(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0053740**

(51)[2022.01] **G06F 16/00**

(13) **B**

(21) 1-2023-02661

(22) 21/04/2023

(45) 25/11/2025 452

(43) 25/08/2023 425A

(73) CÔNG TY TNHH SAMSUNG SDS VIỆT NAM (VN)
Lô CN05, đường YP6, Khu công nghiệp Yên Phong, xã Yên Trung, huyện Yên Phong, tỉnh Bắc Ninh

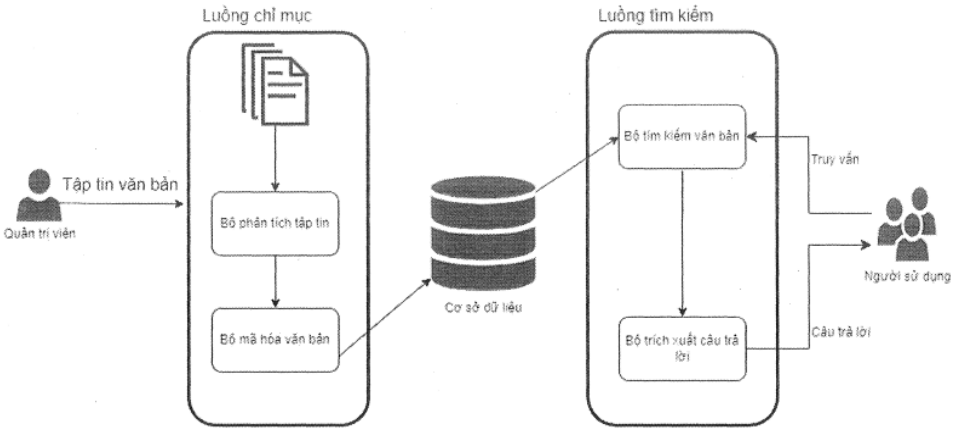
(72) TRẦN NGỌC SƠN (VN); NGUYỄN ANH TÚ (VN); HỒ ĐẮC PHÚC (VN); ĐÀM TRỌNG TUYÊN (VN); LÊ XUÂN BỘ (VN); NGUYỄN THỊ THU (VN); NGUYỄN MINH TRÍ (VN); KIM TAE HYUN (KR).

(74) Công ty TNHH NACILAW (NACILAW)

(54) HỆ THỐNG VÀ PHƯƠNG PHÁP TRUY VẤN TÀI LIỆU VÀ HỎI ĐÁP NGỮ NGHĨA TỰ ĐỘNG TIẾNG VIỆT

(21) 1-2023-02661

(57) Sáng chế đề cập đến hệ thống và phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt cho phép người dùng tìm kiếm thông tin bên trong các tài liệu với nhiều định dạng khác nhau. Hệ thống được thực hiện qua hai luồng chỉ mục và luồng tìm kiếm với các bộ: bộ phân tích tập tin, bộ mã hóa văn bản, bộ tìm kiếm văn bản, bộ trích xuất câu trả lời, bộ cơ sở dữ liệu. Phương pháp được thực hiện qua các bước: bước 1: trích xuất nội dung chữ của tập tin; bước 2: phân loại tài liệu và trích rút thông tin; bước 3: mã hóa nội dung chữ; bước 4: tìm kiếm khi có truy vấn; bước 5: trả lời truy vấn dựa vào các văn bản liên quan; bước 6: thu thập phản hồi.



Hình 2

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến hệ thống và phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt. Cụ thể, hệ thống và phương pháp được đề cập trong sáng chế có thể được áp dụng trong hệ thống trò chuyện tự động và trợ lý ảo, hệ thống tìm kiếm tài liệu, trích xuất thông tin tự động từ kho dữ liệu lớn để khai thác dữ liệu lớn và hệ thống hỏi đáp tự động trên kho tri thức của tổ chức.

Tình trạng kỹ thuật của sáng chế

Những nghiên cứu tiên phong trong việc phát triển hệ thống hỏi đáp được bắt nguồn từ bài toán tìm kiếm thông tin tập trung trong một lĩnh vực nhỏ. Một trong những hệ thống hỏi đáp đầu tiên đó là *Baseball* được thiết kế vào năm 1961 với mục đích trả lời các câu hỏi liên quan đến các trận đấu bóng rổ ở Mỹ bao gồm thời gian diễn ra, địa điểm và tên đội bóng. Trong hệ thống này, tất cả các thông tin được lưu trữ dưới dạng những từ điển được định nghĩa trước và sau đó câu hỏi của người dùng sẽ được dịch thành những câu truy vấn cơ sở dữ liệu bằng những phương thức ngôn ngữ học để có thể trích xuất thông tin từ tập từ điển. Năm 1999, hệ thống hỏi đáp lĩnh vực mở được định nghĩa là hệ thống trích chọn năm văn bản đầu tiên chứa câu trả lời chính xác. So sánh với những hệ thống hỏi đáp trước đó, hệ thống hỏi đáp lĩnh vực mở sử dụng số lượng lớn tài liệu phi cấu trúc, sau đó câu trả lời trong các tài liệu sẽ được trích xuất và trả về cho người dùng.

Kiến trúc hệ thống hỏi đáp truyền thống được thể hiện dưới Hình 1, hệ thống gồm ba phần chính: phân tích câu hỏi, tìm kiếm văn bản và trích chọn câu trả lời. Trong đó:

Phân tích câu hỏi là một bước quan trọng trong hệ thống hỏi đáp truyền thống. Mục tiêu của việc phân tích câu hỏi đó là phân loại câu hỏi của người dùng và chuyển câu hỏi dạng ngôn ngữ tự nhiên thành các câu truy vấn trong cơ sở dữ liệu. Việc phân tích câu hỏi nhằm xác định loại câu hỏi (địa điểm, người, thời gian) và phân tích cú pháp câu hỏi và xác định các yếu tố như động từ, chủ từ và tân ngữ. Sau đó, hệ thống sẽ xác định loại câu hỏi để tìm kiếm các thông tin liên quan. Trong bước này, các phương pháp học máy truyền thống như máy vectơ hỗ trợ (SVM), k láng giềng gần nhất (KNN) được sử dụng cho việc phân tích câu hỏi.

Tìm kiếm văn bản: giai đoạn này nhằm trích chọn một số lượng nhỏ các tài liệu liên quan từ một bộ sưu tập các tài liệu phi cấu trúc. Những tài liệu này chứa câu trả lời chính xác có thể trả về cho người dùng. Việc tìm kiếm và giới hạn các tài liệu liên quan giúp giảm

không gian tìm kiếm để đi đến câu trả lời cuối cùng. Trong những thập kỷ qua, các mô hình truy hồi khác nhau đã được được phát triển cho việc truy xuất tài liệu, trong đó một số mô hình phổ biến là mô hình không gian vectơ, mô hình xác suất, cụ thể:

- Mô hình không gian vectơ: mô hình không gian vectơ biểu diễn câu hỏi và từng tài liệu dưới dạng vectơ từ trong một không gian từ d chiều, trong đó d là số của các từ trong từ vựng. Khi tìm kiếm các tài liệu liên quan đến một câu hỏi nhất định, mức độ liên quan điểm của mỗi tài liệu được tính toán bằng máy tính độ tương tự (ví dụ: độ tương tự cosin) hoặc khoảng cách (ví dụ: khoảng cách euclide) giữa vectơ của nó và vectơ câu hỏi.
- Mô hình xác suất: các mô hình xác suất cung cấp một cách tích hợp các mối quan hệ xác suất giữa từ thành một mô hình. Okapi BM25 là một mô hình xác suất với tần suất thuật ngữ và tài liệu chiều dài, đó là một trong những thành công nhất về mặt thực nghiệm các mô hình truy xuất và được sử dụng rộng rãi trong hệ thống tìm kiếm.

Trích chọn câu trả lời: trong các hệ thống hỏi đáp truyền thống, các câu hỏi thực tế và danh sách các câu hỏi đã được nghiên cứu rộng rãi trong một thời gian dài. Các câu hỏi thực tế (ví dụ: khi nào, ở đâu, ai...) câu trả lời thường là một đoạn văn bản duy nhất trong tài liệu, bao gồm tên thực thể, mã thông báo từ hoặc danh từ, cụm từ. Trong khi danh sách các câu hỏi có câu trả lời là một tập hợp thông tin thực tế xuất hiện trong cùng một tài liệu hoặc được tổng hợp từ các tài liệu khác nhau. Vì vậy loại câu trả lời nhận được từ giai đoạn phân tích câu hỏi có vai trò hết sức quan trọng, đặc biệt cho câu hỏi đã có câu trả lời là các thực thể được đặt tên. Do đó, các hệ thống ban đầu phụ thuộc rất nhiều vào nhận dạng thực thể (NER) để có thể đưa ra câu trả lời cuối cùng.

Trong thập kỷ gần đây, những kĩ thuật về học sâu mạng nơ-ron đã được ứng dụng thành công cho hệ thống hỏi đáp mở. Đối với phân tích câu hỏi, một số công trình phát triển các bộ phân loại mạng nơ-ron để xác định các loại câu hỏi. Đối với truy xuất tài liệu, các phương pháp dựa trên biểu diễn vectơ đã được đề xuất để giải quyết vấn đề các từ đồng nghĩa, là một vấn đề đã tồn tại từ lâu làm ảnh hưởng đến hiệu suất truy xuất. Đây là phương pháp khác hoàn toàn so với các phương pháp truyền thống như TF-IDF và BM25 sử dụng tìm kiếm dựa trên thống kê để vectơ hoá văn bản. Đối với trích xuất câu trả lời, như một giai đoạn quyết định để các hệ thống hỏi đáp đa ngành để có thể đưa ra câu trả lời cuối cùng, các mô hình mạng nơ-ron cũng có thể được áp dụng trích chọn câu trả lời từ một số tài liệu có liên quan cho một câu hỏi nhất định.

Trong khi đó, hiện nay đối với các tổ chức lớn, thông tin được lưu trữ trong nhiều hệ thống dữ liệu phân tán và đa dạng các định dạng tập tin (file). Do đó, việc tìm kiếm và khai

thác dữ liệu là một nhiệm vụ khó khăn. Hiện nay, chỉ có thể tìm kiếm các văn bản thông qua các từ khóa, tiêu đề hoặc các siêu dữ liệu do quản trị viên hệ thống nhập vào. Do đó, người dùng cuối không thể tìm kiếm hoặc trích xuất các thông tin bên trong các văn bản một cách tự động để xử lý các tác vụ nhập liệu vào hệ thống ERP (hệ thống hoạch định tài nguyên doanh nghiệp) và CRM (hệ thống quản lý quan hệ khách hàng), dẫn đến các công việc này đòi hỏi rất nhiều thời gian và công sức để xử lý.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất một hệ thống và phương pháp cho phép người dùng tìm kiếm thông tin bên trong các tài liệu với nhiều định dạng khác nhau: hệ thống cơ sở dữ liệu, các tập tin định dạng DOCX, PPT, XLSX, tài liệu quét và có thể mở rộng theo yêu cầu, trả về cho người dùng câu trả lời trực tiếp cho câu truy vấn và tài liệu tham chiếu. Hệ thống và phương pháp sử dụng công nghệ nhận dạng ký tự quang học để trích xuất văn bản từ ảnh và sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để phân loại, phân cụm và trích xuất các trường thông tin liên kết với văn bản, giúp cho việc tìm kiếm và khai thác thông tin hiệu quả.

Để đạt được mục đích trên, hệ thống và phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được đề xuất. Hệ thống bao gồm hai luồng xử lý thông tin chính cho hai đối tượng người dùng:

- Luồng chỉ mục: quản trị viên hệ thống có thể tải lên hệ thống các loại văn bản với định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét hoặc ảnh chụp văn bản. Các định dạng này có thể mở rộng theo yêu cầu của người dùng. Ngoài ra, hệ thống có các đầu nối tới các hệ thống cơ sở dữ liệu khác như SQL hoặc NoSQL để thu thập dữ liệu từ nhiều nguồn, nhiều phòng ban. Sau đó, các văn bản này đi qua bộ phân tích tập tin để được chuyển đổi về định dạng văn bản, phân loại và trích xuất các thông tin thêm như: tên người, tên công ty, ngày tháng, giá trị hợp đồng, các từ khóa. Tiếp theo, phần văn bản của văn bản sẽ được mã hóa bằng các mô hình ngôn ngữ để ra các vectơ ngữ nghĩa đặc trưng. Toàn bộ các thông tin này sẽ được liên kết với văn bản gốc và lưu trữ vào cơ sở dữ liệu tập trung.
- Luồng tìm kiếm: câu truy vấn của người dùng cuối sẽ được đưa vào bộ tìm kiếm văn bản để mã hóa thành vectơ ngữ nghĩa và tính điểm tương đồng với các vectơ ngữ nghĩa của văn bản và sắp xếp theo thứ tự giảm dần. Danh sách các văn bản này có thể được lọc một lần nữa thông qua các siêu dữ liệu và đưa vào bộ trích xuất câu trả lời để đưa ra câu trả lời liên quan nhất cho người dùng.

Theo đó, hệ thống được đề xuất trong sáng chế bao gồm hai luồng xử lý (luồng chỉ mục, luồng tìm kiếm) với các bộ như sau:

Bộ phân tích tập tin: nhiệm vụ chuyển đổi các tệp tin từ các định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét (scan), ảnh sang văn bản (định dạng DOC), đồng thời phân loại, trích xuất các nhãn (tag), các từ khóa, các thực thể như tên người, tên tổ chức, địa điểm, thời gian. Nếu văn bản dài hơn số từ mà mô hình học sâu có thể xử lý (chẳng hạn 512 từ), bộ phân tích sẽ đề xuất để người dùng chia nhỏ văn bản thành các đoạn ngắn hơn độ dài này;

Bộ mã hóa văn bản: các văn bản được đưa qua các mô hình ngôn ngữ (một mô hình học sâu chuyên xử lý ngôn ngữ dưới dạng văn bản chữ) để mã hóa ra các vectơ ngữ nghĩa với kích thước cố định (768 hoặc 1024). Nếu nằm ngoài khoảng này, kết quả tìm kiếm sẽ thiếu chính xác (nhỏ hơn 768) hoặc tốn nhiều tài nguyên hệ thống (lớn hơn 1024);

Bộ tìm kiếm văn bản: mã hóa các câu truy vấn của người dùng ra các vectơ ngữ nghĩa có cùng kích thước với vectơ văn bản, sau đó tính độ tương đồng với các vectơ văn bản và sắp xếp theo thứ tự giảm dần để đưa ra danh sách k văn bản gần nhất với câu truy vấn về mặt ngữ nghĩa. Phương pháp truyền thống sử dụng mô hình xác suất như BM25 cũng có thể được sử dụng thay cho mô hình học sâu. Phương pháp truyền thống có tốc độ xử lý nhanh hơn mô hình học sâu, tuy nhiên độ chính xác sẽ thấp hơn.

Bộ trích xuất câu trả lời: danh sách k văn bản gần nhất với câu truy vấn được cho vào một mô hình học sâu chuyên biệt cho việc trả lời câu hỏi → tìm ra các đoạn thông tin có trong văn bản liên quan trực tiếp đến câu truy vấn.

Bộ cơ sở dữ liệu: lưu trữ các văn bản gốc, các vectơ ngữ nghĩa và các siêu dữ liệu liên quan đến văn bản.

Với các bộ như trên, phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được thực hiện qua các bước:

Bước 1: trích xuất nội dung chữ của tập tin;

Bước 2: phân loại tài liệu và trích rút thông tin;

Bước 3: mã hóa nội dung chữ;

Bước 4: tìm kiếm khi có truy vấn;

Bước 5: trả lời truy vấn dựa vào các văn bản liên quan;

Bước 6: thu thập phản hồi.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình vẽ kiến trúc hệ thống hỏi đáp truyền thống;

Hình 2: là hình vẽ tổng quan hệ thống truy vấn và hỏi đáp tự động được đề cập trong sáng chế;

Hình 3: là hình vẽ thể hiện luồng truy vấn thông tin của người dùng cuối;

Hình 4: là hình vẽ thể hiện luồng trích xuất thông tin và đánh chỉ mục;

Hình 5: là hình vẽ minh họa kết quả của quá trình lượng tử hóa mô hình học sâu;

Hình 6: là hình vẽ mô tả quá trình chắt lọc tri thức;

Hình 7: là hình vẽ so sánh xử lý tuần tự và xử lý theo khối;

Hình 8: là hình vẽ mô tả việc thu thập phản hồi từ người dùng;

Hình 9: là hình vẽ mô tả quá trình lưu trữ mối quan hệ và trả lời câu hỏi bằng cơ sở đồ thị tri thức.

Mô tả chi tiết sáng chế

Hệ thống truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được đề cập trong sáng chế được thực hiện qua hai lường xử lý bao gồm các bộ, cụ thể như sau:

Luồng chỉ mục: chuyển đổi tệp tin, đánh chỉ mục và mã hóa văn bản; quản trị viên hệ thống có thể tải lên hệ thống các loại văn bản với định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét hoặc ảnh chụp văn bản. Các định dạng này có thể mở rộng theo yêu cầu của người dùng. Ngoài ra, hệ thống có các đầu nối tới các hệ thống cơ sở dữ liệu khác như SQL hoặc NoSQL để thu thập dữ liệu từ nhiều nguồn, nhiều phòng ban. Sau đó, các văn bản này đi qua bộ phân tích tập tin để được chuyển đổi về định dạng văn bản, phân loại và trích xuất các thông tin thêm như: tên người, tên công ty, ngày tháng, giá trị hợp đồng, các từ khóa. Tiếp theo, phần văn bản của văn bản sẽ được mã hóa bằng các mô hình ngôn ngữ để ra các vectơ ngữ nghĩa đặc trưng. Toàn bộ các thông tin này sẽ được liên kết với văn bản gốc và lưu trữ vào cơ sở dữ liệu tập trung;

Luồng tìm kiếm: tìm kiếm và trích xuất câu trả lời cho người dùng; câu truy vấn của người dùng cuối sẽ được đưa vào bộ tìm kiếm văn bản để mã hóa thành vectơ ngữ nghĩa và tính điểm tương đồng với các vectơ ngữ nghĩa của văn bản và sắp xếp theo thứ tự giảm dần. Danh sách các văn bản này có thể được lọc một lần nữa thông qua các siêu dữ liệu và đưa vào bộ trích xuất câu trả lời để đưa ra câu trả lời liên quan nhất cho người dùng.

Bộ phân tích tập tin: nhiệm vụ chuyển đổi các tệp tin từ các định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét, ảnh sang văn bản, đồng thời phân loại, trích xuất các tag, các từ khóa, các thực thể như tên người, tên tổ chức, địa điểm, thời gian. Nếu văn bản dài hơn số từ mà mô hình học sâu có thể xử lý (chẳng hạn 512 từ), bộ phân tích sẽ đề xuất để người dùng chia nhỏ văn bản thành các đoạn ngắn hơn độ dài này;

Bộ mã hóa văn bản: các văn bản được đưa qua các mô hình ngôn ngữ (một mô hình học sâu chuyên xử lý ngôn ngữ dưới dạng văn bản chữ) để mã hóa ra các vectơ ngữ nghĩa với kích thước cố định (768 hoặc 1024). Nếu nằm ngoài khoảng này, kết quả tìm kiếm sẽ thiếu chính xác (nhỏ hơn 768) hoặc tốn nhiều tài nguyên hệ thống (lớn hơn 1024);

Bộ tìm kiếm văn bản: mã hóa các câu truy vấn của người dùng ra các vectơ ngữ nghĩa có cùng kích thước với vectơ văn bản, sau đó tính độ tương đồng với các vectơ văn bản và sắp xếp theo thứ tự giảm dần → danh sách k văn bản gần nhất với câu truy vấn về mặt ngữ nghĩa. Phương pháp truyền thống sử dụng mô hình xác suất như BM25 cũng có thể được sử dụng thay cho mô hình học sâu. Phương pháp truyền thống có tốc độ xử lý nhanh hơn mô hình học sâu, tuy nhiên độ chính xác sẽ thấp hơn.

Bộ trích xuất câu trả lời: danh sách k văn bản gần nhất với câu truy vấn được cho vào một mô hình học sâu chuyên biệt cho việc trả lời câu hỏi → tìm ra các đoạn thông tin có trong văn bản liên quan trực tiếp đến câu truy vấn.

Bộ cơ sở dữ liệu: lưu trữ các văn bản gốc, các vectơ ngữ nghĩa và các siêu dữ liệu liên quan đến văn bản.

Phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được đề cập trong sáng chế bao gồm các bước sau:

Bước 1: trích xuất nội dung chữ của tập tin;

Từ tập tin là tài liệu hoặc hình ảnh mà quản trị viên đưa vào, bộ phân tích tập tin sử dụng công nghệ nhận dạng ký tự quang học, trích xuất nội dung tập tin để lấy nội dung chữ có trong tập tin (có thể là định dạng văn bản dạng chữ docx, pdf, xlsx… hoặc định dạng hình ảnh jpg, png…). Đầu ra của bước này là nội dụng dạng chữ của tập tin.

Bước 2: phân loại tài liệu và trích rút thông tin;

Từ nội dung văn bản, các thông tin như: thể loại văn bản, các thực thể, quan hệ… được tự động trích rút và thêm vào siêu dữ liệu. Các thông tin về các thực thể và quan hệ giữa chúng được sử dụng để xây dựng đồ thị tri thức. Quản trị viên cũng phân cấp văn bản về độ bảo mật để hạn chế quyền xem và tìm kiếm một số văn bản có độ bảo mật cao theo phân quyền người dùng theo cấp bậc, phân cấp dữ liệu, dữ liệu cá nhân và chung.

Cụ thể, việc phân quyền người dùng theo cấp bậc, phân cấp dữ liệu, dữ liệu cá nhân và chung được thực hiện như sau:

Khi tích hợp với thông tin nhân viên của tổ chức, hệ thống có thể chia tập người dùng thành ít nhất ba cấp như bảng 1: quản trị hệ thống, người dùng chung, người dùng đặc biệt.

- Quản trị hệ thống: được phép cấu hình hệ thống, được phép thêm các tài liệu và tìm kiếm các văn bản chung của doanh nghiệp;
- Người dùng chung: được phép thêm tài liệu và tìm kiếm văn bản chung;
- Người dùng đặc biệt: được phép thêm tài liệu, tìm kiếm các văn bản chung và bảo mật, được phép tạo không gian lưu trữ và tìm kiếm riêng;

Bảng 1: cấp độ và quyền của người dùng hệ thống

	Quyền thêm tài liệu	Tìm kiếm và hiển thị văn bản/tài liệu chung	Cấu hình hệ thống	Tìm kiếm và hiển thị các văn bản mật	Tạo không gian lưu trữ riêng
Quản trị hệ thống	x	x	x		
Người dùng chung	x	x			
Người dùng đặc biệt	x	x		x	x

Tương tự đối với các mức bảo mật của văn bản cũng được chia thành ba mức độ:

- Mức bảo mật thông thường: các tài liệu không bị mã hóa và đánh chỉ mục cho việc tìm kiếm và có thể tải về máy tính cá nhân và được phép lưu hành trên các nền tảng lưu trữ khác;
- Mức bảo mật bí mật: chỉ người dùng chung và người dùng đặc biệt có quyền tìm kiếm và hiển thị nhưng không thể tải về được;
- Mức bảo mật tuyệt mật: văn bản tài liệu được mã hóa theo chuẩn mã hóa cao cấp AES-256, yêu cầu người dùng đặc biệt phải nhập mật khẩu để hiển thị tài liệu.

Bảng 2: Các mức bảo mật tài liệu và quyền xem của người dùng

Mức bảo mật	Người dùng đặc biệt	Người dùng chung	Quản trị hệ thống
Tuyệt mật	x		
Bí mật	x	x	
Thông thường	x	x	x

Bước 3: mã hóa nội dung chữ;

Từ văn bản dạng chữ, bộ mã hóa văn bản (bản chất là một mô hình học sâu) sẽ mã hóa văn bản này thành các vectơ ngữ nghĩa và lưu các vectơ này vào trong cơ sở dữ liệu. Bên cạnh việc mã hóa văn bản, văn bản gốc và siêu dữ liệu cũng được lưu vào cơ sở dữ liệu. Cơ sở dữ liệu ở đây có thể là điện toán đám mây hoặc máy chủ vật lý.

Bước 4: tìm kiếm khi có truy vấn;

Khi người sử dụng đầu cuối đưa ra câu hỏi, bộ tìm kiếm văn bản sẽ kết hợp các phương pháp học sâu và truyền thống để tìm ra các văn bản liên quan đến câu hỏi đó trong cơ

sở dữ liệu. Với phương pháp học sâu, câu hỏi của người dùng sẽ được mã hóa thành các vectơ ngữ nghĩa và so sánh với các vectơ ngữ nghĩa trong cơ sở dữ liệu. Các mô hình học sâu được tối ưu theo các phương pháp trong tối ưu tốc độ mô hình học sâu. Phương pháp truyền thống dựa vào mô hình xác suất và các từ trong câu hỏi để tìm kiếm văn bản liên quan. Đầu ra là các văn bản liên quan được đưa tới bộ trích xuất câu trả lời.

Trong đó, tối ưu tốc độ mô hình học sâu được hiểu như sau:

Mô hình học sâu thường có tốc độ xử lý chậm hơn so với các phương pháp thống kê truyền thống. Tốc độ của mô hình học sâu phụ thuộc vào nhiều yếu tố như số lượng tham số, định dạng dấu phẩy động, phần cứng máy tính, cấu trúc mạng nơ-ron, thuật toán xử lý… Trong hệ thống này, mô hình có thể được tối ưu bằng các phương pháp lượng tử hóa, chắt lọc tri thức, xử lý theo khối.

- Lượng tử hóa là việc chuyển định dạng dấu phẩy động từ cao về thấp (ví dụ từ 32-bit xuống 16-bit). Tham số 16-bit trong mạng nơ-ron sẽ chiếm ít phần cứng máy tính hơn so với tham số 32-bit. Quá trình này được mô tả trong Hình 5.
- Chắt lọc tri thức: là quá trình huấn luyện một mô hình học sâu nhỏ hơn từ mô hình lớn hơn. Trong cơ chế huấn luyện này, đầu ra của mô hình lớn sẽ được tính như câu trả lời tiêu chuẩn để mô hình nhỏ học theo. Việc huấn luyện mô hình nhỏ hơn sẽ giúp giảm số lượng tham số phải tính mỗi lần chạy. Quá trình này được minh họa trong Hình 6.
- Xử lý theo khối: khi có nhiều yêu cầu xử lý công việc cùng lúc, hệ thống sẽ gộp các yêu cầu này và chạy xử lý theo khối. Một mô hình học sâu có thể xử lý theo khối lớn hay nhỏ tùy theo cấu trúc mạng nơ-ron. Hình 7 so sánh khác biệt giữa xử lý theo khối và xử lý tuần tự.

Bước 5: trả lời truy vấn dựa vào các văn bản liên quan;

Trong các văn bản liên quan được đưa ra ở bước trên, bộ trích xuất câu trả lời (bản chất là một mô hình học sâu) trả về câu trả lời trực tiếp cho câu truy vấn. Ngoài ra, đồ thị tri thức cũng được sử dụng để đưa ra thêm một câu trả lời nữa thông qua việc phân tích thực thể trong câu truy vấn. Câu trả lời thêm vào này có thể được hiển thị cho người dùng tùy vào độ tin cậy của đồ thị tri thức. Việc này được minh họa trong Hình 9. Đồ thị tri thức và ứng dụng của nó được giới thiệu trong đồ thị tri thức, cụ thể:

Định nghĩa thực thể và quan hệ:

- Dựa vào lĩnh vực áp dụng để có thể định nghĩa các thực thể và quan hệ trong lĩnh vực đó. Ví dụ trong lĩnh vực ngân hàng, tài chính, các thực thể bao gồm ngân hàng, sản

phẩm tài chính (như khoản vay, thẻ tín dụng và đầu tư), khách hàng, giao dịch, khoản đầu tư.

- Xác định các thuộc tính và đặc điểm mô tả của từng thực thể. Ví dụ, thực thể ngân hàng có thể có các thuộc tính như tên, vị trí và tổng tài sản, trong khi sản phẩm tài chính có thể có các thuộc tính như lãi suất, phí và tiêu chí đủ điều kiện.
- Xác định các mối quan hệ kết nối các thực thể trong đồ thị tri thức. Các mối quan hệ thường là động từ hoặc cụm từ mô tả cách thức các thực thể liên quan, chẳng hạn như "cung cấp", "sở hữu", "đầu tư vào", "được quản lý bởi" hoặc "thuộc về chỉ số".
- Thiết kế một đồ thị xác định cấu trúc đồ thị tri thức bao gồm các thực thể, thuộc tính và mối quan hệ. Đồ thị này hoạt động như một bản mẫu để tổ chức và lưu trữ dữ liệu trong đồ thị.

Trích xuất thông tin:

- Với đầu vào là văn bản dạng chữ từ tài liệu, các thông tin như thực thể và quan hệ sẽ được trích xuất bằng cách sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên. Các phương pháp này giúp xác định và phân biệt các thực thể và mối quan hệ trong văn bản, cũng như xác định các thuộc tính của chúng.
- Chuyển đổi thông tin đã trích xuất thành một định dạng có thể dễ dàng nhập vào cơ sở dữ liệu đồ thị. Dữ liệu được biểu diễn dưới dạng nút (thực thể) và cạnh (mối quan hệ) trong đồ thị.

Khởi tạo và cập nhật đồ thị tri thức:

- Lưu trữ thông tin đã trích xuất và xử lý trong cơ sở dữ liệu đồ thị.
- Đảm bảo chất lượng và tính nhất quán của dữ liệu bằng cách triển khai các quy trình làm sạch và kiểm tra lại dữ liệu, chuẩn hóa định dạng, giải quyết trùng lặp và xác minh độ chính xác của thông tin đã trích xuất.
- Định kỳ cập nhật đồ thị tri thức với thông tin mới để đảm bảo độ chính xác cao và tính cập nhật. Hệ thống có các quy trình tự động để xử lý công việc này, chẳng hạn như các đoạn mã thu thập dữ liệu mới trên các trang mạng theo chu kỳ hoặc các yêu cầu API lấy thông tin tài chính mới nhất.

Tích hợp đồ thị tri thức vào hệ thống tìm kiếm:

Việc tích hợp đồ thị tri thức vào hệ thống gồm các thành phần sau:

- Mở rộng câu truy vấn: Sử dụng đồ thị tri thức để xác định các thực thể và mối quan hệ liên quan đến truy vấn của người dùng. Việc sử dụng đồ thị tri thức giúp cải thiện độ chính xác và mức độ liên quan của kết quả tìm kiếm

- Xếp hạng: Tận dụng đồ thị tri thức để cải thiện việc xếp hạng kết quả tìm kiếm. Ưu tiên kết quả dựa trên mức độ liên quan đến truy vấn của người dùng.
- Cá nhân hóa: Tận dụng đồ thị tri thức để có thể trả về kết quả phù hợp với người dùng dựa trên lịch sử tìm kiếm.

Bước 6: thu thập phản hồi;

Khi câu trả lời được đưa ra, hệ thống sẽ có lựa chọn để người dùng phản hồi về chất lượng câu trả lời đó (mức độ liên quan và chính xác). Phản hồi của người dùng sẽ được lưu lại vào cơ sở dữ liệu như trong Hình 8F. Sau đó, các phản hồi này có thể được sử dụng để huấn luyện mô hình học sâu ở Bước 1.

Yêu cầu bảo hộ

1. Hệ thống truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được thực hiện qua hai lường xử lý bao gồm các bộ, cụ thể như sau:

luồng chỉ mục: chuyển đổi tệp tin, đánh chỉ mục và mã hóa văn bản; quản trị viên hệ thống có thể tải lên hệ thống các loại văn bản với định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét hoặc ảnh chụp văn bản; các định dạng này có thể mở rộng theo yêu cầu của người dùng; ngoài ra, hệ thống có các đầu nối tới các hệ thống cơ sở dữ liệu khác như SQL hoặc NoSQL để thu thập dữ liệu từ nhiều nguồn, nhiều phòng ban; sau đó, các văn bản này đi qua bộ phân tích tập tin để được chuyển đổi về định dạng văn bản, phân loại và trích xuất các thông tin thêm như: tên người, tên công ty, ngày tháng, giá trị hợp đồng, các từ khóa; tiếp theo, phần văn bản của văn bản sẽ được mã hóa bằng các mô hình ngôn ngữ để ra các vectơ ngữ nghĩa đặc trưng; toàn bộ các thông tin này sẽ được liên kết với văn bản gốc và lưu trữ vào cơ sở dữ liệu tập trung;

luồng tìm kiếm: tìm kiếm và trích xuất câu trả lời cho người dùng; câu truy vấn của người dùng cuối sẽ được đưa vào bộ tìm kiếm văn bản để mã hóa thành vectơ ngữ nghĩa và tính điểm tương đồng với các vectơ ngữ nghĩa của văn bản và sắp xếp theo thứ tự giảm dần; danh sách các văn bản này có thể được lọc một lần nữa thông qua các siêu dữ liệu và đưa vào bộ trích xuất câu trả lời để đưa ra câu trả lời liên quan nhất cho người dùng;

bộ phân tích tập tin: nhiệm vụ chuyển đổi các tệp tin từ các định dạng khác nhau như DOCX, PPT, EXCEL, văn bản quét, ảnh sang văn bản (định dạng DOC), đồng thời phân loại, trích xuất các tag, các từ khóa, các thực thể như tên người, tên tổ chức, địa điểm, thời gian; nếu văn bản dài hơn số từ mà mô hình học sâu có thể xử lý (chẳng hạn 512 từ), bộ phân tích sẽ đề xuất để người dùng chia nhỏ văn bản thành các đoạn ngắn hơn độ dài này;

bộ mã hóa văn bản: các văn bản được đưa qua các mô hình ngôn ngữ (một mô hình học sâu chuyên xử lý ngôn ngữ dưới dạng văn bản chữ) để mã hóa ra các vectơ ngữ nghĩa với kích thước cố định (768 hoặc 1024); nếu nằm ngoài khoảng này, kết quả tìm kiếm sẽ thiếu chính xác (nhỏ hơn 768) hoặc tốn nhiều tài nguyên hệ thống (lớn hơn 1024);

bộ tìm kiếm văn bản: mã hóa các câu truy vấn của người dùng ra các vectơ ngữ nghĩa có cùng kích thước với vectơ văn bản, sau đó tính độ tương đồng với các vectơ văn bản và sắp xếp theo thứ tự giảm dần → danh sách k văn bản gần nhất với câu truy vấn về mặt ngữ nghĩa. Phương pháp truyền thống sử dụng mô hình xác suất như BM25 cũng có thể được sử dụng thay cho mô hình học sâu; phương pháp truyền thống có tốc độ xử lý nhanh hơn mô hình học sâu, tuy nhiên độ chính xác sẽ thấp hơn;

bộ trích xuất câu trả lời: danh sách k văn bản gần nhất với câu truy vấn được cho vào

một mô hình học sâu chuyên biệt cho việc trả lời câu hỏi → tìm ra các đoạn thông tin có trong văn bản liên quan trực tiếp đến câu truy vấn;

bộ cơ sở dữ liệu: lưu trữ các văn bản gốc, các vectơ ngữ nghĩa và các siêu dữ liệu liên quan đến văn bản.

2. Phương pháp truy vấn tài liệu và hỏi đáp ngữ nghĩa tự động tiếng Việt được đề cập trong sáng chế bao gồm các bước sau:

bước 1: trích xuất nội dung chữ của tập tin;

từ tập tin là tài liệu hoặc hình ảnh mà quản trị viên đưa vào, bộ phân tích tập tin sử dụng công nghệ nhận dạng ký tự quang học, trích xuất nội dung tập tin để lấy nội dung chữ có trong tập tin (có thể là định dạng văn bản dạng chữ docx, pdf, xlsx… hoặc định dạng hình ảnh jpg, png…); đầu ra của bước này là nội dụng dạng chữ của tập tin;

bước 2: phân loại tài liệu và trích rút thông tin;

từ nội dung văn bản, các thông tin như: thể loại văn bản, các thực thể, quan hệ… được tự động trích rút và thêm vào siêu dữ liệu; các thông tin về các thực thể và quan hệ giữa chúng được sử dụng để xây dựng đồ thị tri thức; quản trị viên cũng phân cấp văn bản về độ bảo mật để hạn chế quyền xem và tìm kiếm một số văn bản có độ bảo mật cao theo phân quyền người dùng theo cấp bậc, phân cấp dữ liệu, dữ liệu cá nhân và chung;

bước 3: mã hóa nội dung chữ;

từ văn bản dạng chữ, bộ mã hóa văn bản (bản chất là một mô hình học sâu) sẽ mã hóa văn bản này thành các vectơ ngữ nghĩa và lưu các vectơ này vào trong cơ sở dữ liệu; bên cạnh việc mã hóa văn bản, văn bản gốc và siêu dữ liệu cũng được lưu vào cơ sở dữ liệu; cơ sở dữ liệu ở đây có thể là điện toán đám mây hoặc máy chủ vật lý;

bước 4: tìm kiếm khi có truy vấn;

khi người sử dụng đầu cuối đưa ra câu hỏi, bộ tìm kiếm văn bản sẽ kết hợp các phương pháp học sâu và truyền thống để tìm ra các văn bản liên quan đến câu hỏi đó trong cơ sở dữ liệu; với phương pháp học sâu, câu hỏi của người dùng sẽ được mã hóa thành các vectơ ngữ nghĩa và so sánh với các vectơ ngữ nghĩa trong cơ sở dữ liệu; các mô hình học sâu được tối ưu theo các phương pháp trong tối ưu tốc độ mô hình học sâu; phương pháp truyền thống dựa vào mô hình xác suất và các từ trong câu hỏi để tìm kiếm văn bản liên quan; đầu ra là các văn bản liên quan được đưa tới bộ trích xuất câu trả lời;

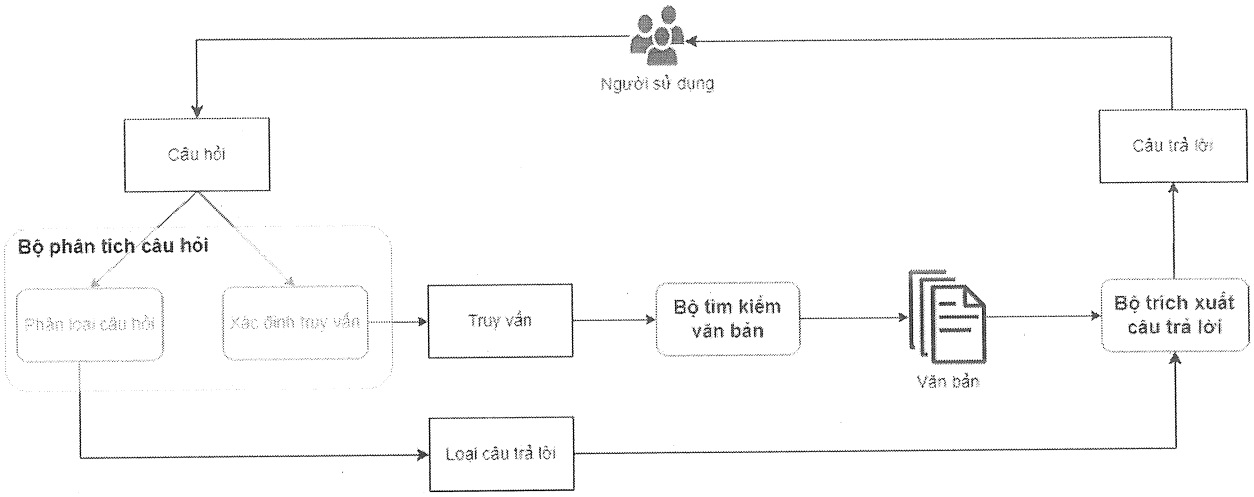
bước 5: trả lời truy vấn dựa vào các văn bản liên quan;

trong các văn bản liên quan được đưa ra ở bước trên, bộ trích xuất câu trả lời (bản chất là một mô hình học sâu) trả về câu trả lời trực tiếp cho câu truy vấn; ngoài ra, đồ thị tri thức cũng được sử dụng để đưa ra thêm một câu trả lời nữa thông qua việc phân tích thực thể trong

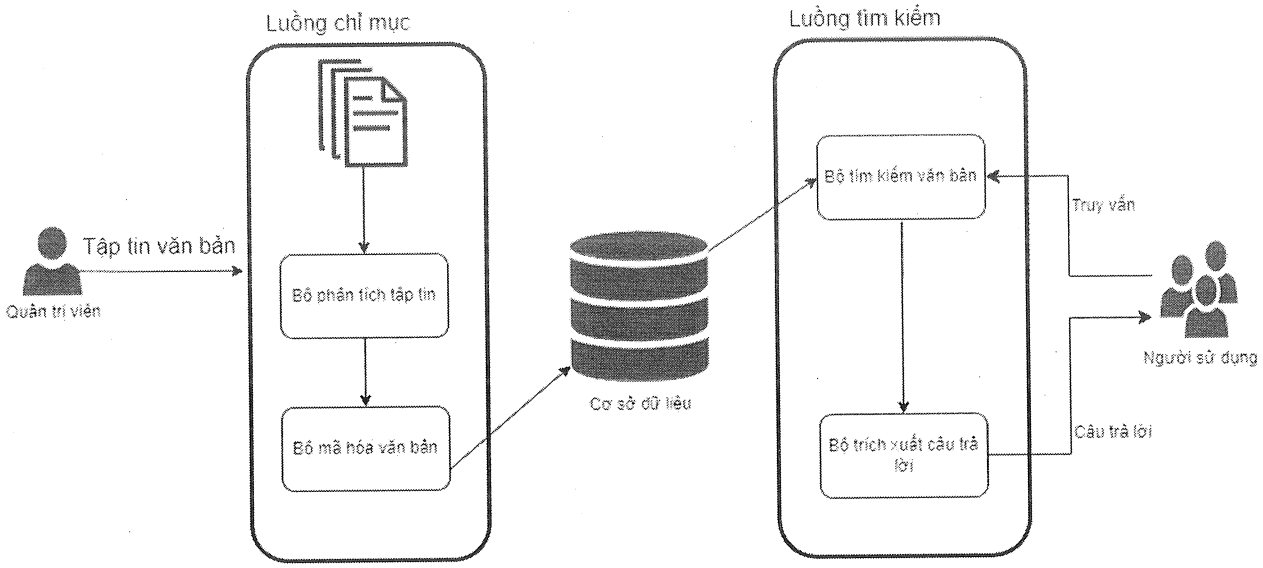
câu truy vấn; câu trả lời thêm vào này có thể được hiển thị cho người dùng tùy vào độ tin cậy của đồ thị tri thức;

bước 6: thu thập phản hồi;

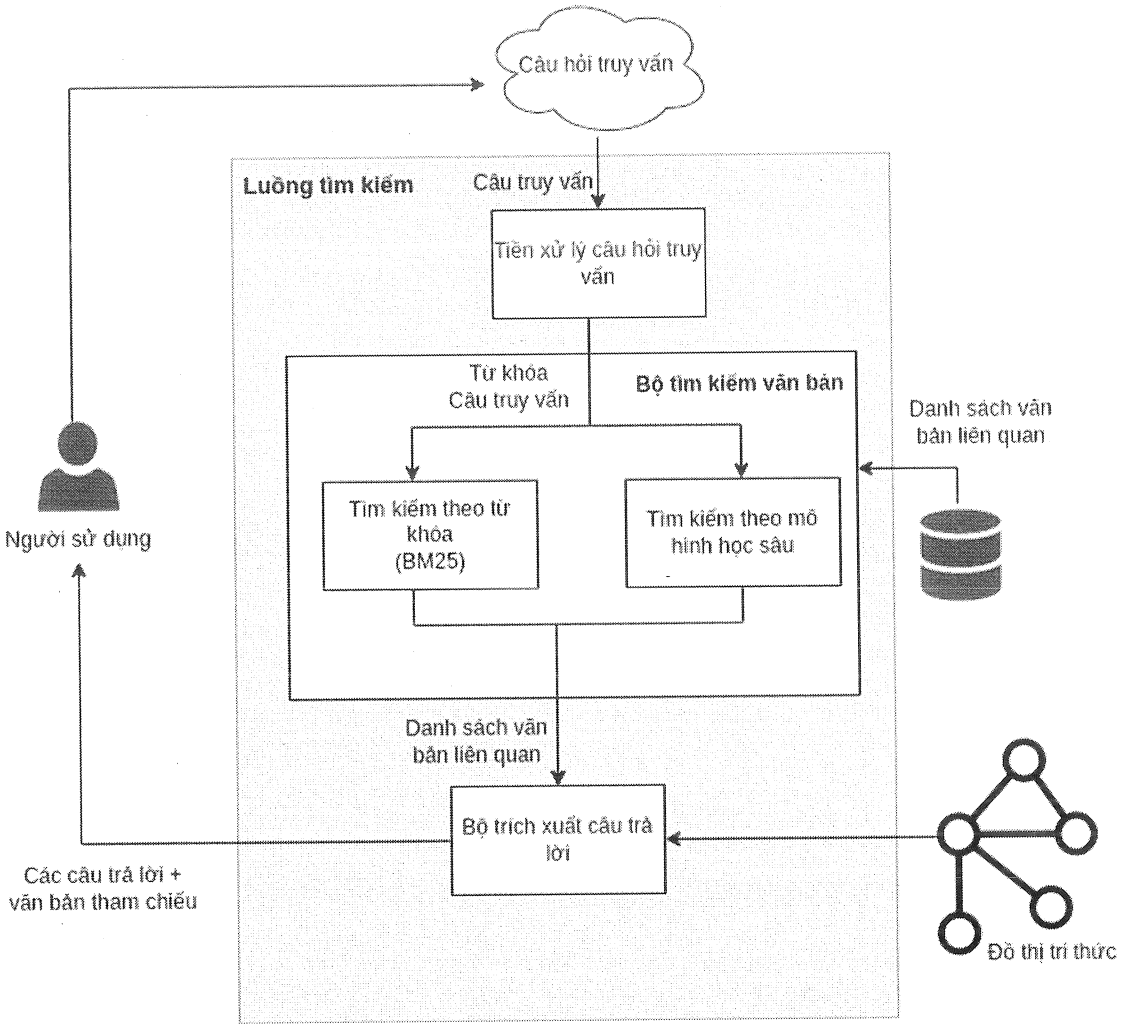
khi câu trả lời được đưa ra, hệ thống sẽ có lựa chọn để người dùng phản hồi về chất lượng câu trả lời đó (mức độ liên quan và chính xác); phản hồi của người dùng sẽ được lưu lại vào cơ sở dữ liệu; sau đó, các phản hồi này có thể được sử dụng để huấn luyện mô hình học sâu ở bước 1.


Người sử dụng
Câu hỏi
Câu trả lời
Bộ phân tích câu hỏi
Phân loại câu hỏi
Xác định truy vấn
Truy vấn
Bộ tìm kiếm văn bản
Văn bản
Bộ trích xuất câu trả lời
Loại câu trả lời

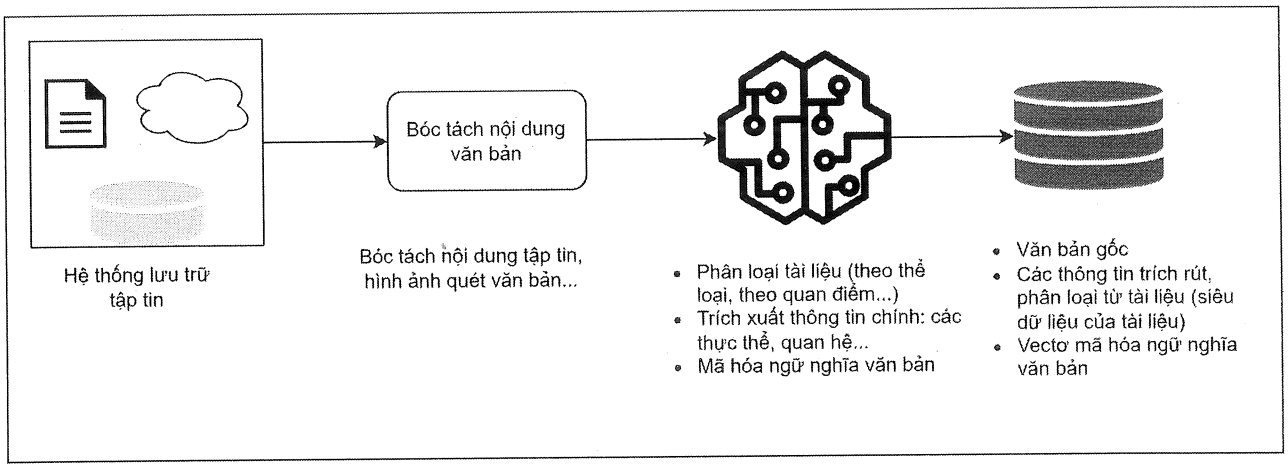
Hình 1


Luồng chỉ mục
Luồng tìm kiếm
Tập tin văn bản
Quản trị viên
Bộ phân tích tập tin
Bộ mã hóa văn bản
Cơ sở dữ liệu
Bộ tìm kiếm văn bản
Truy vấn
Người sử dụng
Bộ trích xuất câu trả lời
Câu trả lời

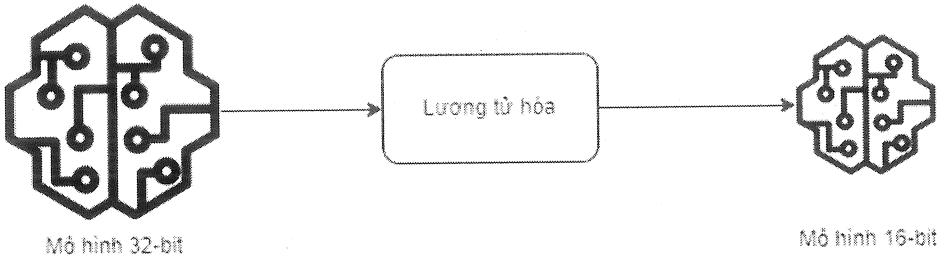
Hình 2


Câu hỏi truy vấn
Luồng tìm kiếm
Câu truy vấn
Tiền xử lý câu hỏi truy vấn
Từ khóa
Câu truy vấn
Bộ tìm kiếm văn bản
Danh sách văn bản liên quan
Tìm kiếm theo từ khóa (BM25)
Tìm kiếm theo mô hình học sâu
Người sử dụng
Danh sách văn bản liên quan
Bộ trích xuất câu trả lời
Các câu trả lời + văn bản tham chiếu
Đồ thị tri thức

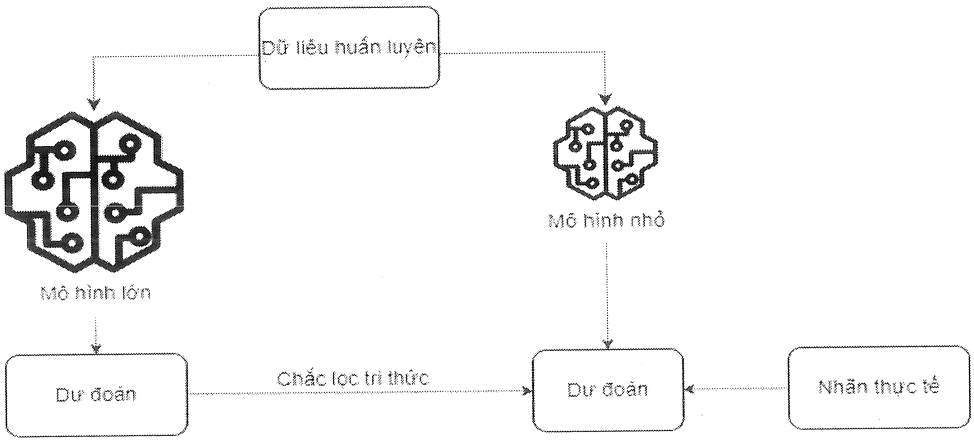
Hình 3


Bóc tách nội dung văn bản
Hệ thống lưu trữ tập tin
Bóc tách nội dung tập tin, hình ảnh quét văn bản...
Phân loại tài liệu (theo thể loại, theo quan điểm...)
Trích xuất thông tin chính: các thực thể, quan hệ...
Mã hóa ngữ nghĩa văn bản
Văn bản gốc
Các thông tin trích rút, phân loại từ tài liệu (siêu dữ liệu của tài liệu)
Vectơ mã hóa ngữ nghĩa văn bản

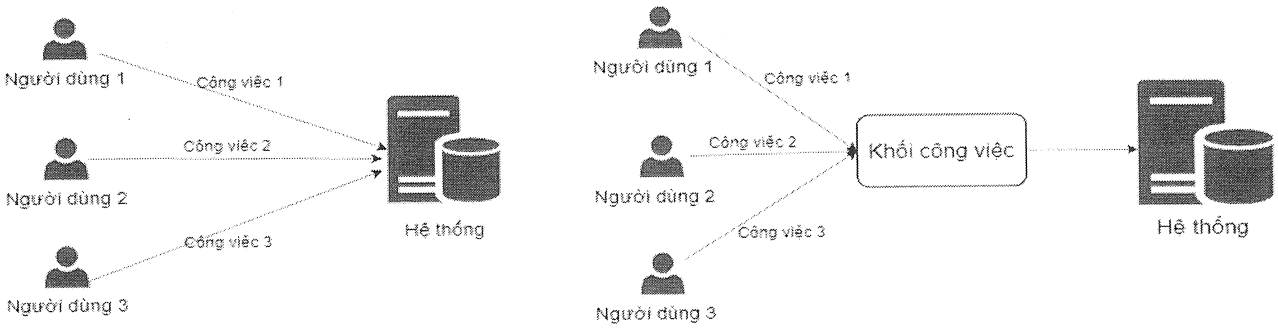
Hình 4



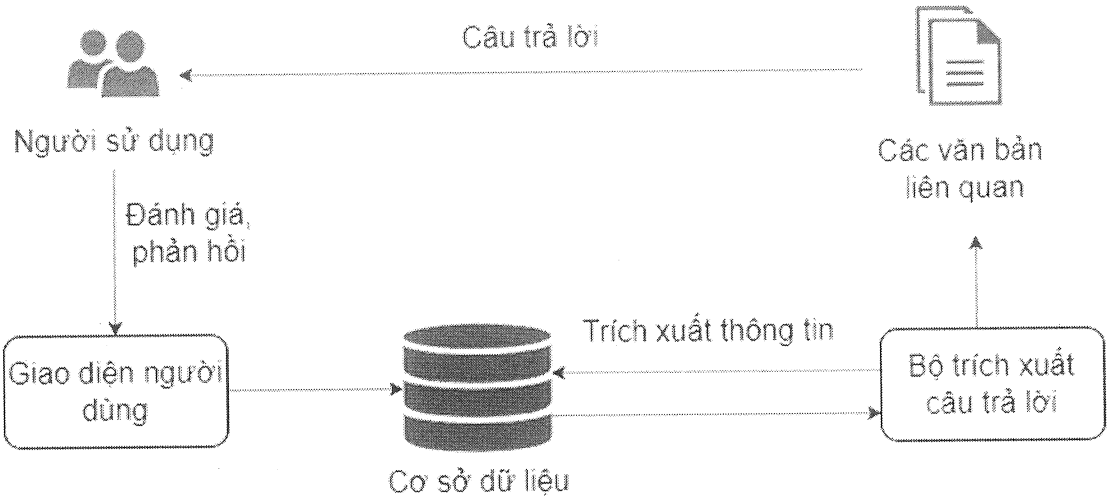
Hình 5



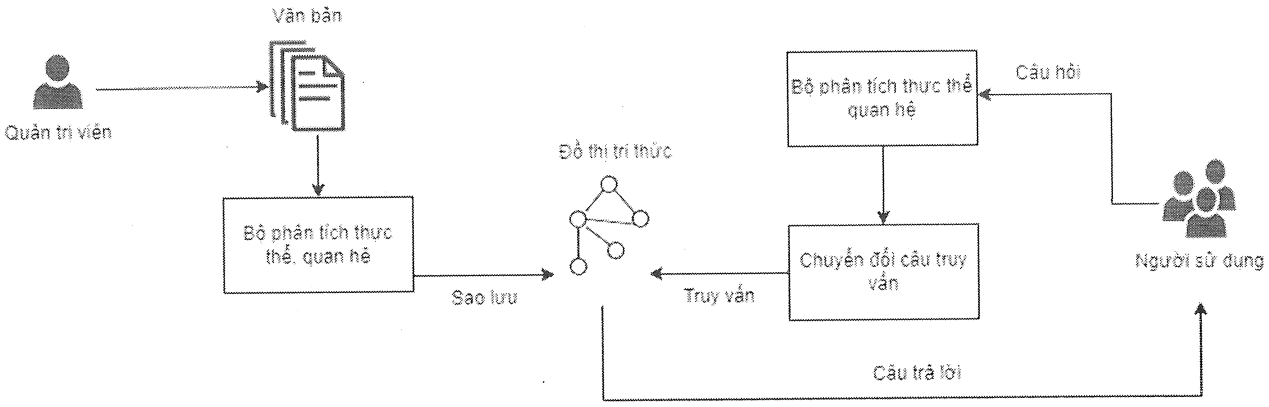
Hình 6



Hình 7


Câu trả lời
Người sử dụng
Các văn bản liên quan
Đánh giá, phản hồi
Giao diện người dùng
Trích xuất thông tin
Bộ trích xuất câu trả lời
Cơ sở dữ liệu

Hình 8


Văn bản
Quản trị viên
Bộ phân tích thực thể, quan hệ
Sao lưu
Đồ thị tri thức
Truy vấn
Bộ phân tích thực thể quan hệ
Câu hỏi
Chuyển đổi câu truy vấn
Người sử dụng
Câu trả lời

Hình 9