International Journal of Innovative
Computing, Information and Control
Volume 4, Number 3, March 2008
 ISSN 1349-4198
pp. 681–688

A DETECTION METHOD FOR THE ILLEGAL COPYING OF DIGITAL DOCUMENTS

Xu Li[1], Guohua Liu[1], Huidong Ma[1] and Lei Wang[2]

[1]College of Information Science and Engineering
[2]College of Mechanical Engineering
Yanshan University
Qinhuangdao 066004, P. R. China
lixu102@sohu.com; ghliu@ysu.edu.cn; huidongma@126.com; sf1300@sina.com

Received January 2007; revised June 2007

Abstract. *Easy accessibility to digital documents via the internet makes it easy for many users to share information. However, it also leads to a perplexing problem concerning intellectual property security. To address the problem, a detection method for the illegal copying of digital documents is proposed in this paper, which can automate to detect the partial or whole overlaps between electronic documents. It is a powerful tool to protect the author's intellectual property and to improve the efficiency of information retrieval. We describe the representing method of a document and define the corresponding overlap measure. An experiment verifies the efficiency of the proposed method and compares it with the word-frequency-based method and the sentence-based method in different data sets. The experimental results show that the proposed method is superior, and it can accurately identify any matching text of a certain length.*
Keywords: Digital document, Copy detection, Feature extraction, Overlap

1. **Introduction.** Digital documents are more easily copied and distributed illegally in the internet environment, therefore the author's intellectual property is not effectively protected. If the problem is not resolved, it will make authors reluctant to share their documents and will reduce the chances for users to access valuable information. Duplication of digital document content also arises for other reasons besides plagiarism, such as documents may be mirrored at different sites or may be unknowingly submitted multiple times. The identical web documents are common. Not only do the copied web documents create redundant information, which take longer to filter unique information and cause additional storage space, but they also degrade the efficiency of information retrieval.

To address these problems, many document copy prevention technologies have been introduced, such as digital watermarking technology [4,15] . It is a solution to the problem of intellectual property protection. However, the scheme can easily be defeated by destroying the watermarks or removing the formatting of the documents, and it can not prevent or detect the partial copy of a document.

A huge amount of digital documents is made public day to day on the Internet. Most of them are not supported by copy prevention technology. This increases the necessity of copy detection technology. Document copy detection focuses on detecting automatically the partial or whole overlaps between a query document and those saved in the database, and returning an overlap by using the corresponding overlap measure. It is a powerful tool to protect the author's intellectual property and to improve the efficiency of information retrieval.

The most popular approach of copy detection for digital documents uses a word frequency vector to represent a document [3,6,9, 14,15,16]. Cosine function and dot product