(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) **1-0053766**

(51)[2022.01] **G06F 16/35**

(13) **B**

(21) 1-2023-06880

(22) 03/10/2023

(45) 25/11/2025 452

(43) 25/01/2024 430A

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội.

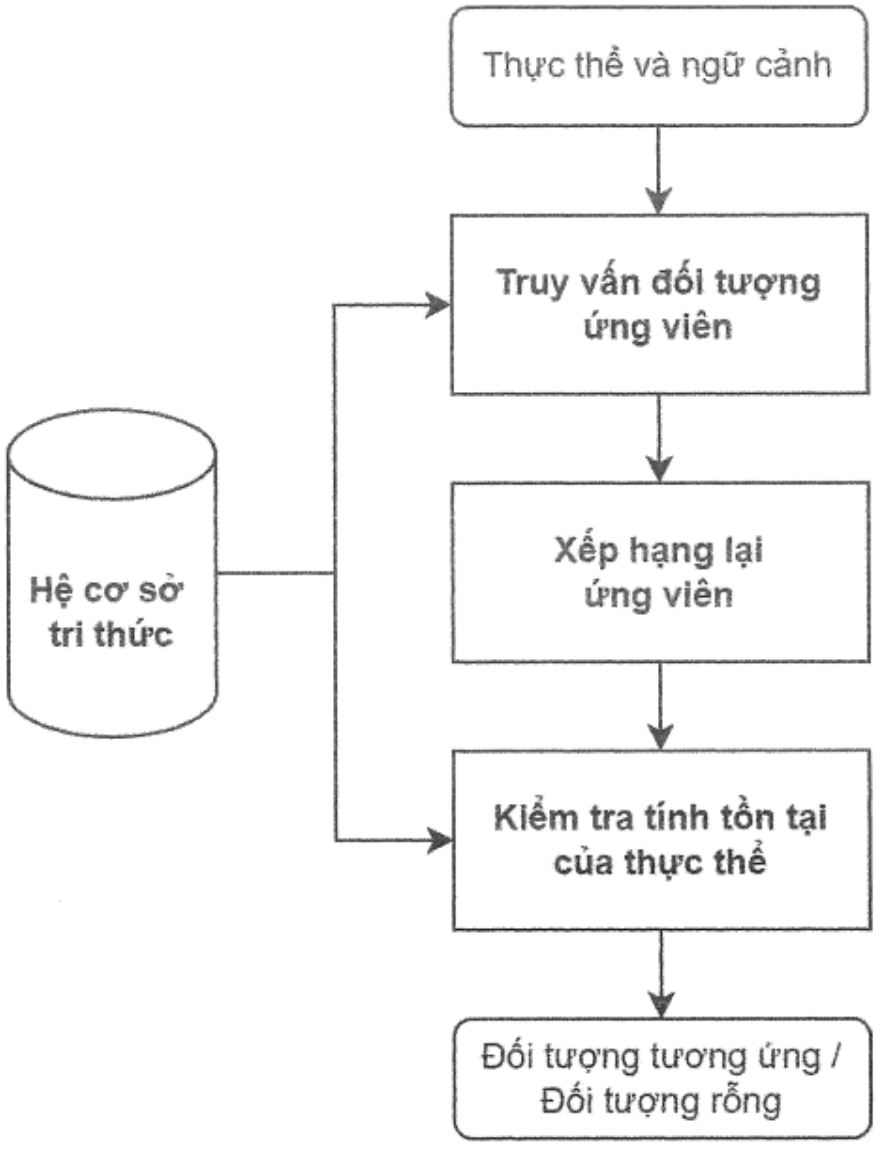
(72) THÁI PHÁT TRIỂN (VN); HUỲNH MINH TRÍ (VN); NGUYỄN VĂN TUẤN (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP HỢP NHẤT CÁC THỰC THỂ TRONG VĂN BẢN ỨNG DỤNG HỌC SÂU

(21) 1-2023-06880

(57) Sáng chế đề cập đến phương pháp hợp nhất các thực thể trích xuất từ văn bản và đồng tham chiếu đến một đối tượng, gồm các bước: bước 1: xây dựng hệ cơ sở tri thức; bước 2: truy vấn đối tượng ứng viên; bước 3: xếp hạng lại các ứng viên; bước 4: kiểm tra tính tồn tại của thực thể trong hệ cơ sở tri thức. Phương pháp tạo cơ sở cho các bài toán xử lý ngôn ngữ tự nhiên khi phân tích dữ liệu văn bản, trích xuất thông tin hay liên kết các sự kiện của một đối tượng dễ dàng hơn, nhằm phục vụ cho các hệ thống khai phá và phân tích dữ liệu từ báo chí hoặc mạng xã hội.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp hợp nhất các thực thể trong văn bản ứng dụng học sâu. Cụ thể, phương pháp được đề cập trong sáng chế hợp nhất các thực thể được trích xuất từ văn bản và cùng tham chiếu đến một đối tượng, và được ứng dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Tình trạng kỹ thuật của sáng chế

Trong thời đại phát triển mạnh mẽ của mạng viễn thông toàn cầu, lượng thông tin ngày càng trở nên dồi dào, đặc biệt là thông tin dạng văn bản. Nguồn dữ liệu khổng lồ này đòi hỏi được khai phá và phân tích cho nhiều mục đích khác nhau. Trong đó, trích xuất thực thể và xác định đối tượng tương ứng của chúng trong thực tế là những nhiệm vụ cơ bản và quan trọng nhất. Tuy nhiên, với nhiều nguyên nhân, dù cùng một đối tượng nhưng chúng lại được đề cập với các tên gọi không thống nhất. Điều này gây khó khăn trong việc phân tích thông tin của đối tượng. Một phương pháp giúp định danh các tên gọi khác nhau của cùng một đối tượng sẽ tạo thuận lợi cho quá trình xử lý.

Một văn bản được tạo ra thường xoay quanh một hay nhiều thực thể nào đó với nhiều cách diễn đạt. Với yêu cầu tránh lặp từ, tác giả thường dùng nhiều từ ngữ khác nhau khi đề cập đến một đối tượng nào đó. Ví dụ các cụm từ như "Sài Gòn", "thành phố mang tên Bác", "TP HCM" thường được dùng để thay thế cho "thành phố Hồ Chí Minh". Bên cạnh đó, việc trùng lặp tên riêng cũng tạo nên sự nhập nhằng cho máy tính trong việc phân tích văn bản, ví dụ điển hình là tên "huyện Châu Thành" đã được sử dụng cho khoảng mười một huyện ở miền Nam Việt Nam. Ngoài ra, văn bản từ các nguồn không chính thống như mạng xã hội hay diễn đàn còn có thể dùng những từ viết tắt, lỗi chính tả hay ký tự đặc biệt để thể hiện cho các đối tượng. Vì vậy, để tạo điều kiện cho máy tính khi phân tích dữ liệu dạng văn bản, giảm thiểu sự nhập nhằng và dư thừa dữ liệu, một phương pháp để hợp nhất các thực thể cùng tham chiếu đến một đối tượng là cần thiết trong các hệ thống tự động xử lý, phân tích và khai phá thông tin báo chí, mạng xã hội.

Trong quá trình phát triển của xử lý ngôn ngữ tự nhiên, vấn đề này đã được giải quyết với nhiều hướng tiếp cận khác nhau. Tuy nhiên, các phương pháp từ truyền thống

như tra từ điển, phân loại dựa trên độ tương đồng về ký tự, cho đến sử dụng các mô hình học máy đã bộc lộ nhược điểm khi mức độ đa dạng của từ ngữ ngày càng tăng cao. Trong sáng chế này, giải pháp kỹ thuật được đề xuất với phương pháp định danh các thực thể không chỉ dựa trên các đặc trưng về từ ngữ mà còn sử dụng ngữ cảnh của chúng. Do đó, giải pháp có thể khắc phục được các nhược điểm còn tồn tại và dễ dàng thích nghi với sự phát triển của ngôn từ.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp hợp nhất các thực thể cùng tham chiếu đến một đối tượng thực tế trong văn bản, thông qua việc liên kết chúng với các định danh tương ứng. Phương pháp bao gồm bốn bước chính được thực hiện như sau:

Bước 1: xây dựng hệ cơ sở tri thức. Tại bước này, tập hợp của các đối tượng khác nhau trong thực tế. Mỗi đối tượng đều có một mã định danh riêng biệt, tên chính thức, các tên gọi khác và đoạn văn mô tả nó.

Bước 2: truy vấn đối tượng ứng viên. Đầu tiên, dữ liệu đầu vào là tên thực thể và ngữ cảnh xung quanh được tiền xử lý bao gồm tách từ, số hóa từ, sắp xếp các từ mã hóa theo thứ tự hợp lý và véc-tơ hóa bằng mô hình học sâu. Sau đó, các véc-tơ đặc trưng đầu vào lần lượt được so sánh với đặc trưng của các đối tượng trong hệ tri thức bằng phép so sánh khoảng cách véc-tơ. Các ứng viên được lựa chọn là những đối tượng có độ tương đồng cao với thực thể đầu vào.

Bước 3: xếp hạng lại các đối tượng ứng viên. Bước này nhận đầu vào là thực thể đầu vào và các ứng viên từ kết quả của bước 2. Tương tự, quá trình tiền xử lý dữ liệu được áp dụng bao gồm tách từ, số hóa từ và sắp xếp các từ mã hóa theo thứ tự hợp lý. Ở bước sắp xếp, các véc-tơ số hóa của thực thể đầu vào và ứng viên được nối với nhau. Sau đó chúng được đưa qua mô hình học sâu và mạng nơ-ron để tính toán giá trị tương đồng cho mỗi cặp thực thể - đối tượng đang xét. Kết quả trả về của bước này là một bảng xếp hạng từ cao xuống thấp các đối tượng theo độ tương đồng với thực thể đang xét.

Bước 4: kiểm tra tính tồn tại của thực thể trong cơ sở tri thức. Vì hệ thống chỉ tập trung vào một số đối tượng nhất định được định nghĩa sẵn, thực thể được nhắc đến trong văn bản có thể không thuộc hệ cơ sở tri thức. Do đó, quá trình kiểm tra tính tồn tại của nó là cần thiết. Mô hình này nhận đầu vào là các thông tin của thực thể được xếp hạng cao

nhất từ bước 3, bao gồm điểm và các tên gọi khác của thực thể đó, nhằm sinh ra một trong hai kết quả phân loại: nhãn "1" tức thực thể đó có tồn tại trong hệ cơ sở tri thức và kết quả cuối cùng là định danh của ứng viên được xếp hạng cao nhất; nhãn "0" là ngược lại và kết quả toàn bộ quá trình là rỗng (không tìm được đối tượng tương ứng).

Trong sáng chế này, từ đầu vào là tên thực thể được trích xuất từ văn bản và ngữ cảnh là các câu văn xung quanh nó, phương pháp tập trung truy vấn định danh thích hợp nhất cho đối tượng này từ hệ cơ sở tri thức. Trong đó, bước xây dựng hệ cơ sở tri thức là quan trọng với yêu cầu cả về chất lượng và độ đa dạng. Vì số lượng thực thể trong thực tế là rất lớn, bước truy vấn các ứng viên giúp giới hạn không gian tìm kiếm và kết quả tính toán. Bước xếp hạng ứng viên có chức năng so sánh tương đồng giữa các thực thể khi đặt véc-tơ mã hóa của chúng cạnh nhau, nhằm tăng cường ý nghĩa về ngữ cảnh và độ chính xác cho kết quả. Tham số của các mô hình học sâu được cập nhật trong quá trình huấn luyện, giúp tối ưu quá trình tạo đặc trưng cho dữ liệu. Đầu ra của phương pháp là một định danh cụ thể cho mỗi thực thể đầu vào; hoặc định danh rỗng, đồng nghĩa với thực thể đó không tồn tại trong hệ cơ sở tri thức. Từ đó, mục tiêu ban đầu của phương pháp có thể được thực hiện, đó là hợp nhất các thực thể đồng tham chiếu đến một đối tượng, thông qua việc nhóm các thực thể dựa trên mã định danh vừa tìm được.

Mô tả vắn tắt các hình vẽ

Hình 1: hình vẽ tổng quát toàn bộ phương pháp hợp nhất các thực thể đồng tham chiếu đến một đối tượng trong văn bản;

Hình 2: hình vẽ miêu tả quá trình truy vấn các đối tượng ứng viên từ hệ cơ sở tri thức;

Hình 3: hình vẽ miêu tả quá trình xếp hạng lại các đối tượng ứng viên;

Hình 4: hình vẽ mô tả quá trình kiểm tra tính tồn tại của thực thể trong hệ cơ sở tri thức.

Mô tả chi tiết sáng chế

Phương pháp được đề xuất trong sáng chế giúp hợp nhất các thực thể được trích xuất từ văn bản và đồng tham chiếu đến một đối tượng. Quá trình tổng quát được mô tả ở Hình 1. Các bước thực hiện chi tiết được trình bày cụ thể như sau:

Bước 1: xây dựng hệ cơ sở tri thức.

Hệ cơ sở tri thức là một cơ sở dữ liệu, lưu trữ thông tin về các thực thể. Với mục đích có được một nền tảng cho việc định danh thực thể, hệ tri thức được xây dựng với yêu cầu phong phú về số lượng và chất lượng. Mỗi phần tử trong hệ tương ứng với một đối tượng trong thực tế hay trong các văn bản. Các thành phần này được đại diện bởi một mã định danh và tên đại diện. Bên cạnh đó, để giúp phân biệt các từ đồng âm nhưng khác nghĩa, cũng như tăng cường thông tin cho việc truy vấn, một đoạn mô tả về các đối tượng cũng sẽ được lưu trữ đồng thời với tên định danh. Với mô tả này, người xử lý cũng như máy tính sẽ hiểu rõ hơn về đối tượng và phân biệt chúng dễ dàng hơn. Đồng thời, những tên gọi khác của mỗi đối tượng cũng được xây dựng và bổ sung vào hệ cơ sở tri thức để phục vụ cho các quá trình phân loại phía sau.

Các thực thể trong hệ cơ sở tri thức được khởi tạo bằng cách thu thập từ các nguồn mở có tính đa dạng và độ uy tín cao, ví dụ như *wiki*. Ở những nguồn này, một thực thể thường được lưu trữ bởi tên và các thông tin liên quan như mô tả hay khái niệm, giải thích ý nghĩa, lịch sử, tiểu sử, địa lý. Trong phương pháp này, tên đại diện, các tên gọi khác và đoạn mô tả của thực thể từ các nguồn mở được thu thập để xây dựng nên hệ cơ sở tri thức. Ví dụ, với thực thể "Thành phố Hồ Chí Minh" sẽ được lưu trữ trong hệ với các thông tin như sau:

{

"ma_dinh_danh": "59",

"ten": "Thành phố Hồ Chí Minh ",

"mo_ta": "Thành phố Hồ Chí Minh (viết tắt TP.HCM), còn gọi bằng tên cũ phổ biến là Sài Gòn, là thành phố lớn nhất Việt Nam và là một siêu đô thị trong tương lai gần. Đây còn là trung tâm kinh tế, chính trị, văn hóa, giải trí và giáo dục tại Việt Nam. Thành phố Hồ Chí Minh là thành phố trực thuộc trung ương thuộc loại đô thị đặc biệt của Việt Nam.[8] Nằm trong vùng chuyển tiếp giữa Đông Nam Bộ và Tây Nam Bộ, thành phố này hiện có 16 quận, 1 thành phố và 5 huyện, tổng diện tích 2.095 km2 (809 dặm vuông Anh)."

"ten_goi_khac": ["HCMC", "TP.HCM", "Sài Gòn"]

}

Bước 2: truy vấn các đối tượng ứng viên.

Bước này có chức năng tìm các đối tượng ứng viên gần giống với đối tượng đầu vào từ một lượng lớn phần tử có trong hệ cơ sở tri thức. Quá trình nhận đầu vào là tên và ngữ cảnh của đối tượng cần định danh. Kết quả là danh sách K đối tượng hàng đầu được truy vấn từ hệ tri thức. Để so sánh được các thực thể với nhau, sau các giai đoạn tiền xử lý, một mô hình học sâu được sử dụng để tạo các véc-tơ đặc trưng. Các véc-tơ này được tính độ tương đồng bằng phép tính tích vô hướng. Các bước thực hiện được mô tả trong Hình 2, cụ thể:

Ngữ cảnh của thực thể đầu vào là các câu văn đứng trước và sau nó trong văn bản. Đầu tiên, thực thể và ngữ cảnh được đi qua mô hình tách từ và mã hóa từ, dựa trên một bộ từ vựng có sẵn. Bộ từ vựng này bao gồm các từ đã xuất hiện của một lượng lớn văn bản và mã số tương ứng cho mỗi phần tử. Mô hình tách từ hoạt động theo nguyên tắc sao cho từ được tách ra là từ con dài nhất của từ đang xét và tồn tại trong bộ từ vựng. Phép biến đổi chữ hoa thành chữ thường cũng được áp dụng ở bước này. Sau đó, các từ đã được tách sẽ đi qua bộ mã hóa thành dạng số giúp máy tính tính toán được.

Xét ví dụ cụm từ : *"Ho Chi Minh City (abbreviated HCMC; Vietnamese: Thành phố Hồ Chí Minh)"* sẽ được tách thành các từ *"ho", "chi", "minh", "city", "(", "abbreviated", "hc", "##mc", ";", "vietnamese", ":", "than", "##h", "ph", "##o", "ho", "chi", "minh", ')'*.

Với từ *"hcmc"* có từ con dài nhất xuất hiện trong bộ từ vựng là *"hc"*, nên nó được chia ra thành *"hc"* và *"##mc"*, do đó có kết quả giống như ví dụ trên. Xét từ *"##mc"*, vì có tồn tại trong bộ từ vựng nên kết quả của việc số hóa là *[16731, 12458]* tương ứng với *["hc", "##mc"]*. Nếu các từ con không xuất hiện trong bộ từ vựng, chúng sẽ được số hóa với mã số *[100]* và ký tự *"[UNK]"*.

Tiếp theo, một bước sắp xếp các từ dưới dạng mã hóa được áp dụng nhằm xây dựng đầu vào cho mô hình học sâu, các ký tự đặc biệt được thêm vào sau đây cũng đã được số hóa:

Đối với thực thể và ngữ cảnh đầu vào, chúng được sắp xếp theo công thức: $[CLS]\ ngu\_canh\_trai\ [Ms]\ thuc\_the\ [Me]\ ngu\_canh\_phai\ [SEP]$. Trong đó *[CLS]* và *[SEP]* lần lượt là ký tự đánh dấu bắt đầu và kết thúc câu. ngu_canh_trai,

ngu_canh_phai lần lượt là các từ của câu văn nằm bên trái và bên phải của thực thể, $thuc\_the$ là các từ của thực thể đó, *[Ms]* và *[Me]* lần lượt là ký tự đánh dấu bắt đầu và kết thúc của thực thể.

Đối với các thực thể và mô tả trong hệ tri thức, chúng được sắp xếp theo công thức: $[CLS]\ doi\_tuong\ [ENT]\ mo\_ta\ [SEP]$. Trong đó, *[CLS]* và *[SEP]* lần lượt là ký tự đánh dấu bắt đầu và kết thúc câu, doi_tuong là các từ của tên đối tượng, *[ENT]* là ký hiệu giúp phân biệt thực thể và đoạn mô tả, mo_ta là các từ của đoạn mô tả.

Sau khi sắp xếp, kết quả có được là véc-tơ mã số v_e cho thực thể đầu vào và véc-tơ mã số v_c cho đối tượng ứng viên. Vì giải pháp đề xuất sử dụng ngữ cảnh để biểu diễn các thực thể, một mô hình học sâu T_1 được áp dụng giúp trích xuất các thông tin ngữ nghĩa dưới dạng các véc-tơ đặc trưng. Các véc-tơ mã số được đưa qua mô hình học sâu thu được kết quả tương ứng $y_e = T_1(v_e)$ và $y_c = T_1(v_c)$. Trong quá trình huấn luyện, tham số của mô hình T_1 sẽ được cập nhật để phù hợp cho dữ liệu đang có.

Sau khi qua mô hình học sâu, phép tích vô hướng được sử dụng để tính toán khoảng cách giữa các véc-tơ đặc trưng của thực thể đầu vào và các ứng viên từ hệ cơ sở tri thức. Khoảng cách giữa các véc-tơ càng nhỏ chứng tỏ thực thể và đối tượng này càng giống nhau. Công thức tính được thể hiện như sau:

$$s(v_e, v_c) = y_e \cdot y_c$$

Trong đó: y_e và y_c lần lượt là các véc-tơ đặc trưng của thực thể đầu vào và véc-tơ đặc trưng của ứng viên từ hệ cơ sở tri thức.

Trong quá trình huấn luyện, hàm mất mát L_1 được sử dụng cho kết quả của việc tính toán khoảng cách giữa các véc-tơ, giúp tối ưu các tham của mô hình học sâu.

$$L_1(v_e, v_c) = -s(v_e, v_c) + log \sum_{j=1}^{B} \exp\left(s(v_e, v_{c_i})\right)$$

Trong đó, B là kích thước của khối đang huấn luyện.

Sau bước này, phương pháp dự đoán được K ứng viên phù hợp với thực thể đầu vào, với K là một siêu tham số được tùy chỉnh thích hợp khi huấn luyện. Lấy "TPHCM" trong câu "TPHCM là một đô thị lớn ở Việt Nam" làm ví dụ, kết quả của bước này có thể là một danh sách các ứng viên: ["Chủ tịch Hồ Chí Minh", "Thành phố Hồ Chí Minh", "Đường

mòn Hồ Chí Minh"]. Mặc dù bước này có điểm và xếp hạng các đối tượng, nhưng các giá trị và thứ tự này vẫn chưa đủ tin cậy và cần phải xếp hạng lại một lần nữa.

Bước 3: xếp hạng lại các đối tượng ứng viên.

Quá trình xếp hạng lại ứng viên có chức năng sắp xếp K ứng viên có được từ bước 2 theo mức độ tương đồng với thực thể đang xét nhằm tìm ra ứng viên phù hợp nhất. Đầu vào là các ứng viên và đối tượng đang xét. Các véc-tơ số hóa của thực thể và ứng viên được nối với nhau trong bước này và được đưa qua mô hình học sâu, giúp so sánh tốt hơn thông tin đặc trưng giữa các ngữ cảnh và đoạn mô tả. Kết quả của bước này là điểm của từng đối tượng ứng viên được sắp xếp từ cao xuống thấp. Quá trình thực hiện được miêu tả trong Hình 3, cụ thể:

Tương tự ở bước 2, kỹ thuật tiền xử lý được áp cho thực thể cùng ngữ cảnh và mô tả của chúng. Khi đã có được các từ dưới dạng số hóa, phương pháp tiến hành sắp xếp chúng theo quy tắc sau để xây dựng đầu vào cho mô hình học sâu:

$$[CLS]\ ngu\_canh\_trai\ [Ms]\ thuc\_the\ [Me]\ ngu\_canh\_phai\ [SEP]$$
$$doi\_tuong\ \ [ENT]\ mo\_ta\ [SEP]$$

Trong đó, $[CLS]$ và $[SEP]$ là các ký tự đánh dấu bắt đầu và kết thúc câu; ngu_canh_trai, ngu_canh_phai: là các từ của các câu văn bên trái và bên phải của thực thể, $[Ms]$ và $[Me]$ là cá ký hiệu đánh dấu bắt đầu và kết thúc của thực thể, $thuc\_the$ là các từ của thực thể cần định danh, doi_tuong là các từ của đối tượng ứng viên, $[ENT]$ là ký tự phân biệt giữa thực thể và đoạn mô tả, mo_ta là các từ của đoạn mô tả ứng viên.

Cách sắp xếp này cho phép mô hình học sâu có sự chú ý chéo giữa ngữ cảnh và mô tả của các thực thể, nhằm tăng cường đặc trưng ngữ nghĩa. Kết quả quá trình sắp xếp của một cặp thực thể e và ứng viên c là một véc-tơ số hóa, ký hiệu $\tau_{e,c}$.

Một mô hình học sâu T_2 được sử dụng để rút trích đặc trưng cho véc-tơ mã số, kết quả thu được là: $f = T_2(\tau_{e,c})$. Tham số của mô hình T_2 cũng được cập nhật trong quá trình huấn luyện.

Để tính độ tương đồng cho ứng viên, véc-tơ đặc trưng f được cho qua một lớp nơ-ron tuyến tính W, thu được kết quả: $s = W(f)$. Lớp W này có tác dụng kết hợp toàn bộ đặc trưng từ đầu ra của mô hình học sâu. Nếu đầu vào có K ứng viên, lớp tuyến tính W sẽ

có kích thước đầu ra là K. Danh sách giá trị trả về $s = [s_1, s_2, \dots s_K]$ cũng đồng thời là điểm của các ứng viên. Cuối cùng, các đối tượng được xếp hạng từ cao xuống thấp dựa trên điểm, với ý nghĩa rằng điểm càng cao thì thực thể càng có khả năng là đối tượng đó.

Trong quá trình huấn luyện mô hình T_2, một hàm mất mát hỗn loạn chéo L_2 được áp dụng nhằm tối ưu hóa tham số cho mô hình học sâu.

$$L_2(s) = -\sum_{i}^{K} y_i \,.\, log(1 - y_i)$$

Với K là kích thước của khối huấn luyện, y_i là "1" nếu ứng viên i là định danh cần tìm, và "0" nếu không phải.

Để minh họa cho quá trình này, ta tiếp tục lấy "TPHCM" trong câu "TPHCM là một đô thị lớn ở Việt Nam" làm ví dụ. Kết quả của bước truy vấn là một danh sách các ứng viên: ["Chủ tịch Hồ Chí Minh", "Thành phố Hồ Chí Minh", "Đường mòn Hồ Chí Minh"]. Tuy nhiên, sau bước xếp hạng này, thứ tự các ứng viên sẽ có thể được thay đổi thành ["Thành phố Hồ Chí Minh", "Chủ tịch Hồ Chí Minh", "Đường mòn Hồ Chí Minh"].

Bước 4: kiểm tra tính tồn tại của thực thể trong hệ cơ sở tri thức.

Số đối tượng trong hệ cơ sở tri thức là hữu hạn, do đó một thực thể được đề cập đến trong văn bản có thể không tương ứng với đối tượng nào đã được định nghĩa trước. Để nhận biết tính chất này, một mô hình bổ sung được thêm vào, bao gồm hai bước xử lý nhỏ: tính đặc trưng và phân loại. Trong đó, mô hình này nhận đầu vào là điểm của ứng viên có xếp hạng cao nhất từ bước 3 và các tên gọi khác của nó, để trích xuất một véc-tơ đặc trưng ba chiều nhằm đưa vào mô hình phân loại có giám sát. Quá trình phân loại sinh ra một trong hai kết quả: "1" tương ứng với thực thể đang xét có tồn tại trong hệ cơ sở trí thức; "0" là ngược lại. Nếu có tồn tại, phương pháp có thể kết luận thực thể đang xét là tương ứng với đối tượng đứng đầu bảng xếp hạng. Nếu không, phương pháp chỉ trả về một định danh rỗng, tức thực thể này không có trong danh sách những đối tượng đã được định nghĩa sẵn. Bước này được mô tả ở Hình 4 với những quá trình như sau:

Đầu tiên, một véc-tơ được xây dựng chứa ba đặc trưng cho ứng viên có điểm cao nhất: một là điểm tương đồng lớn nhất từ bước 2 là s_{max}, hai là điểm *Levenshtein* (sau đây được ký hiệu là s_{lev}), ba là điểm *Dice* (sau đây được ký hiệu là s_{dice}). Trong đó, điểm

Levenshtein và điểm *Dice* đặc trưng cho độ tương đồng về chuỗi ký tự của tên thực thể đầu vào và tên của đối tượng ứng viên tương ứng. Các điểm s_{lev} và s_{dice} được tính toán dựa trên việc so sánh giữa thực thể và từng tên gọi khác của đối tượng, sau đó chọn giá trị lớn nhất, với cách tính lần lượt là:

Khoảng cách *Levenshtein* $lev(a, b)$ giữa hai chuỗi ký tự a và b được tính theo theo công thức:

$$lev(a,b) = \begin{cases} |a|, & nếu\ |b| = 0 \\ |b|, & nếu\ |a| = 0 \\ lev\big(tail(a), tail(b)\big), & nếu\ a[0] = b[0] \\ 1 + min\begin{cases} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b)) \end{cases}, & còn\ lại \end{cases}$$

Trong đó, $lev(a, b)$ là khoảng cách *Levenshtein* của hai chuỗi ký tự a và b, $|x|$ là chiều dài chuỗi ký tự x, $tail(x)$ là chuỗi ký tự x nhưng không chứa ký tự đầu tiên, $x[0]$ là ký tự đầu tiên của chuỗi x.

Điểm tương đồng *Levenshtein* chuẩn hóa s_{lev} giữa thực thể có tên N_{tt} và đối tượng có danh sách gồm V tên $[N_1, N_2, \dots, N_V]$, với $|x|$ là chiều dài chuỗi ký tự x; $lev(a, b)$ là khoảng cách *Levenshtein* của hai chuỗi ký tự a và b; và η là một hằng số dương nhỏ để tránh việc chia 0:

$$s_{lev} = \max_{i=0 \to V} \left(\frac{lev(N_{tt}, N_i)}{\max\left(|N_{tt}|, |N_i|\right) + \eta}\right)$$

Điểm *Dice* chuẩn hóa s_{dice} giữa thực thể có tên N_{tt} và đối tượng có danh sách gồm V tên $[N_1, N_2, \dots, N_V]$ được tính như sau:

$$s_{dice} = \max_{i=0 \to V} \left(\frac{2 \times |N_{tt} \cup N_i|}{|N_{tt}| + |N_i| + \eta}\right)$$

Trong đó, $a \cup b$ là phần giao giữa hai chuỗi ký tự a và b, $|x|$ là chiều dài chuỗi ký tự x, η là một hằng số dương nhỏ để tránh việc chia 0.

Trong các đặc trưng trên, s_{max} là số thực tầm giá trị không giới hạn và nó được cho là cao khi mang giá trị dương. Đối với điểm *Levenshtein* chuẩn hóa s_{lev} và điểm *Dice* chuẩn hóa s_{dice}, tầm giá trị của chúng là từ 0 đến 1. Các đặc trưng trên có giá trị càng lớn

nghĩa là độ tương đồng càng cao, và đó cũng là dấu hiệu để nhận biết thực thể đó có tồn tại trong hệ cơ sở tri thức hay không.

Tiếp theo, một véc-tơ ba chiều $[s_{max}; s_{lev}; s_{dice}]$ được đưa vào một mô hình học có giám sát đã được huấn luyện từ trước để đưa ra kết luận cuối cùng cho kết quả cho toàn bộ quá trình. Mô hình học có thể là mô hình máy véc-tơ hỗ trợ (Support Vector Machine – SVM) hoặc cây quyết định (Decision Tree).

Giả sử thực thể "Đại học Bình Dương" không tồn tại trong hệ cơ sở tri thức của hệ thống. Tuy nhiên, sau khi thực hiện đến bước 3, hệ thống vẫn sinh ra một bảng xếp hạng với thứ tự lần lượt là ["Bình Dương", "Becamex Bình Dương", "Thành phố mới Bình Dương"] với số điểm tương ứng $[-20.67; -42.89, -87.12]$. Phần tử xếp hạng cao nhất là "Bình Dương" được tiếp tục xét trong bước 4, với đặc trưng s_{max}, s_{lev}, s_{dice} lần lượt là $[-20.67; 0.35; 0.55]$. Với các phần tử trong véc-tơ đặc trưng đều thấp, bộ phân loại được mong đợi rằng sẽ dự đoán kết quả là 0. Ngược lại, đối với ví dụ thực thể "TPHCM" đã được đề cập ở các bước trước, véc-tơ đặc trưng tính toán được là $[15.3; 0.95; 1]$ đều chứa các phần tử có giá trị lớn, nên mô hình học máy có thể khẳng định thực thể này tồn tại trong hệ cơ sở tri thức.

Yêu cầu bảo hộ

1. Phương pháp hợp nhất các thực thể trong văn bản ứng dụng học sâu bao gồm bốn bước:

bước 1: xây dựng hệ cơ sở tri thức;

các đối tượng trong hệ cơ sở tri thức được khởi tạo bằng cách thu thập từ các nguồn mở có tính đa dạng và độ uy tín cao; ở những nguồn này, một thực thể thường được lưu trữ bởi tên và các thông tin liên quan như mô tả hay khái niệm, giải thích ý nghĩa, lịch sử, tiểu sử, địa lý; trong phương pháp này, tên đại diện, các tên gọi khác và đoạn mô tả của thực thể từ các nguồn mở được thu thập để xây dựng nên hệ cơ sở tri thức;

bước 2: truy vấn các đối tượng ứng viên;

bước này có chức năng tìm các đối tượng ứng viên gần giống với đối tượng đầu vào từ một lượng lớn phần tử có trong hệ cơ sở tri thức; quá trình nhận đầu vào là tên và ngữ cảnh của đối tượng cần định danh; kết quả là danh sách K đối tượng hàng đầu được truy vấn từ hệ tri thức; quá trình này bao gồm các giai đoạn:

đầu tiên, thực thể và ngữ cảnh xung quanh nó được đi qua mô hình tách từ và mã hóa từ, dựa trên một bộ từ vựng có sẵn; bộ từ vựng này bao gồm các từ đã xuất hiện của một lượng lớn văn bản và mã số tương ứng cho mỗi phần tử; mô hình tách từ hoạt động theo nguyên tắc sao cho từ được tách ra là từ con dài nhất của từ đang xét và tồn tại trong bộ từ vựng; phép biến đổi chữ hoa thành chữ thường cũng được áp dụng ở bước này; sau đó, các từ đã được tách sẽ đi qua bộ mã hóa thành dạng số giúp máy tính tính toán được;

tiếp theo, một bước sắp xếp các từ dưới dạng mã hóa được áp dụng nhằm xây dựng đầu vào cho mô hình học sâu, các ký tự đặc biệt được thêm vào sau đây cũng đã được số hóa:

đối với thực thể và ngữ cảnh đầu vào, chúng sắp xếp theo công thức: $[CLS]\ ngu\_canh\_trai\ [Ms]\ thuc\_the\ [Me]\ ngu\_canh\_phai\ [SEP]$; trong đó $[CLS]$ và $[SEP]$ lần lượt là ký tự đánh dấu bắt đầu và kết thúc câu; ngu_canh_trai, ngu_canh_phai lần lượt là các từ của câu văn nằm

bên trái và bên phải của thực thể, $thuc\_the$ là các từ của thực thể đó, $[Ms]$ và $[Me]$ lần lượt là ký tự đánh dấu bắt đầu và kết thúc của thực thể;

đối với các đối tượng và mô tả trong hệ tri thức, chúng sắp xếp theo công thức: $[CLS]\ doi\_tuong\ [ENT]\ mo\_ta\ [SEP]$; trong đó, $[CLS]$ và $[SEP]$ lần lượt là ký tự đánh dấu bắt đầu và kết thúc câu, doi_tuong là các từ của tên đối tượng, *[ENT]* là ký hiệu giúp phân biệt thực thể và đoạn mô tả, mo_ta là các từ của đoạn mô tả;

sau khi sắp xếp, kết quả có được là véc-tơ mã số v_e cho thực thể đầu vào và véc-tơ mã số v_c cho đối tượng ứng viên; sau đó, một mô hình học sâu T_1 được áp dụng giúp trích xuất các thông tin ngữ nghĩa dưới dạng các véc-tơ đặc trưng; các véc-tơ mã số được đưa qua mô hình học sâu thu được kết quả tương ứng $y_e = T_1(v_e)$ và $y_c = T_1(v_c)$;

sau khi qua mô hình học sâu, phép tích vô hướng được sử dụng để tính toán khoảng cách giữa các véc-tơ đặc trưng của thực thể đầu vào và các ứng viên từ hệ cơ sở tri thức; khoảng cách giữa các véc-tơ càng nhỏ chứng tỏ cặp thực thể - đối tượng này càng giống nhau; công thức tính được thể hiện như sau:

$$s(v_e, v_c) = y_e \cdot y_c$$

trong đó: y_e và y_c lần lượt là các véc-tơ đặc trưng của thực thể đầu vào và véc-tơ đặc trưng của ứng viên từ hệ cơ sở tri thức;

trong quá trình huấn luyện, hàm mất mát L_1 được sử dụng cho kết quả của việc tính toán khoảng cách giữa các véc-tơ, giúp tối ưu các tham của mô hình học sâu;

$$L_1(v_e, v_c) = -s(v_e, v_c) + log \sum_{j=1}^{B} \exp\left(s(v_e, v_{c_i})\right)$$

trong đó, B là kích thước của khối đang huấn luyện;

bước 3: xếp hạng lại các đối tượng ứng viên;

quá trình xếp hạng lại ứng viên có chức năng sắp xếp K ứng viên từ bước 2 theo mức độ tương đồng với thực thể đang xét nhằm tìm ra ứng viên phù hợp nhất; đầu vào là

các ứng viên và đối tượng đang xét; kết quả của bước này là điểm của từng đối tượng ứng viên được sắp xếp từ cao xuống thấp; các giai đoạn trong bước này bao gồm:

kỹ thuật tiền xử lý tương tự như ở bước 2 được áp cho thực thể cùng ngữ cảnh và mô tả của chúng; khi đã có được các từ dưới dạng số hóa, phương pháp tiến hành sắp xếp chúng theo quy tắc sau để xây dựng đầu vào cho mô hình học sâu:

$[CLS]\ ngu\_canh\_trai\ [Ms]\ thuc\_the\ [Me]\ ngu\_canh\_phai\ [SEP]$

$doi\_tuong\ [ENT]\ mo\_ta\ [SEP]$

trong đó, $[CLS]$ và $[SEP]$ là các ký tự đánh dấu bắt đầu và kết thúc câu; ngu_canh_trai, ngu_canh_phai: là các từ của các câu văn bên trái và bên phải của thực thể, $[Ms]$ và $[Me]$ là cá ký hiệu đánh dấu bắt đầu và kết thúc của thực thể, $thuc\_the$ là các từ của thực thể cần định danh, doi_tuong là các từ của đối tượng ứng viên, $[ENT]$ là ký tự phân biệt giữa thực thể và đoạn mô tả, mo_ta là các từ của đoạn mô tả ứng viên;

kết quả quá trình sắp xếp của một cặp thực thể e và ứng viên c là một véc-tơ số hóa, ký hiệu $\tau_{e,c}$;

sau đó, một mô hình học sâu T_2 được sử dụng để rút trích đặc trưng cho véc-tơ mã số, kết quả thu được là: $f = T_2(\tau_{e,c})$; tham số của mô hình T_2 cũng được cập nhật trong quá trình huấn luyện;

để tính độ tương đồng cho ứng viên, véc-tơ đặc trưng f được cho qua một lớp nơ-ron tuyến tính W, thu được kết quả: $s = W(f)$; lớp W này có tác dụng kết hợp toàn bộ đặc trưng từ đầu ra của mô hình học sâu. Nếu đầu vào có K ứng viên, lớp tuyến tính W sẽ có kích thước đầu ra là K; danh sách giá trị trả về $s = [s_1, s_2, \dots s_K]$ cũng đồng thời là điểm của các ứng viên; cuối cùng, các đối tượng được xếp hạng từ cao xuống thấp dựa trên điểm, với ý nghĩa rằng điểm càng cao thì thực thể càng có khả năng là đối tượng đó;

trong quá trình huấn luyện mô hình T_2, một hàm mất mát hỗn loạn chéo L_2 được áp dụng nhằm tối ưu hóa tham số cho mô hình học sâu;

$$L_2(s) = -\sum_{i}^{K} y_i \,.\, log(1 - y_i)$$

với K là kích thước của khối huấn luyện, y_i là "1" nếu ứng viên i là định danh cần tìm, và "0" nếu không phải;

bước 4: kiểm tra tính tồn tại của thực thể trong hệ cơ sở tri thức;

số đối tượng trong hệ cơ sở tri thức là hữu hạn, do đó một thực thể được đề cập đến trong văn bản có thể không tương ứng với đối tượng nào đã được định nghĩa trước; để nhận biết tính chất này, một mô hình bổ sung được thêm vào, bao gồm hai bước xử lý nhỏ: tính đặc trưng và phân loại; trong đó, mô hình này nhận đầu vào là điểm của ứng viên có xếp hạng cao nhất từ bước 3 và các tên gọi khác của nó, để trích xuất một véc-tơ đặc trưng ba chiều nhằm đưa vào mô hình phân loại có giám sát; quá trình phân loại sinh ra một trong hai kết quả: "1" tương ứng với thực thể đang xét có tồn tại trong hệ cơ sở trí thức; "0" là ngược lại; nếu có tồn tại, phương pháp có thể kết luận thực thể đang xét là tương ứng với đối tượng đứng đầu bảng xếp hạng; nếu không, phương pháp chỉ trả về một định danh rỗng, tức thực thể này không có trong danh sách những đối tượng đã được định nghĩa sẵn; bước này bao gồm các quá trình sau:

đầu tiên, một véc-tơ được xây dựng chứa ba đặc trưng cho ứng viên có điểm cao nhất: một là điểm tương đồng lớn nhất từ bước 2 là s_{max}, hai là điểm *levenshtein* (sau đây được ký hiệu là s_{lev}), ba là điểm *dice* (sau đây được ký hiệu là s_{dice}); trong đó, điểm *levenshtein* và điểm *dice* đặc trưng cho độ tương đồng về chuỗi ký tự của tên thực thể đầu vào và tên của đối tượng ứng viên tương ứng; các điểm s_{lev} và s_{dice} được tính toán dựa trên việc so sánh giữa thực thể và từng tên gọi khác của đối tượng, sau đó chọn giá trị lớn nhất, với cách tính lần lượt là:

khoảng cách *levenshtein* $lev(a, b)$ giữa hai chuỗi ký tự a và b được tính theo theo công thức:

$$lev(a,b) = \begin{cases} |a|, & nếu\ |b| = 0 \\ |b|, & nếu\ |a| = 0 \\ lev\big(tail(a), tail(b)\big), & nếu\ a[0] = b[0] \\ 1 + min\begin{cases} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b)) \end{cases}, & còn\ lại \end{cases}$$

trong đó, $lev(a, b)$ là khoảng cách *levenshtein* của hai chuỗi ký tự a và b, $|x|$ là chiều dài chuỗi ký tự x, $tail(x)$ là chuỗi ký tự x nhưng không chứa ký tự đầu tiên, $x[0]$ là ký tự đầu tiên của chuỗi x;

điểm tương đồng *levenshtein* chuẩn hóa s_{lev} giữa thực thể có tên N_{tt} và đối tượng có danh sách gồm V tên $[N_1, N_2, \dots, N_V]$, với $|x|$ là chiều dài chuỗi ký tự x; $lev(a, b)$ là khoảng cách *levenshtein* của hai chuỗi ký tự a và b; và η là một hằng số dương nhỏ để tránh việc chia 0:

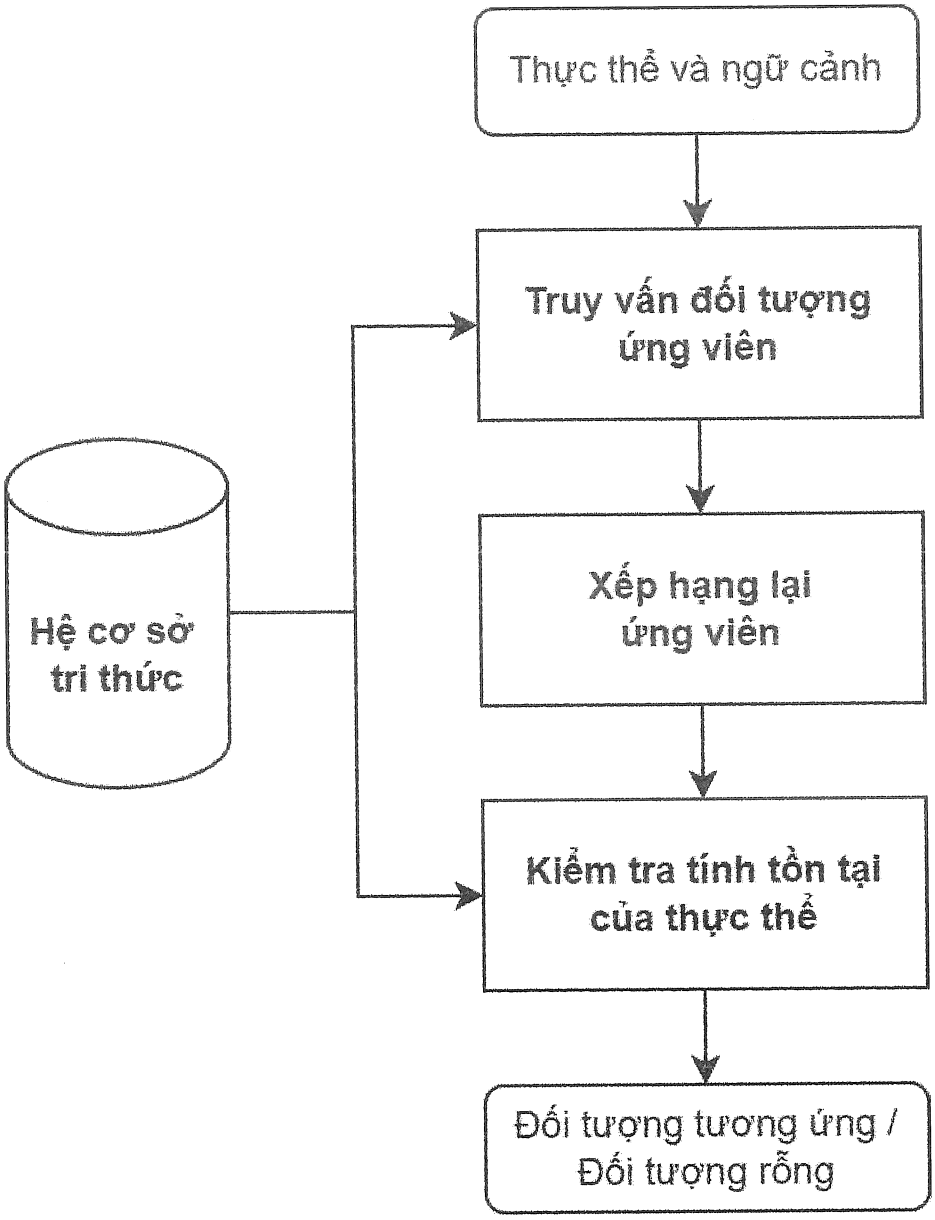
$$s_{lev} = \max_{i=0 \to V} \left(\frac{lev(N_{tt}, N_i)}{\max(|N_{tt}|, |N_i|) + \eta}\right)$$

điểm *dice* chuẩn hóa s_{dice} giữa thực thể có tên N_{tt} và đối tượng có danh sách gồm V tên $[N_1, N_2, \dots, N_V]$ được tính như sau:

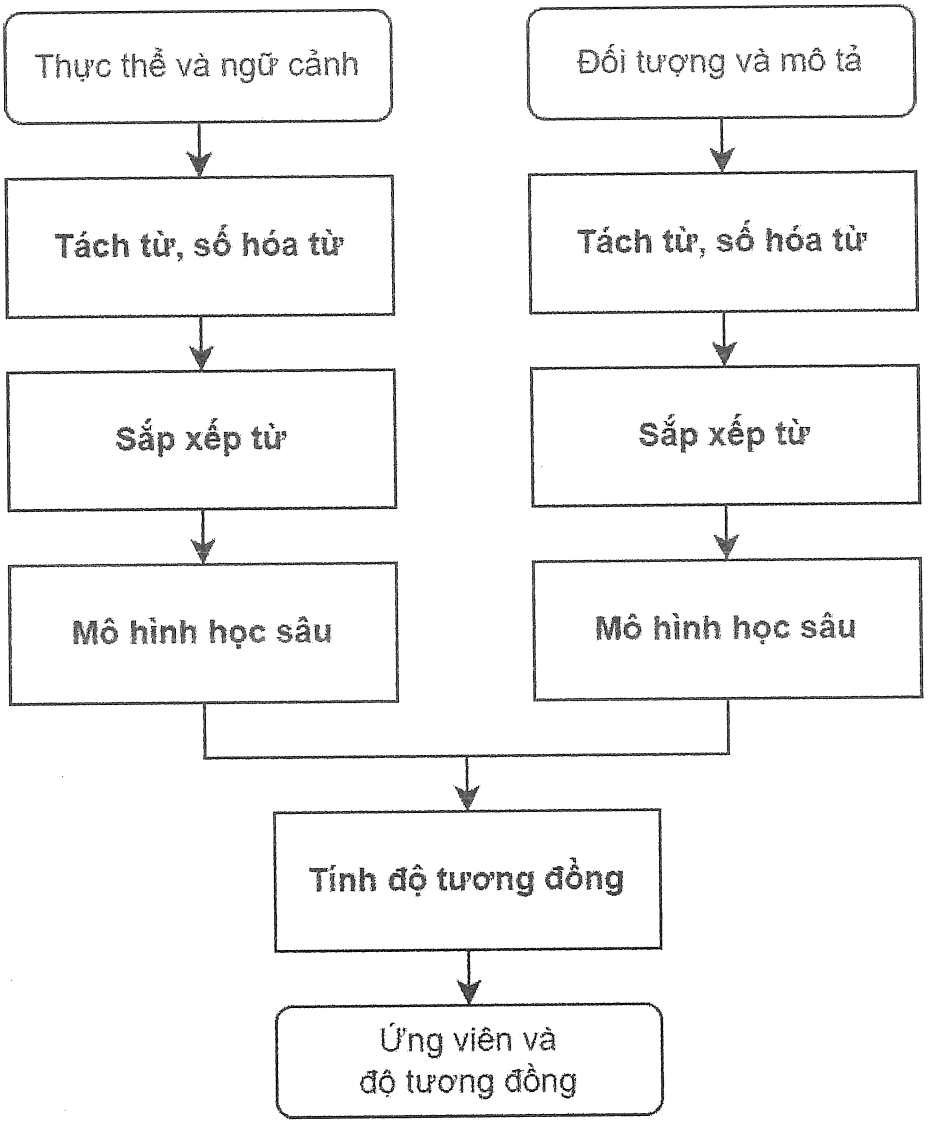
$$s_{dice} = \max_{i=0 \to V} \left(\frac{2 \times |N_{tt} \cup N_i|}{|N_{tt}| + |N_i| + \eta}\right)$$

trong đó, $a \cup b$ là phần giao giữa hai chuỗi ký tự a và b, $|x|$ là chiều dài chuỗi ký tự x, η là một hằng số dương nhỏ để tránh việc chia 0;

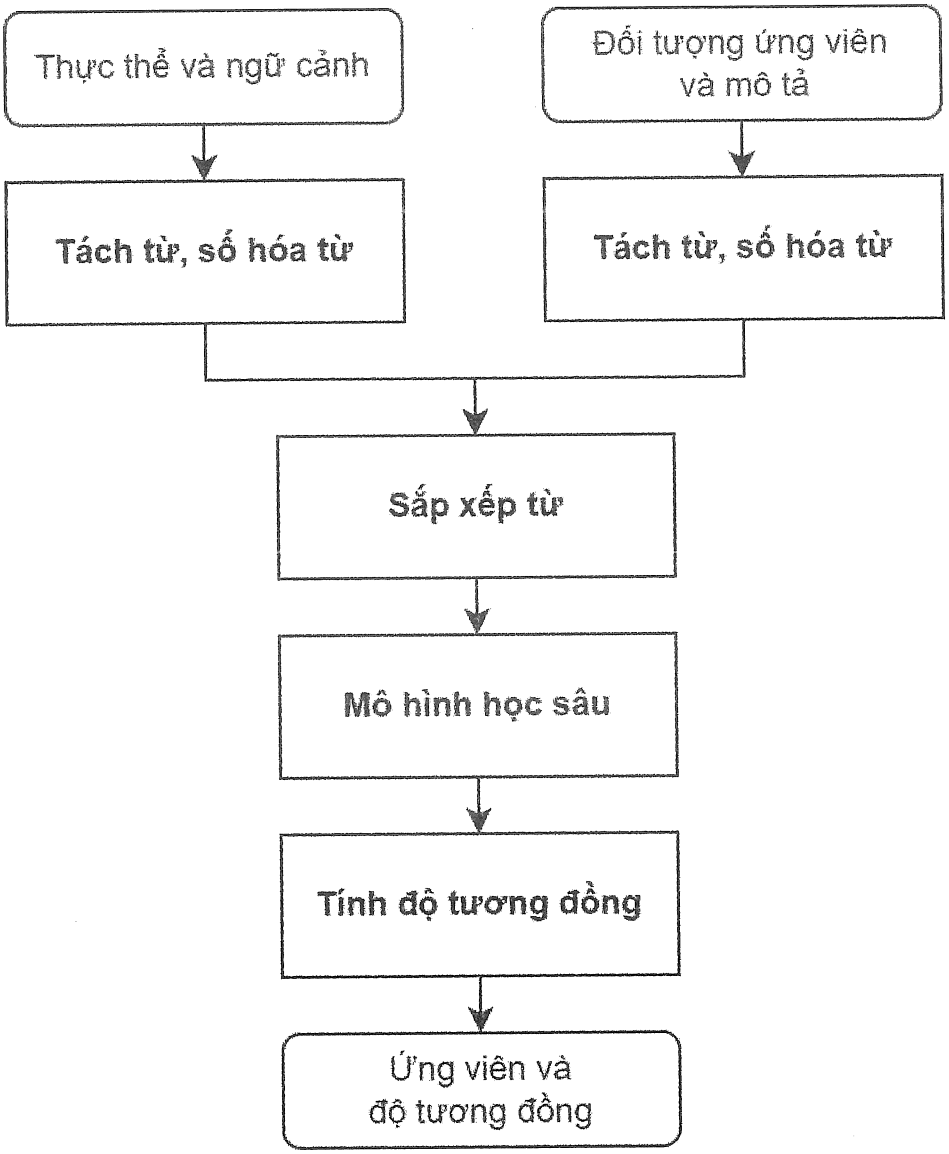
tiếp theo, một véc-tơ ba chiều $[s_{max}; s_{lev}; s_{dice}]$ được đưa vào một mô hình học có giám sát đã được huấn luyện từ trước để đưa ra kết luận cuối cùng cho kết quả cho toàn bộ quá trình; mô hình học có thể là mô hình máy véc-tơ hỗ trợ (support vector machine – SVM) hoặc cây quyết định (decision tree).


Thực thể và ngữ cảnh
Truy vấn đối tượng ứng viên
Xếp hạng lại ứng viên
Kiểm tra tính tồn tại của thực thể
Đối tượng tương ứng / Đối tượng rỗng
Hệ cơ sở tri thức

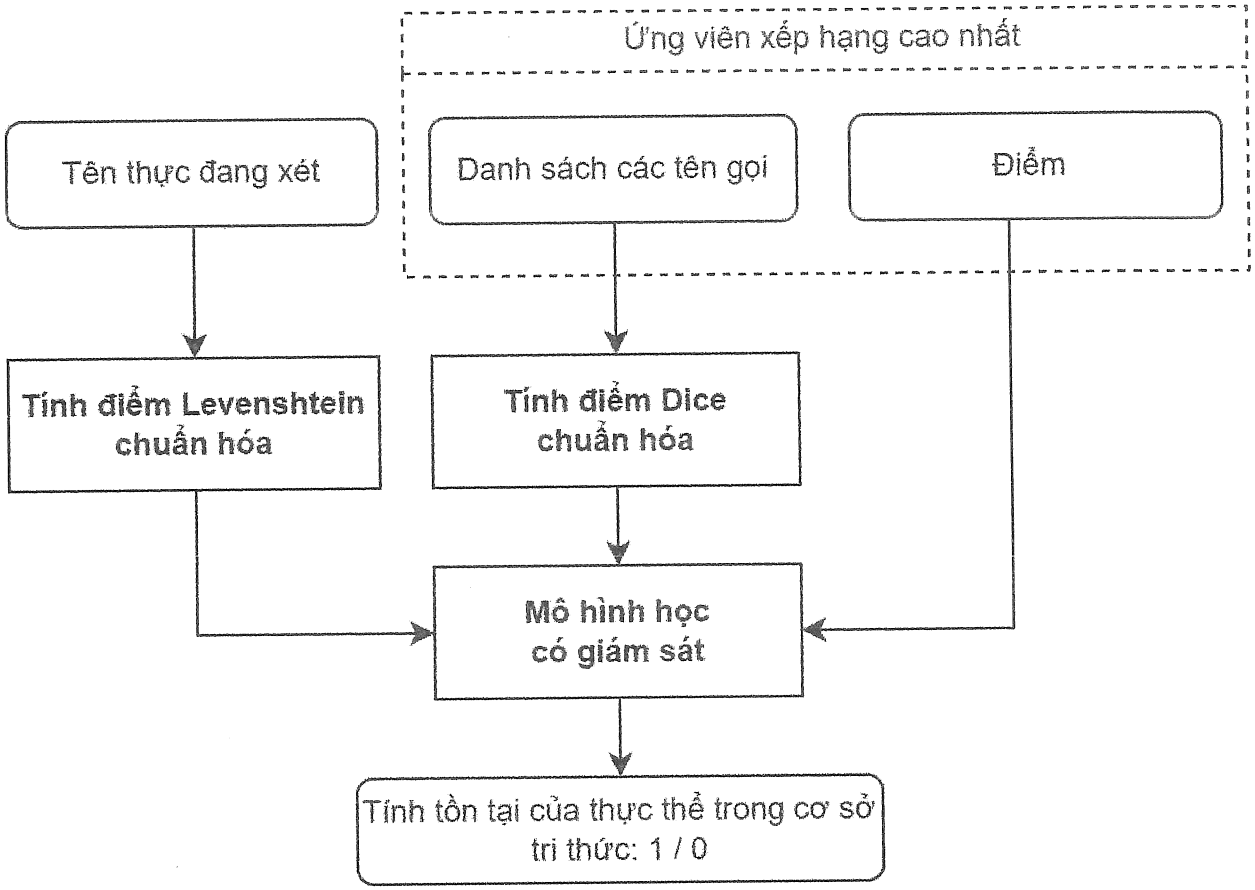
Hình 1


Thực thể và ngữ cảnh
Đối tượng và mô tả
Tách từ, số hóa từ
Tách từ, số hóa từ
Sắp xếp từ
Sắp xếp từ
Mô hình học sâu
Mô hình học sâu
Tính độ tương đồng
Ứng viên và
độ tương đồng

Hình 2


Thực thể và ngữ cảnh
Đối tượng ứng viên và mô tả
Tách từ, số hóa từ
Tách từ, số hóa từ
Sắp xếp từ
Mô hình học sâu
Tính độ tương đồng
Ứng viên và độ tương đồng

Hình 3


Ứng viên xếp hạng cao nhất
Tên thực đang xét
Danh sách các tên gọi
Điểm
Tính điểm Levenshtein chuẩn hóa
Tính điểm Dice chuẩn hóa
Mô hình học có giám sát
Tính tồn tại của thực thể trong cơ sở tri thức: 1 / 0

Hình 4