Giải pháp xây dựng phần mềm phát hiện đạo văn tiếng Việt trong các đề tài nghiên cứu khoa học quân sự

Hà Trung Hải[1*], Trần Ngọc Anh[2], Nguyễn Nhật An[3], Lê Mạnh Cường[4]

[1]Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ quân sự;
[2]Ban Khoa học quân sự, BTL 86;
[3]Viện Vật lý Y sinh học, Viện Khoa học và Công nghệ quân sự;
[4]Cục Khoa học quân sự.
*Email: hatrunghai1982@gmail.com.
Nhận bài ngày 30/11/2022; Hoàn thiện ngày 10/01/2022; Chấp nhận đăng ngày 10/4/2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.78.2022.166-169

TÓM TẮT

Đạo văn hiện nay là một trong những vấn nạn trong môi trường nghiên cứu khoa học và giáo dục đào tạo. Với sự phát triển nhanh chóng của Internet và các thiết bị Công nghệ thông tin, việc sao chép các nội dung từ các tài liệu khác là vô cùng 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 trên không gian mạng được chia sẻ như thư viện số, các tạp chí. Trong bài báo này, chúng tôi sẽ trình bày một giải pháp xây dựng phần mềm để phát hiện đạo văn tiếng Việt dựa trên các bài toán xử lý ngôn ngữ tự nhiên tiếng Việt như tách câu, tách từ, gán nhãn từ loại, sinh tập từ khóa phục vụ cho việc phát hiện sự trùng lặp về nội dung, sản phẩm nghiên cứu phục vụ công tác quản lý đề tài khoa học quân sự.

Từ khóa: Đạo văn; Hệ thống sao chép; Xử lý ngôn ngữ tự nhiên tiếng Việt.

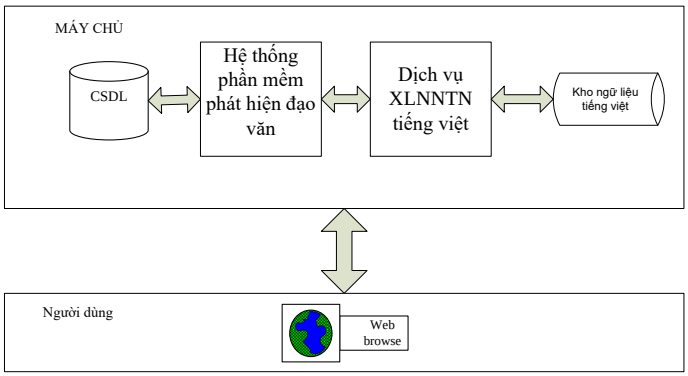
1. ĐẶT VẤN ĐỀ

Hiện nay công tác quản lý nghiên cứu khoa học trong quân đội đã có những quy định, chế tài nghiêm ngặt về việc sao chép các nội dung nghiên cứu khoa học, tuy nhiên, để nâng cao chất lượng trong công tác nghiên cứu cần có công cụ hiệu quả để đánh giá, ngăn chặn các hình thức đạo văn. Qua nghiên cứu tìm hiểu trên thị trường hiện nay đã có một số phần mềm phát hiện đạo văn trong các ngôn ngữ thông dụng trên thế giới đặc biệt là tiếng Anh như Phần mềm kiểm tra đạo văn Plagiarism-checker.me [1], phần mềm Turnitin [2], ngoài ra thì còn có một số hệ thống phát hiện đạo văn khác đã và đang được áp dụng thực tế như iThenticate, Viper, Dupli checker, Copy leaks, Paperrater, Plagium, Plagiarisma, Plagscan [3, 4]. Trong nước thì đã có một số nhóm xây dựng và triển khai phần mềm phát hiện đạo văn cho văn bản tiếng việt như: Nhóm tác giả Trần Cao Đệ và các cộng sự [5] thuộc Đại học Cần Thơ đã xây dựng một hệ thống phát hiện đạo văn với cơ sở dữ liệu ban đầu là 3.000 tài liệu; Sản phẩm của Trường Đại học Công nghệ, Đại học quốc gia với tên gọi DoIT (Document improvement Tool)[6]; Phần mềm Coopy[7] của Viện công nghệ thông tin và Truyền thông/Đại học Bách khoa Hà Nội. Tuy nhiên, đối với các phần mềm nước ngoài bên cạnh mức giá cao thì chưa có minh chứng làm việc tốt trên tiếng Việt, đối với các sản phẩm trong nước thì phải dùng trực tuyến việc thực hiện kiểm tra sao chép 1 tài liệu từ kho tài liệu trên internet, tức là kiểm tra với tài liệu nguồn từ internet. Trong khi đó, dữ liệu cần kiểm tra đối với các đơn vị trong Quân đội là kho dữ liệu đóng chính vì vậy nhóm tác giả trình bày một phương pháp để xây dựng phần mềm phát hiện đạo văn tiếng Việt sử dụng nguồn dữ liệu đóng ứng dụng trong các đơn vị quản lý đề tài trong Quân đội.

2. GIẢI PHÁP XÂY DỰNG PHẦN MỀM PHÁT HIỆN ĐẠO VĂN TIẾNG VIỆT

Phần mềm được xây dựng với mục đích là kiểm tra phát hiện đạo văn sao chép và phát hiện đạo văn ý tưởng. Với dữ liệu đầu vào là một file định dạng pdf hoặc định dạng word, hệ thống sẽ tự động kiểm tra trong kho cơ sở dữ liệu đóng và phát hiện ra các tài liệu bị sao chép, hiển thị kết quả lên cho người dùng.

2.1. Kiến trúc thành phần của hệ thống phần mềm



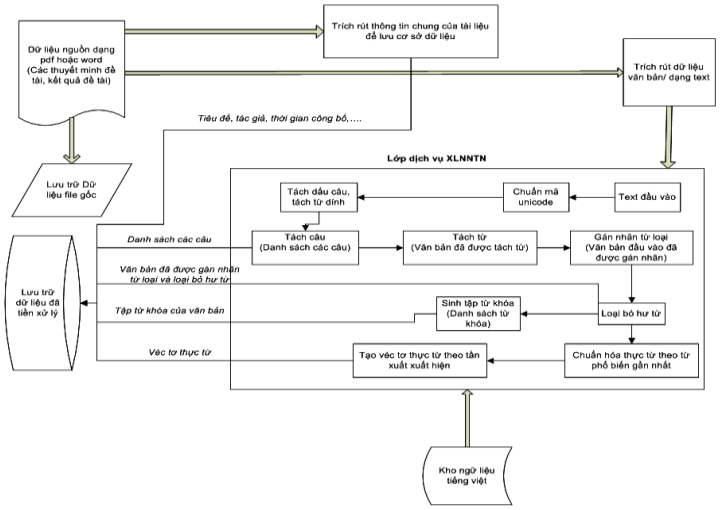
Hình 1. *Kiến trúc thành phần của hệ thống phần mềm.*

Hệ thống phần mềm phát hiện đạo văn bao gồm các thành phần sau:

- Cơ sở dữ liệu: Bao gồm cơ sở dữ liệu lưu toàn bộ thông tin của hệ thống phần mềm phát hiện đạo văn như dữ liệu người dùng được phép khai thác phần mềm; dữ liệu về các tài liệu của đề tài nhiệm vụ khoa học
- Kho ngữ liệu tiếng Việt: Chính là kho ngữ liệu tri thức văn bản tiếng Việt, từ điển tiếng Việt đã được xử lý phục vụ cho các lớp dịch vụ xử lý ngôn ngữ tự nhiên.
- Hệ thống phần mềm: gồm các chức năng cho phép người dùng có thể mở rộng kho dữ liệu đóng, các chức năng kiểm tra phát hiện đạo văn.
- Dịch vụ xử lý ngôn ngữ tự nhiên tiếng Việt: Là lớp dịch vụ chạy ngầm cung cấp các module về XLNN phục vụ cho việc tiền xử lý dữ liệu, phát hiện đạo văn sao chép, đạo văn ý tưởng.

2.2. Quy trình tiền xử lý dữ liệu và tạo lập kho cơ sở dữ liệu

Phần này chúng tôi sẽ trình bày chi tiết các bước thực hiện để xây dựng kho dữ liệu, dữ liệu được lưu trữ đã được tiền xử lý để tăng tốc độ tìm kiếm phát hiện đạo văn.

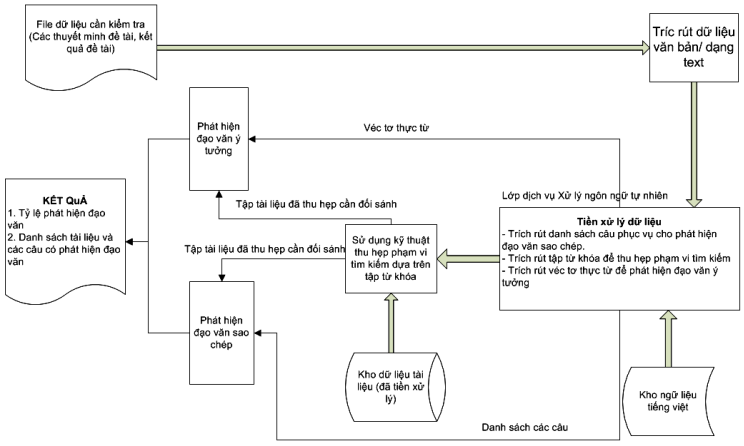


Hình 2. *Quy trình tiền xử lý và tạo lập kho cơ sở dữ liệu.*

Ngoài những dữ liệu chung của file tài liệu như Tên đề tài (tiêu đề), các tác giả, ngày tháng năm công bố tài liệu, lĩnh vực của đề tài, mục tiêu đề tài, nội dung nghiên cứu của đề tài thì cần lưu trữ các dữ liệu đã được tiền xử lý phục vụ cho bài toán phát hiện đạo văn. Dữ liệu được lưu trong kho bao gồm danh sách câu, văn bản đã được gán nhãn từ loại và loại bỏ hư từ, tập từ khóa của văn bản và véc tơ thực từ.

2.3. Quy trình xây dựng chức năng phát hiện đạo văn

Phần này chúng tôi sẽ trình bày các bước để tiến hành phát hiện đạo văn sao chép và đạo văn ý tưởng.



Hình 3. *Quy trình phát hiện đạo văn.*

Giải thích quy trình:

+ File tài liệu cần kiểm tra sẽ được trích rút dữ liệu văn bản/ dạng text làm đầu vào cho lớp tiền xử lý dữ liệu để được kết quả đầu ra gồm: (1) tập từ khóa của tài liệu (2) danh sách các câu của tài liệu (3) véc tơ thực từ của tài liệu.

+ Tập từ khóa của tài liệu được sử dụng để thu hẹp phạm vi tìm kiếm trong kho cơ sở dữ liệu, dựa trên tập từ khóa của các tài liệu đã được tiền xử lý trong kho cơ sở dữ liệu.

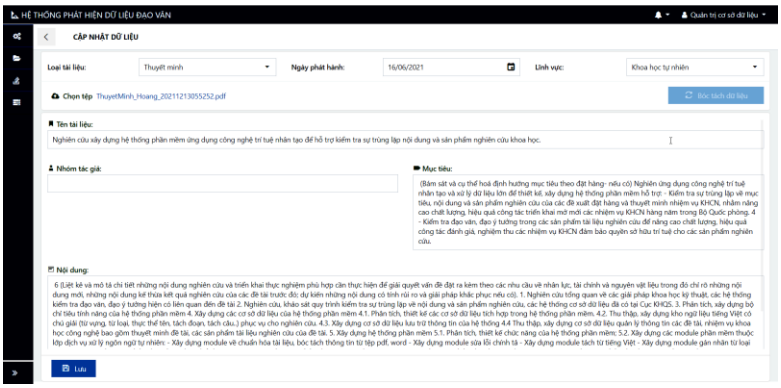
+ Danh sách các câu của tài liệu sẽ được sử dụng để đối sánh với tài liệu trong danh sách tài liệu gốc (đã được thu hẹp phạm vi tìm kiếm) để phát hiện đạo văn sao chép.

+ Vecto thực từ của tài liệu sẽ được sử dụng để đánh giá với các véc tơ thực từ của tài liệu trong danh sách tài liệu gốc (đã được thu hẹp phạm vi tìm kiếm) để phát hiện đạo văn ý tưởng.

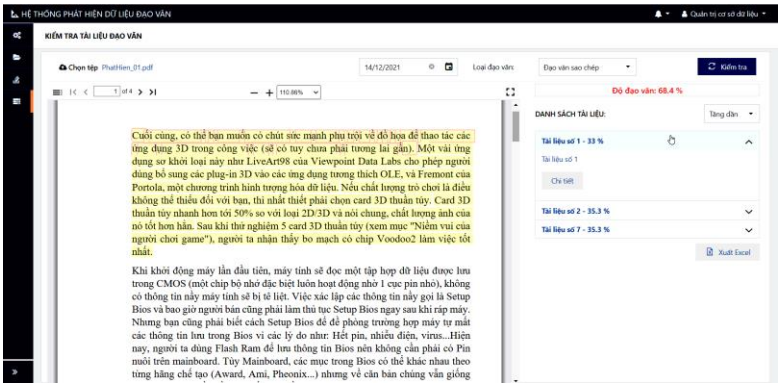
2.4. Giao diện một số chức năng của phần mềm

Người dùng sẽ chọn file dữ liệu đầu vào thuộc đề xuất, thuyết minh hay kết quả nghiên cứu. Sau khi dữ liệu được trích rút dạng văn bản sẽ tự động đưa vào các trường dữ liệu, tiền xử lý dữ liệu vào lưu vào kho cơ sở dữ liệu.

Kết quả hiển thị phía bên trái sẽ là file dữ liệu cần kiểm tra; bên phải là danh sách các tài liệu có phát hiện bị đạo văn và độ đạo văn. Khi người dùng chọn đến một tài liệu nguồn thì tài liệu kiểm tra sẽ được bôi màu vàng các đoạn văn bản có phát hiện đạo văn.



Hình 4. *Giao diện chức năng xây dựng kho cơ sở dữ liệu.*



Hình 5. *Giao diện chức năng kiểm tra đạo văn.*

3. KẾT LUẬN

Bài báo đã trình bày một phương pháp tiếp cận của chúng tôi để phát triển hệ thống phần mềm phát hiện đạo văn gồm cả đạo văn sao chép và đạo văn ý tưởng sử dụng các bài toán xử lý ngôn ngữ tự nhiên chuyên sâu như tách câu, tách từ, gán nhãn từ loại, loại bỏ hư từ, tạo véc tơ thực từ để tiền xử lý dữ liệu, thu hẹp phạm vi tìm kiếm qua đó tăng tốc độ tìm kiếm. Phần mềm đã được đưa vào chạy thử nghiệm tại Cục khoa học Quân sự/Bộ Quốc phòng.

TÀI LIỆU THAM KHẢO

[1]. Plagiarism-checker, Website https://Plagiarism-checker.me
[2]. Turnitin, https://www.turnitin.com/
[3]. Naik, Ramesh R., Maheshkumar B. Landge, and C. Namrata Mahender. "*A review on plagiarism detection tools*." International Journal of Computer Applications 125.11 (2015).
[4]. Ali, Asim M. El Tahir, Hussam M. Dahwa Abdulla, and Vaclav Snasel. "*Overview and Comparison of Plagiarism Detection Tools.*" Dateso. 2011.
[5]. De, T. C. "*Developing plagiarism detection system for Vietnamese University.*" 12th Vietnam—Japan International Joint Symposium, Can Tho. 2014.
[6]. DoIt, http://doit.uet.vnu.edu.vn/
[7]. Coopy, http://coopy.soict.ai/

ABSTRACT

Solutions building software detecting Vietnamese plagiarism in military science research topics

Plagiarism is currently one of the problems in scientific research, education, and the training environment. With the rapid development of the Internet and Information Technology devices, it is effortless to copy content from other documents. Violators have many means to find and steal other people's content or ideas because the research and ideas are almost readily available in data warehouses, libraries, etc. In this paper, we will present a model to build software to detect Vietnamese plagiarism based on Vietnamese natural language processing problems such as sentence separation, word separation, word classification, generation of words, etc. serve to detect duplication of research content and products for the management of military science topics.

Keywords: Plagiarism; Natural language Vietnamese processing.