

PHẦN MỀM PHÁT HIỆN SAO CHÉP LUẬN VĂN ĐẠI HỌC CỬU LONG

Trần Lam Quyền*, Trần Thị Thúy**

TÓM TẮT

Trong bài báo này chúng tôi đề xuất một giải pháp phát hiện sao chép luận văn dựa trên các kỹ thuật của lĩnh vực tìm kiếm thông tin. Việc kiểm chứng luận văn có được sao chép hay không dựa trên hai nguồn là cơ sở dữ liệu (CSDL) cục bộ và trực tuyến. Việc kiểm tra được thực hiện ở mức câu vì đây là hình thức phổ biến nhất. Đối với việc kiểm tra trực tuyến, chúng tôi thực hiện truy vấn bộ máy tìm kiếm google để tìm các tài liệu có thể bị sao chép và sử dụng độ đo Jaccard để tính điểm sự giống nhau giữa hai câu. Đối với CSDL cục bộ, chúng tôi sử dụng hệ quản trị CSDL MongoDB để lưu chỉ mục nghịch đảo nguồn CSDL cục bộ và sử dụng độ đo Cosine để xác định điểm tương đồng giữa hai câu. Hệ thống có khả năng tìm được các câu có khả năng sao chép từ nguồn khác và đồng thời loại bỏ được câu sao chép với câu trích dẫn. Kết quả thực nghiệm cho thấy giải pháp mà chúng tôi đề xuất mang lại hiệu quả cao trong thực tế.

ABSTRACT

This paper proposes a plagiarism detection software, which based on the information retrieval techniques. The thesis plagiarism is commonly verified by two basic resources: local and online databases. The copy will be checked at sentence level because this is the most common copied form. To detect the plagiarism online, the system will initially run a query on the searching engine Google to find the documents where plagiarism may derive from; and then the Jaccard measure is used to compute the similarity between two sentences. In terms of local resources, the MongoDB database is used to store the inverse index and the Cosine measure is simultaneously used to compute the similarity between two sentences. The study results showed that the proposed software is effectively applied in reality.

1. Cơ sở khoa học và thực tiễn

Tình trạng sao chép hiện nay đang là một vấn nạn, có một bộ phận xem việc sao chép là bình thường. Tại hội nghị “*Liên chính học*

thuật” tổ chức tại trường Đại học Hoa Sen vào ngày 29/5/2015, hành vi về đạo văn, quay cóp, gian lận đã được chia sẻ và đưa ra biện pháp ngăn ngừa. Theo báo Dân trí tại hội nghị này, trường Đại học Duy Tân đã thực hiện khảo sát sinh viên của trường và kết quả cho thấy việc sao chép đang ở tình trạng đáng báo động; với kết quả 16% sinh viên không thực hiện, 9,5% sinh viên ít thực hiện, 49% sinh viên thực hiện,

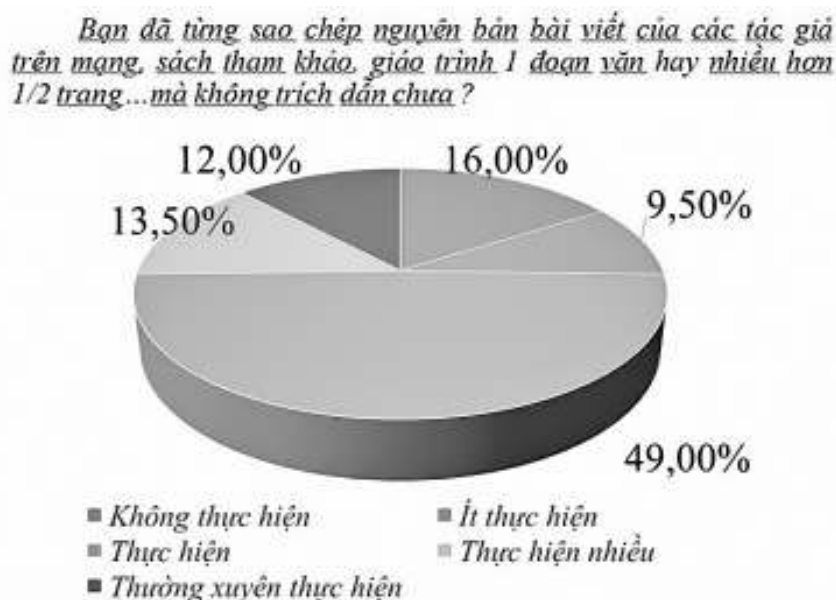
* Thạc sĩ, Phòng Quản lý Khoa học-Sau đại học & Hợp tác Quốc tế

** Thạc sĩ, Khoa Kỹ thuật – Công Nghệ

13,5% sinh viên thực hiện nhiều và 12% sinh viên thường xuyên thực hiện (dantri.com.vn, 30/5/2015). Biểu đồ như Hình 1.

Mặt khác hiện nay, cũng có một số trang web cũng như công cụ cho phép kiểm tra sao chép như: turnitin.com, ithenticate.com, writecheck.com, plagtracker.com...

(neoedu.fpt.edu.vn, 27/12/2014; litado.com, 12/12/2014). Một số công cụ bắt buộc phải chạy trên Windows, một số khác thì đòi hỏi người dùng phải trả phí và cơ sở dữ liệu tiếng Việt dành cho các lĩnh vực kỹ thuật chưa phong phú nên vẫn còn hiện tượng bỏ sót khi kiểm tra đạo văn.



Hình 1. Biểu đồ tình trạng sao chép hiện nay (dantri.com.vn, 30/5/2015)

2. Nghiên cứu có liên quan

Bài toán kiểm tra sao chép tài liệu thu hút nhiều sự quan tâm, nghiên cứu của các nhà khoa học trong và ngoài nước kể từ những năm 90 của thế kỷ trước. Có rất nhiều hệ thống được đề xuất để kiểm tra sao chép đối với tài liệu tiếng Anh và bước đầu đạt được thành công nhất định.

- Đầu tiên phải kể đến là hệ thống COPS [2], hệ thống này được phát triển trong dự án xây dựng thư viện điện tử tại đại học Stanford, hệ thống đánh giá sao chép giữa tài liệu đã đăng ký và tài liệu cần kiểm tra ở mức câu. Hệ thống cho kết quả chính xác nếu như câu được

sao chép hoàn toàn nhưng lại tỏ ra không hiệu quả khi câu chỉ trùng lặp một phần.

- SCAM [6], được phát triển để giải quyết hạn chế của COPS, hệ thống dùng tần suất xuất hiện của từ trong toàn văn bản để phát hiện sao chép, nhưng do thực hiện so khớp ở mức tài liệu nên hai tài liệu khác nhau vẫn được đánh giá là sao chép khi khả năng có nhiều từ chung là rất lớn.

- Cũng dựa theo ý tưởng kiểm tra sao chép chính xác, MDR [4], xây dựng hậu tố để phát hiện sao chép, kết quả thực hiện phát hiện sao chép rất chính xác nhưng việc xây dựng

cây hậu tố rất phức tạp và hạn chế với những sao chép có chỉnh sửa.

- Đi theo một cách tiếp cận khác, CHECK [7] đề xuất sử dụng cấu trúc thông tin và từ khóa được rút trích từ nội dung của tài liệu để phát hiện sao chép, nhưng hệ thống lại chỉ phù hợp với những văn bản có cấu trúc rõ ràng.

- Không dừng lại ở việc phát hiện sao chép đạo văn một cách đơn giản, SSK [1] phân tích và rút trích chuỗi ngữ nghĩa biểu diễn tài liệu và dùng hàm nhân (kernel function) để tính độ tương đồng giữa các tài liệu. Thử nghiệm cho thấy hệ thống phù hợp với mục đích phát hiện sao chép ngữ nghĩa.

Hiện nay, nghiên cứu về phát hiện sao chép văn bản tiếng Việt cũng được thực hiện bởi một số nhóm nghiên cứu nhưng cũng chỉ dừng lại ở mức độ nghiên cứu khoa học mà không hình thành sản phẩm cụ thể cũng như chưa quan tâm đến tính đặc thù (thông tin ngữ nghĩa) của tài liệu cần kiểm tra là luận văn/luận án. Tính hiệu quả của việc quan tâm đến thông tin ngữ nghĩa của tài liệu sẽ được chúng tôi phân tích chi tiết hơn ở phần sau. Trong bối cảnh đó, các trường đại học tại Việt Nam hiện nay phần lớn đều sử dụng các hệ thống (website) kiểm tra sao chép tài liệu nói chung để kiểm tra sao chép luận văn. Có thể liệt kê ra đây các hệ thống nổi bật được các trường đại học tại Việt Nam sử dụng để kiểm tra sao chép bao gồm:

- Hệ thống Turnitin (<http://Turnitin.com>) có hỗ trợ bổ sung nguồn tài liệu khoa học tiếng Việt;

- Plagiarisma (<http://Plagiarisma.net>) hỗ trợ khoảng 190 ngôn ngữ;

- PlagScan (<http://PlagScan.com>) với giao diện người dùng phức tạp.

Phần lớn các hệ thống trên đều có nhược điểm chung là chưa thật sự hiệu quả với các tài liệu tiếng Việt, giao diện người dùng khó sử dụng. Các phiên bản miễn phí thì hạn chế về số lượt sử dụng trong khi phiên bản đầy đủ tính năng thì thu phí sử dụng khá cao.

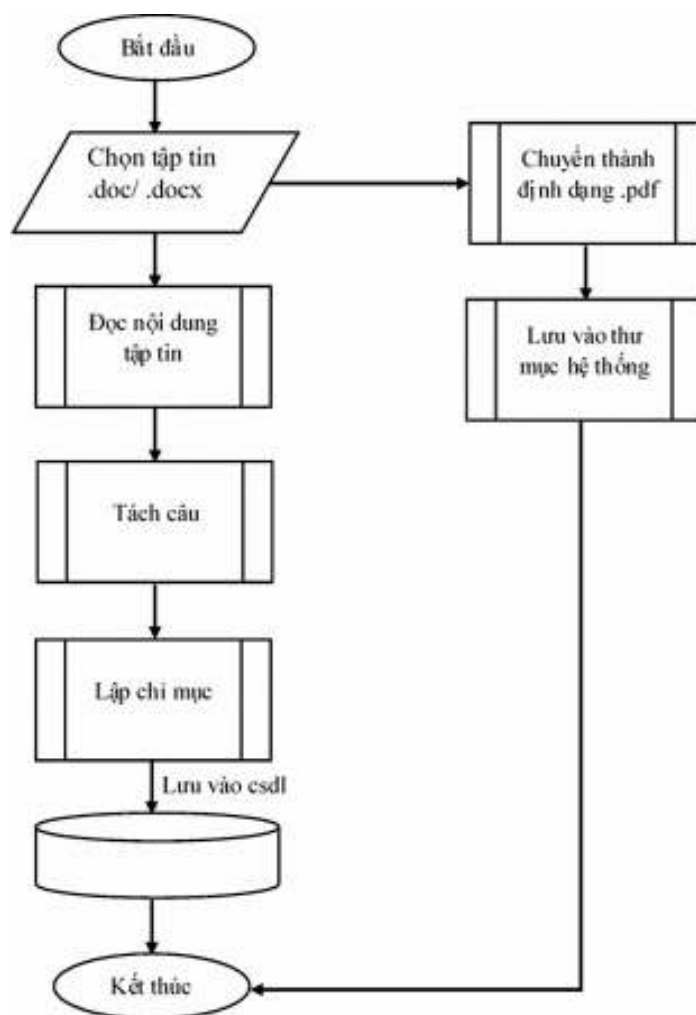
3. Tổng quan hệ thống đề xuất

Hệ thống có các phân hệ như tiền xử lý văn bản, lập chỉ mục, phân hệ kiểm tra sao chép CSDL cục bộ và phân hệ kiểm tra sao chép trên internet.

3.1. Mô-đun tiền xử lý văn bản

Tiền xử lý văn bản là một bước khá quan trọng của việc tìm kiếm nội dung và kiểm tra sao chép. Hệ thống lưu trữ nhận các tập tin đầu vào với định dạng .doc hoặc .docx, đọc nội dung, thực hiện tách câu và tách từ, bắt đầu lập chỉ mục. Bên cạnh đó sẽ chuyển tập tin có dạng .doc hoặc .docx thành tập tin có định dạng .pdf để lưu trữ và hiển thị cho người dùng.

Đối với hệ thống này sẽ chọn đơn vị văn bản kiểm tra là mức câu. Vì, thực tế khi một người thực hiện sao chép sẽ không hoặc ít thay đổi một số từ không quan trọng của câu và mức câu sẽ không gây nhiễu thông tin khi thực hiện tìm kiếm. Với đoạn thì có nhiều từ, khi thực hiện tìm kiếm sẽ cho nhiều kết quả không như mong muốn, mặt khác người sao chép có thể thay đổi vị trí câu. Khi ấy sẽ không cho kết quả tốt khi kiểm tra sao chép vì không khớp với đoạn chỉ mục. Mặt khác, khi người dùng thực hiện tìm kiếm tài liệu theo nội dung thì câu truy vấn ở mức câu.



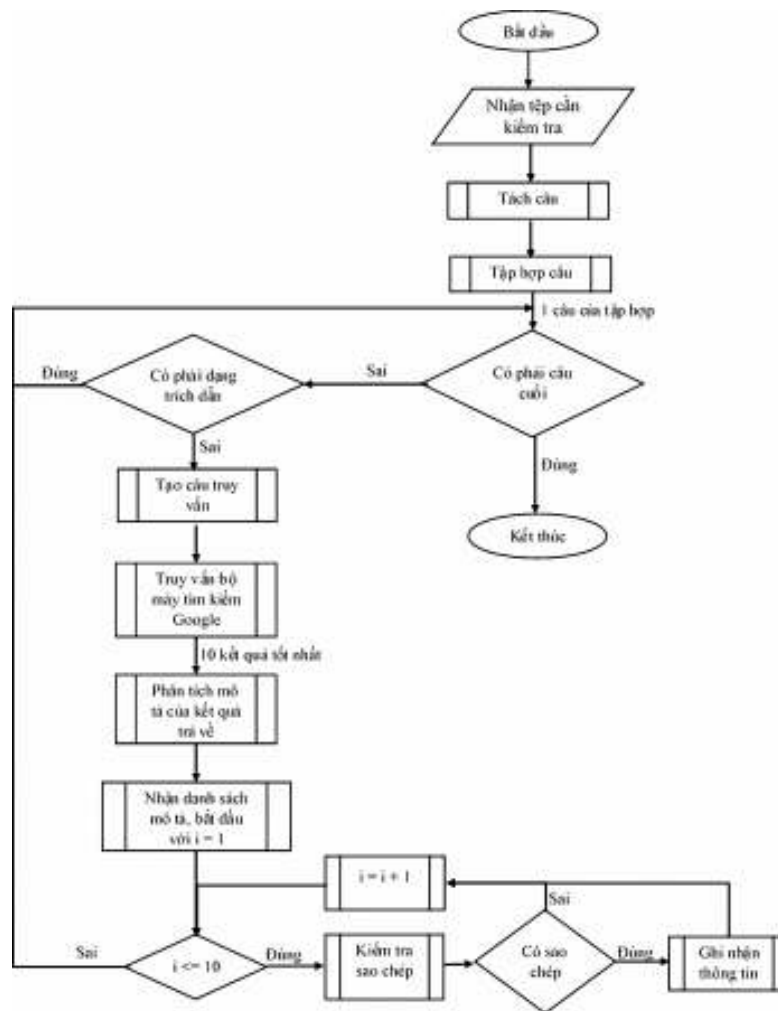
Hình 2. Các bước tiền xử lý văn bản

3.2. Mô-đun kiểm tra sao chép trên Internet

Hiện nay, tài liệu trên Internet là một con số khổng lồ, đa số tài liệu của nhiều lĩnh vực đều có trên Internet. Vì thế mà việc đọc và tìm hiểu trở nên dễ dàng, kể cả việc sao chép văn bản cũng thế. Nhưng với con số tài liệu nhiều như thế thì không thể nào truy cập trực tiếp đến nội dung của tất cả các tài liệu. Nên việc kiểm tra sao chép sẽ chuyển hóa thành tìm kiếm tài liệu có liên quan đến nội dung của tài liệu cần kiểm tra. Mô-đun này sẽ hỗ trợ người dùng kiểm tra sao chép tập tin luận

văn, không đưa ra kết luận có sao chép hay không? Quyết định này do người dùng quyết định sau khi xem kết quả kiểm tra sao chép của hệ thống.

Hệ thống sẽ thực hiện các bước: tách câu và kiểm tra xem câu có thuộc dạng trích dẫn tài liệu hay không, nếu không thì xây dựng câu truy vấn tìm kiếm, gửi đến máy chủ tìm kiếm web và nhận kết quả bao gồm mô tả của kết quả tìm kiếm, kiểm tra lại kết quả và trả về kết quả có độ tương đồng cao nhất so với câu truy vấn.



Hình 3. Các bước kiểm tra sao chép của một tập tin

3.3. Mô-đun kiểm tra với CSDL cục bộ

Hệ thống sử dụng MongoDB để lưu trữ nội dung của tài liệu vì MongoDB là hệ quản trị CSDL hướng tài liệu cung cấp cấu trúc lưu trữ thích hợp với cấu trúc chỉ mục nghịch đảo. Dữ liệu của MongoDB lưu trữ dạng phi cấu trúc, không có tính ràng buộc và dễ dàng mở rộng lưu trữ. Tài liệu cũng được lập chỉ mục mức câu nhưng có quan tâm đến vị trí của từ trong câu. Thông tin vị trí của từ trong câu sẽ được tích hợp (ví dụ như tổng khoảng cách về phương diện vị trí của các từ) vào hàm tính độ tương đồng để việc phát hiện sao

chép được chính xác hơn. Hệ thống sử dụng hai collection để lưu trữ tập chỉ mục của hệ thống, một collection dùng để lưu cấu trúc chỉ mục nghịch đảo – myIndex, một collection dùng để lưu thông tin của tài liệu – myDoc.

Đối với collection myDoc, mỗi một document có cấu trúc như sau:

Id	id tự sinh của MongoDB
docID	mã tài liệu
Content	tóm tắt
Link	vị trí lưu tài liệu

Đối với collection myIndex, mỗi một document có cấu trúc như sau:

Id	id tự sinh của MongoDB
IndexTerm	từ chỉ mục
Mã tài liệu 1	số lần xh vị trí 1 vị trí 2...
Mã tài liệu 2	số lần xh vị trí 1 vị trí 2...
...	

Khi có báo cáo luận văn cần kiểm tra, hệ thống sẽ căn cứ vào các thông tin thuộc tính có liên quan đến luận văn, cụ thể sẽ dựa vào ngành học, chủ đề của luận văn để xác định tập các báo cáo lưu trữ trong hệ thống cần phải kiểm tra. Cũng giống như phân hệ kiểm tra sao chép trên internet, phương pháp đồ thị được sử dụng để ước lượng độ tương đồng giữa hai câu, trong đó có quan tâm đến vị trí xuất hiện của từ trong câu.

3.4. Dạng trích dẫn tài liệu

Các phương pháp trích dẫn: gọi cước chú, gọi hậu chú, Vancouver, Harvard và hỗn hợp thứ tự số - chữ cái thì các dạng trích dẫn đều nằm ở cuối câu và có các dạng như sau:

- Dấu ngoặc đơn, bên trong là số hoặc tên tác giả - năm.

- Dấu ngoặc vuông, bên trong là số.

3.5. Tìm kiếm với máy chủ tìm kiếm web

Hiện nay, bộ máy tìm kiếm mạnh nhất trên Internet có thể nói là Google.com nên hệ thống sẽ chọn bộ máy này để thực hiện tìm kiếm trên Internet. Mặt khác không thể dùng GoogleSearchAPI vì giới hạn miễn phí là 100 lần mỗi ngày, trong khi đó một tài liệu có khi lên đến vài trăm câu. Vì thế phải xây dựng một

mô-đun truy vấn và phân tích kết quả tìm kiếm của Google.com.

Mô-đun truy vấn này sẽ sử dụng cấu trúc tìm kiếm:

https://www.google.com/search?q=CauTruyVan&num=10

Trong đó, **https://www.google.com/search?q=** là địa chỉ google dùng cho việc tìm kiếm; **CauTruyVan** là câu muốn truy vấn, cụ thể ở đây là câu cần kiểm tra sao chép sẽ thay thế cho **CauTruyVan**; **num=10** là nhận về 10 kết quả tốt nhất, đặt charset = UTF-8.

Ngoài ra, khi thực tìm kiếm thì một từ đơn sẽ được đặt trong dấu ngoặc kép “” để có thể tìm ra chính xác tài liệu chứa những từ này và thay dấu cách – khoảng trắng – bằng dấu +.

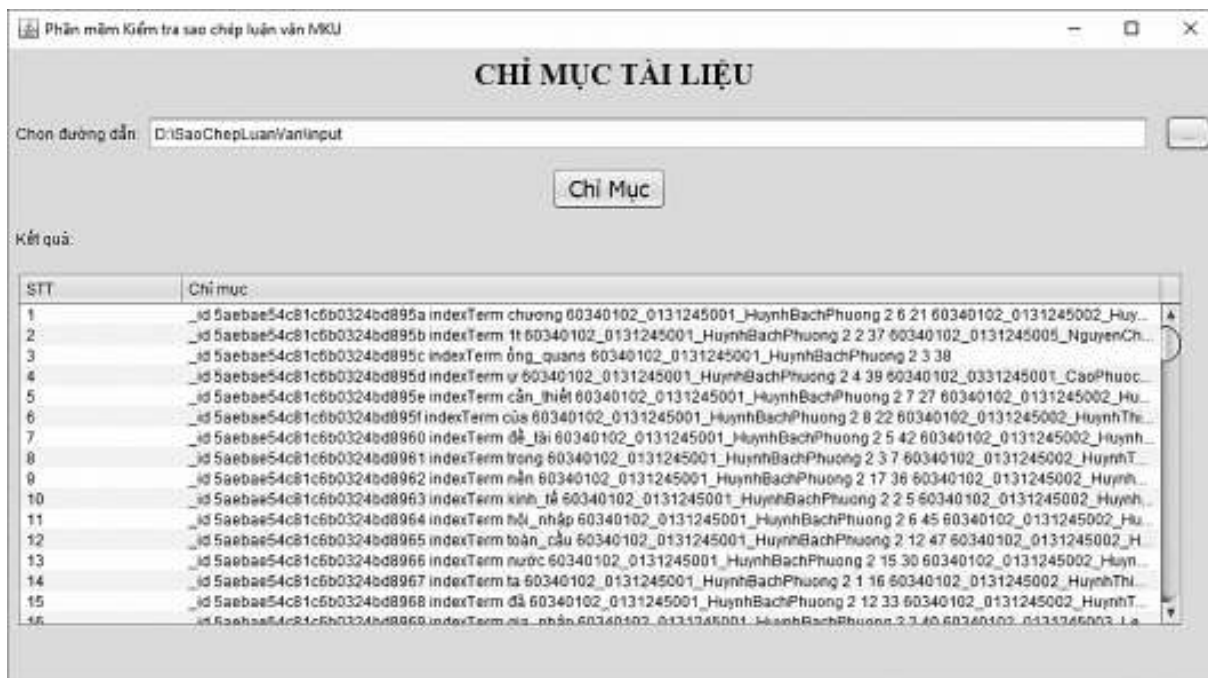
3.6. Kiểm tra kết quả trả về

Khi Google trả về 10 kết quả tốt nhất được cho là có nội dung giống với câu truy vấn hay câu kiểm tra sao chép, mỗi kết quả trả về có kèm theo một đoạn mô tả để minh chứng cho kết quả đó. Ở những mô tả minh chứng này sẽ chứa những từ trong câu truy vấn, những từ này có thể nằm liên tục hoặc rải rác. Vì thế phải thực hiện việc kiểm tra độ tương đồng giữa câu truy vấn với từng mô tả minh chứng (ở đây có 10 mô tả) để xác định câu truy vấn có thực sự sao chép hay không. Sau khi tính độ tương đồng, hệ thống sẽ chọn ra mô tả minh chứng có độ tương đồng cao nhất để hiển thị ra cho người dùng.

Việc kiểm tra từng câu sẽ giúp cho việc xác định giá trị sao chép của toàn tài liệu được chính xác hơn.

4. Kết quả đạt được

4.1. Chức năng chỉ mục luận văn



Hình 4. Giao diện chỉ mục luận văn

4.2. Chức năng kiểm tra sao chép internet

Chức năng này giúp cho giảng viên kiểm tra một luận văn của học viên có sao chép hay không và nếu có thì sao chép từ địa chỉ web nào.

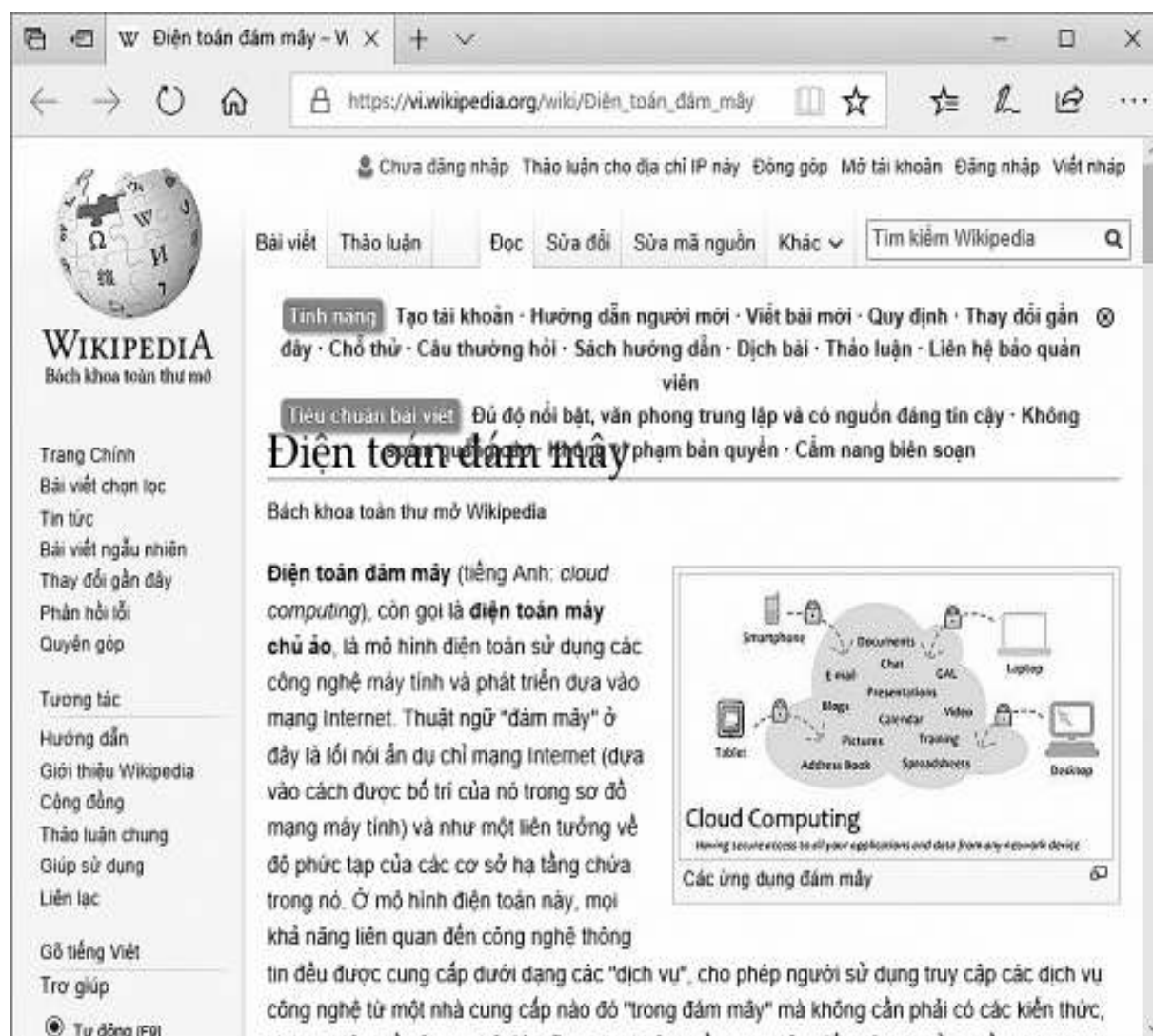


Hình 5. Giao diện kiểm tra sao chép trên internet

Kết quả sau khi kiểm tra, hệ thống sẽ trả về thông tin bao gồm: số thứ tự câu bị phát hiện sao chép, câu bị phát hiện sao chép, địa chỉ sao chép và chỉ số sao chép. Câu là từng câu của tập tin luận văn, minh chứng sao chép sẽ là phần mô tả sau khi tìm kiếm và in đậm những từ trùng khớp với từ trong câu của tập tin luận văn, địa chỉ sao chép chính là địa chỉ

của trang web chứa phần minh chứng sao chép. Giảng viên nhấp chuột vào địa chỉ web để có thể mở trực tiếp web chứa nội dung minh chứng sao chép.

Khi nhấp chuột vào 1 dòng của table hiển thị kết quả thì trình duyệt sẽ được mở với địa chỉ tương ứng với liên kết phát hiện bị sao chép.



Hình 6. Kết quả khi nhấp vào liên kết phát hiện bị sao chép

4.3. Chức năng kiểm tra sao chép cục bộ

Chức năng này giúp cho giảng viên kiểm tra một luận văn của học viên có sao chép các luận văn trong CSDL luận văn của trường hay không và nếu có thì từ tài liệu luận văn nào.



Hình 7. Giao diện kiểm tra sao chép trên CSDL cục bộ

Khi nhấn chuột vào 1 dòng của table hiển thị kết quả thì tài liệu bị sao chép tương ứng sẽ được mở để giảng viên kiểm tra.



Hình 8. Kết quả khi nhấn vào báo cáo phát hiện bị sao chép

5. Kết luận và hướng phát triển

5.1. Kết luận

Hệ thống hỗ trợ tìm kiếm luận văn trên cơ sở dữ liệu cục bộ, đặc biệt là tìm kiếm luận

văn theo nội dung của luận văn. Đồng thời, hệ thống hỗ trợ thành viên hội đồng bảo vệ luận văn kiểm tra sao chép luận văn, có loại trừ một số trường hợp trích dẫn kiểu *Vancouver*; *Harvard*, *hỗn hợp thứ tự số - chữ cái*, nghĩa

là loại trừ các trường hợp trích dẫn sử dụng dấu ngoặc đơn () hoặc dấu ngoặc vuông [] ở cuối câu.

5.2. Hướng phát triển của đề tài

- Xây dựng thêm cơ sở dữ liệu để ứng dụng cho quản lý luận văn, niên luận và đề tài nghiên cứu khoa học của sinh viên.

- Xây dựng hệ thống gợi ý giáo viên hướng dẫn phù hợp cho học viên, sinh viên.

TÀI LIỆU THAM KHẢO

1. Bao Jun-Peng, Shen Jun-Yi, Liu Xiao-Dong, Liu Hai-Yan, Zhang Xiao-Di. — Document Copy Detection Based On Kernel Method, In 2003 International Conference on Natural Language Processing and Knowledge Engineering Proceedings.
2. Brin, S., Davis, J., and Garcia-Molina, H, 1995. Copy Detection Mechanisms for Digital Documents. In Proceedings of ACM SIGMOD Annual Conference, San Jose, CA.
3. Hawkins, Tim; Plugge, Eelco; Membrey, Peter (September 26, 2010), The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing (1st ed.), Apress, p. 350, ISBN 978-1-4302-3051-9.
4. Krisztian Monostori, Arkady Zaslavsky, Heinz Schmidt —Document Overlap Detection System for Distributed Digital Libraries, Proceedings of the fifth ACM conference on Digital libraries, 2000, pp. 226 – 227.
5. Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho, 2008. A hybrid approach to word segmentation of Vietnamese texts. Springer-Verlag Berlin Heidelberg, Berlin.
6. Shivakumar, N. and Garcia-Monlina, H. —SCAM: A Copy Detection Mechanisms for Digital Documents. In Proceedings of International Conference on Theory and Practice of Digital Libraries, Austin, Texas. June 1995.
7. Si, A., Leong, H., and Lau, R. —CHECK: A Document Plagiarism Detection System. In Proceedings of ACM Symposium for Applied Computing, pp. 70-77, Feb 1997.
8. Quoc-Dinh Truong, Taoufiq Dkaki, Josiane Mothe, Pierre-Jean Charrel. GVC: a graph-based Information Retrieval Model. Conférence francophone en Recherche d'Information et Applications (CORIA 2008), Trégastel (France), 12/03/08-14/03/08, CNRS, p. 337-351, 2008

Ngày nhận bài: 1/6/2018

Ngày gửi phản biện: 14/6/2018