ỨNG DỤNG CÁC KỸ THUẬT TÌM KIẾM THÔNG TIN CHO HỆ THỐNG CHIA SẺ TÀI NGUYÊN HỌC TẬP

APPLICATION OF INFORMATION RETRIEVAL TECHNIQUES FOR A LEARNING RESOURCE SHARING SYSTEM

NGUYỄN THỊ HỒNG CẨM[1,a], TRƯƠNG QUỐC ĐỊNH[2],
TRƯƠNG MỸ THU THẢO[1]
[1]*Trường Đại học Sư phạm Kỹ thuật Vĩnh Long*
[2]*Trường Công nghệ thông tin và truyền thông, Trường Đại học Cần Thơ*
[a]*Tác giả liên hệ: nthcam.c3hoaninh@vinhlong.edu.vn*

Nhận bài (Received): 06/12/2024; Phản biện (Reviewed): 28/12/2024; Chấp nhận (Accepted): 05/01/2025

TÓM TẮT

Nhu cầu chia sẻ và truy cập tài nguyên học tập ngày càng tăng trong giáo dục hiện đại, đặc biệt là trong bối cảnh học trực tuyến và tự học, đòi hỏi một hệ thống hiệu quả để quản lý nội dung chất lượng và đáp ứng các nhu cầu tìm kiếm phức tạp từ người dùng. Nghiên cứu này phát triển một hệ thống chia sẻ tài nguyên học tập trên mô hình Client-Server, giải quyết hai vấn đề chính: (1) quản lý nội dung với tính năng kiểm tra trùng lặp và (2) hỗ trợ tìm kiếm tài nguyên và các chủ đề thảo luận. Phân quyền chặt chẽ được thiết kế để quản trị viên và người dùng có thể thực hiện các tác vụ khác nhau, phù hợp với vai trò của họ. Để đảm bảo chất lượng nội dung, nghiên cứu sử dụng độ đo tương đồng Jaccard cải tiến, giúp phát hiện trùng lặp tài liệu với độ chính xác cao, đặc biệt trong các trường hợp có từ khóa không đồng nhất. Về tính năng tìm kiếm, hệ thống tích hợp mở rộng truy vấn với PhoBERT để tự động bổ sung các từ khóa liên quan, mở rộng đáng kể phạm vi tìm kiếm và mang lại nhiều lựa chọn hơn cho người dùng. Kết quả kiểm thử cho thấy hệ thống đạt hiệu quả cao trong quản lý nội dung, với độ chính xác vượt trội của Jaccard cải tiến so với Jaccard thông thường trong các kịch bản phức tạp. Tính năng tìm kiếm, qua 50 mẫu thử, thể hiện khả năng trả về kết quả phù hợp, đáp ứng tốt nhu cầu người dùng. Tuy nhiên, mở rộng truy vấn vẫn còn một số hạn chế về độ chính xác của từ khóa gợi ý. Nghiên cứu này không chỉ góp phần nâng cao hiệu quả truy xuất thông tin mà còn hỗ trợ người học trong việc tiếp cận và khai thác tài nguyên học tập một cách hiệu quả nhất.

Từ khóa: hệ thống chia sẻ tài nguyên học tập, kiểm tra trùng lặp, tìm kiếm tài nguyên, đo độ tương đồng Jaccard cải tiến, mở rộng truy vấn.

ABSTRACT

The demand for sharing and accessing learning resources is increasing in modern education, especially in the context of online learning and self-study. This necessitates an effective system for managing quality content and meeting users' complex search needs. This research develops a learning resource sharing system based on a Client-Server model, addressing two main issues: (1) content management with a duplication checking feature and (2) supporting resource search and discussion topics. A strict role-based

access control system is designed to enable administrators and users to perform different tasks according to their roles. To ensure content quality, the study utilizes an improved Jaccard similarity measure, which helps detect document duplicates with high accuracy, especially in cases with heterogeneous keywords. Regarding search functionality, the system integrates query expansion with PhoBERT to automatically add related keywords, significantly expanding the search scope and providing users with more options. Testing results show that the system performs highly effectively in content management, with the improved Jaccard measure outperforming the conventional Jaccard in complex scenarios. The search functionality, tested with 50 sample queries, demonstrates the system's ability to return relevant results that meet user needs. However, the query expansion feature still has some limitations in the accuracy of suggested keywords. This study not only improves information retrieval efficiency but also aids learners in accessing and utilizing learning resources more effectively.

Keywords: *learning resource sharing system, duplication checking, resource search, Improved Jaccard similarity measure, query expansion.*

1. MỞ ĐẦU

1.1. Giới thiệu

Trong thời đại công nghệ thông tin phát triển mạnh mẽ, giáo dục đang trải qua những thay đổi đáng kể. Công nghệ hiện đại không chỉ hỗ trợ giảng dạy truyền thống mà còn mở ra phương pháp học tập linh hoạt và hiệu quả hơn. Các hệ thống chia sẻ tài nguyên học tập số đã trở thành công cụ thiết yếu, hỗ trợ học sinh và giáo viên trong việc tìm kiếm và chia sẻ kiến thức.

Theo Chương trình Giáo dục Phổ thông 2018, phát triển khả năng tự học là một năng lực cốt lõi. Học sinh có thể rèn luyện kỹ năng tự học và mở rộng kiến thức qua việc tìm kiếm tài liệu học tập. Tuy nhiên, thư viện truyền thống còn nhiều hạn chế về số lượng, chi phí và tính đa dạng tài liệu, trong khi tài liệu số giúp khắc phục những hạn chế này với khả năng truy cập dễ dàng. Mặt khác, tài liệu số cũng đặt ra thách thức trong việc đánh giá tính chính xác của thông tin, khi một nghiên cứu gần đây cho thấy nhiều học sinh gặp khó khăn trong việc xác định nguồn gốc đáng tin cậy trực tuyến, dẫn đến nguy cơ sử dụng thông tin sai lệch.

Tài liệu số mang lại sự tiện lợi với khả năng truy cập dễ dàng qua internet, nhưng cũng đặt ra thách thức trong việc đánh giá tính chính xác của thông tin. Một nghiên cứu gần đây cho thấy nhiều học sinh gặp khó khăn trong việc xác định nguồn gốc chính xác của thông tin trực tuyến, dẫn đến nguy cơ sử dụng thông tin sai lệch.

Trong bối cảnh đó, hệ thống chia sẻ tài nguyên học tập số dành cho giáo viên và học sinh Trung học Phổ thông là giải pháp hiệu quả. Nghiên cứu này không chỉ xây dựng một hệ thống chia sẻ tài nguyên học tập mà còn giới thiệu hai cải tiến nổi bật: (1) áp dụng Jaccard cải tiến có xét đến vị trí từ, giúp nâng cao độ chính xác trong việc phát hiện nội dung trùng lặp, và (2) tích hợp PhoBERT trong mở rộng truy vấn ngữ nghĩa, hỗ trợ tìm kiếm tài liệu liên quan ngay cả khi từ khóa không hoàn toàn trùng khớp. Đây là những điểm mới góp phần khắc phục các hạn chế của các hệ thống hiện có.

1.2. Những nghiên cứu liên quan

Trong những năm gần đây, nhiều hệ thống chia sẻ tài nguyên học tập đã ra đời

nhằm đáp ứng nhu cầu tìm kiếm, chia sẻ và quản lý tài liệu học tập của người dùng. Các hệ thống này cung cấp nhiều chức năng hữu ích như tìm kiếm từ khóa, chia sẻ tài liệu, bình luận và đánh giá, tạo điều kiện cho học sinh và giáo viên dễ dàng truy cập và trao đổi kiến thức. Một số nền tảng tiêu biểu bao gồm các thư viện học liệu mở, diễn đàn học thuật, và các trang web chia sẻ tài liệu trực tuyến (123.doc.org; vndoc.vn; violet; tailieuhoctap.net; tailieu.vn…).

Tuy nhiên, các hệ thống hiện tại còn tồn tại một số hạn chế. Trước hết, chúng thường thiếu tính năng kiểm tra và loại bỏ nội dung trùng lặp, dẫn đến tình trạng lặp lại không cần thiết trong kho tài liệu, làm giảm hiệu quả tìm kiếm và gây khó khăn cho người dùng trong việc xác định nguồn thông tin đáng tin cậy. Thêm vào đó, nhiều hệ thống chỉ hỗ trợ tìm kiếm từ khóa đơn giản mà không tích hợp các phương pháp tìm kiếm theo nội dung, khiến người dùng khó tiếp cận thông tin toàn diện và chính xác. Đặc biệt, các hệ thống này chưa được thiết kế riêng cho đối tượng học sinh THPT theo Chương trình Giáo dục Phổ thông 2018, vốn tập trung phát triển năng lực tự học và kỹ năng nghiên cứu của học sinh, nên khó hỗ trợ phù hợp các nội dung theo yêu cầu chương trình này. Những hạn chế trên đã tạo động lực cho các nghiên cứu nhằm nâng cao khả năng tìm kiếm tài liệu theo nội dung và kiểm tra mức độ trùng lặp, đồng thời phát triển các nền tảng chuyên biệt cho nhu cầu của học sinh Trung học Phổ thông.

Vì những lý do trên, một hệ thống chia sẻ tài nguyên học tập được xây dựng nhằm khắc phục những hạn chế hiện tại và đáp ứng nhu cầu học tập đặc thù của học sinh THPT. Hệ thống này bước đầu tập trung vào hai bài toán cụ thể: kiểm tra mức độ trùng lặp nội dung giữa các tài liệu số và tìm kiếm tài liệu theo nội dung. Đây là hai yếu tố cốt lõi, giúp đảm bảo tính chính xác và sự tiện dụng của tài liệu học tập, đồng thời hỗ trợ người dùng tìm kiếm và sử dụng tài nguyên một cách hiệu quả hơn.

Kỹ thuật tìm kiếm thông tin (IR) là một yếu tố quan trọng quyết định đến hiệu quả sử dụng các hệ thống chia sẻ tài nguyên học tập. Mục tiêu của IR là cung cấp cho người dùng các thông tin liên quan và hữu ích dựa trên nhu cầu của người dùng.

Một số công trình nghiên cứu liên quan đến IR trên thế giới và tại Việt Nam hiện nay bao gồm:

+ Nhóm tác giả ([1]) đã đề xuất mô hình truy xuất thông tin mới mang tên GVC (Graph Vertices Comparison), dựa trên việc so sánh các đỉnh đồ thị. Mô hình này sử dụng phương pháp đo độ tương đồng để so sánh các tài liệu và truy vấn của người dùng dựa trên sự khớp đồ thị.

+ Nhóm tác giả ([2]) đã sử dụng độ tương đồng cosine để tìm kiếm các tài liệu cụ thể về một chủ đề đã được chỉ định. Quá trình này bao gồm các bước tiền xử lý văn bản, tính toán trọng số từ khóa và đo lường sự tương đồng văn bản.

+ Tác giả ([3]) đã áp dụng kỹ thuật tìm kiếm thông tin để xác định các tin thật có nội dung gần giống với tin cần kiểm tra. Sau đó, họ sử dụng độ đo cosine để đánh giá khả năng của bản tin kiểm tra có thể là tin giả hay không.

+ Trong một nghiên cứu khác, nhóm tác giả ([4]) đã sử dụng phương pháp tìm kiếm thông tin để đề xuất giải pháp phát hiện sao chép trong luận văn. Họ kiểm chứng các báo cáo luận văn có sao chép từ hai nguồn tài nguyên: cơ sở dữ liệu cục bộ và nguồn dữ liệu trực tuyến.

+ Nhóm tác giả ([5]) đã phát triển một

hệ thống chia sẻ các khóa học trực tuyến và tài liệu học tập. Hệ thống này cung cấp các tính năng cập nhật các khóa học cùng với các bài kiểm tra và đánh giá. Học sinh có thể dễ dàng tra cứu tài liệu bài giảng của giáo viên bằng cách sử dụng kỹ thuật tìm kiếm toàn văn.

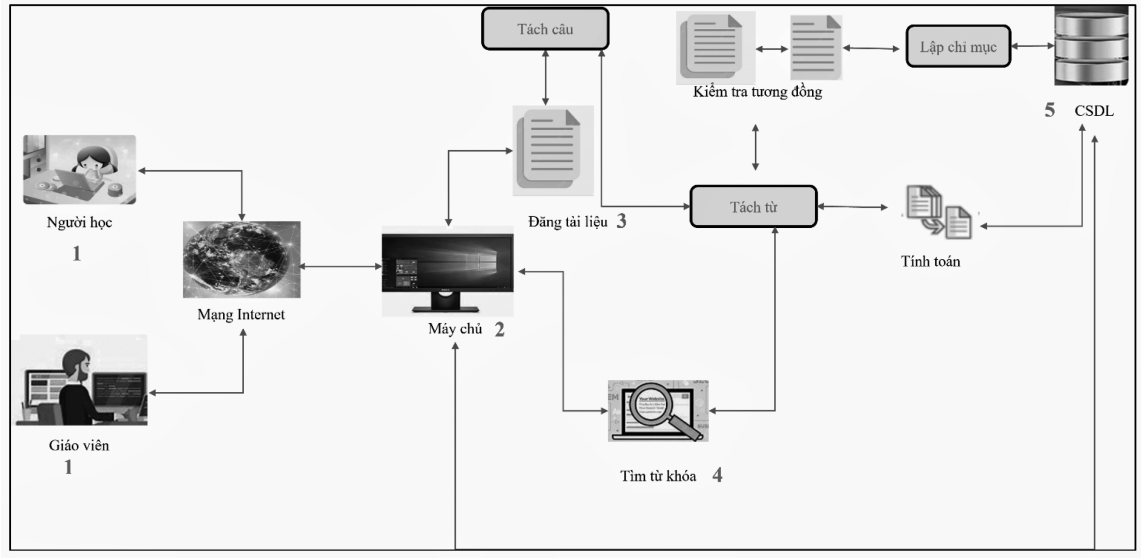
Mặc dù các nghiên cứu trước đã áp dụng các phương pháp như độ đo tương đồng Jaccard và tìm kiếm từ khóa đơn giản, tuy nhiên, chúng chưa xem xét yếu tố vị trí từ trong việc phát hiện nội dung trùng lặp, dẫn đến độ chính xác không cao trong các trường hợp nội dung tương tự nhưng trật tự từ khác nhau. Nghiên cứu này đã cải tiến phương pháp Jaccard bằng cách tích hợp yếu tố vị trí từ để cải tiến phương pháp xác định trùng lặp về mặt nội dung. Bên cạnh đó, việc ứng dụng PhoBERT trong mở rộng truy vấn mang lại hiệu quả vượt trội so với tìm kiếm từ khóa truyền thống, đặc biệt trong ngữ cảnh hệ thống tìm kiếm tài nguyên học tập.

Tuy nhiên, những công trình trên đã tạo nền tảng quan trọng cho việc phát triển các giải pháp tìm kiếm tài liệu và phát hiện nội dung trùng lặp, tuy nhiên vẫn còn cần nhiều cải tiến để đáp ứng các yêu cầu đặc thù của đối tượng học sinh THPT theo chương trình giáo dục mới. Dựa trên cơ sở này, nghiên cứu áp dụng các kỹ thuật IR tiên tiến cùng với phương pháp phát hiện trùng lặp để xây dựng hệ thống tài nguyên đáp ứng hiệu quả hơn nhu cầu học tập và tự học của học sinh THPT.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Kiến trúc hệ thống

Hệ thống chia sẻ tài nguyên học tập dạng diễn đàn được xây dựng trên kiến trúc Client-Server nhằm tối ưu hóa hiệu quả quản lý và tìm kiếm tài nguyên (Hình 1).



Hình 1. Mô hình kiến trúc tổng quát hệ thống (Mô hình Client-Server)

Trong đó, Client đóng vai trò giao diện người dùng, cho phép học sinh và giáo viên (1) gửi yêu cầu đăng tải nội dung hoặc tìm kiếm tài liệu. Server (2) là thành phần cốt lõi, chịu trách nhiệm xử lý các yêu cầu của Client. Cụ thể, server thực hiện các tác vụ như kiểm tra quyền hạn người dùng, xử lý bài toán đăng tài liệu (kiểm tra trùng lặp nội

dung – 3) và xử lý tìm kiếm tài nguyên học tập (4) dựa trên truy vấn từ người dùng. Cơ sở dữ liệu MongoDB được sử dụng để lưu trữ CSDL (5).

2.2. Nội dung và Giải pháp cho Hai Bài Toán trong Hệ thống Chia sẻ Tài nguyên Học tập

Trong hệ thống chia sẻ tài nguyên học tập, hai bài toán chính cần giải quyết là quản lý nội dung và tìm kiếm tài nguyên. Cả hai bài toán đều yêu cầu ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên và truy xuất thông tin nhằm tối ưu hóa chất lượng thông tin và cải thiện trải nghiệm người dùng.

2.2.1. Bài toán quản lý nội dung

a) Nội dung bài toán

Quản lý nội dung là yếu tố cần thiết để duy trì tính chính xác và độ tin cậy của hệ thống. Bài toán này bao gồm hai thành phần là kiểm duyệt nội dung và phát hiện trùng lặp (sử dụng các kỹ thuật kiểm tra độ trùng lặp giữa các tài liệu để loại bỏ các tài liệu giống nhau, giúp tiết kiệm không gian lưu trữ và giảm độ phức tạp của hệ thống).

b) Giải pháp thực hiện

Phần kỹ thuật kiểm duyệt nội dung được thực hiện bởi các Admin, những người có chuyên môn phù hợp do Super Admin phân quyền.

Phát hiện trùng lặp nội dung:

Độ đo Jaccard ([6]) là một trong những chỉ số phổ biến nhất dùng để đo lường sự tương đồng giữa hai tập hợp. Chỉ số này được định nghĩa là tỷ lệ giữa phần giao của hai tập hợp và phần hợp của chúng. Nó có những hạn chế trong việc xử lý các văn bản có cấu trúc khác nhau hoặc sự thay đổi vị trí của các từ. Để cải thiện độ chính xác và tính ứng dụng, nhiều phương pháp cải tiến đã được phát triển ([7]) như Jaccard có xét đến vị trí từ, Jaccard theo trọng số, và Jaccard sử dụng văn bản ngữ nghĩa ([8], [9].

- Phương pháp Jaccard cải tiến được phát triển với khả năng xét đến vị trí từ trong văn bản. Điều này cải thiện độ chính xác của phép đo, đặc biệt trong các trường hợp văn bản có nội dung tương tự nhưng trật tự từ khác nhau.

Độ đo tương đồng Jaccard cải tiến được xác định theo công thức sau:

$$Sim_{J-cải\ tiến}(a,b) = \frac{|w \in a \cap b \ || \ pos_a(w) - pos_b(w) \ | \leq mark|}{|a \cup b|} \quad (1)$$

Trong đó:

a và b: Hai văn bản được so sánh

$|a \cup b|$: Tập hợp hợp của hai văn bản a và b.

$intersection = \{w \in a \cap b \ || \ pos_a(w) - pos_b(w) \ | \leq mark\}$: Phần giao có xét đến vị trí từ. Một từ w được xem là thuộc phần giao nếu:

+ Tồn tại trong cả hai văn bản.

+ Khoảng cách giữa vị trí của từ đó trong hai văn bản không lớn hơn giá trị ***mark*** (một ngưỡng được định trước).

$pos_a(w)$ và $pos_a(w)$: Vị trí của từ w trong văn bản a và b

- Quy trình tính toán Jaccard cải tiến có xét đến vị trí từ:

Bước 1: Tiền xử lý văn bản

+ Tách văn bản thành danh sách các từ.

+ Ghi lại vị trí của từng từ trong văn bản.

Bước 2: Tính tập hợp hợp không cần xét đến vị trí của từ:

Bước 3: Tính phần giao có xét đến vị trí của từ

+ Xét qua từng từ (w) trong danh sách từ của văn bản a

+ Kiểm tra xem từ đó (w) có tồn tại trong b hay không

+ Nếu từ tồn tại trong cả hai văn bản và $pos_a(w) - pos_b(w) | \leq mark$, thêm từ đó vào tập giao.

khoảng cách giữa vị trí của từ đó trong hai văn bản không lớn hơn mark, từ đó được xem là một phần của giao

Bước 4: Tính độ đo Jaccard cải tiến theo công thức (1).

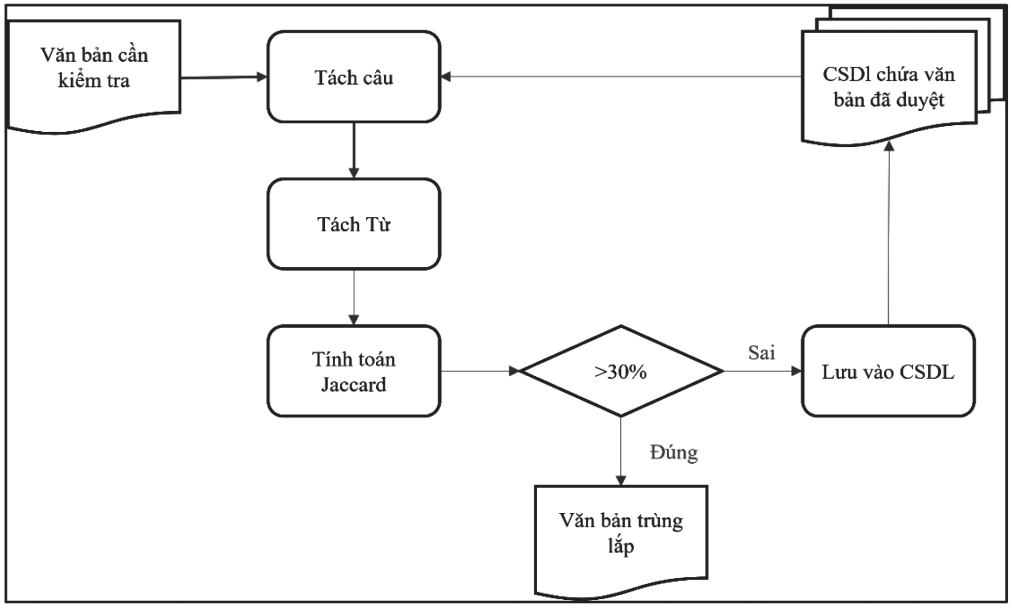
- Quy trình phát hiện trùng lặp: được thực hiện qua hai bước: (1) Kiểm tra tiêu đề: Tiêu đề tài liệu được so sánh với các tiêu đề đã có trong cơ sở dữ liệu. Hệ thống sẽ cảnh báo nếu phát hiện trùng lặp. (2) Kiểm tra nội dung: Nội dung tài liệu được kiểm tra ở cấp độ câu (quy trình được minh họa trong Hình 2).

+ Đầu tiên, văn bản được chia thành các câu nhỏ để dễ dàng xử lý.

+ Sau đó, từng câu trong văn bản sẽ được tách thành các từ riêng lẻ.

+ Dựa trên các từ đã tách, hệ thống sẽ tính toán độ tương đồng giữa văn bản cần kiểm tra và các văn bản đã duyệt trong cơ sở dữ liệu, sử dụng chỉ số Jaccard cải tiến để đo lường mức độ tương đồng này. Kết quả tính toán chỉ số Jaccard cải tiến được so sánh với ngưỡng 30%. Nếu độ tương đồng lớn hơn 30%, văn bản được xác định là trùng lặp và sẽ không được kiểm duyệt. Ngược lại, nếu độ tương đồng nhỏ hơn hoặc bằng 30%, văn bản được coi là không trùng lặp và sẽ được lưu vào cơ sở dữ liệu.

+ Cơ sở dữ liệu chứa các văn bản đã duyệt là nơi lưu trữ những tài liệu đã qua kiểm duyệt, được sử dụng để so sánh trùng lặp với các văn bản mới.



Hình 2. Mô hình chức năng kiểm tra trùng lặp nội dung

Toàn bộ quá trình này giúp đảm bảo rằng các tài liệu được đăng tải lên hệ thống là duy nhất và không bị trùng lặp, từ đó tiết kiệm tài nguyên và đảm bảo chất lượng thông tin chia sẻ.

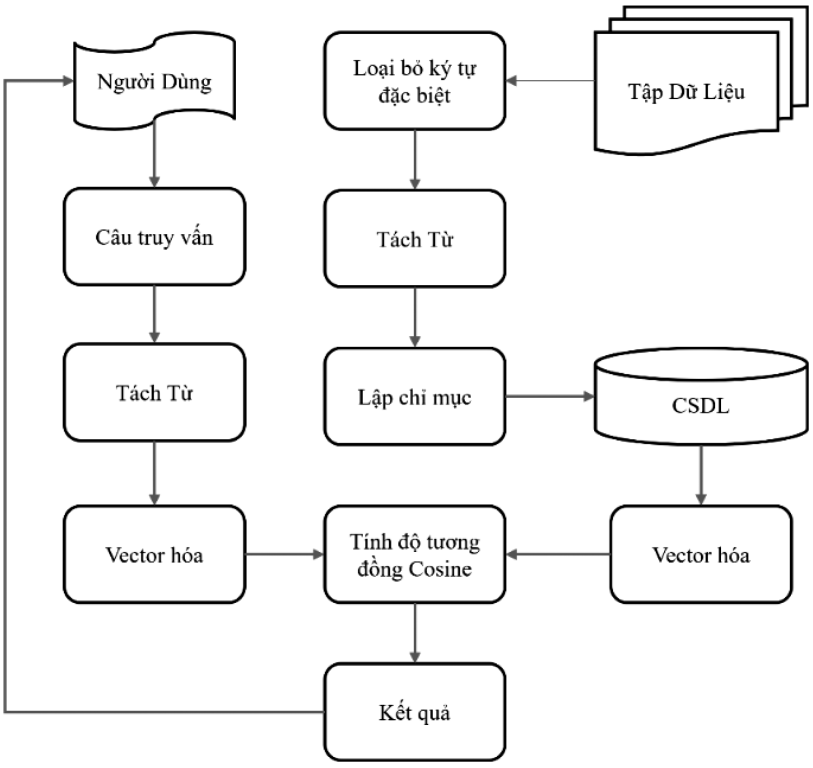
2.2.2. Bài toán tìm kiếm tài nguyên, chủ đề thảo luận

a) Nội dung bài toán

Tìm kiếm tài nguyên học tập và chủ đề thảo luận là một bài toán cơ bản nhưng phức tạp trong hệ thống chia sẻ tài nguyên. Người dùng nhập câu truy vấn để tìm kiếm thông tin hoặc tài liệu phù hợp. Hệ thống cần xác định các tài liệu liên quan dựa trên ngữ nghĩa của truy vấn, nhằm cung cấp các kết quả chính xác và hữu ích nhất.

b) Giải pháp thực hiện

- Quy trình tìm kiếm trong hệ thống: bắt đầu từ câu truy vấn của người dùng, qua quá trình xử lý và so sánh với các tài liệu trong cơ sở dữ liệu để tìm ra các kết quả phù hợp nhất. Đầu tiên, câu truy vấn sẽ được tách từ sử dụng công cụ VnCoreNLP ([10]) nhằm loại bỏ các ký tự đặc biệt và chuẩn hóa các từ khóa. Sau đó, tập dữ liệu cũng được chuẩn bị bằng cách tách từ và lập chỉ mục. Các tài liệu sau khi được lập chỉ mục sẽ lưu trữ trong cơ sở dữ liệu dưới dạng chỉ mục nghịch đảo được lưu trong MongoDB để tối ưu hóa việc truy xuất thông tin. Tiếp theo, cả câu truy vấn và các tài liệu đều được vector hóa, chuyển đổi thành các đại diện vector, giúp định lượng và so sánh chúng thông qua độ đo Cosine Similarity. Công cụ này giúp xác định mức độ tương đồng về ngữ nghĩa giữa câu truy vấn và các tài liệu trong cơ sở dữ liệu (Hình 3).



Hình 3. Mô hình chức năng tìm kiếm

PhoBERT ([11], [12]) là một mô hình học sâu dựa trên Transformer, được huấn luyện đặc biệt cho ngữ liệu tiếng Việt. Trong bài toán mở rộng truy vấn, PhoBERT giúp bổ sung các từ hoặc cụm từ liên quan đến từ khóa gốc, cải thiện khả năng tìm kiếm tài liệu phù hợp ngay cả khi từ khóa và tài liệu không hoàn toàn trùng khớp.

- Quy trình Giải thuật Mở rộng Truy vấn:

Bước 1. Nhập đầu vào:

+ Truy vấn *Q*: Một cụm từ khóa.

+ Mô hình PhoBERT: Đã được huấn luyện trên dữ liệu tiếng Việt.

+ Ngưỡng tương đồng: Để xác định mức độ liên quan của các từ mở rộng.

Bước 2: Tiền xử lý truy vấn

+ Tách truy vấn *Q* thành các từ hoặc cụm từ sử dụng công cụ VnCoreNLP.

Bước 3: Sử dụng PhoBERT để mở rộng truy vấn

+ Với mỗi từ hoặc cụm từ trong *Q*, sử dụng PhoBERT để tìm các từ liên quan (từ đồng xuất hiện cùng ngữ cảnh).

Bước 4: Hiển thị ra hệ thống tất cả các từ mở rộng truy vấn để người dùng lựa chọn.

Chức năng mở rộng truy vấn giúp hệ thống tìm thấy các tài liệu liên quan nhưng sử dụng ngôn ngữ khác so với từ khóa ban đầu, từ đó làm giảm nguy cơ bỏ sót tài liệu liên quan.

3. KẾT QUẢ NGHIÊN CỨU

3.1. Giao diện hệ thống

Hệ thống chia sẻ tài nguyên học tập đã được xây dựng thành công với nhiều chức năng hữu ích đáp ứng nhu cầu của người dùng (Hình 4). Hệ thống không chỉ cho phép người dùng đăng tải và chia sẻ các tài liệu học tập mà còn hỗ trợ các chức năng kiểm duyệt nội dung và kiểm tra sự trùng lặp, đảm bảo tính duy nhất và chất lượng của các tài liệu được chia sẻ.

Hình 4. Giao diện chính của hệ thống chia sẻ tài nguyên học tập

3.2. Các kịch bản kiểm thử

3.2.1. Kịch bản kiểm thử cho bài toán Quản lý Nội dung

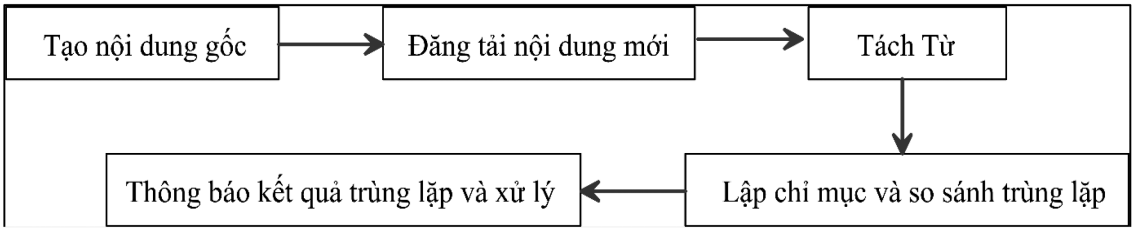
a) Mục tiêu kiểm thử

Mục tiêu của kịch bản kiểm thử này là đánh giá khả năng của hệ thống trong việc quản lý và kiểm duyệt nội dung do người dùng đăng tải, đặc biệt là khả năng phát hiện các nội dung trùng lặp với tài liệu có sẵn. Quá trình kiểm thử sẽ đánh giá độ chính xác và hiệu quả của hệ thống trong việc phát hiện và xử lý các nội dung trùng lặp, từ đó giúp giảm thiểu việc đăng tải các nội dung không phù hợp hoặc lặp lại.

Để đánh giá ưu điểm của độ đo Jaccard cải tiến so với Jaccard thông thường, các quy trình kiểm thử đều được tiến hành dựa trên cả hai phương pháp và so sánh hiệu quả của nhau.

b) Nội dung kiểm thử

Quy trình kiểm thử nhằm đánh giá hiệu quả của hệ thống trong việc phát hiện và quản lý các nội dung trùng lặp bao gồm các bước sau (Hình 5): (1) Tạo nội dung gốc, 2) Đăng tải nội dung mới, (3) Phân tích và tách từ, (4) Lập chỉ mục và so sánh trùng lặp dựa trên Jaccard cải tiến và Jaccard thông thường, và (5) Thông báo kết quả trùng lặp và xử lý nội dung.



Hình 5. Quy trình chi tiết kịch bản kiểm thử bài toán quản lý nội dung

Dựa trên báo cáo trùng lặp, quản trị viên có thể quyết định chấp nhận hoặc từ chối để đảm bảo chất lượng và tính độc quyền của nội dung trên hệ thống.

Chi tiết của các nội dung kiểm thử được trình bày trong Bảng 1.

Bảng 1. Các kịch bản kiểm thử cho bài toán quản lý nội dung

STT	Số hiệu kịch bản	Nội dung kịch bản	Số lượng mẫu thử	Ký hiệu mẫu thử
1	KBBT1-01	Đăng bài viết hoàn toàn mới về nội dung. Duyệt và đăng thành công.	40	KBBT1-01(01) → KBBT1-01(40)
2	KBBT1-02	Kịch bản đăng bài viết giống hoàn toàn với bài trong CSDL	40	KBBT1-02(01) → KBBT1-02(40)
3	KBBT1-03	Kịch bản văn bản mới có nội dung giống hoàn toàn một phần nội dung của văn bản gốc (*giống nhau từ 60% trở lên*)	40	KBBT1-03(01) → KBBT1-03(40)
3	KBBT1-04	Kịch bản văn bản mới có nội dung giống hoàn toàn một phần nội dung của văn bản gốc (*giống nhau từ 30% - 60%*)	40	KBBT1-04(01) → KBBT1-04(05)

c) Kết quả

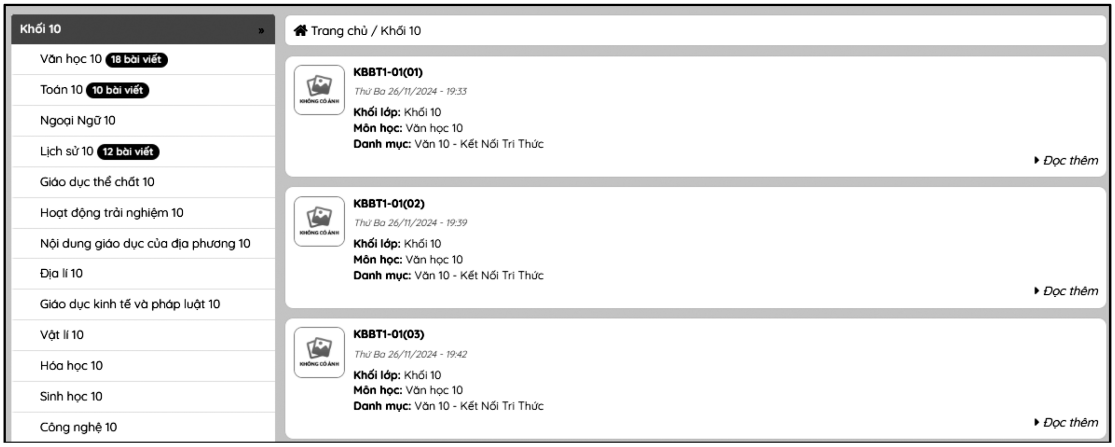
Nội dung một số mẫu thử KBBT1-01 được trình bày trong Bảng 2.

Bảng 2. Nội dung một số mẫu thử thuộc KBBT1-01

Ký hiệu mẫu thử	Nội dung mẫu thử
KBBT1-01(01)	Trong thời kỳ hiện nay, sự tiến bộ của công nghệ trong lĩnh vực giáo dục mang lại các phương thức thuận tiện hơn thông qua sự đa dạng trong việc hỗ trợ quá trình học tập, chia sẻ tập tin, giao bài tập và đánh giá.

Ký hiệu mẫu thử	Nội dung mẫu thử
KBBT1-01(02)	Vì thế, việc ứng dụng kỹ thuật tìm kiếm thông tin cho hệ thống chia sẻ tài nguyên học tập số là rất cần thiết cho một trường THPT trong bối cảnh đổi mới toàn diện của nền giáo dục hiện đại. Hệ thống này được thiết kế dành riêng cho đối tượng người dùng là giáo viên và học sinh bậc THPT.
KBBT1-01(03)	Bên cạnh đó, việc ứng dụng kỹ thuật tìm kiếm giúp người sử dụng hệ thống tìm kiếm tài liệu đơn giản, nhanh chóng, chuẩn xác. Điều này phù hợp với kết quả khảo sát về mức độ cần thiết của một hệ thống chia sẻ tài nguyên học tập đã được thực hiện tại trường THPT Hòa Ninh.

Kết quả của các mẫu thử KBBT1-01(01) đến KBBT1-01(03) được thể hiện trong Hình 6. Trong đó, bài viết được người dùng soạn thảo đăng tải thành công và được quản trị viên duyệt đăng.

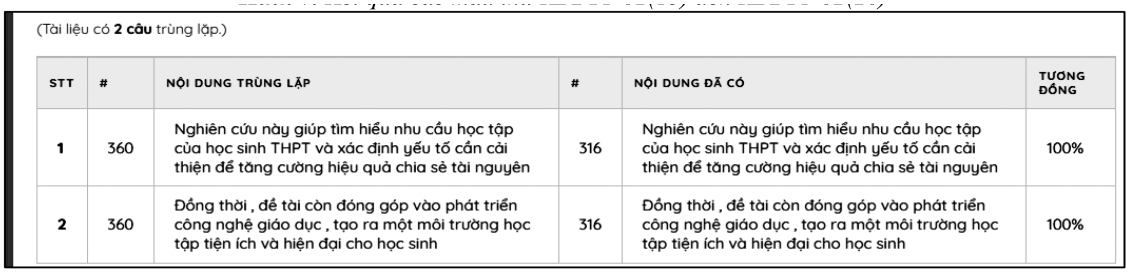


Hình 6. Kết quả một số mẫu thử thuộc kịch bản KB-BT1-01

Trong kịch bản KBBT1-02, 40 bài viết có nội dung giống hoàn toàn với 40 bài viết thuộc các mẫu trong KBBT1-01 đã được đăng tải. Hệ thống xác định mức độ trùng lặp là 100% và gửi thông báo chi tiết cho quản trị viên về sự trùng lặp hoàn toàn này, như minh họa trong (Hình 7). Ví dụ, kết quả kiểm tra trùng lặp 100% của mẫu thử KBBT1-02(05) được hiển thị trong (Hình 8), trong đó hệ thống phát hiện 2 câu trùng khớp hoàn toàn với bài viết đã có trên hệ thống.

MÃ	TRẠNG THÁI	TIÊU ĐỀ	NGÀY ĐĂNG	TÁC GIẢ	TỶ LỆ TRÙNG LẶP (JACCARD)	TỶ LỆ TRÙNG LẶP (JACCARD CẢI TIẾN)
381	CHỜ DUYỆT	KBBT1-02(20)	26-11-2024 (22:03:00)	Nguyễn Thị Hồng Cẩm	100% (KBBT1-01(20)) Xem chi tiết	100% (KBBT1-01(20))Xem chi tiết
380	CHỜ DUYỆT	KBBT1-02(19)	26-11-2024 (22:02:31)	Nguyễn Thị Hồng Cẩm	100% (KBBT1-01(19)) Xem chi tiết	100% (KBBT1-01(19))Xem chi tiết 32.26% (KBBT1-01(27))Xem chi tiết
379	CHỜ DUYỆT	KBBT1-02(18)	26-11-2024 (22:01:57)	Nguyễn Thị Hồng Cẩm	100% (KBBT1-01(18)) Xem chi tiết	100% (KBBT1-01(18))Xem chi tiết
377	CHỜ DUYỆT	KBBT1-02(17)	26-11-2024 (21:59:58)	Nguyễn Thị Hồng Cẩm	100% (KBBT1-01(17)) Xem chi tiết	23.81% (KBBT1-01(11))Xem chi tiết 100% (KBBT1-01(17))Xem chi tiết
376	CHỜ DUYỆT	KBBT1-02(16)	26-11-2024 (21:59:02)	Nguyễn Thị Hồng Cẩm	33% (KBBT1-01(12)) Xem chi tiết 100% (KBBT1-01(16)) Xem chi tiết	100% (KBBT1-01(16))Xem chi tiết
375	CHỜ DUYỆT	KBBT1-02(15)	26-11-2024 (21:58:35)	Nguyễn Thị Hồng Cẩm	100% (KBBT1-01(15)) Xem chi tiết	100% (KBBT1-01(15))Xem chi tiết

Hình 7. Kết quả các mẫu thử KBBT1-02(15) đến KBBT1-02(20)



(Tài liệu có **2 câu** trùng lặp.)

STT	#	NỘI DUNG TRÙNG LẶP	#	NỘI DUNG ĐÃ CÓ	TƯƠNG ĐỒNG
1	360	Nghiên cứu này giúp tìm hiểu nhu cầu học tập của học sinh THPT và xác định yếu tố cần cải thiện để tăng cường hiệu quả chia sẻ tài nguyên	316	Nghiên cứu này giúp tìm hiểu nhu cầu học tập của học sinh THPT và xác định yếu tố cần cải thiện để tăng cường hiệu quả chia sẻ tài nguyên	100%
2	360	Đồng thời , đề tài còn đóng góp vào phát triển công nghệ giáo dục , tạo ra một môi trường học tập tiện ích và hiện đại cho học sinh	316	Đồng thời , đề tài còn đóng góp vào phát triển công nghệ giáo dục , tạo ra một môi trường học tập tiện ích và hiện đại cho học sinh	100%

Hình 8. Hiển thị chi tiết kết quả trùng lặp của các câu trong mẫu thử KBBT1-02(05)

Trong kịch bản KBBT1-03, nội dung các mẫu thử được xây dựng bằng cách thay thế ngẫu nhiên 10 đến 30% số từ trong từng mẫu thuộc KBBT1-01. Ví dụ cho mẫu thử KBBT1-03(01) đươc trình bày trong Bảng 3.

Bảng 3. Nội dung mẫu thử KBBT1-03(01) thuộc KBBT1-03

Ký hiệu mẫu thử	Nội dung văn bản gốc	Nội dung mẫu thử	Độ tương đồng (Jaccard)	Độ tương đồng (Jaccard cải tiến)
KBBT1-03(01)	Trong thời kỳ hiện nay, sự tiến bộ của công nghệ trong lĩnh vực giáo dục mang lại các phương thức thuận tiện hơn thông qua sự đa dạng trong việc hỗ trợ quá trình học tập, chia sẻ tập tin, giao bài tập và đánh giá.	Trong thời đại ngày nay, sự phát triển của công nghệ trong ngành giáo dục mang lại những phương pháp tiện lợi hơn thông qua sự phong phú trong việc hỗ trợ quá trình học tập, chia sẻ tài liệu, giao bài tập và đánh giá.	63%	62.6%

Sự khác biệt giữa hai văn bản thuộc mẫu thử KBBT1-01(01) và KBBT1-03(01) bao gồm việc thay thế các cụm từ như "thời kỳ hiện nay" thành "thời đại ngày nay", "sự tiến bộ" thành "sự phát triển", và "tập tin" thành "tài liệu" .... Mặc dù ngữ nghĩa tổng thể không thay đổi đáng kể, sự thay thế này làm giảm độ tương đồng xuống còn 63% (độ đo Jaccard) và 62.6% (độ đo Jaccard cải tiến).

Độ tương đồng giữa các mẫu thử thuộc KBBT1-03 và các mẫu gốc dao động từ 63% đến 94%, tùy thuộc vào mức độ thay đổi nội dung. Kết quả một số mẫu thử khác thuộc kịch bản KBBT1-03 được trình bày trong Hình 9.

MÃ	TRẠNG THÁI	TIÊU ĐỀ	NGÀY ĐĂNG	TÁC GIẢ	TỶ LỆ TRÙNG LẶP (JACCARD)	TỶ LỆ TRÙNG LẶP (JACCARD CẢI TIẾN)
448	CHỜ DUYỆT	**KBBT1-03(40)**	26-11-2024 (23:49:01)	Nguyễn Thị Hồng Cẩm	**31%** (KBBT1-01(39)) Xem chi tiết **66%** (KBBT1-01(40)) Xem chi tiết	**54.55%** (KBBT1-01(40))Xem chi tiết
447	CHỜ DUYỆT	**KBBT1-03(39)**	26-11-2024 (23:47:52)	Nguyễn Thị Hồng Cẩm	**33%** (KBBT1-01(32)) Xem chi tiết **74%** (KBBT1-01(39)) Xem chi tiết **33%** (KBBT1-01(40)) Xem chi tiết	**73.08%** (KBBT1-01(39))Xem chi tiết
446	CHỜ DUYỆT	**KBBT1-03(38)**	26-11-2024 (23:46:58)	Nguyễn Thị Hồng Cẩm	**86%** (KBBT1-01(38)) Xem chi tiết	**78.13%** (KBBT1-01(38))Xem chi tiết
445	CHỜ DUYỆT	**KBBT1-03(37)**	26-11-2024 (23:46:04)	Nguyễn Thị Hồng Cẩm	**65%** (KBBT1-01(37)) Xem chi tiết	**64%** (KBBT1-01(37))Xem chi tiết

Hình 9. Kết quả các mẫu thử KBBT1-03(37) đến KBBT1-03(40)

Trong kịch bản KBBT1-04, nội dung các mẫu thử được tạo ra bằng cách thay thế 30 đến 60% số từ trong từng mẫu thuộc KBBT1-01. Hệ thống sử dụng độ đo tương đồng Jaccard cải tiến đã đánh giá mức độ trùng lặp của các mẫu thử trong kịch bản KBBT1-04. Độ tương đồng giữa các mẫu thử và các mẫu gốc dao động từ 30% đến 58%, tùy thuộc vào mức độ thay đổi nội dung của từng mẫu thử. Kết quả chi tiết của một số mẫu thử khác được thể hiện trong Bảng 4 và minh họa trong Hình 10.

Bảng 4. Nội dung mẫu thử KBBT1-04(01) và KBBT1-04(02) thuộc KBBT1-04

Ký hiệu mẫu thử	Nội dung văn bản gốc	Nội dung mẫu thử	Độ tương đồng (Jaccard)	Độ tương đồng (Jaccard cải tiến)
KBBT1-04(01)	Trong thời kỳ hiện nay, sự tiến bộ của công nghệ trong lĩnh vực giáo dục mang lại các phương thức thuận tiện hơn thông qua sự đa dạng trong việc hỗ trợ quá trình học tập, chia sẻ tập tin, giao bài tập và đánh giá.	Hiện nay, những tiến bộ trong công nghệ giáo dục mang đến nhiều cách thức tiện lợi hơn, bao gồm sự phong phú trong hỗ trợ học tập, trao đổi tài liệu, nộp bài tập và kiểm tra đánh giá.	41%	15%

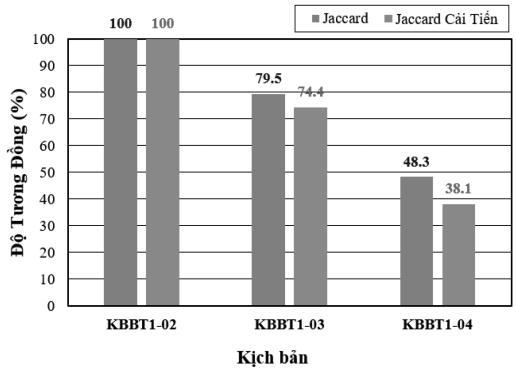
MÃ	TRẠNG THÁI	TIÊU ĐỀ	NGÀY ĐĂNG	TÁC GIẢ	TỶ LỆ TRÙNG LẶP (JACCARD)	TỶ LỆ TRÙNG LẶP (JACCARD CẢI TIẾN)
506	CHỜ DUYỆT	KBBT1-04(40)	27-11-2024 (12:03:16)	Nguyễn Thị Hồng Cẩm	55% (KBBT1-01(40)) Xem chi tiết	13.04% (KBBT1-01(30))Xem chi tiết 12% (KBBT1-01(38))Xem chi tiết 10.64% (KBBT1-01(39))Xem chi tiết 42.11% (KBBT1-01(40))Xem chi tiết
505	CHỜ DUYỆT	KBBT1-04(39)	27-11-2024 (12:02:06)	Nguyễn Thị Hồng Cẩm	43% (KBBT1-01(39)) Xem chi tiết	11.76% (KBBT1-01(32))Xem chi tiết 10.42% (KBBT1-01(33))Xem chi tiết 45.71% (KBBT1-01(39))Xem chi tiết 12.77% (KBBT1-01(40))Xem chi tiết
504	CHỜ DUYỆT	KBBT1-04(38)	27-11-2024 (11:48:46)	Nguyễn Thị Hồng Cẩm	46% (KBBT1-01(38)) Xem chi tiết	14.58% (KBBT1-01(30))Xem chi tiết 35.71% (KBBT1-01(38))Xem chi tiết 14.89% (KBBT1-01(39))Xem chi tiết
502	CHỜ DUYỆT	KBBT1-04(37)	27-11-2024 (11:46:48)	Nguyễn Thị Hồng Cẩm	37% (KBBT1-01(37)) Xem chi tiết	14.71% (KBBT1-01(37))Xem chi tiết
501	CHỜ DUYỆT	KBBT1-04(36)	27-11-2024 (11:45:26)	Nguyễn Thị Hồng Cẩm	51% (KBBT1-01(36)) Xem chi tiết	13.11% (Văn hóa là gì?)Xem chi tiết 12.24% (KBBT1-01(35))Xem chi tiết 47.62% (KBBT1-01(36))Xem chi tiết

Hình 10. Kết quả một số mẫu thử KBBT1-04(36) đến KBBT1-04(40) thuộc kịch bản KBBT1-04

Hình 11 trình bày kết quả so sánh độ tương đồng trung bình của các mẫu thử trong các kịch bản KBBT1-02, KBBT1-03, và KBBT1-04 theo hai phương pháp: Jaccard thông thường và Jaccard cải tiến. Đối với kịch bản KBBT1-02, cả hai phương pháp đều cho kết quả tương đồng tuyệt đối (100%), cho thấy rằng trong trường hợp các văn bản giống nhau hoàn toàn, việc xét đến vị trí từ trong Jaccard cải tiến không tạo ra sự khác biệt. Tuy nhiên, trong kịch bản KBBT1-03, phương pháp Jaccard cải tiến đạt độ tương đồng thấp hơn (74.4%) so với Jaccard thông thường (79.5%). Sự khác biệt này có thể được giải thích bởi tiêu chí xét đến khoảng cách vị trí từ trong Jaccard cải tiến, làm giảm số lượng từ thuộc tập giao khi các vị trí chênh lệch vượt quá giá trị ngưỡng.

Đáng chú ý nhất là ở kịch bản KBBT1-04, phương pháp Jaccard cải tiến chỉ đạt 38.1%, thấp hơn đáng kể so với Jaccard thông thường là 48.3%. Điều này chứng minh rằng trong trường hợp các từ trong văn bản có sự thay đổi đáng kể về vị trí, Jaccard cải tiến có xu hướng phản ánh mức độ tương đồng chặt chẽ hơn dựa trên vị trí của từ.

Kết quả phân tích cho thấy phương pháp Jaccard cải tiến phù hợp với các bài toán yêu cầu đánh giá chính xác dựa trên vị trí từ, trong khi Jaccard thông thường lại thích hợp hơn khi xét tương đồng tổng quát mà không cần xét vị trí.



Hình 11. So sánh giá trị trung bình độ tương đồng (%) của các mẫu thử thuộc các kịch bản KBBT1-02, KBBT1-03, và KBBT1-04 theo hai phương pháp Jaccard và Jaccard cải tiến

Nhìn chung, kết quả kiểm thử các kịch bản quản lý nội dung (KBBT1-01, KBBT1-02, KBBT1-03, KBBT1-04) đã chứng minh hệ thống hoạt động chính xác, ổn định và đáp ứng các yêu cầu đề ra.

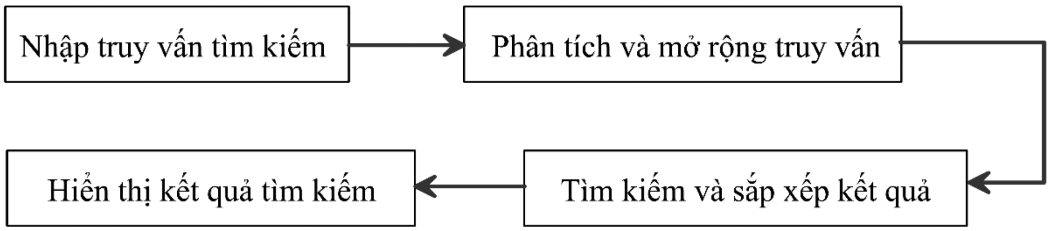
3.2.2. Kịch bản kiểm thử cho bài toán Tìm kiếm tài nguyên

a) Mục tiêu kiểm thử

Mục tiêu của kịch bản kiểm thử này là đánh giá tính chính xác và hiệu quả của hệ thống trong việc tìm kiếm tài nguyên dựa trên truy vấn của người dùng.

b) Nội dung kiểm thử

Hình 12 mô tả chi tiết nội dung kiểm thử đối với bài toán tìm kiếm tài nguyên, với mục tiêu kiểm tra tính chính xác và hiệu quả của hệ thống trong việc truy xuất thông tin dựa trên truy vấn của người dùng.



Hình 12. Quy trình chi tiết kịch bản kiểm thử bài toán tìm kiếm tài nguyên

Chi tiết của các nội dung kiểm thử được trình bày trong Bảng 5.

Bảng 5. Các kịch bản kiểm thử cho bài toán tìm kiếm tài nguyên

STT	Số hiệu kịch bản	Nội dung kịch bản	Số lượng mẫu thử	Ký hiệu mẫu thử
1	KBBT2-01	Tìm kiếm từ khóa đơn giản (từ đơn hoặc từ ghép)	40	KBBT2-01(01) → KBBT2-01(4)
2	KBBT2-02	Tìm kiếm từ khóa là cụm từ	10	KBBT1-02(01) → KBBT1-02(02)

Để đánh giá ưu điểm của phương pháp tìm kiếm trong hệ thống, nghiên cứu đã so sánh hiệu quả giữa hai phương pháp: (1) tìm kiếm truyền thống bằng hàm LIKE, và (2) tìm kiếm nâng cao kết hợp Cosine với PhoBERT. Sự so sánh này giúp làm nổi bật ưu điểm vượt trội của các phương pháp hiện đại trong việc cải thiện độ chính xác, độ bao phủ, và khả năng xử lý các truy vấn ngữ nghĩa phức tạp.

c) Kết quả

Kịch bản KBBT2-01 (40 mẫu thử từ KBBT2-01(01) đến KBBT2-01(40) được thiết lập để kiểm tra chức năng tìm kiếm của hệ thống khi người dùng sử dụng các từ khóa đơn giản. Số lượng bài viết tìm được dao động từ 0 đến 10, phụ thuộc vào mức độ phổ biến và nội dung của từ khóa. Các kết quả chi tiết một số mẫu thử được trình bày trong Bảng 6. Đối với từ khóa đơn giản, cả hai phương pháp tìm kiếm truyền thống và tìm kiếm nâng cao đều cho ra cùng số lượng bài viết.

Bảng 6. Kết quả thực hiện một số mẫu thử thuộc kịch bản KBBT2-01 – tìm kiếm từ khóa đơn giản

Số hiệu mẫu thử	Nội dung tìm kiếm	Số bài viết	Số hiệu mẫu thử	Nội dung tìm kiếm	Số bài viết
KBBT2-01(14)	Hà Nội	2	KBBT2-01(20)	Đất nước	3
KBBT2-01(15)	Phục Hưng	2	KBBT2-01(21)	Cổ đại	3
KBBT2-01(16)	Hồ Chí Minh	2	KBBT2-01(22)	Học sinh	3
KBBT2-01(17)	Nguyễn Tuân	2	KBBT2-01(23)	Việt Nam	3
KBBT2-01(18)	Thần	3	KBBT2-01(24)	Máy tính	4
KBBT2-01(19)	Nhà Nước	3	KBBT2-01(25)	Giáo dục	4

Minh họa chi tiết cho truy vấn "Việt Nam" được trình bày trong Hình 13, nơi giao diện diễn đàn hiển thị danh sách các tập tin chứa từ khóa, và kết quả được hiển thị trong phần mềm MongoDB.

Hình 13. Kết quả được hiển thị trong phần mềm MongoDB

(a) Kết quả trả về khi tìm kiếm từ khóa "Việt Nam",

(b). Từ khóa "Việt Nam" được lưu trong CSDL, và

(c) Kết quả idtimkiem được hiển thị trong MongoDB

Kịch bản KBBT2-02 được thiết lập đã kiểm tra khả năng của hệ thống trong việc xử lý các từ khóa phức tạp và trả về kết quả tìm kiếm chính xác. Từ khóa phức tạp trong kịch bản bao gồm các cụm từ ghép hoặc từ khóa và yêu cầu mở rộng truy vấn để tìm kiếm các tài liệu liên quan (Bảng 7).

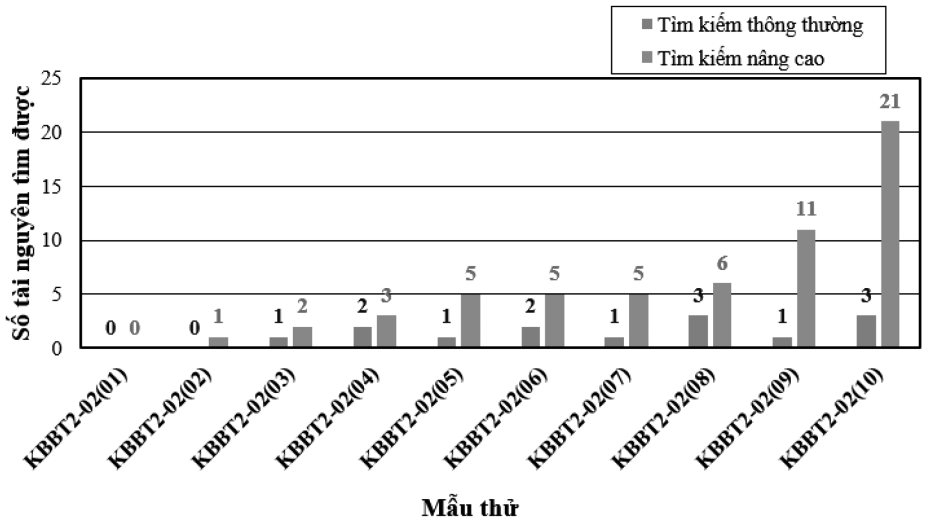
Bảng 7. Kết quả kiểm thử cho bài toán tìm kiếm tài nguyên, các mẫu thử thuộc nhóm KBBT2-02 - Tìm kiếm từ khóa phức tạp

Số hiệu mẫu thử	Nội dung tìm kiếm	Tìm kiếm thông thường	Tìm kiếm nâng cao		
			Tách từ	Từ đồng nghĩa	Số bài viết
KBBT2-02(01)	Tin học chuyên ngành	0	"tin học", "chuyên ngành"	"thông tin chuyên ngành", "tin học ."	0
KBBT2-02(02)	Phương pháp cải tiến	0	"phương pháp", "cải tiến"	"phụ nữ cải tiến", "phương pháp ."	1
KBBT2-02(03)	Mở rộng truy vấn	1	"mở rộng", "truy vấn"	"câu truy vấn", "mở rộng ."	2
KBBT2-02(04)	Quản lý nội dung	2	"quản lý", "nội dung"	"thể thao nội dung", "quản lý ."	3
KBBT2-02(05)	Văn minh phương Tây	1	"văn minh", "phương Tây"	"văn hóa phương tây", "văn minh phương Đông"	5
KBBT2-02(06)	Kết nối vạn vật	2	"kết nối", "vạn vật"	"con người vạn vật", "kết nối ."	5
KBBT2-02(07)	Chiến tranh thế giới	1	"chiến tranh", "thế giới"	"chính trị thế giới", "chiến tranh ."	5
KBBT2-02(08)	thiết bị thông minh	3	"thiết bị", "thông minh"	"Robot thông minh" "thiết bị."	6
KBBT2-02(09)	Kỹ thuật tìm kiếm thông tin	1	"kỹ thuật", "tìm kiếm", "thông tin"	"Người tìm kiếm thông tin", "Kỹ thuật an toàn thông tin", "Kỹ thuật tìm kiếm ."	11
KBBT2-02(10)	Lịch sử Việt Nam	3	"lịch sử", "Việt Nam"	"lịch sử ."	21

Ví dụ cho mẫu thử KBBT2-02(09), từ khóa tìm kiếm là "kỹ thuật tìm kiếm thông tin". Hệ thống tiến hành phân tách cụm từ thành các từ khóa nhỏ hơn, bao gồm "kỹ thuật", "tìm kiếm", và "thông tin". Sau đó, chức năng mở rộng truy vấn (PhoBERT)

được áp dụng để gợi ý thêm một số cụm từ liên quan như "người tìm kiếm thông tin", "kỹ thuật an toàn thông tin", và "kỹ thuật tìm kiếm .". Hệ thống tiếp tục thực hiện tìm kiếm trên tất cả các từ khóa này và trả về 11 tài liệu có nội dung chứa các từ khóa liên quan.

Hình 14 trình bày sự khác biệt rõ rệt giữa kết quả tìm kiếm thông thường và tìm kiếm nâng cao (bao gồm tách từ và mở rộng truy vấn) qua các mẫu thử của kịch bản KBBT2-02. Từ đó, các ưu điểm của phương pháp tìm kiếm nâng cao được thể hiện một cách rõ rệt.



Hình 14. So sánh số lượng bài viết hệ thống trả về khi thực hiện tìm kiếm theo phương pháp thông thường và phương pháp nâng cao

Cụ thể, ở tất cả các mẫu thử, số lượng tài nguyên tìm kiếm được bằng phương pháp nâng cao luôn vượt trội hơn so với tìm kiếm thông thường. Đặc biệt, các mẫu thử KBBT2-02(09) và KBBT2-02(10) (các từ khóa phức tạp như Kỹ thuật tìm kiếm thông tin và Lịch sử Việt Nam) cho thấy mức tăng trưởng ấn tượng, lần lượt từ 1 bài viết trong tìm kiếm thông thường lên đến 11 và 21 bài viết trong tìm kiếm nâng cao. Điều này chứng minh khả năng của hệ thống trong việc xử lý hiệu quả các cụm từ phức tạp và mở rộng ý nghĩa tìm kiếm. Các cải tiến, như tách từ và mở rộng từ đồng xuất hiện, giúp hệ thống nhận diện được các từ khóa liên quan trong cơ sở dữ liệu, từ đó tăng cường phạm vi tìm kiếm. Tìm kiếm nâng cao không chỉ gia tăng số lượng tài nguyên mà còn mở rộng phạm vi bao phủ của thông tin tìm kiếm. Việc hệ thống nhận diện được các tài nguyên có liên quan thông qua các từ đồng xuất hiện cho thấy tính linh hoạt và thông minh của thuật toán.

Các kịch bản kiểm thử KBBT2-01 (40 mẫu thử) đến KBBT2-02 (10 mẫu thử) đã được triển khai nhằm đánh giá khả năng hoạt động của chức năng tìm kiếm tài nguyên và chủ đề thảo luận. Qua từng kịch bản, cho thấy hệ thống có khả năng tìm kiếm từ khóa và lọc tài liệu hiệu quả. Kết quả so sánh cho thấy rõ ràng rằng tìm kiếm nâng cao với sự kết hợp của tách từ và mở rộng truy vấn mang lại hiệu quả vượt trội so với tìm kiếm thông thường. Hệ thống không chỉ cải thiện số lượng kết quả tìm kiếm mà còn giúp người dùng tiếp cận thông tin một cách toàn diện và chính xác hơn. Điều này đặc biệt hữu ích khi xử lý các từ khóa phức tạp hoặc cụm từ ghép.

4. KẾT LUẬN

Bài báo này đã trình bày quá trình thiết kế, phát triển và kiểm thử một hệ thống chia sẻ tài nguyên học tập trực tuyến dành cho học sinh và giáo viên, với trọng tâm là hai bài toán chính: quản lý nội dung và tìm kiếm tài nguyên học tập. Hệ thống được xây dựng trên nền tảng client-server, ứng dụng các kỹ thuật tiên tiến như chỉ mục nghịch đảo MongoDB, trọng số TF-IDF, độ đo tương đồng Jaccard cải tiến, độ đo tương đồng Cosine, và mô hình PhoBERT nhằm nâng cao hiệu quả quản lý và tìm kiếm tài nguyên.

Các mẫu kiểm thử (160 mẫu) cho bài toán quản lý nội dung đã minh chứng hiệu quả cao trong việc phát hiện trùng lặp nội dung. Kết quả so sánh trên các kịch bản thử nghiệm cho thấy phương pháp Jaccard cải tiến đạt độ trùng lặp thấp hơn so với Jaccard thông thường trong các trường hợp phức tạp. Đặc biệt, ở các kịch bản có sự xuất hiện của nhiều từ khóa không đồng nhất, Jaccard cải tiến đã chứng minh tính hiệu quả nhờ khả năng xử lý linh hoạt và đánh giá tốt mức độ tương đồng giữa các tài liệu liên quan. Điều này khẳng định vai trò của các cải tiến thuật toán trong việc tối ưu hóa hiệu suất so sánh và tìm kiếm.

Qua 50 mẫu thử tìm kiếm tài liệu và thảo luận, hệ thống đã thể hiện khả năng trả về kết quả phù hợp, đáp ứng nhu cầu người dùng. Tìm kiếm nâng cao, nhờ tích hợp các kỹ thuật như mở rộng truy vấn dựa trên PhoBERT, đã mở rộng đáng kể phạm vi tìm kiếm, mang lại nhiều lựa chọn hơn cho người dùng. Tuy nhiên, một số hạn chế vẫn tồn tại, đặc biệt về độ chính xác của từ khóa gợi ý.

Với các tính năng trên, hệ thống không chỉ cung cấp giải pháp hiệu quả cho các vấn đề về quản lý và tìm kiếm tài nguyên học tập mà còn tạo nền tảng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực giáo dục trực tuyến. Trong tương lai, hệ thống có thể được mở rộng để tích hợp thêm các công nghệ xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo, nhằm tối ưu hóa trải nghiệm người dùng và đáp ứng tốt hơn nhu cầu ngày càng tăng của học sinh và giáo viên.

TÀI LIỆU THAM KHẢO

[1] Q. D. Truong, T. Dkaki, J. Mothe, and P.-J. Charrel, "GVC: a graph-based Information Retrieval Mode.," in *CORIA*, 2008, pp. 337–351. Accessed: Nov. 21, 2023. [Online]. Available: https://asso-aria.org/coria/2008/337.pdf

[2] D. Gunawan, C. A. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," in *Journal of physics: conference series*, IOP Publishing, 2018, p. 012120. Accessed: Nov. 21, 2023. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/978/1/012120/meta

[3] T. H. Y. Trần, "Ứng dụng các kỹ thuật tìm kiếm thông tin vào hệ thống tìm kiếm ảnh dựa trên nội dung," PhD Thesis, Trường Đại học Bách khoa Hà Nội, 2013. Accessed: Nov. 21, 2023. [Online]. Available: https://dlib.hust.edu.vn/handle/HUST/17102

[4] Trần T. T., Trần T. N. T., and Trương Q. Đ., "Ứng dụng các kỹ thuật tìm kiếm thông tin cho bài toán kiểm tra sao chép luận văn," in *Hội thảo toàn quốc về công nghệ thông tin 2017, Cần Thơ, Việt Nam*, 2017.

[5] T. L. Vũ, T. H. Nguyễn, and T. T. H. Trần, "Xây dựng ứng dụng web để chia sẻ tài liệu học tập cho sinh viên ngành Công nghệ thông tin - Học viện nông nghiệp Việt Nam," *Tạp Chí Khoa Học Nông Nghiệp Việt Nam*, vol. 19, no. 4, pp. 507–519, 2020.

[6] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, 2013, pp. 380–384. Accessed: Dec. 12, 2024. [Online]. Available: https://www.iaeng.org/publication/IMECS2013/IMECS2013_pp380-384.pdf

[7] M.-C. Kim and K.-S. Choi, "A comparison of collocation-based similarity measures in query expansion," *Inf. Process. Manag.*, vol. 35, no. 1, pp. 19–30, 1999.

[8] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Inf. Sci.*, vol. 483, pp. 53–64, 2019.

[9] A. A. Amer and L. Nguyen, "Combinations of Jaccard with Numerical Measures for Collaborative Filtering Enhancement: Current Work and Future Proposal," Nov. 24, 2021, *arXiv*: arXiv:2111.12202. doi: 10.48550/arXiv.2111.12202.

[10] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 56–60. doi: 10.18653/v1/N18-5012.

[11] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," Oct. 05, 2020, *arXiv*: arXiv:2003.00744. Accessed: Aug. 30, 2024. [Online]. Available: http://arxiv.org/abs/2003.00744

[12] K. Quoc Tran, A. Trong Nguyen, P. G. Hoang, C. D. Luu, T.-H. Do, and K. Van Nguyen, "Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 573–594, Jan. 2023, doi: 10.1007/s00521-022-07745-w.