Document Copy Detection System based on Plagiarism Patterns*

NamOh Kang, 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.

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

* 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.