect of the sentence access pattern to some extent.

5 Conclusion

The study elaborates on the infringement of intellectual property rights and the need of techniques to prevent them as well as on the types of natural language piracy, detection techniques and domestic and foreign natural language piracy detection software. Currently, further study and development are needed in this field and domestically, the utilization of piracy detection software is insufficient. Without solving the problem of the protection of intellectual property rights, we will not be able to leap into an information power in the future and for this, therefore, the protection of intellectual property rights will be essential.

The study suggests a technique focusing on detecting pirated documents most reasonably and quickly by selectively extracting information most likely to be pirated. The study also considered the minimization of redundant complexities. Based on these detectors, piracy will be detected effectively by detecting pirated documents quickly among numerous documents.

The core detector access pattern was also proved to be one of the ways to get over the positive and negative defects of the word access pattern and sentence access pattern.

In DICOM, there is an element, which utilizes keywords in the core detector creation process. In the future, more study on new detection methods for documents without definite keywords such as news script,

editorial, essay, drama script, etc. and complement works are necessary to be done.

References

[1] Computer Program Deliberation & Mediation Committee: http://www.pdmc.or.kr

[2] Mi-Nyeong Hwang, Eun-Mi Kang, Kee-Duck Han "Applying genomic sequence alignment methodology for source codes plagiarism detection"

[3] Computer Program Deliberation & Mediation Committee " Research on Comparison and Application of S/W Assessment Tools"

[4] http://www.findsame.com

[5] http://www.caNexus.com

[6] http://www.turnitin.com

[7] http://www.CopyCatch.freeserve.co.uk

[8] http://www.WordCHECKsystems.com

[9] http://www.gyosuclub.com

[10] http://ropas.kaist.ac.kr/n/clonechecker

[11] http://jade.cs.pusan.ac.kr/ ~ hgcho

[12] S. Brin, J. Davis, and H. Garcia-Molina, "Copy detection mechanisms digital documents." In Proceeding of the ACM SIGMOD Annual Conference, CA, May 1995

[13] Narayanan Shivakumar, HectorGarcia-molina, "Building a Scalable and Accurate Copy Detection Mechanism"

[14] Made Sig Ager,Oliver Dancy,and Henning Korsholm Rohde "Fast Partial Evaluation of Pattern Matching in Strings" Department of Computer Science University of Aarhus.