

XÂY DỰNG KHO DỮ LIỆU PHỤC VỤ HỆ THỐNG PHÁT HIỆN SAO CHÉP

BUILDING A DATA BANK FOR DUPLICATION DETECTION SYSTEM

Tác giả: Châu Thùy Dương, **Võ Trung Hùng***, **Hồ Phan Hiếu**

Tóm tắt bằng tiếng Việt:

Trong bài báo này, chúng tôi trình bày kết quả xây dựng kho dữ liệu sẽ được sử dụng trong hệ thống phát hiện sao chép từ các nguồn tài liệu của Đại học Đà Nẵng. Kho dữ liệu này bao gồm các tài liệu gốc, cơ sở dữ liệu thông tin chung về tài liệu và dữ liệu đã được trích xuất từ các tài liệu lưu trữ dưới định dạng XML. Chúng tôi đã đề xuất cấu trúc lưu trữ và các chương trình tương ứng để dễ dàng lưu trữ, cập nhật kho dữ liệu và xử lý các dữ liệu trên kho. Chúng tôi đã tiến hành thử nghiệm và lưu trữ trên kho này với hơn 100 tài liệu mỗi loại cho báo cáo tốt nghiệp của sinh viên ngành công nghệ thông tin, luận văn tốt nghiệp cao học và báo cáo tổng kết đề tài nghiên cứu. Kết quả nghiên cứu này là tiền đề để xây dựng một hệ thống phát hiện tự động việc sao chép trái phép trên các tài liệu khoa học, góp phần hạn chế nạn “đạo văn” đang diễn ra phổ biến hiện nay.

Từ khóa: phát hiện sao chép; kho dữ liệu; đạo văn; chuyển định dạng; học liệu

Tóm tắt bằng tiếng Anh:

This paper presents the results of building a data bank to be used in the duplication detection system from learning resources of the University of Danang. This data bank includes original documents, abstract information database about the original documents and the data which has been extracted from the documents to store in XML format. We have proposed storage structure and the corresponding programs to easily store, update and manage data in data bank. We have experimented and stored in this data bank over 300 documents such as course papers by IT students, master theses and reports of research projects. Results of this study imply a prerequisite for building an automated system to detect the duplication in the scientific documents, contributing to controlling "plagiarism".

Key words: duplication detection; data bank; plagiarism; format converting; learning resource