(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) **1-0053769**

(51)[2022.01] **G06F 40/166**

(13) **B**

(21) 1-2023-06877

(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.

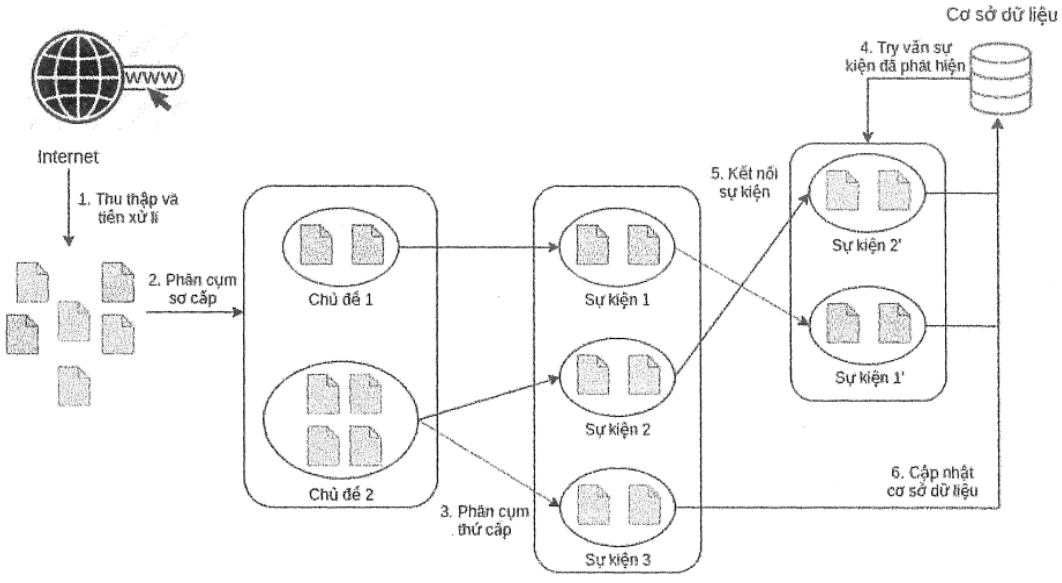
(72) LÂM DUY KHANG (VN); NGUYỄN VĂN TUẤN (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP TỰ ĐỘNG PHÁT HIỆN SỰ KIỆN NỔI BẬT TỪ DỮ LIỆU VĂN BẢN BÁO CHÍ THEO THỜI GIAN THỰC

(21) 1-2023-06877

(57) Sáng chế đề cập phương pháp phát hiện sự kiện nổi bật từ văn bản báo chí theo thời gian thực, gồm các bước: bước 1: thu thập bài báo và tiền xử lý; bước 2: phân cụm sơ cấp; bước 3: phân cụm thứ cấp; bước 4: phát hiện các sự kiện mới và cập nhật các sự kiện cũ. Phương pháp được ứng dụng vào các hệ thống thu thập, theo dõi và phân tích một số lượng lớn các văn bản báo chí, cung cấp một cái nhìn tổng quan về các sự kiện nổi bật đang diễn ra trên thế giới.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp tự động phát hiện sự kiện nổi bật từ văn bản báo chí theo giời gian thực. Cụ thể, phương pháp được đề cập trong sáng chế ứng dụng trong các hệ thống thu thập, theo dõi và phân tích thông tin, giúp người dùng có một cái nhìn khái quát về các sự kiện đang diễn ra trên không gian mạng một cách có hệ thống.

Tình trạng kỹ thuật của sáng chế

Hiện nay, trong một xã hội hiện đại với nhịp sống nhanh, cùng với sự bùng nổ của công nghệ thông tin, các trang báo điện tử ra đời nhằm đáp ứng nhu cầu tiếp nhận thông tin của con người và đã dần thay thế các trang báo truyền thống. Các trang báo điện tử có nhiều ưu điểm hơn so với các trang báo truyền thống trên nhiều khía cạnh như sức lan tỏa, dễ tiếp cận đến người đọc, dễ dàng trong việc xuất bản, đăng bài. Trên thực tế, có một số lượng lớn các bài báo điện tử được đăng tải liên tục mỗi ngày. Tuy nhiên, điều này vô tình lại gây nhiều khó khăn cho người đọc khi phải xử lý một lượng lớn thông tin gồm nhiều chủ đề, sự kiện khác nhau và chứa những thông tin trùng lặp, dư thừa. Bên cạnh đó, người đọc phải tiếp nhận một lượng thông tin lớn, có thể chứa các thông tin nhiễu, trong khi không nhận được thông tin về các sự kiện mà họ quan tâm, cũng như khó nắm bắt được các thông tin nào đang được đề cập bởi nhiều trang báo điện tử khác nhau. Do đó, nhu cầu cấp thiết đặt ra là xây dựng một phương pháp tự động thu thập các bài báo điện tử trên mạng internet, sau đó tiến hành phát hiện các sự kiện nổi bật đang xảy ra. Với cách sắp xếp thông tin như vậy, người đọc có thể có một cái nhìn tổng quan nhất về các sự kiện đang được diễn ra, đồng thời có thể theo dõi những sự kiện mà họ quan tâm.

Về phương diện kĩ thuật, các nghiên cứu trước đây về nhận diện sự kiện nổi bật thường được chuyển thành bài toán phân cụm văn bản với ý tưởng những bài báo nói về một sự kiện sẽ có mối quan hệ với nhau và tạo thành một cụm trong không gian đặc trưng. Cụ thể, các văn bản báo chí được thu thập sẽ được trích xuất đặc trưng về nội dung bằng các mô hình ngôn ngữ, sau đó phân chia các bài báo vào các cụm bằng các thuật toán phân cụm như K-means, DBSCAN,... khi đó mỗi cụm tương ứng với một sự kiện. Tuy nhiên, trong thực tế cụm bài báo có thể chứa nhiều sự kiện, ví dụ, một cụm có thể chứa hai sự kiện "Triều Tiên phóng

thử tên lửa" và "Triều Tiên tiến hành duyệt binh" vì hai sự kiện này đều liên quan đến tình hình quân sự ở Triều Tiên. Do đó sáng chế này đề xuất phương pháp phân cụm phân cấp, giúp các sự kiện được phát hiện một cách chính xác hơn.

Phương pháp tự động phát hiện sự kiện nổi bật này được thiết kế để có thể triển khai theo thời gian thực. Các bài báo liên tục được thu thập theo một mốc thời gian cố định, bài báo mới sẽ được phân vào các cụm, các sự kiện được liên tục cập nhật. Việc thiết kế theo thời gian thực cho phép cập nhật thông tin một cách nhanh chóng, việc xử lý dữ liệu cũng tối ưu hơn vì số lượng bài báo xử lý trong một lần là nhỏ hơn rất nhiều so với việc xử lý toàn bộ bài báo, đồng thời dòng thời gian cũng được cập nhật một cách nhất quán thuận tiện cho việc theo dõi các dòng sự kiện. Trong phạm vi sáng chế, dữ liệu được sử dụng là tiếng Anh, ngoài ra, phương pháp này có thể mở rộng cho các loại báo chí được viết với các ngôn ngữ khác nhau.

Bản chất kỹ thuật của sáng chế

Mục tiêu của sáng chế là đề xuất một phương pháp tự động nhận diện các sự kiện đang diễn ra và xây dựng các dòng sự kiện từ nguồn báo chí.

Để đạt được mục đích trên, giải pháp kỹ thuật được đề xuất trong sáng chế bao gồm bốn bước:

Bước 1: thu thập bài báo và tiền xử lý. Mục đích của bước này là thu thập bài báo từ nhiều nguồn báo chí khác nhau và tiền xử lý văn bản báo chí, loại bỏ những kí tự, thông tin không cần thiết, chuẩn bị cho quá trình trích xuất đặc trưng bằng mạng học sâu. Đầu tiên, tiến hành thu thập văn bản báo chí từ các nguồn đã được xác định sẵn, văn bản báo chí thô sau khi thu thập sẽ đi qua mô đun tiền xử lý. Đầu ra của bước này là các văn bản báo chí đã được tiền xử lý.

Bước 2: phân cụm sơ cấp. Mục đích của bước này là trích xuất đặc trưng ngữ nghĩa của văn bản, sau đó dùng đặc trưng này để phân cụm các bài báo theo các chủ đề khác nhau. Phương pháp phân cụm được sử dụng là thuật toán phân cụm trực tuyến, liên tục nhận bài báo mới và thực hiện phân cụm trên không gian vec-tơ đặc trưng ngữ nghĩa đã được trích xuất bởi một mạng học sâu. Đầu ra là các cụm văn bản báo chí, mỗi cụm có một chủ đề khác nhau.

Bước 3: phân cụm thứ cấp. Mục đích của bước này là nhận diện các sự kiện nổi bật trên từng cụm bài báo thuộc một chủ đề nhất định. Đầu tiên, xét theo từng cụm chủ đề, tiến hành trích xuất thông tin của bài báo về từ khóa, tiêu đề, và vec-tơ ngữ nghĩa. Sau đó, xây

dựng đồ thị quan hệ của các bài báo trong chủ đề. Cuối cùng, dùng thuật toán phát hiện cộng đồng để phát hiện các cụm bài báo có mật độ liên kết lớn với nhau, từ đó phát hiện sự kiện nổi bật. Đầu ra là các sự kiện nổi bật, mỗi sự kiện được biểu diễn thông qua một cụm bài báo cùng nói đến sự kiện đó.

Bước 4: phát hiện sự kiện mới và cập nhật sự kiện cũ. Mục đích của bước này là đảm bảo tính thống nhất của các sự kiện. Một sự kiện tại thời điểm hiện tại nếu không tìm được bất kỳ liên kết nào đến các sự kiện trước đó thì được xem là một sự kiện mới và được lưu vào cơ sở dữ liệu. Ngược lại, thực hiện cập nhật lại các sự kiện đã có trong cơ sở dữ liệu.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình vẽ minh họa toàn bộ quá trình thu thập, xử lý và phát hiện sự kiện nổi bật từ các nguồn báo chí;

Hình 2: là hình vẽ mô tả thuật toán phát hiện cộng đồng trong phân cụm thứ cấp.

Mô tả chi tiết sáng chế

Sáng chế đề xuất phương pháp phát hiện sự kiện nổi bật từ các nguồn báo chí, từ đó xây dựng dòng sự kiện một cách tự động như mô tả ở Hình 1. Phạm vi sáng chế này sẽ tập trung chủ yếu vào các sự kiện được đề cập đến trong văn bảo báo chí như kinh tế, chính trị, thể thao, ... Chi tiết các bước thực hiện của phương pháp được trình bày cụ thể như sau:

Bước 1: thu thập bài báo và tiền xử lý;

Bước đầu tiên trong phương pháp phát hiện sự kiện nổi bật từ các nguồn báo chí là thu thập dữ liệu và tiền xử lý. Quá trình thu thập dữ liệu cần đảm bảo hai yếu tố: một là các nguồn báo phải là các nguồn uy tín, chính thống giúp hệ thống phát hiện đúng các sự kiện nổi bật, tránh nhận diện các sự kiện không chính xác, sự kiện giả; hai là dữ liệu báo chí phải luôn được cập nhật liên tục giúp phát hiện các sự kiện nổi bật một cách nhanh chóng.

Trong phạm vi phương pháp mà sáng chế đề cập, dữ liệu sử dụng được thu thập từ các trang báo điện tử tiếng Anh, cụ thể được thu thập từ các trang báo chí điện tử chính thống như: Reuters, CNN, AFP, StraitsTimes, ...

Các văn bản báo chí thô được thu thập bằng sử dụng các công cụ mã nguồn mở dựa trên phân tích cấu trúc trang nội dung các bài báo. Dữ liệu mỗi trang báo được thu thập gồm các trường thông tin như nội dung bài viết, các chú thích hình ảnh, và các thông tin mô tả bài báo như tiêu đề, nguồn báo, thời gian đăng bài, ... Những trường thông tin này sẽ được lưu lại trong cơ sở dữ liệu của hệ thống.

Dữ liệu các bài báo sau khi được thu thập sẽ được tiền xử lý thông qua một số bước sau đây:

- Chuyển văn bản đầu vào sang dạng viết thường và phân tách văn bản thành các từ. Phương pháp tách từ được sử dụng là tách theo biểu thức chính quy và kết hợp tách từ theo khoảng trắng. Ví dụ, câu "Britain's royal family rushed to be with Queen Elizabeth after doctors said they were concerned about the health of the 96-year-old monarch" sẽ được tách thành một danh sách các từ đơn: 'britain', 's', 'royal', 'family', 'rushed', 'to', 'be', 'with', 'queen', 'elizabeth', 'after', 'doctors', 'said', 'they', 'were', 'concerned', 'about', 'the', 'health', 'of', 'the', '96', 'year', 'old', 'monarch'.

- Biến đổi các từ đơn về từ gốc thông qua bổ đề ngôn ngữ bằng cách loại bỏ những tiền tố và hậu tố của từ để đưa về dạng nguyên mẫu. Ví dụ các từ "walks", "walking", "walked" sẽ được biến đổi về từ gốc là "walk". Tương tự với bổ đề ngôn ngữ, chuyển từ đồng nghĩa là một cách khác để đưa các từ liên quan đến nhau về một từ gốc. Ví dụ "canadienne", "canadian" sẽ được chuyển về "canada".

- Loại bỏ các từ dừng. Từ dừng được định nghĩa là các từ phổ biến trong các ngôn ngữ, ví dụ như "the", "a", "an", "and", "of", "to", ... Những từ này có tần suất xuất hiện nhiều nhưng lại không mang quá nhiều ý nghĩa về mặt nội dung. Các từ dừng có thể được xem là một loại nhiễu vì nó có thể gây khó khăn trong việc nắm bắt được những từ, nội dung quan trọng và làm ảnh hưởng đến các mô hình ngôn ngữ. Số lượng từ dừng nhiều cũng góp phần làm tăng chi phí tính toán.

Sau khi thực hiện các bước tiền xử lý dữ liệu thì thu được văn bản báo chí có các từ được chuyển về dạng viết thường, loại từ được chuyển về dạng nguyên mẫu và loại bỏ các từ dừng. Văn bản này sẽ được sử dụng cho bước phân cụm sơ cấp.

Bước 2: phân cụm sơ cấp;

Ở bước này các bài báo được trích xuất đặc trưng ngữ nghĩa và được biểu diễn dưới dạng các vec-tơ ngữ nghĩa trong không gian đặc trưng. Sau đó, một thuật toán phân cụm theo hướng học không giám sát được áp dụng lên không gian đặc trưng nhằm phân các bài báo vào các cụm (sau đây gọi là cụm sơ cấp), mỗi cụm sơ cấp gồm các bài báo cùng nói đến một chủ đề nào đó. Xét một ví dụ với hai bài báo có tiêu đề lần lượt là "Pakistan chìm trong biển nước sau cơn lũ khi có thêm 18 người chết" (8/9/2022) và "Người đứng đầu Liên Hợp Quốc đến Pakistan, nơi đang bị ảnh hưởng nặng nề do khí hậu" (10/9/2022), mô hình sẽ thực hiện

phân hai bài báo này vào cùng một cụm sơ cấp, mặc dù chúng đề cập đến hai sự kiện hoàn toàn khác nhau và diễn ra ở hai thời điểm cách khá xa nhau. Trong trường hợp này, thuật toán theo hướng học không giám sát không trả về một chủ đề cụ thể mà hai bài báo muốn nhắc đến nhưng thông qua tiêu đề và nội dung của chúng, khả năng cao chủ đề mà chúng đề cập là tình hình ở Pakistan sau cơn lũ.

Trong phạm vi sáng chế này, nhóm tác giả đề xuất sử dụng mô hình ngôn ngữ Sentence-BERT cho việc trích xuất véc-tơ ngữ nghĩa cho cả văn bản. Cụ thể, một bài báo sau khi đã được tiền xử lý sẽ được mô hình Sentence-BERT xử lý để thu được véc-tơ ngữ nghĩa tương ứng của văn bản đó. Quá trình trích xuất đặc trưng bằng Sentence-BERT có thể được mô tả dưới dạng công thức như sau:

$$e_j = S(d_j), d_j \in D \quad (1)$$

Trong đó: e_j là véc-tơ ngữ nghĩa biểu diễn cho bài báo d_j trong tập n bài báo $D = \{d_j | j \in [1, n]\}$.

Các véc-tơ ngữ nghĩa, được trích xuất từ văn bản đầu vào, sẽ được biến đổi về một miền không gian khác có số chiều ít hơn thông qua thuật toán phép chiếu và xấp xỉ đa tạp đồng nhất (sau đây viết tắt là UMAP). Thuật toán này được hiện thực dựa trên ý tưởng xây dựng một đồ thị liền kề trên không gian gốc của dữ liệu và tìm một đồ thị tương tự trên một không gian ít chiều hơn. Ưu điểm của UMAP nằm ở tốc độ xử lý nhanh, khả năng mở rộng tốt về kích thước tập dữ liệu và cả số chiều nhưng vẫn bảo toàn cấu trúc tổng thể của dữ liệu. Quá trình giảm số chiều của véc-tơ đặc trưng bằng UMAP có thể được biểu diễn như sau:

$$v_j = U(e_j) \quad (2)$$

Trong đó: e_j là véc-tơ ngữ nghĩa được trích xuất từ bài báo, v_j là véc-tơ ngữ nghĩa đã được giảm số chiều thông qua thuật toán giảm số chiều UMAP được ký hiệu là $U: \mathbb{R}^m \rightarrow \mathbb{R}^n, n < m$.

Một mô hình phân cụm được sử dụng để phân các văn bản đầu vào thành nhiều cụm, mỗi cụm gồm các văn bản cùng nói về một chủ đề. Trong sáng chế này, nhóm tác giả đề xuất sử dụng mô hình giảm số lần lặp và phân cụm cân bằng dựa trên phương pháp phân cấp (sau đây viết tắt là mô hình BIRCH). Một mặt, mô hình BIRCH cho thấy khả năng phân cụm hiệu quả ngay cả khi kích thước của tập dữ liệu tăng lên. Mặt khác, thông qua kỹ thuật học trực tuyến, mô hình BIRCH có thể liên tục được cập nhật lại trọng số khi có dữ liệu mới đến mà không cần phải huấn luyện lại trên toàn bộ dữ liệu. Cụ thể, với đầu vào là véc-tơ ngữ nghĩa

đã được giảm số chiều v_j của văn bản j, mô hình BIRCH trả về chủ đề t_j mà văn bản đó thuộc về. Quá trình phân cụm bằng BIRCH được biểu diễn thông qua công thức sau đây:

$$t_j = \mathcal{B}(v_j), t_j \in [1, m] \quad (3)$$

Trong đó: m là tổng số chủ đề được dự đoán tính đến thời điểm đang xét; $\mathcal{B}$ là mô hình phân cụm BIRCH.

Một cách tổng quát, có thể mô hình hóa toàn bộ quá trình phân cụm sơ cấp theo công thức dưới đây:

$$T = \mathcal{B} \circ U \circ S(D) \quad (4)$$

Trong đó, đầu vào là một tập n các bài báo $D = \{d_1, d_2, \dots, d_n\}$, đầu ra $T = \{t_i | i \in [1, n]\}$ là tập các chủ đề được dự đoán của mỗi bài báo, còn $\mathcal{B}, U, S$ lần lượt là mô hình trích xuất véc-tơ đặc trưng ngữ nghĩa, mô hình giảm chiều dữ liệu và mô hình phân cụm sơ cấp. Khi đó, một cụm sơ cấp $C_i = \{d_j | t_j = i, d_j \in D\}$ với $i \in [1, m]$ được nhóm tác giả mô tả là một tập các bài báo thuộc cùng chủ đề i.

Sau bước phân cụm sơ cấp, các văn bản đầu vào được phân vào các cụm sơ cấp tương ứng với chủ đề mà chúng thể hiện.

Bước 3: phân cụm thứ cấp;

Sau khi đã phân cụm các bài báo vào các chủ đề, một thuật toán phân cụm thứ cấp được áp dụng lên các cụm bài báo để trích xuất được các sự kiện nổi bật.

Các bài báo cùng nói về một sự kiện sẽ có mối quan hệ chặt chẽ với nhau và tạo thành một cộng đồng. Từ quan sát này, nhóm tác giả đề xuất sử dụng phương pháp phát hiện sự kiện bằng cách mô hình hóa các bài báo trong một chủ đề thành một đồ thị liền kề vô hướng, sau đó dùng thuật toán nhận diện cộng đồng để phát hiện ra tập các bài báo có mối quan hệ mật thiết với nhau, cùng nói về một sự kiện, từ đó xác định được các sự kiện nổi bật.

Cụ thể, xét một cụm chủ đề $C_i = \{d_j | j \in [1, n]\}$ được trích xuất từ bước phân cụm sơ cấp. Đồ thị liền kề mô tả mối quan hệ giữa các bài báo cùng chủ đề là một đồ thị vô hướng với mỗi nút là một bài báo d_j, mối quan hệ giữa hai bài báo d_p và d_q được biểu diễn là một cạnh nối hai nút và có trọng số tương quan là $w_{p,q}$ (tham chiếu Hình 2).

Trong sáng chế này, trọng số tương quan quan giữa hai bài báo (trọng số cạnh nối hai nút trong đồ thị) được xét trên các yếu tố: từ khóa, ngữ nghĩa nội dung, tiêu đề. Cụ thể, mỗi bài báo khi đưa vào đồ thị sẽ được trích xuất một số đặc trưng:

- Vec-tơ đặc trưng ngữ nghĩa e_p là một vec-tơ mang thông tin về mặt ngữ nghĩa, nội

dung của văn bản và được trích xuất bởi mô hình Sentence-BERT.

- Từ khóa là từ hoặc cụm từ quan trọng trong một văn bản, thể hiện nội dung của bài viết. Từ các từ khóa, ta có thể biết được nội dung và chủ đề thể hiện trong bài viết, do đó từ khóa là một đặc trưng quan trọng để so sánh sự tương đồng. Sáng chế này sử dụng mô hình KeyBERT để trích xuất các từ khóa quan trọng trong bài báo. Ý tưởng của KeyBERT là chia nhỏ văn bản thành các từ đơn và từ ghép, sau đó tính độ tương đồng giữa các từ với véc-tơ đặc trưng ngữ nghĩa của văn bản thông qua phép tính cô-sin được biểu diễn như sau:

$$cos(w_i, w_j) = \frac{w_i . w_j}{\|w_i\|\|w_j\|} \quad (5)$$

Trong đó, w_i, w_j lần lượt là véc-tơ biểu diễn của từ và véc-tơ đặc trưng ngữ nghĩa của văn bản.

Kết quả thu được là một danh sách K từ có tương đồng cao (từ khóa) với đặc trưng ngữ nghĩa của văn bản. Tập K từ khóa của bài báo d_p được biểu diễn bằng $k_p = \{k_1, k_2, \dots, k_K\}$.

- Tiêu đề bài báo là một câu văn ngắn tóm tắt bài báo, giúp người đọc biết được nội dung và chủ đề của bài báo. Do đó, sử dụng tiêu đề bài báo làm tăng độ chính xác khi so sánh tương quan các bài báo. Tiêu đề bài báo được phân tách thành một tập các từ $t_p = \{w_1, w_2, \dots, w\}$.

Từ những đặc trưng bài báo trên, các độ tương quan giữa hai bài báo d_p và d_q được nhóm tác giả định nghĩa như sau:

- Tương đồng ngữ nghĩa $f(d_p, d_q)$ được định nghĩa là độ tương đồng cosin của vec-tơ đặc trưng ngữ nghĩa e_p và e_q của d_p và d_q

- Tương đồng từ khóa $k(d_p, d_q)$ được định nghĩa là tỉ lệ giữa số lượng từ khóa trùng nhau của hai văn bản báo chí trên số lượng từ khóa.

- Tương đồng về tiêu đề $t(d_p, d_q)$ được định nghĩa là tỉ lệ số từ trùng nhau của hai tiêu đề t_p và t_q trên độ dài của tiêu đề dài hơn.

- Độ tương đồng giữa hai bài báo được nhóm tác giả định nghĩa bằng tổng các độ tương đồng thành phần theo công thức:

$$w_{p,q} = \sigma\left(\alpha f(d_p, d_q) + \beta k(d_p, d_q) + \gamma t(d_p, d_q)\right) \quad (6)$$

Với $w_{p,q}$ là xác suất mà hai bài báo d_p và d_q cùng nói về một sự kiện, σ là hàm sigmoid và α, β, γ lần lượt là các trọng số quyết định sự quan trọng của các độ tương quan. Các trọng số này được học thông qua một mô hình hồi quy tuyến tính nhằm xác định xem hai bài báo đầu vào có cùng nói đến một sự kiện hay không.

Từ đồ thị liền kề giữa các bài báo trong cùng một cụm sơ cấp, nhóm tác giả sử dụng thuật toán nhận diện cộng đồng Edmot để xác định các cụm thứ cấp, tương ứng với các sự kiện khác nhau. Các bài báo trong cùng một cụm thứ cấp sẽ có mật độ liên kết dày đặc hơn so với phần còn lại của đồ thị. Cụ thể, thuật toán nhận đầu vào là đồ thị liền kề giữa các bài báo (ký hiệu là G) và trả về danh sách các cụm thứ cấp tương ứng:

$$\mathcal{E} = F(G) \tag{7}$$

Trong đó: F là thuật toán nhận diện cộng đồng Edmot; $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_m\}$ là tập m cụm thứ cấp trong một cụm sơ cấp, mỗi cụm thứ cấp là một tập các bài báo xoay quanh một sự kiện nào đó $\mathcal{E}_j = \{d_1, d_2, \dots, d\}$.

Sau bước phân cụm thứ cấp, các văn bản thuộc cùng một cụm sơ cấp tiếp tục được phân về các cụm thứ cấp tương ứng với sự kiện chính mà cụm thứ cấp đề cập đến.

Bước 4: phát hiện sự kiện mới và cập nhật sự kiện cũ;

Các bước 1 đến 3 đều được thực hiện trên một tập các bài báo được thu thập trong một khoảng thời gian cố định. Tuy nhiên, các sự kiện xảy ra liên tục và các bài báo cũng liên tục được thu thập và cập nhật. Do đó, cần phải có một bước để cập nhật lại các sự kiện đã có bằng cách thêm các bài báo liên quan mới được thu thập, hoặc tạo thêm một sự kiện mới và lưu vào cơ sở dữ liệu.

Xét hai khoảng thời gian liên tiếp nhau t và $t+1$ lần lượt có tập các sự kiện $\mathcal{E}^t = \{\mathcal{E}_1^t, \mathcal{E}_2^t, \dots, \mathcal{E}_k^t\}$ và $\mathcal{E}^{t+1} = \{\mathcal{E}_1^{t+1}, \mathcal{E}_2^{t+1}, \dots, \mathcal{E}_l^{t+1}\}$. Độ tương đồng giữa hai sự kiện được tính bằng độ đo tỉ lệ giữa phần giao và phần hội IoU (thường được dùng trong bài toán phát hiện đối tượng trong xử lý ảnh) thông qua công thức:

$$IoU(\mathcal{E}_k^t, \mathcal{E}_l^{t+1}) = \frac{\mathcal{E}_k^t \cap \mathcal{E}_l^{t+1}}{\mathcal{E}_k^t \cup \mathcal{E}_l^{t+1}} \tag{8}$$

Giải pháp kỹ thuật sử dụng thuật toán Hungarian để kết nối các sự kiện tại hai thời điểm khác nhau sao cho tỉ lệ giữa phần giao và hội IoU giữa chúng là lớn nhất. Nếu tồn tại một liên kết giữa một sự kiện mới với một sự kiện đã có, thực hiện gộp các bài báo có trong hai sự kiện này vào cùng một cụm thứ cấp theo một định danh cho trước. Trong trường hợp sự kiện mới không tìm được kết nối đến bất kỳ sự kiện nào đã có thì một cụm thứ cấp mới được tạo ra với một định danh duy nhất. Kết quả cuối cùng thu được là các cụm thứ cấp, với mỗi cụm thứ cấp có một định danh duy nhất chứa các bài báo cùng đề cập đến một sự kiện nổi bật.

Yêu cầu bảo hộ

1. Phương pháp tự động phát hiện sự kiện nổi bật từ dữ liệu văn bản báo chí theo thời gian thực bao gồm:

bước 1: thu thập bài báo và tiền xử lý; trong bước này thực hiện các công việc sau:

thu thập các văn bản báo chí thô bằng sử dụng các công cụ mã nguồn mở dựa trên phân tích cấu trúc trang nội dung các bài báo; dữ liệu sau khi thu thập sẽ được xử lý để giữ lại những thông tin cần thiết như nội dung bài viết, các chú thích hình ảnh, và các thông tin mô tả bài báo như tiêu đề, nguồn báo, thời gian đăng bài, ... những dữ liệu này sẽ được lưu lại trong cơ sở dữ liệu của hệ thống;

chuyển văn bản thành chữ thường và phân tách văn bản thành các từ; phương pháp tách từ được sử dụng là tách theo biểu thức chính quy và kết hợp tách từ theo khoảng trắng;

biến đổi về từ gốc hay thực hiện bổ đề ngôn ngữ bằng cách loại bỏ những tiền tố và hậu tố của từ đến đưa từ về nguyên mẫu;

loại bỏ các từ dừng bằng cách sử dụng một bộ các từ dừng đã được thu thập trước đó để loại bỏ chúng trong văn bản;

bước 2: phân cụm sơ cấp; cụ thể:

với đầu vào là một bài báo trong tập n bài báo được thu thập trong một khung thời gian cho trước $D = \{d_j | j \in [1, n]\}$, mô hình sentence-BERT, ký hiệu là S thực hiện trích xuất véc-tơ ngữ nghĩa e_i:

$$e_j = S(d_j), d_j \in D \quad (1)$$

trong đó: e_j là véc-tơ ngữ nghĩa biểu diễn cho bài báo d_j trong tập n bài báo $D = \{d_j | j \in [1, n]\}$;

các véc-tơ ngữ nghĩa sau đó sẽ được biến đổi về một miền không gian khác có số chiều ít hơn thông qua phép chiếu và xấp xỉ đa tạp đồng nhất, viết tắt là UMAP:

$$v_j = U(e_j) \quad (2)$$

trong đó: e_i là véc-tơ ngữ nghĩa được trích xuất từ bài báo d_j; v_j là véc-tơ ngữ nghĩa đã được giảm số chiều thông qua thuật toán giảm số chiều UMAP được ký hiệu là $U: \mathbb{R}^m \rightarrow \mathbb{R}^n, n < m$;

thuật toán giảm số lần lặp và phân cụm cân bằng dựa trên phương pháp phân cấp BIRCH được sử dụng để phân các văn bản đầu vào thành nhiều cụm, mỗi cụm gồm các văn bản cùng nói về một chủ đề. Với đầu vào là véc-tơ ngữ nghĩa đã được giảm số chiều v_j của

văn bản j, mô hình phân cụm trả về chủ đề t_j mà văn bản đó thuộc về:

$$t_j = \mathcal{B}(v_j), t_j \in [1, m] \quad (3)$$

trong đó: m là tổng số chủ đề được dự đoán tính đến thời điểm đang xét; $\mathcal{B}$ là mô hình phân cụm sơ cấp BIRCH;

một cách tổng quát, ta có thể mô hình hóa toàn bộ quá trình phân cụm sơ cấp này dưới dạng công thức:

$$T = \mathcal{B} \circ U \circ S(D) \quad (4)$$

trong đó, đầu vào là một tập n các bài báo $D = \{d_1, d_2, \dots, d_n\}$, đầu ra $T = \{t_i | i \in [1, n]\}$ là tập các chủ đề được dự đoán của mỗi bài báo, còn $\mathcal{B}, U, S$ lần lượt là mô hình trích xuất véc-tơ đặc trưng ngữ nghĩa, mô hình giảm chiều dữ liệu và mô hình phân cụm sơ cấp. Khi đó, một cụm sơ cấp $C_i = \{d_j | t_j = i, d_j \in D\}$ với $i \in [1, m]$ được nhóm tác giả mô tả là một tập các bài báo thuộc cùng chủ đề i;

bước 3: phân cụm thứ cấp; tại bước này, sáng chế đề xuất phương pháp phát hiện sự kiện bằng cách mô hình hóa các bài báo trong một chủ đề thành một đồ thị vô hướng, sau đó dùng thuật toán nhận diện cộng đồng để phát hiện ra tập các bài báo có mối quan hệ mật thiết với nhau, cùng nói về một chủ đề, từ đó xác định được các sự kiện nổi bật;

xét một cụm chủ đề $C_i = \{d_j | j \in [1, n]\}$, đồ thị liền kề mô tả mối quan hệ giữa các bài báo cùng chủ đề là một đồ thị vô hướng với mỗi nút là một bài báo d_j, mối quan hệ giữa hai bài báo d_p và d_q được biểu diễn là một cạnh nối hai nút và có trọng số tương quan là $w_{p,q}$;

trọng số liên quan giữa hai bài báo (trọng số cạnh nối hai nút trong đồ thị) được xét trên các yếu tố: từ khóa, ngữ nghĩa nội dung, tiêu đề; mỗi bài báo khi đưa vào đồ thị được trích xuất một số đặc trưng:

vec-tơ đặc trưng ngữ nghĩa e_p là một vec-tơ mang thông tin về mặt ngữ nghĩa, nội dung của văn bản và được trích xuất bởi Sentence-BERT ở bước 1;

từ khóa là từ hoặc cụm từ quan trọng trong một văn bản, thể hiện nội dung của bài viết; sáng chế này sử dụng mô hình KeyBERT để trích xuất các từ khóa quan trọng trong bài báo; tập K từ khóa của bài báo d_p được biểu diễn bằng $k_p = \{k_1, k_2, \dots, k_K\}$;

tiêu đề bài báo là một câu văn ngắn tóm tắt bài báo, giúp người đọc biết được nội dung và chủ đề của bài báo; tiêu đề bài báo được phân tách thành một tập các từ $t_p = \{w_1, w_2, \dots, w\}$;

từ những đặc trưng bài báo trên, ta tiến hành tính các độ tương quan giữa hai bài báo d_p và d_q;

tương đồng ngữ nghĩa $f(d_p, d_q)$ được định nghĩa là độ tương đồng cosin của vec-tơ đặc trưng ngữ nghĩa e_p và e_q của d_p và d_q;

tương đồng từ khóa $k(d_p, d_q)$ được định nghĩa là tỉ lệ giữa số lượng từ khóa trùng nhau của 2 văn bản báo chí trên số lượng từ khóa;

tương đồng về tiêu đề $t(d_p, d_q)$ được định nghĩa là tỉ lệ số từ trùng nhau của 2 tiêu đề t_p và t_q trên độ dài của tiêu đề dài hơn;

độ tương đồng giữa hai bài báo được tính là tổng các độ tương đồng theo công thức:

$$w_{p,q} = \sigma\left(\alpha f(d_p, d_q) + \beta k(d_p, d_q) + \gamma t(d_p, d_q)\right) \quad (5)$$

với $w_{p,q}$ là xác suất mà 2 bài báo d_p và d_q cùng nói về một sự kiện, σ là hàm sigmoid và α, β, γ lần lượt là các trọng số quyết định sự quan trọng của các độ tương quan; các trọng số này được học thông qua một mô hình hồi quy tuyến tính nhằm xác định xem hai bài báo đầu vào có cùng nói đến một sự kiện hay không;

từ đồ thị liền kề giữa các bài báo trong cùng một cụm sơ cấp, nhóm tác giả sử dụng thuật toán nhận diện cộng đồng edmot để xác định các cụm thứ cấp, tương ứng với các sự kiện khác nhau; các bài báo trong cùng một cụm thứ cấp sẽ có mật độ liên kết dày đặc hơn so với phần còn lại của đồ thị; cụ thể, thuật toán nhận đầu vào là đồ thị liền kề giữa các bài báo (ký hiệu là G) và trả về danh sách các cụm thứ cấp tương ứng:

$$\mathcal{E} = F(G) \quad (6)$$

trong đó: F là thuật toán nhận diện cộng đồng edmot; $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_m\}$ là tập m cụm thứ cấp trong một cụm sơ cấp, mỗi cụm thứ cấp là một tập các bài báo xoay quanh một sự kiện nào đó $\mathcal{E}_j = \{d_1, d_2, \dots, d\}$;

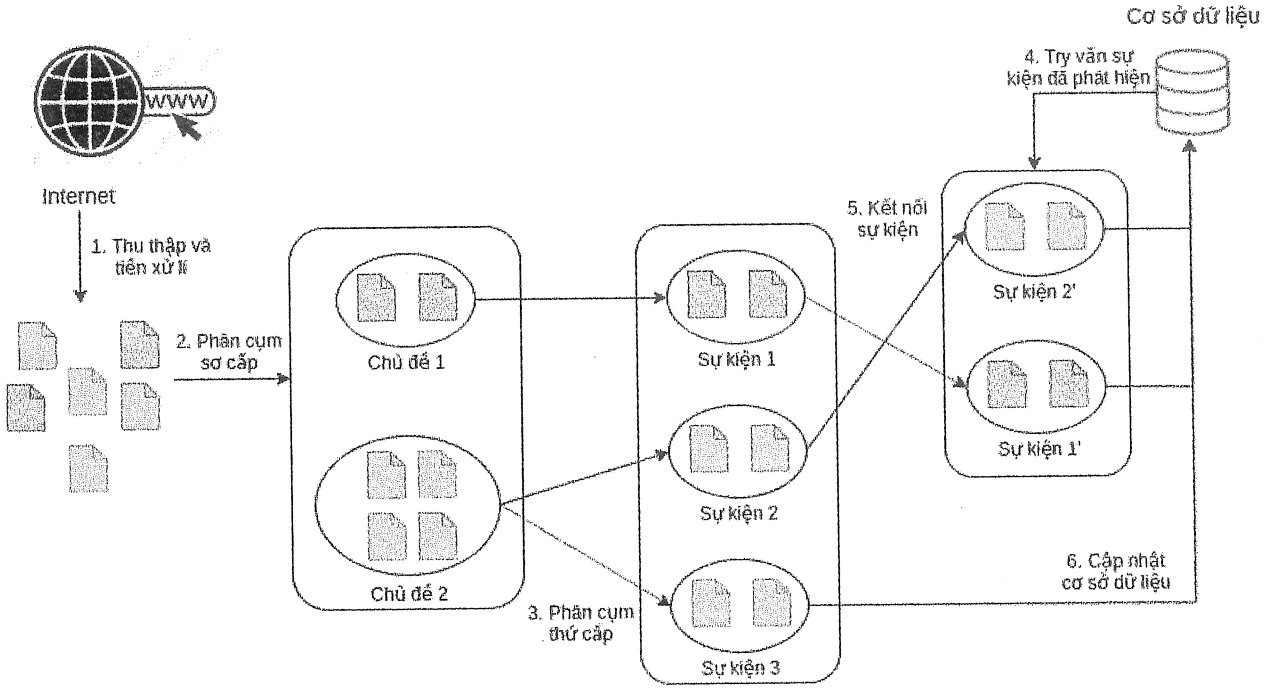
bước 4: phát hiện sự kiện mới và cập nhật sự kiện cũ; tại bước này tiến hành:

xét hai khoảng thời gian liên tiếp nhau t và $t+1$ lần lượt có tập các sự kiện $\mathcal{E}^t = \{\mathcal{E}_1^t, \mathcal{E}_2^t, \dots, \mathcal{E}_k^t\}$ và $\mathcal{E}^{t+1} = \{\mathcal{E}_1^{t+1}, \mathcal{E}_2^{t+1}, \dots, \mathcal{E}_l^{t+1}\}$; ta tính sự tương đồng giữa hai sự kiện bằng độ đo trùng lặp bài báo IoU:

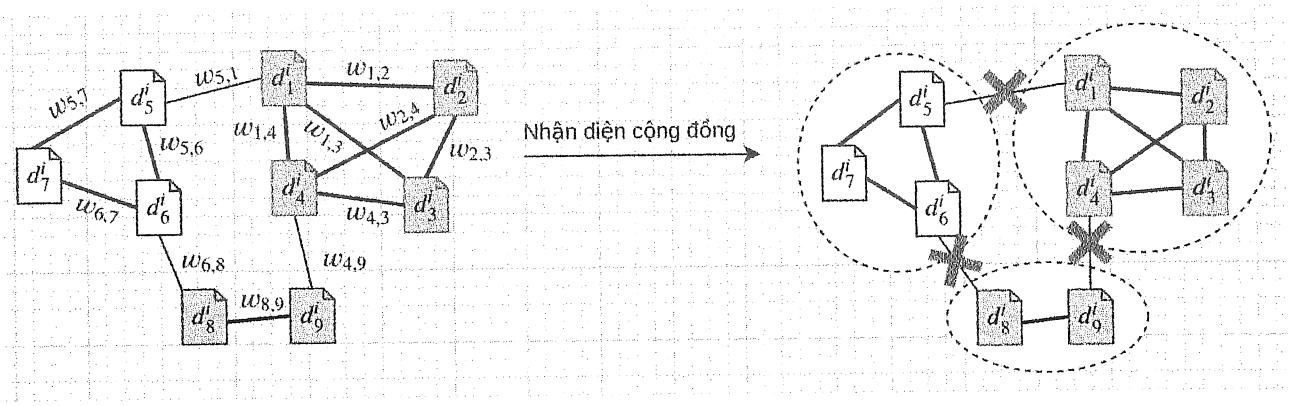
$$IoU(\mathcal{E}_k^t, \mathcal{E}_l^{t+1}) = \frac{\mathcal{E}_k^t \cap \mathcal{E}_l^{t+1}}{\mathcal{E}_k^t \cup \mathcal{E}_l^{t+1}} \quad (7)$$

nhóm tác giả sử dụng thuật toán hungarian để kết nối các sự kiện tại hai thời điểm khác nhau sao cho tổng độ trùng lặp sự kiện IoU là lớn nhất; nếu tồn tại một liên kết giữa một sự

kiện mới với một sự kiện đã có, thực hiện cập nhật lại sự kiện bằng cách thêm các bài báo trong sự kiện mới vào cụm thứ cấp tương ứng với sự kiện đã có; trong trường hợp sự kiện mới không tìm được kết nối đến bất kỳ sự kiện nào đã có thì được khởi tạo và lưu vào cơ sở dữ liệu.


Cơ sở dữ liệu
4. Trợ văn sự kiện đã phát hiện
Internet
1. Thu thập và tiền xử lí
2. Phân cụm sơ cấp
Chủ đề 1
Chủ đề 2
3. Phân cụm thứ cấp
Sự kiện 1
Sự kiện 2
Sự kiện 3
5. Kết nối sự kiện
Sự kiện 2'
Sự kiện 1'
6. Cập nhật cơ sở dữ liệu

Hình 1


Nhận diện cộng đồng

Hình 2