Document Copy Detection System Based on Plagiarism Patterns*

NamOh Kang and SangYong Han**

School of Computer Science and Engineering,
ChungAng University, Seoul, Korea
kang@archi.cse.cau.ac.kr, hansy@cau.ac.kr

Abstract. Document copy detection is a very important tool for protecting author's copyright. We present a document copy detection system that calculates the similarity between documents based on plagiarism patterns. Experiments were performed using CISI document collection and show that the proposed system produces more precise results than existing systems.

1 Introduction

For protecting author's copyright, many kinds of intellectual property protection techniques have been introduced; copy prevention, signature and content based copy detection, etc. Copy protection and signature-based copy detection can be very useful to prevent or detect copying of a whole document. However, these techniques have some drawbacks that they make it difficult for users to share information and can not prevent copying of the document in partial parts [1].

Huge amount of digital documents is made public day to day in Internet. Most of the documents are not supported by either copy prevention technique or signature based copy detection technique. This situation increases the necessity in content based copy detection techniques. So far, many document copy detection (DCD) systems based on content based copy detection technique have been introduced, for example COPS [2], SCAM [1], CHECK [3], etc. However, most DCD systems mainly focus on checking the possibility of copy between original documents and a query document. They do not give any evidence of plagiaristic sources to user. In this paper, we propose a DCD system that provides evidence of plagiarism style to the user.

[*] This research was supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

[**] Corresponding author.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 571 – 574, 2006.

2 Comparing Unit and Overlap Measure Function

DCD system divides documents efficiently in comparing unit (chunking unit) for checking the possibility of copy. In this paper, we select the comparing unit as a sentence because the similarity comparison between sentences becomes a good norm to calculate the local similarity and can provide plagiarism pattern information between them.

Overlap measure function is used to get copy information of the comparing units extracted from documents. Traditionally, many DCD systems use vector space model or cosine similarity model. It is no problem to calculate the similarity between two objects but it is not enough to calculate the degree of copy. In this research, we suggest the overlap measure function which can quantify the overlap between comparing units and give information about plagiarism.

Let S_o come from an original document and S_c from a query document. The $Sim(S_o, S_c)$ can be calculated as follows.

$$S_o = \{w_1, w_2, w_3 \ldots\ldots\ w_n\} \qquad S_c = \{w_1, w_2, w_3 \ldots\ldots\ w_m\}$$

$$Comm(S_o, S_c) = S_o \cap S_c \qquad Diff(S_o, S_c) = S_o - S_c$$

$$Syn(w) = \{\text{The synonym of w}\}$$

$$SynWord(S_o, S_c) = \{w_i \mid w_i \in Diff(S_c, S_o) \cap Syn(w_i) \in S_o\}$$

$$WordOverlap(S_o, S_c) = \frac{|S_o|}{|Comm(S_o, S_c)| + 0.5 \times |SynWord(S_o, S_c)|}$$

$$SizeOverlap(S_o, S_c) = \sqrt{|Diff(S_o, S_c)| + |Diff(S_c, S_o)|}$$

$$Sim(S_o, S_c) = |S_o| \times \left(\frac{1}{WordOverlap(S_o, S_c) + SizeOverlap(S_o, S_c)}\right)$$

Calculation of $Sim(S_o, S_c)$ gives not only similarity between S_o and S_c but also the plagiarism information. The following table 1 shows how to decide plagiarism patterns.

Table 1. Plagiarism patterns and their decision parameters

Plagiarism pattern	Decision parameters
Sentence copy exactly	$WordOverLap(S_o, S_c) = 1$, $SizeOverlap(S_o, S_c) = 0$
Word insertion	$SizeOverlap(S_o, S_c) \neq 0$, $Diff(S_o, S_c) > 1$
Word remove	$SizeOverlap(S_o, S_c) \neq 0$, $Diff(S_c, S_o) > 1$
Changing word	$1 < WordOverLap(S_o, S_c) < \infty$, $SizeOverlap(S_o, S_c) = 0$
Changing sentence	$WordOverLap(S_o, S_c) = 1$, $SizeOverlap(S_o, S_c) = 0$

3 System Design and Algorithm

All original documents are stored in document data base. When the query document is input, the system divides the query document and the original documents into comparing units - sentences. The divided sentences are then used to calculate the overlap and the plagiarism information in local_similarity_extractor by using the overlap measure function defined in section 2. The extracted information is used to calculate the degree of copy in original documents from each other, and the ordered information is supplied to user. The algorithm of the proposed system is followed.

```
Algorithm
Input:
    $Document\_DB = \{D_1, D_2, D_3, ....., D_n\}$ and each
                 $D_i = \{S_{i1}, S_{i2}, S_{i3}, ....., S_{im}\}$
    $QueryDocument = \{QS_1, QS_2, QS_3, ....., QS_t\}$

Output:
   Decreasing ordered document list in document
   similarity value

for i = 1 to n
   for j = 1 to t
      localsimilarity[1..j] = 0
      for k = 1 to m
         if $|Comm(S_{ik}, QS_j)| \geq \frac{|S_{ik}|}{2}$ then
            localsimilarity[j] = max {localsimilarity[j],
                                       $Sim(S_{ik}, QS_j)$ }
   documentsimilarity[i] = $\sum_j localsimilarity[j]$
return sort(documentsimilarity)
```

4 Experiment and Discussion

We generated the test document set from CISI as follow.

1. 11 relevant documents related to a specific query are selected from CISI document set.
2. One document is selected as an original document. The others 10 documents are selected as candidate document for plagiarism.
3. A partial part extracted from the original document is transformed (exact copy, changing synonym, changing sentence structure) and it is inserted into the candidate documents for plagiarism to make plagiarized document.
4. The plagiarized documents are returned into the CISI document set. Selected original document is removed from the CISI document set and becomes the query document.

For comparison with the proposed system (P_System), we made DCD system based on word similarity of document (WD_System) and of sentence (WS_System). For performance checking, we chose R-precision as the evaluation norm and R is set to 10.

Table 2. Copy detection test (R = 10)

	WD_System	WS_System	P_System
Exact copy	2	6	8
Synonym	2	6	8
Changing structure	1	4	4

The experimental results show that the proposed P_System produces more precise results in exact copy and changing synonym. It shows that in the proposed method overlap measure function is more useful to check the copy of document than the normalized comparison value like cosine similarity. And if user decides the copy of document with the consideration of plagiarism pattern information produced in comparison, the more precise decision could be made.

References

1. Shivakumar, N. and Garcia-Monlina, H. SCAM: A Copy Detection Mechanisms for Digital Documents. In Proceedings of International Conference on Theory and Practice of Digital Libraries, Austin, Texas. June 1995.
2. Brin, S., Davis, J., and Garcia-Molina, H. Copy Detection Mechanisms for Digital Documents. In Proceedings of ACM SIGMOD Annual Conference, San Jose, CA, 1995
3. Si, A., Leong, H., and Lau, R. CHECK: A Document Plagiarism Detection System. In Proceedings of ACM Symposium for Applied Computing, pp. 70-77, Feb 1997.
4. Bao Jun-Peng, Shen Jun-Yi, Liu Xiao-Dong, Liu Hai-Yan, Zhang Xiao-Di. Document Copy Detection Based On Kernel Method. In Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering.
5. Karen Fullam, J. Park. Improvements for Scalable and Accurate Plagiarism Detection in Digital Documents. 2002