Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

PHÁT TRIỂN HỆ THỐNG PHÁT HIỆN ĐẠO VĂN CHO TRƯỜNG ĐẠI HỌC VIỆT NAM

Trần Cao Đệ[1], Lê Văn Lâm[1], Bùi Võ Quốc Bảo[1], Nguyễn Gia Hưng[1] và Trần Cao Trị[1]
[1]*Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 22/10/2014
Ngày chấp nhận: 29/12/2014

Title:
Developing plagiarism detection system for Vietnamese university

Từ khóa:
Đạo văn, phát hiện sao chép, hệ thống phân tán, tính toán hiệu năng cao, chỉ mục nghịch đảo

Keywords:
Plagiarism, plagiarism detection system, plagiarism detector

ABSTRACT

Plagiarism is known as a serious concern in academic environment. Beside strict policy applied to plagiarist, there could be some kind of tools to help both educators and students prevent it. There are commercial products produced to detect plagiarism. However, these products are too expensive to educators in Vietnam and they have not given any proof whether these products work well in Vietnamese. Moreover, there are some detection methods available that could be a good choice to work in Vietnamese academic environment. However, these products have their own detection methods and these methods could not be changed. In addition, scalability is also an important feature for a plagiarism detection system because the number of documents in database is very large and increases rapidly. In this paper, we present a plagiarism detection system to detect plagiarism that have three above features: working on one's own database, flexibility, and scalability.

TÓM TẮT

Đạo văn được biết đến như một vấn nạn trong môi trường học thuật. Bên cạnh các chế tài nghiêm ngặt cho người đạo văn, cần có những công cụ hiệu quả để ngăn chặn, không để xảy ra tình trạng đạo văn trong trường đại học và trong sinh viên. Đã có nhiều ứng dụng được xây dựng để phát hiện đạo văn. Tuy nhiên, các sản phẩm này thường là khá đắt đỏ và chưa được kiểm chứng có thực sự phù hợp với nguồn tài liệu tiếng Việt hay không. Chúng thường dựa trên các thuật toán phát hiện đạo văn của riêng mình và thường thì không thể bổ sung hay tùy biến nhằm phù hợp với môi trường và ngôn ngữ tiếng Việt. Ngoài ra, khả năng mở rộng cũng là một tính năng quan trọng đối với một hệ thống phát hiện đạo văn vì số lượng tài liệu trong cơ sở dữ liệu là rất lớn và tăng lên nhanh chóng. Trong bài báo này, chúng tôi trình bày một hệ thống phát hiện sao chép để phát hiện đạo văn với các tính năng quan trọng: làm việc trên một cơ sở dữ liệu riêng, lớn của một tổ chức như trường đại học; linh hoạt, dễ mở rộng; đáp ứng hiệu năng tính toán mong đợi. Chúng tôi đề xuất giải pháp sử dụng hệ thống phân tán, sử dụng công nghệ NoSQL, lập chỉ mục nghịch đảo với công nghệ Hyperdex. Việc tính toán xử lí trong hệ thống là tính toán song song được trên nền tảng công nghệ JPPF.

1 GIỚI THIỆU

Đạo văn là một trong những vấn nạn trong môi trường học thuật. Với sự phát triển nhanh chóng của Internet và các thiết bị Công nghệ thông tin (CNTT), việc đạo văn gần đây đã được thực hiện rất dễ dàng. Người vi phạm có nhiều phương tiện để tìm kiếm và ăn cắp nội dung hay ý tưởng của người khác bởi vì những nghiên cứu và ý tưởng gần như có sẵn rất nhiều trên mạng Internet. Hơn nữa, họ cũng tận dụng kỹ thuật của CNTT để dấu việc đạo văn của họ. Ở Việt Nam, đạo văn là một trong những mối quan tâm đặc biệt trong hầu hết các trường đại học. Mỗi trường đại học có chính sách riêng về đạo văn của mình để ngăn chặn sinh viên đạo luận văn, tài liệu học thuật. Tuy nhiên, đạo văn vẫn còn tồn tại và có chiều hướng gia tăng trong học đường ở Việt Nam.

Có một loạt các phương pháp tiếp cận, giải pháp và sản phẩm có sẵn để phát hiện đạo văn trong các ngôn ngữ thông dụng trên thế giới đặc biệt là tiếng Anh. Những giải pháp như các ứng dụng riêng lẻ hoặc các dịch vụ dường như không thể được sử dụng trong giáo dục Việt Nam vì một số lý do nhất định. Thứ nhất, giá sản phẩm quá đắt đối với các cơ sở giáo dục Việt Nam. Thứ hai, không có minh chứng rõ ràng cho thấy các sản phẩm hay dịch vụ đó có thể làm việc tốt trên tiếng Việt và môi trường học thuật Việt Nam. Thứ ba, hầu hết các luận văn tốt nghiệp và bài báo khoa học từ các trường đại học Việt Nam đang được lưu trữ cục bộ trong cơ sở dữ liệu thư viện các trường đại học. Vì vậy, ứng dụng phát hiện đạo văn phải cung cấp tính năng làm việc được trên tập cơ sở dữ liệu "riêng tư" để phát hiện đạo văn.

Đại học Cần Thơ (ĐHCT) có hệ thống cơ sở dữ liệu lưu trữ luận án nghiên cứu sinh và các bài báo khoa học. Chúng được lưu trữ cục bộ tại cơ sở dữ liệu của Trường và có thể được truy cập bởi các sinh viên và giảng viên. Đạo văn là một trong những vấn đề được quan tâm đặc biệt tại Đại học Cần Thơ. Căn cứ vào các nghiên cứu hiện tại và phương thức hoạt động của những hệ thống phát hiện đạo văn hiện hữu cũng như nhu cầu cấp thiết của Đại học Cần Thơ trong phát hiện đạo văn, chúng tôi đề xuất một hệ thống phát hiện đạo văn cho Đại học Cần Thơ. Hệ thống phát hiện đạo văn của chúng tôi có thể được áp dụng cho các trường đại học khác. Nó cũng có thể được coi là hệ thống phát hiện đạo văn đầu tiên cho các trường đại học tại Việt Nam.

2 ĐẠO VĂN

Phần này cung cấp một cách nhìn tổng quan về đạo văn bao gồm: định nghĩa về đạo văn và đạo văn trong môi trường học thuật.

2.1 Đạo văn trong môi trường học đường

Theo Meuschke và Gipp (Meuschke and Gipp, 2013), đạo văn là việc sử dụng các ý tưởng của người khác, mà không đưa ra lời xác nhận và tài liệu tham khảo phù hợp. Người phạm tội trình bày ý tưởng hay lời nói của người khác như là của riêng của họ. Meuschke và Gipp nói rằng một số nhà nghiên cứu mô tả đạo văn học văn học như trộm cắp, ăn cắp ý tưởng hay lời nói từ những người khác (Ercegovac and Richardson, 2004; Park, 2003).

Tình trạng đạo văn học trên thế giới đã được thảo luận trong (Gipp, 2014). Nó cho thấy rằng đạo văn xảy ra trên toàn thế giới và trở thành một vấn đề chưa được giải quyết. Một nghiên cứu được tiến hành trên 80.000 sinh viên trong ba năm ở Mỹ và Canada 2002-2005 (McCabe, 2005) cho thấy 38% sinh viên đại học và 25% sinh viên sau đại học đã sao chép hoặc diễn giải các câu văn mà không đưa ra nguồn gốc. Các nghiên cứu khác bên ngoài Mỹ và Canada cũng cho thấy tỷ lệ đạo văn rất cao trong môi trường học tập. Một số hệ thống phát hiện đạo văn đã được thực hiện và họ phát hiện 20% hoặc nhiều tài liệu có nội dung đáng ngờ (Barrett and Malcolm, 2006; Culwin, 2006). Dựa trên những số liệu này, Gipp và Bela kết luận rằng đạo văn trong môi trường học thuật là một vấn đề nghiêm trọng.

Ở Việt Nam, đạo văn học đã thực sự được quan tâm trong xã hội. Có rất nhiều cuộc thảo luận, hội thảo, hội nghị tập trung vào đạo văn trong học đường. Tuy nhiên, có rất ít nghiên cứu về đạo văn trong học thuật được xuất bản gần đây. Hầu như tất cả các trường hợp đạo văn được đưa tin trên các tờ báo như Thanh Niên, Tuổi Trẻ,... Những tờ báo này mô tả đạo văn xảy ra khá phổ biến trong cả hai chương trình đại học và sau đại học. Họ đề nghị các trường đại học Việt Nam phải chống đạo văn nghiêm ngặt, nghiêm túc hơn. Hơn nữa, ứng dụng CNTT để phát hiện đạo văn cũng được đề cập đến như một trong những cách thức hiệu quả để giảm đạo văn. Các trường đại học có thể xây dựng một số hệ thống phát hiện đạo văn để giúp cả sinh viên và giảng viên kiểm tra đạo văn.

2.2 Các hình thức đạo văn

Meuschke và Gipp (Meuschke and Gipp, 2013) phân loại các hình thức đạo văn học như sau:

Đạo văn hoàn toàn: được mô tả như là một loại sao chép gần như không thay đổi so với tài liệu nguồn. Nó bao gồm các hình thức "sao chép và dán" (Maurer, Kappe *et al.,* 2006) và "trộn và dán" (Weber-Wulff, 2010). "sao chép và dán" là hình thức sao phép hoàn toàn nội dung mà không có một sự thay đổi nào. "trộn và dán" là hình thức sao chép có một vài thay đổi rất nhỏ so với tài liệu nguồn.

Giả tạo đạo văn: được mô tả như là một loại diễn giải, ngụy trang kỹ thuật, hay dịch từ ngôn ngữ này sang ngôn ngữ khác.

Đạo văn cấu trúc và ý tưởng: đề cập một loại sử dụng cấu trúc của người khác, khái niệm rộng hơn mà không đưa ra trích dẫn nguồn phù hợp.

Tự đạo văn: đề cập đến một loại tái sử dụng câu hay đoạn văn của của riêng mình mà không ghi nguồn phù hợp.

Theo những quan sát của chúng tôi, đạo văn theo dạng "sao chép và dán" xảy ra khá phổ biến. Đây là loại đạo văn xảy ra trong cả hai chương trình đại học và sau đại học. Nghiêm trọng hơn, có một số trường hợp trong đó sinh viên sao chép một số chương, hay thậm chí toàn bộ nội dung luận văn của người khác. Các loại khác của đạo văn hiếm khi được phát hiện và ghi nhận. Điều đó không có nghĩa là không xảy ra tại Việt Nam. Lý do những loại đạo văn này khó phát hiện ra vì các trường đại học Việt Nam không có bất kỳ hệ thống phát hiện đạo văn nào.

2.3 Những cách tiếp cận phát hiện đạo văn

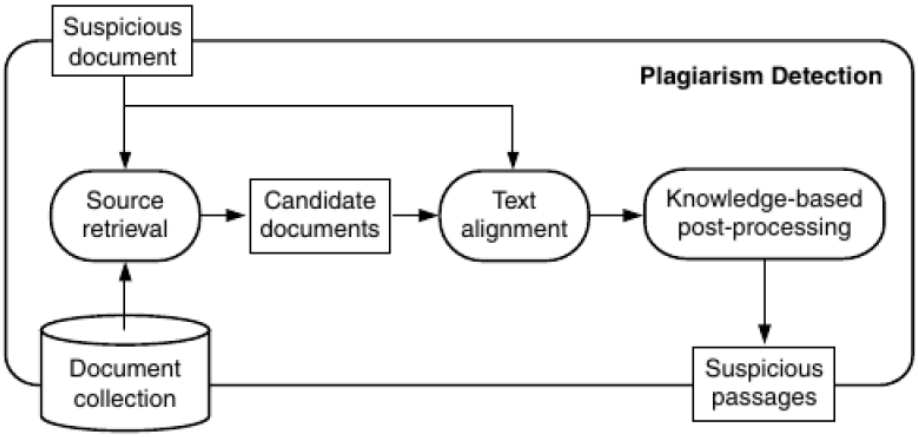
Meuschke và Gipp (Meuschke and Gipp 2013) phân loại các phương pháp phát hiện đạo văn thành hai nhóm: so sánh tương tự cục bộ và so sánh tương tự toàn cục. So sánh tương tự cục bộ quan tâm đến tính tương tự giữa các phân đoạn văn bản, trong khi đó so sánh tương tự toàn cục quan tâm đến sự tương đồng giữa văn bản dài hoặc toàn bộ tài liệu. Trong bài báo này, chúng tôi sử dụng một trong những phương pháp đánh giá tương tự cục bộ để phân tích đạo văn. Phương pháp chúng tôi lựa chọn là Kasprzak (Kasprzak and Brandejs, 2010), xuất hiện trên tốp 10 phương pháp phát hiện đạo văn trong các cuộc thi quốc tế về phát hiện đạo văn. Theo phương pháp này, để phát hiện đạo văn trước hết phân chia một tài liệu cần kiểm tra thành một danh sách các từ n-gram. Sau đó, so sánh từng từ n-gram trong tài liệu cần kiểm tra với các từ n-gram của tất cả các tài liệu trong tập dữ liệu luận văn đang lưu trữ. Chi tiết của phương pháp này sẽ được trình bày trong phần tiếp theo.

3 HỆ THỐNG PHÁT HIỆN ĐẠO VĂN CHO TRƯỜNG ĐẠI HỌC CẦN THƠ

Trong phần này chúng tôi trình bày hệ thống phát hiện đạo văn tổng quát và sau đó đề xuất một hệ thống phát hiện đạo văn cho Đại học Cần Thơ.

3.1 Hệ thống phát hiện đạo văn tổng quát

Hình 1 trình bày quá trình xử lý chung để phát hiện đạo văn (Potthast, Hagen *et al.,* 2013, Stein, zu Eissen *et al.,* 2007). Với một tài liệu cần kiểm tra nào đó, quá trình tìm kiếm để phát hiện đạo văn sẽ phải tìm kiếm trên một tập dữ liệu rất lớn. Quá trình này bao gồm ba bước chính. Ở bước thứ nhất, do số lượng tài liệu trong bộ sưu tập là rất lớn vì vậy bước này sẽ chọn một nhóm nhỏ các tài liệu ứng cử viên từ tập tài liệu lớn. Các tài liệu ứng cử viên là các tài liệu được xác định có khả năng cao là nguồn của đạo văn liên quan đến tài liệu cần kiểm tra. Bước thứ hai thực hiện việc liên kết văn bản, so sánh các tài liệu ứng cử viên và các tài liệu cần kiểm tra đạo văn, và trích xuất các đoạn tương tự từ cả hai. Bước thứ ba, dựa trên tri thức cho trước, hệ thống trình bày các tài liệu cần kiểm tra đạo văn theo một thể thức nhất định nhằm giúp cho người sử dụng có thể xử lý các tác vụ về sau.



Hình 1: Mô hình xử lý dữ liệu tổng quát phát hiện đạo văn (Potthast, Hagen *et al.*, 2013, Stein, zu Eissen *et al.*, 2007)

Ngoài ra, một hệ thống phát hiện đạo văn thường cần tạo chỉ mục của tất cả tài liệu trong tập tài liệu nguồn. Điều này giúp cải thiện hiệu suất hoạt động của hệ thống phát hiện đạo văn trên yếu tố thời gian tính toán. Hơn nữa, tất cả các tài liệu (tài liệu nguồn và các tài liệu cần kiểm tra) phải được tiền xử lý và lưu trữ dưới một hình thức được xác định.

3.2 Hệ thống phát hiện đạo văn cho Đại học Cần Thơ

Dựa trên hệ thống phát hiện sao chép, đạo văn tổng quát được trình bày trong phần trước, chúng tôi đề xuất một hệ thống phát hiện đạo văn cho Trường Đại học Cần Thơ (ĐHCT) với những điểm chính yếu sau:

– Sử dụng phương pháp phát hiện đạo văn từ Kasprzak (Kasprzak and Brandejs 2010) với một số thay đổi để nó làm việc tốt hơn trong môi trường tiếng Việt. Những thay đổi bao gồm chiều dài từ (2 ký tự thay vì 3 ký tự), chiều dài của n-gram (4-gram thay vì 5-gram). Các hiệu chỉnh này dựa trên kết quả thực nghiệm mà chúng tôi thực hiện trên cả 2 tập dữ liệu PAN và dữ liệu luận văn tiếng Việt tại ĐHCT.

– Sử dụng JPPF (Java Parallel Processing Framework) để tính toán song song nhằm đạt hiệu năng về thời gian tính toán mong đợi. JPPF cung cấp các giải pháp để phân chia công việc thành những phần nhỏ hơn có thể được thực hiện đồng thời trên các máy khác nhau. JPPF cũng làm cho hệ thống phát hiện đạo văn được đề xuất có khả năng mở rộng dễ dàng hơn.

– Sử dụng giao diện web để tương tác với người sử dụng và các dịch vụ web để giao tiếp giữa máy chủ web và các ứng dụng web. Điều này làm cho hệ thống phát hiện đạo văn của chúng tôi linh hoạt hơn, dễ dàng thay đổi sau này.

3.2.1 Tiền xử lý các tài liệu

Cả hai tài liệu cần kiểm tra và tài liệu nguồn đều được tiền xử lý như sau:

– Xác định từ vựng: Một tập tin văn bản được chia thành các từ có độ dài ít nhất 2 ký tự. Thông tin về vị trí bắt đầu và kết thúc của các từ được lưu trữ để sử dụng sau này.

– Xác định các đoạn từ kết hợp: Từ danh sách các từ của mỗi tài liệu, chúng tôi hình thành các đoạn từ 4-gram, sắp xếp các đoạn từ và tính toán giá trị băm MD5 cho các đoạn từ. Giá trị băm MD5 được sử dụng như định danh của đoạn từ. Các vị trí của các ký tự đầu tiên và cuối cùng trong đoạn cũng được lưu trữ.

3.2.2 Lập chỉ mục tài liệu nguồn

Để tăng tốc độ hoạt động của hệ thống, các tài liệu nguồn được phân tích và lập chỉ mục theo dạng chỉ mục nghịch đảo. Cụ thể, chúng tôi ánh xạ định danh đoạn 4-gram vào danh sách các cấu trúc (định danh tài liệu, vị trí của ký tự đầu tiên của đoạn từ, vị trí của ký tự cuối cùng của đoạn từ).

3.2.3 Tìm kiếm các tài liệu tiềm năng

Số lượng tài liệu nguồn thường là rất lớn vì vậy cần phải hạn chế số lượng tài liệu tiềm năng tìm kiếm đạo văn. Chỉ có tài liệu có ít nhất 20 đoạn chung 4-gram với các tài liệu cần kiểm tra sẽ được coi là tài liệu tiềm năng. Trên thực tế, không cần

phải quan tâm tới tất cả tài liệu tiềm năng vì vậy chúng tôi chỉ chọn 100 tài liệu đầu trong danh sách các tài liệu có số 4-gram chung được xếp thứ tự giảm dần.

3.2.4 So sánh tài liệu cần kiểm tra với tài liệu tiềm năng

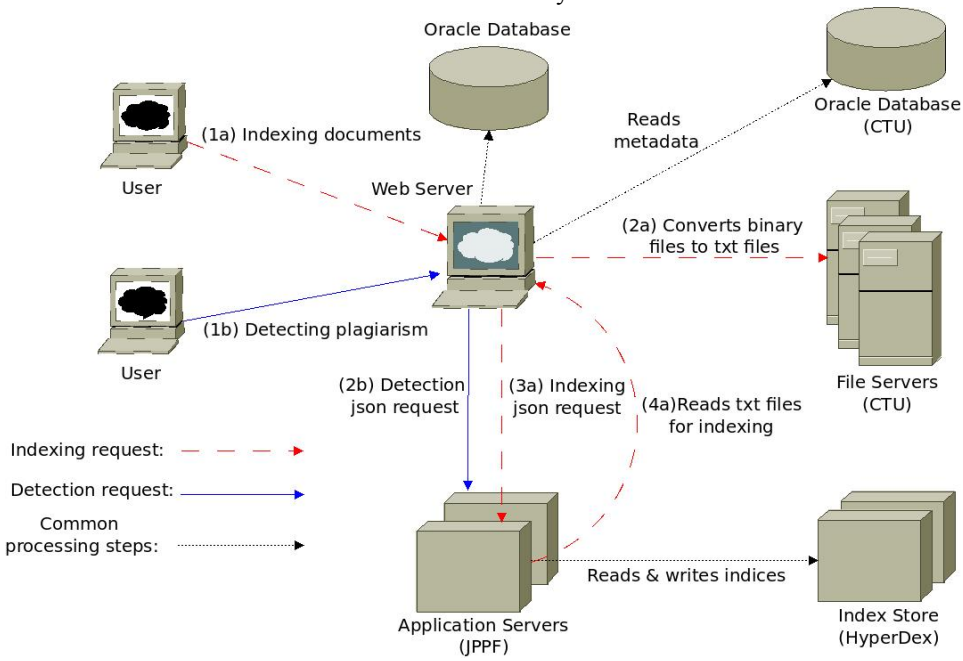
Tài liệu cần kiểm tra được so sánh với mỗi tài liệu tiềm năng. Đối với mỗi cặp (một tài liệu cần kiểm tra và một tài liệu tiềm năng), trước hết kiểm tra xem có một số đoạn chung của cả hai tạo thành một hoặc nhiều đoạn tài liệu hợp lệ. Một đoạn tài liệu hợp lệ được định nghĩa là đoạn tài liệu có ít nhất 20 đoạn từ chung và khoảng cách giữa hai đoạn từ chung lân cận không dài quá 150 ký tự.

3.2.5 Lọc kết quả

Các đoạn tài liệu hợp lệ được coi là đoạn đạo văn. Tuy nhiên, có thể có một số phát hiện chồng chéo vì vậy chúng ta cần phải loại bỏ chúng bằng cách chỉ giữ lại một đoạn dài nhất trong các cặp chồng chéo. Hơn nữa, chúng tôi sử dụng tỷ lệ giữa chiều dài của đoạn nghi ngờ và chiều dài của đoạn nguồn là ngưỡng để lựa chọn đoạn hợp lệ. Dựa trên thực nghiệm, chúng tôi chọn giá trị ngưỡng là 0.25.

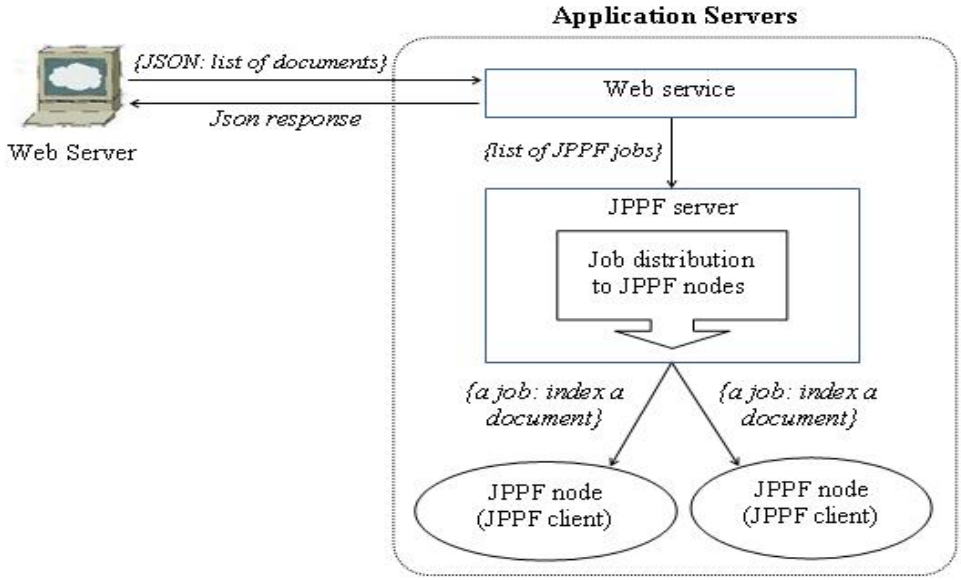
3.3 Sử dụng JPPF để tăng hiệu suất hệ thống

Dựa trên Hình 1, chúng tôi xác định hai công việc chính đòi hỏi rất nhiều thời gian tính toán: lập chỉ mục tài liệu và kiểm tra đạo văn. Chúng tôi sử dụng JPPF để tăng hiệu suất hệ thống phát hiện đạo văn như Hình 2. Hệ thống đáp ứng hai yêu cầu chính của người sử dụng. Yêu cầu đầu tiên là lập chỉ mục tài liệu - đòi hỏi hệ thống phát hiện đạo văn đọc siêu dữ liệu của các tài liệu nguồn từ cơ sở dữ liệu ĐH Cần Thơ, chuyển đổi tài liệu ở định dạng nhị phân từ các máy chủ của Đại học Cần Thơ sang định dạng văn bản, và sau đó lưu trữ chúng trong hệ thống tập tin cục bộ. Yêu cầu thứ hai là kiểm tra đạo văn. Hệ thống phát hiện đạo văn đọc các tài liệu cần kiểm tra cho trước, chuyển đổi chúng sang dạng văn bản, sau đó lưu chúng trong một thư mục tạm thời để sử dụng về sau. Cả hai yêu cầu trên đều được chuyển sang các yêu cầu JSON đến các máy chủ ứng dụng (JPPF) để xử lý các yêu cầu.



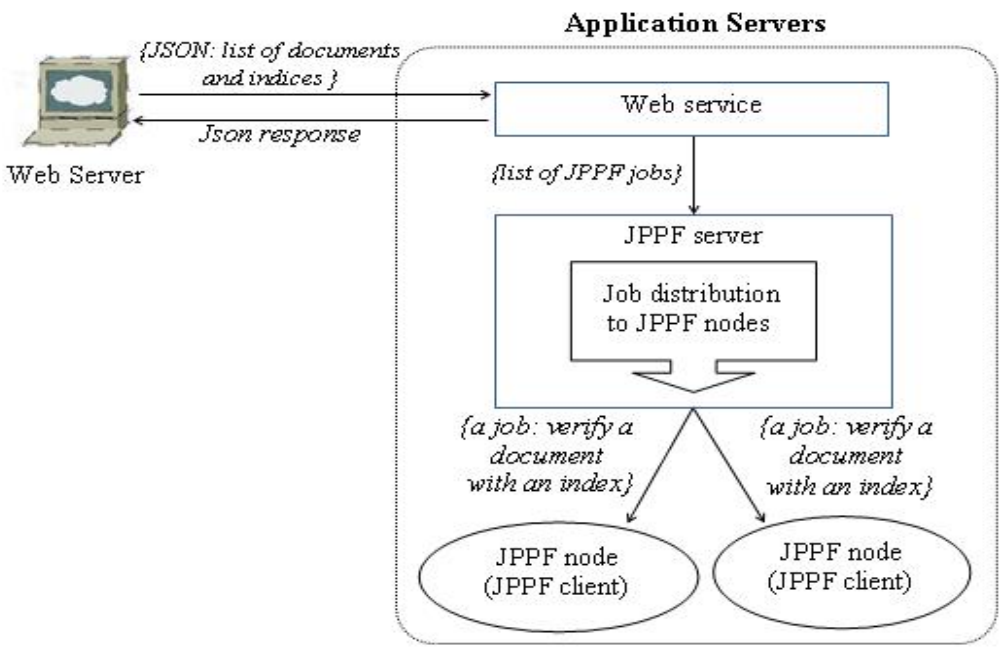
Hình 2: Kiến trúc của hệ thống phát hiện đạo văn

Yêu cầu lập chỉ mục được xử lý bởi các máy chủ ứng dụng như Hình 3. Người sử dụng tương tác với các máy chủ web để yêu cầu lập chỉ mục một danh sách các tài liệu. Các máy chủ web tạo thành một yêu cầu JSON gửi đến các máy chủ ứng dụng. Các máy chủ ứng dụng đọc các tài liệu trong danh sách, phân tích từ vựng, tạo ra các 4-gram, và lưu trữ các 4-gram tại nơi lưu trữ chỉ mục (Hyperdex). Mỗi công việc lập chỉ mục tài liệu được xử lý như một công việc độc lập và được giao cho một trong các nút JPPF thực hiện. Kết quả lập chỉ mục tài liệu được trả về cho máy chủ web theo định dạng JSON.



Hình 3: Các máy chủ ứng dụng xử lý một yêu cầu lập chỉ mục

Tương tự như vậy, yêu cầu kiểm tra đạo văn được xử lý bởi các máy chủ ứng dụng như Hình 4. Người sử dụng tương tác với các máy chủ web để yêu cầu kiểm tra đạo văn một tài liệu cần kiểm tra. Các máy chủ ứng dụng sử dụng các thuật toán phát hiện đạo văn để xác định khả năng tài liệu cần kiểm tra được đạo văn từ một trong những tài liệu trong tập chỉ mục được lưu trữ. Mỗi công việc kiểm tra đạo văn của một tài liệu là một công việc độc lập và được giao cho một trong các nút JPPF. Kết quả của công việc kiểm tra đạo văn của một tài liệu được trả về cho các máy chủ web theo định dạng JSON.



Hình 4: Các máy chủ ứng dụng xử lý yêu cầu kiểm tra đạo văn

4 ĐÁNH GIÁ

Để đánh giá hệ thống phát hiện đạo văn như đề xuất, chúng tôi triển khai một hệ thống phát hiện đạo văn như Hình 2. Chúng tôi sử dụng 4 máy tính với bộ xử lý Intel ® Core ™ i3 3.4GHz, bộ nhớ 4GB và hệ điều hành Ubuntu 12.04 để triển khai các máy chủ ứng dụng và lưu trữ chỉ mục. Chức năng của các máy tính được mô tả như sau:

– Máy tính 1 làm việc như máy chủ ứng dụng web và chạy Hyperdex coordinator.

– Máy tính 2 hoạt động như một nút JPPF và chạy Hyperdex daemon

– Máy tính 3 hoạt động như một nút JPPF và chạy Hyperdex daemon

– Máy tính 4 hoạt động như một trình điều khiển JPPF

Ngoài ra còn có một máy tính khác chạy các ứng dụng web và cho phép người dùng tương tác với hệ thống.

Sự vận hành của các máy tính dựa trên mô tả được trình bày trong Hình 2 và phần 3.1. Chúng tôi thực hiện hai loại thí nghiệm: một để đo độ chính xác và một để đo thời gian tính toán của hệ thống.

4.1 Độ chính xác

Để kiểm tra hệ thống ở yếu tố độ chính xác, chúng tôi tạo ra một số tài liệu cần kiểm tra từ 145 tài liệu nguồn. Bảng 1 trình bày cách tạo ra các tài liệu cần kiểm tra và số lượng tài liệu cần kiểm tra.

Các kết quả thử nghiệm trên tập dữ liệu được thể hiện trong Bảng 2. Chúng tôi đo 4 yếu tố đánh giá PAN đã được sử dụng để đánh giá một hệ thống phát hiện đạo văn trong PAN (Kasprzak and Brandejs 2010). Những yếu tố này bao gồm plagdet, precision, recall, and granularity. Nhìn chung, hệ thống của chúng tôi đã cho kết quả rất tốt trong tất cả các yếu tố đánh giá PAN.

Bảng 1: Tạo các tài liệu cần kiểm tra

Cách thức tạo tài liệu cần kiểm tra	Tỷ lệ thay đổi so với tài liệu nguồn	Số tài liệu nghi ngờ
Chép và dán	0%	580
Chép và dán với thay đổi ít	10-15%	580
Chép và dán có thay đổi lớn	30-45%	579
Tổng	0-45%	1739

Bảng 2: Kết quả đo chỉ số đánh giá PAN

Tập dữ liệu	Plagdet	Precision	Recall	Granularity
Chép và dán	0.9639	0.9355	0.9940	1.0000
Chép và dán với thay đổi ít	0.9189	0.9138	0.9319	1.0057
Chép và dán có thay đổi lớn	0.7961	0.8958	0.7395	1.0246
Tổng	**0.8951**	0.9151	0.8886	1.0101

4.2 Thời gian tính toán

Để đánh giá thời gian tính toán, chúng tôi sử dụng hệ thống phát hiện đạo văn được triển khai ở phần trên để thực hiện hai tác vụ: lập chỉ mục và kiểm tra đạo văn cho các tài liệu cần kiểm tra trên hai tập dữ liệu: một tập từ cuộc thi quốc tế lần thứ 5 về phát hiện đạo văn (Potthast, Hagen *et al*., 2013) và một tập từ Trường đại học Cần Thơ (cơ sở dữ liệu luận án của sinh viên).

Bảng 3 trình bày thời gian tính toán trong tác vụ lập chỉ mục tài liệu nguồn. Khi số lượng tập tin tăng lên, thời gian thực hiện tác vụ lập chỉ mục cũng tăng nhưng thấp hơn giá trị tuyến tính theo số lượng tập tin. Điều này chứng tỏ tính hiệu quả của việc sử dụng JPPF trong việc xử lý công việc song song tại nhiều nút khác nhau. Tuy nhiên, giá trị thời gian thực thi là khá lớn. Do đó, chúng tôi cấu hình hệ thống phát hiện đạo văn lập chỉ mục chỉ khi có tài liệu nguồn mới phát sinh trong cơ sở dữ liệu Đại học Cần Thơ. Thường thì dữ liệu luận văn của sinh viên chỉ phát sinh hai lần trong năm.

Bảng 3: Thời gian tính toán của tác vụ lập chỉ mục

Tập kiểm tra của PAN		CSDL ĐHCT	
Số lượng tập tin	Thời gian thực thi (giây)	Số lượng tập tin	Thời gian thực thi (giây)
1	0.9	1	9.8
2	1.33	2	12
10	9.65	10	31
50	17.75	20	86.9
200	86		
500	227		
1000	433		

Bảng 4: Thời gian thực hiện kiểm tra đạo văn

CSDL ĐHCT (3000 tài liệu nguồn)	
Số lượng tập tin cần kiểm tra	Thời gian thực thi (giây)
1	2.6
2	2.8
3	4.5
4	4.8
5	8.2
6	9.0

Bảng 4 trình bày các kết quả thí nghiệm của chúng tôi để đo thời gian tính toán khi thực hiện tác vụ phát hiện đạo văn cho các tài liệu cần kiểm tra. Khi số lượng các tài liệu cần kiểm tra tăng lên, thời gian thực thi tăng không nhiều (chỉ tăng gần tuyến tính). Kiến trúc hệ thống phát hiện đạo văn sử dụng JPPF để thực hiện nhiều công việc cùng một lúc cho thấy hiệu quả của nó trong trường hợp này.

5 KẾT LUẬN

Đạo văn là một vấn nạn trong môi trường học thuật Việt Nam. Đến nay, vấn đề đạo văn vẫn chưa được giải quyết triệt để. Bên cạnh những chế tài nghiêm ngặt được áp dụng, các trường đại học Việt Nam cần có công cụ để ngăn chặn tình trạng đạo văn. Các công cụ có thể giúp cả giảng viên và sinh viên phát hiện và ngăn ngừa đạo văn, giúp giảm đạo văn trong môi trường học thuật Việt Nam. Trong bài báo này, chúng tôi trình bày phương pháp tiếp cận của chúng tôi để phát triển một hệ thống phát hiện đạo văn cho các cơ sở đại học Việt Nam, lấy Đại học Cần Thơ là nơi thực nghiệm mô hình. Phương thức chúng tôi sử dụng để phát triển hệ thống phát hiện đạo văn là sử dụng phương pháp của Kasprzak và JPPF. Hệ thống phát hiện đạo văn của chúng tôi có thể làm việc trên cơ sở dữ liệu định sẵn, linh hoạt và có khả năng mở rộng. Chúng tôi sửa đổi một số tính năng từ phương pháp của Kasprzak để làm cho nó làm việc tốt trong môi trường tiếng Việt. Trong khi đó, giải pháp JPPF giúp hệ thống của chúng tôi cải thiện thời gian tính toán. Các kết quả thử nghiệm trên cả hai tập dữ liệu (PAN và CSDL ĐHCT) cho thấy rằng hệ thống phát hiện đạo văn của chúng tôi có kết quả khá tốt trong cả hai thông số: thời gian tính toán và độ chính xác. Trong tương lai, chúng tôi tiếp tục nghiên cứu tích hợp phương pháp ngữ nghĩa tiềm ẩn vào việc lọc các tài liệu tiềm năng để có thể cải tiến hơn nữa hiệu năng của hệ thống. Ngoài ra, sử dụng Google để tìm kiếm tài liệu tiềm năng cũng được xác định như là một hướng phát triển của đề tài nhằm mở rộng phạm vi phát hiện đạo văn.

TÀI LIỆU THAM KHẢO

1. Barrett, R. and J. Malcolm, 2006. Embedding plagiarism education in the assessment process. International Journal for Educational Integrity 2(1).
2. Culwin, F., 2006. An active introduction to academic misconduct and the measured demographics of misconduct. Assessment & Evaluation in Higher Education 31(2): 167-182.
3. Ercegovac, Z. and J. V. Richardson, 2004. Academic Dishonesty, Plagiarism Included, in the Digital Age: A Literature Review. College & Research Libraries 65(4): 301-318.
4. Gipp, B., 2014. Plagiarism Detection. Citation-based Plagiarism Detection, Springer Fachmedien Wiesbaden: 9-42.
5. Kasprzak, J. and M. Brandejs, 2010. Improving the reliability of the plagiarism detection system. Lab Report for PAN at CLEF: 359-366.
6. Maurer, H. A., F. Kappe and B. Zaka, 2006. Plagiarism-A Survey. J. UCS 12(8): 1050-1084.
7. McCabe, D. L., 2005. Cheating among college and university students: A North American perspective. International Journal for Educational Integrity 1(1).
8. Meuschke, N. and B. Gipp, 2013. State-of-the-art in detecting academic plagiarism.

International Journal for Educational Integrity 9(1).

9. Meuschke, N. and B. Gipp, 2013. State of the Art in Detecting Academic Plagiarism. International Journal for Educational Integrity 9(1): 50-71.
10. Park, C., 2003. In Other (People's) Words: Plagiarism by university students--literature and lessons. Assessment & Evaluation in Higher Education 28(5): 471-488.
11. Potthast, M., M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos and S. Benno, 2013. Overview of the 5th International Competition on Plagiarism Detection in. CLEF (Online Working Notes/Labs/Workshop).
12. Stein, B., S. M. zu Eissen and M. Potthast, 2007. Strategies for retrieving plagiarized documents in. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.
13. Weber-Wulff, D., 2010. Test cases for plagiarism detection software in. Proceedings of the 4th International Plagiarism Conference.