Đo độ tương tự ngữ nghĩa tiềm ẩn để phát hiện việc sao chép tài liệu

Trần Cao Đệ
Khoa Công nghệ Thông tin & Truyền thông, Đại học Cần Thơ
tcde@cit.ctu.edu.vn

Việc xác định một văn bản có sao chép từ các văn bản khác hay không là một vấn đề khá nhạy cảm. Nó giúp kiểm tra, phát hiện từ việc sao chép lại đồ án, luận văn đến vi phạm tác quyền, vi phạm sở hữu trí tuệ.

Đã có một số nghiên cứu đề xuất các phương pháp khác nhau để xác định xem một đoạn văn bản của một tài liệu có nằm trong một tài liệu nào khác hay không. Các phương pháp này chủ yếu dựa trên tìm kiếm và so khớp chuỗi (string matching). Tuy nhiên, các phương pháp so khớp chuỗi chỉ có hiệu quả nếu việc sao chép là "nguyên văn". Nó không thể phát hiện các sao chép có sửa đổi đôi chút như thay thế một số từ bằng từ đồng nghĩa hay thay đổi một ít trong thứ tự các câu trong văn bản.

Bài viết này đề xuất một giải pháp xác định xem một tài liệu có được sao chép từ các tài liệu khác hay không bằng việc đo độ tương tự ngữ nghĩa của tài liệu đang xét với các tài liệu khác. Độ tương tự ngữ nghĩa ở đây là ngữ nghĩa tiềm ẩn (latent semantic). Kỹ thuật ngữ nghĩa tiềm ẩn (latent semantic indexing) sẽ được dùng để xác định mức độ tương tự ngữ nghĩa của hai đoạn văn bản.

Các thực nghiệm cho thấy rằng, giải pháp đưa ra có thể giải quyết được hạn chế của phương pháp so khớp chuỗi. Một văn bản được sao chép từ một văn bản khác rồi sửa đổi đôi chút như thay từ đồng nghĩa, xáo trộn một ít trong trật tự các câu đều có thể bị phát hiện.

Từ khóa: so sánh chuỗi, độ tương tự ngữ nghĩa, chỉ mục ngữ nghĩa tiềm ẩn, tìm kiếm theo ngữ nghĩa, phát hiện sao chép.

I. Giới thiệu

Trong thời đại công nghệ số như hiện nay, các nguồn tài liệu là vô cùng phong phú. Việc "sao chép tài liệu" theo nghĩa tiêu cực như đạo văn, sao chép các luận án, luận văn, đồ án trở nên phổ biến và đang là vấn nạn. Ở qui mô rộng hơn, các thư viện điện tử ngày càng nhiều, một tài liệu có thể được phát hành trên internet nhiều lần trong những thư viện điện tử khác nhau, trên các trang web khác nhau.

Làm thế nào để phát hiện sự sao chép tài liệu theo nghĩa tiêu cực? Làm thế nào ngăn chặn việc sao chép trái phép, đạo văn, đạo nhạc, đạo luận văn, đồ án?

Bài viết này sẽ tập trung vào chủ đề phát hiện sao chép tài liệu văn bản. Chủ đề này đã được nghiên cứu từ khoảng 10 năm qua. Hiện tại, đã có một số giải pháp cho việc phát hiện sao chép và một vài công cụ phần mềm cho phép phát hiện một tài liệu (gọi là *văn bản kiểm tra*) có sao chép từ một tập hợp các *tài liệu nguồn* hay không. Tập hợp các tài liệu nguồn có thể là đóng- tức là các tài liệu tập hợp trước trong một thư viện điện tử- hoặc là mở, chẳng hạn như tập các tài liệu văn bản trên internet.