(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) **1-0042170**

(51)2020.01 **G06F 40/20**

(13) **B**

(21) 1-2021-08499 (22) 30/12/2021

(45) 25/12/2024 441 (43) 25/03/2022 408

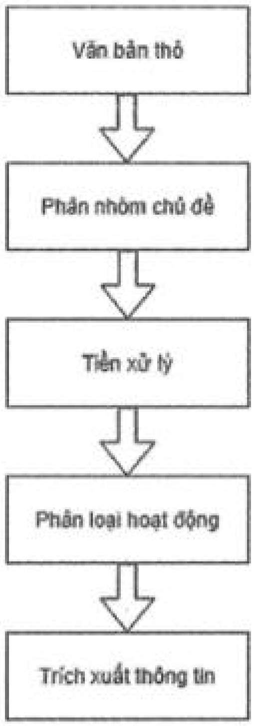
(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

(72) Nguyễn Văn Tuấn (VN); Nguyễn Gia Thịnh (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP THEO DÕI HOẠT ĐỘNG CỦA MỤC TIÊU TRÊN BIỂN DỰA TRÊN PHÂN TÍCH DỮ LIỆU DẠNG VĂN BẢN

(57) Sáng chế đề cập phương pháp theo dõi hoạt động của mục tiêu trên biển dựa trên phân tích dữ liệu dạng văn bản bao gồm các bước: bước 1: gom nhóm các bản tin cần phân tích theo nhóm chủ đề được định nghĩa trước; bước 2: thực hiện tiền xử lý văn bản tùy theo loại dữ liệu; bước 3: phát hiện và phân loại các hoạt động được đề cập trong văn bản; bước 4: phân loại và trích xuất các thông tin chi tiết về đối tượng liên quan đến hoạt động đã được phân loại; bước 5: hợp nhất, theo dõi các mục tiêu được trích xuất theo thời gian, địa điểm. Phương pháp được ứng dụng vào các hệ thống tự động thu thập, phân tích, giám sát, cảnh báo sớm về tình huống trên biển từ dữ liệu được thu thập trên không gian mạng.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp theo dõi hoạt động của mục tiêu trên biển dựa trên phân tích dữ liệu dạng văn bản. Cụ thể, phương pháp được đề cập trong sáng chế được ứng dụng vào các hệ thống tự động thu thập, phân tích, giám sát, cảnh báo sớm về tình huống trên biển từ dữ liệu được thu thập trên không gian mạng.

Tình trạng kỹ thuật của sáng chế

Hiện nay, với sự phát triển mạnh mẽ của các hệ thống thông tin như mạng xã hội, trang tin điện tử trực tuyến, lượng thông tin mà mỗi người có thể tiếp cận là rất lớn. Trong đó, dữ liệu văn bản được xem là có số lượng nhiều nhất và phổ biến nhất hiện nay, do đó, đối với một lĩnh vực cụ thể, các nhà chức trách có thể cần phải có công cụ theo dõi tự động các đối tượng được đề cập trên mạng xã hội, báo chí điện tử để từ đó nắm bắt sớm thông tin và đưa ra các phương án giải quyết kịp thời.

Đối với dữ liệu dạng văn bản, các phương pháp xử lý ngôn ngữ tự nhiên được sử dụng nhằm mục đích biến đổi dữ liệu từ dạng phi cấu trúc là các văn bản thô sang dạng dữ liệu có cấu trúc, từ đó phân tích, phát hiện các thông tin tiềm ẩn có giá trị trong văn bản. Với số lượng lớn về nguồn dữ liệu cũng như tốc độ gia tăng dữ liệu, việc theo dõi và trích xuất thủ công bởi con người sẽ dẫn đến mất rất nhiều thời gian, công sức và chi phí bỏ ra, từ đó cần có một phương pháp hoặc hệ thống tự động có thể thu thập, trích xuất, phân tích các thông tin từ văn bản trên không gian mạng. Hiện nay, các phương pháp trích xuất thông tin từ dữ liệu văn bản như thông tin sự kiện, thông tin về quan hệ và thông tin về các thực thể có tên trong văn bản đã được nghiên cứu và đạt được các kết quả nhất định. Tuy nhiên, đa phần các phương pháp dừng lại ở mức nghiên cứu riêng lẻ và chưa thực sự được tích hợp vào một hệ thống cụ thể, đồng thời phạm trù và lĩnh vực của các phương pháp này mang tính tổng quát, bao hàm các thông tin mang tính phổ biến trong đời sống xã hội mà không đi sâu vào trong một lĩnh vực cụ thể, nên không thể có nhiều ứng dụng trong thực tế.

Với sự phát triển về hiệu năng, tốc độ tính toán của các công nghệ về mặt phần cứng như các bộ phận xử lý đồ họa (gọi tắt là GPU), các hướng tiếp cận dựa trên trí tuệ nhân tạo với các kỹ thuật, mô hình học sâu đang dần chiếm ưu thế và đạt được các kết quả đáng kể so với các phương pháp truyền thống. Sáng chế này mô tả phương pháp phân tích và trích xuất các thông tin liên quan đến các đối tượng trên biển, từ đó tự động hợp nhất, theo dõi các đối tượng, mục tiêu hoạt động trên biển được đề cập trên không gian mạng. Dữ liệu văn bản thô được thu thập từ các hệ thống mạng xã hội, trang tin tức trực tuyến sẽ được chuyển đổi thành các thông tin có ý nghĩa như loại hoạt động (được định nghĩa trước), đối tượng, thời gian, địa điểm diễn ra và một số thành phần khác ứng với từng loại hoạt động cụ thể.

Bản chất kỹ thuật của sáng chế

Mục tiêu của sáng chế là đưa ra phương pháp theo dõi các đối tượng, mục tiêu hoạt động trên khu vực biển thông qua việc phân tích và trích xuất các thông tin liên quan đến đối tượng từ dữ liệu văn bản thô. Phương pháp bao gồm bốn bước chính được thực hiện như sau:

Bước 1: gom nhóm các bản tin cần phân tích theo các nhóm chủ đề định nghĩa trước. Từ các bản tin dữ liệu dạng văn bản được thu thập, thông qua danh sách các chủ đề, từ khóa được định nghĩa trước, các văn bản không phù hợp với chủ đề cần phân tích, theo dõi sẽ được loại bỏ nhằm giảm thiểu số lượng tính toán, tối ưu tốc độ thực thi của phương pháp.

Bước 2: thực hiện tiền xử lý văn bản tùy theo loại dữ liệu. Với đặc thù dữ liệu văn bản đến từ nhiều nguồn, các bước tiền xử lý nhằm chuẩn hóa, lọc bỏ các yếu tố nhiễu, thống nhất về chung định dạng của đầu vào trước khi đưa vào các bước huấn luyện, kiểm định các mô hình ở các bước tiếp theo.

Bước 3: phát hiện và phân loại các hoạt động được đề cập trong văn bản. Dữ liệu văn bản sau khi đã được tiền xử lý sẽ được huấn luyện nhằm mục đích phát hiện và phân loại các loại hoạt động của đối tượng cùng với dấu hiệu nhận biết của từng loại hoạt động được thể hiện bằng các từ kích hoạt. Từ các loại hoạt động của đối tượng sẽ được sử dụng để kết hợp với các thông tin khác ở các bước tiếp theo phục vụ cho mục đích cuối cùng là theo dõi đối tượng được đề cập trên không giang mạng.

Bước 4: phân loại và trích xuất các thông tin chi tiết về đối tượng liên quan đến hoạt động đã được phân loại. Đầu vào là các văn bản kèm theo các loại hoạt động và từ kích hoạt được trích xuất ở bước trước, đầu ra là các thông tin, thành phần liên quan đến loại hoạt động được trích xuất.

Bước 5: hợp nhất, theo dõi các mục tiêu được trích xuất theo thời gian, địa điểm. Đầu vào của bước này là các văn bản kèm theo các loại hoạt động cùng thông tin tương ứng đã trích xuất được. Từ đó xây dựng phương pháp hợp nhất để đưa ra các bản tin cần theo dõi.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình vẽ minh họa sơ đồ tổng quát quy trình của phương pháp theo dõi và phân tích thông tin từ dữ liệu văn bản;

Hình 2: là hình vẽ minh họa quy trình gom nhóm các bản tin theo chủ đề;

Hình 3: là hình vẽ minh họa quy trình thực hiện phát hiện và phân loại hoạt động;

Hình 4: là hình vẽ minh họa quy trình thực hiện phân tích thông tin trong văn bản.

Mô tả chi tiết sáng chế

Sáng chế đề xuất phương pháp theo dõi hoạt động của các mục tiêu trên biển dựa trên phân tích dữ liệu dạng văn bản được mô tả qua quy trình tổng quát ở Hình 1. Một cách tổng quan, hoạt động của một mục tiêu trên biển được định nghĩa cũng như phụ thuộc vào rất nhiều yếu tố khác nhau như khu vực, các loại hình hoạt động, các sự kiện tác động vào mục tiêu. Do đó, phạm vi sáng chế này sẽ tập trung vào việc theo dõi các hoạt động của mục tiêu trên biển như: hoạt động của các tàu, giàn khoan, các sự kiện liên quan đến các đảo, đá ngầm, các loại vũ khí được lắp đặt, triển khai trên biển. Qua đó, chi tiết các bước thực hiện của phương pháp được trình bày cụ thể như sau:

Bước 1: gom nhóm các bản tin cần phân tích theo các nhóm chủ đề định nghĩa trước;

Dữ liệu trong thực tế được thu thập từ nhiều nguồn khác nhau, bao hàm nhiều chủ đề khác nhau được đề cập. Điều này dẫn đến việc gia tăng lượng bản tin phải xử lý, bao gồm các bản tin có chủ đề ngoài vùng quan tâm. Chính vì vậy, việc thực hiện gom nhóm trước khi xử lý và phân tích dữ liệu là cần thiết để tập trung hóa các chủ đề cần được quan tâm và chú trọng, đồng thời một phần giảm sai số của các bước phân loại cũng như phân tích ở các bước tiếp theo. Phương pháp mô hình hóa chủ đề được sử dụng nhằm mục đích tìm ra các từ khóa chính đại diện cho một chủ đề nhất định. Các chủ đề được định nghĩa trước, bao gồm các chủ đề cần quan tâm và chủ đề khác. Thông qua các từ khóa tương ứng với các chủ đề, các bản tin được phân nhóm theo các chủ đề được định nghĩa. Hình 2 mô tả quy trình khởi tạo và gom nhóm chủ đề.

Về phương pháp, đầu vào là một tập dữ liệu văn bản và một tập các chủ đề được định nghĩa trước, đầu ra là các từ khóa tìm được ứng với từng loại chủ đề, và dựa trên đó thực hiện phân loại chủ đề. Với một tập dữ liệu văn bản D và tập các chủ đề $C = \{c_1, \dots, c_n\}$, mô hình thực hiện trích xuất một tập từ khóa $S_i = \{w_1, \dots, w_m\}$ từ tập dữ liệu văn bản D tương ứng với từng loại chủ đề c_i.

Ban đầu, các từ khóa đại diện cho các chủ đề được mặc định là bằng chính các chủ đề. Trong quá trình huấn luyện, tập S_i dần dần tích hợp thêm các từ đại diện khác sao cho bộ từ khóa đại diện cho chủ đề được định nghĩa trở nên chính xác và mang nhiều ý nghĩa hơn. Quá trình mã hóa thông tin chủ đề riêng biệt được tối ưu thông qua hàm lỗi (1), và tín hiệu giám sát yếu được truyền qua các từ khác thông qua công thức (2) và (3). Hàm lỗi (1) là hàm lỗi phân kỳ Kullback-Leibler được sử dụng để tính toán độ hỗn loạn tương đối giữa hai phân bố xác suất. Theo góc nhìn về quá trình học bộ nhúng, hàm lỗi (1) thúc đẩy các bộ nhúng của từng chủ đề trở thành các điểm riêng biệt trong không gian nhúng sao cho chúng nằm xa nhau, đồng thời được vây quanh bởi các từ tượng trưng cho chủ đề đó. Công thức (2) và (3) mô tả sự kết hợp của một hệ số học k_w cho mỗi từ w nhằm tối ưu quá trình huấn luyện đồng thời của bộ từ nhúng và xác định tính cụ thể của phân phối xác suất các từ một cách không giám sát.

$$\mathcal{L}_{topic} = \sum_{c_i \in C} \sum_{w \in S_i} KL(l_w || p_w) \quad (1)$$

trong đó $\boldsymbol{l}_w$ là một véc-tơ với một giá trị là 1, còn lại là 0, $\boldsymbol{p}_w$ là phân bố xác suất của một từ w trên tất cả các lớp chủ đề. $\boldsymbol{p}_w$ có dạng $[p(c_1|w) \dots p(c_n|w)]^T$ với $p(c_i|w)$ là xác suất từ w thuộc vào lớp c_i.

$$p(w_i|d) = \frac{\exp\left(k_{w_i} u_{w_i}^T d\right)}{\sum_{d' \in D} \exp\left(k_{w_i} u_{w_i}^T d'\right)} \quad (2)$$

$$p\left(w_{i+j}|w_i\right) = \frac{\exp\left(k_{w_i} u_{w_i}^T v_{w_{i+j}}\right)}{\sum_{w' \in V} \exp\left(k_{w_i} u_{w_i}^T v_{w'}\right)} \quad (3)$$

trong đó $\boldsymbol{u}_w$ là véc-tơ biểu diễn đầu vào của một từ khóa w (thông thường là bộ nhúng từ), $\boldsymbol{v}_w$ là véc-tơ biểu diễn đầu ra của từ khóa w có xét yếu tố ngữ cảnh, $\boldsymbol{d}$ là bộ nhúng tài liệu (một tài liệu bao gồm nhiều từ khóa), $\boldsymbol{c}_i$ là bộ nhúng của chủ đề.

Bước 2: thực hiện tiền xử lý văn bản tùy theo loại dữ liệu;

Ở bước tiền xử lý chung, các kỹ thuật tiền xử lý được áp dụng bao gồm chuyển từ viết hoa thành viết thường và tách từ. Quá trình tách từ thực hiện phân tách văn bản thành các đơn vị nhỏ hơn là mức từ hoặc từ con. Thực nghiệm cho thấy việc chuyển các từ viết hoa thành viết thường đem lại kết quả tốt hơn so với việc chỉ tách từ. Đối với dữ liệu mạng xã hội, cần thực hiện thêm các kỹ thuật chuẩn hóa như xóa bỏ các kí tự đặc biệt khác trong văn bản, xóa bỏ các đường liên kết trang mạng (website).

Đối với dữ liệu báo chí, để giảm bớt khối lượng tính toán cho các bước sau, cần sử dụng thêm mô hình nhận dạng thực thể có tên, mô hình gán nhãn phần hội thoại để xác định các loại từ quan trọng. Sử dụng mô hình nhận dạng thực thể để loại bỏ đi các từ thuộc vào dạng thực thể có tên. Sử dụng mô hình gán nhãn phần hội thoại để xác định các loại từ như danh từ, động từ trong văn bản. Hình 3 mô tả quy trình tiền xử lý dữ liệu cho từng loại văn bản.

Xét một ví dụ cụ thể như sau: “*#China’s hypersonic weapon test in July over the #SouthChinaSea the firing of a missile as it approached its target travelling at least 5x* *the speed of sound — a capability no country has previously demonstrated; Pentagon caught off guard by advance. https://t.co/hADIIZGthu*".

Kết quả sau khi đi qua từng bước xử lý dữ liệu:

- Xóa bỏ liên kết: "*#China's hypersonic weapon test in July over the #SouthChinaSea the firing of a missile as it approached its target travelling at least 5x the speed of sound — a capability no country has previously demonstrated; Pentagon caught off guard by advance.*"
- Tách từ: "#", "*China*", "'", "*s*", "*hypersonic*", "*weapon*", "*test*", …
- Viết thường từ: "#", "*china*", "'", "*s*", "*hypersonic*", "*weapon*", "*test*", …

Bước 3: phát hiện và phân loại các hoạt động được đề cập trong văn bản;

Sau khi qua bước tiền xử lý dữ liệu, văn bản được xử lý tiếp tục được đưa qua mô hình phát hiện và phân loại hoạt động. Mô hình này hoạt động theo kiến trúc nhiều-một, nghĩa là đầu vào là số nhiều và đầu ra chỉ có một. Đây là dạng mô hình phổ biến thường được sử dụng cho bài toán phân loại văn bản trong các phương pháp xử lý ngôn ngữ tự nhiên sử dụng mạng học sâu. Ngoài ra, mô hình còn sử dụng cơ chế tập trung thông qua kiến trúc máy biến hình nhằm tăng độ chính xác cho quá trình phân loại khi kết hợp yếu tố ngữ nghĩa và yếu tố ngữ cảnh. Giống như với chủ đề, các hoạt động cần phân loại cũng được định nghĩa trước, và phải phù hợp với các chủ đề quan tâm.

Mô hình được huấn luyện trước trên tập dữ liệu văn bản vô cùng lớn, và trọng số của mô hình này tiếp tục được sử dụng để khởi tạo cho quá trình huấn luyện tiếp theo trên lượng dữ liệu ít hơn và là dữ liệu mục tiêu. Dữ liệu mục tiêu là các bản tin nằm trong các chủ đề được quan tâm, đồng thời mô tả đến các hoạt động liên quan đến khu vực biển. Mô hình được huấn luyện và tối ưu sử dụng phương pháp Adam. Sau khi huấn luyện, mô hình được sử dụng để phát hiện và phân loại các hoạt động của mục tiêu được đề cập trong văn bản.

Mục tiêu của mô hình này là xác định từ trọng tâm trong câu mà mô tả được một cách rõ nhất hoạt động được định nghĩa trước. Trong trường hợp này, tuy đầu vào là một câu văn bản, nhưng mong muốn là thực hiện phân loại trên mức từ thay vì mức câu. Chính vì vậy, trong quá trình dự đoán cũng như huấn luyện, thực hiện chèn thêm một số kí tự đặc biệt để đánh dấu vị trí của từ cần xét. Việc làm này giúp xác định được đâu là từ cần xác định để phân loại, đồng thời kết hợp được thông tin ngữ cảnh của cả câu văn bản. Cụ thể, cho một câu S bao gồm 1 chuỗi các từ $w_1 w_2 \dots w_i \dots w_n$ với $i \in \{1, \dots, n\}$ ứng với các thành phần của câu đã được tách, thực hiện thêm hai kí tự đặc biệt bao quanh từ mục tiêu

$$S = w_1 < special > w_2 < special > w_3 w_4 w_5 \dots \qquad (4)$$

Mô hình sử dụng kiến trúc máy biến hình, cụ thể là kiến trúc biểu diễn bộ mã hóa (sau đây gọi là Encoder) hai chiều từ máy biến hình, để mã hóa và ánh xạ thông tin đầu vào là các từ đã tách qua vùng không gian mới mang nhiều ý nghĩa hơn. Công thức (5) biểu diễn quá trình mã hóa, với h_i là giá trị mã hóa ứng với từ thứ i, $encoder$ là bộ mã hóa:

$$\{h_1, \dots, h_n\} = encoder(w_1, \dots, t, \dots, w_n) \qquad (5)$$

Sau khi có được bộ mã hóa, thực hiện cơ chế đa tổng hợp động trên các bộ mã hóa, từ đó có được véc-tơ tổng hợp x với $[.]_j$ là thành phần thứ j của một véc-tơ và i là vị trí của từ kích hoạt.

$$[\overleftarrow{x}]_j = \max\{[h_1]_j, \dots, [h_i]_j\}$$

$$[\vec{x}]_j = \max\{[h_{i+1}]_j, \dots, [h_n]_j\} \qquad (6)$$

$$x = [\overleftarrow{x}; \vec{x}]$$

Đầu ra của mô hình là phân bố xác suất của các loại hoạt động, chọn loại hoạt động có xác suất cao nhất. Quá trình dự đoán bắt ta phải thực hiện chạy qua từng từ thành phần của câu, tuy nhiên số lượng đã được giảm thiểu qua bước tiền xử lý ở bước trước. Với véc-tơ x tổng hợp được, thực hiện tính toán xác suất cho từng loại hoạt động sử dụng công thức SOFTMAX với $\sigma(\vec{z})_i$ là xác suất hoạt động thứ i xảy ra:

$$z = W.x$$

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{7}$$

Hàm lỗi cần tối ưu để cập nhật trọng số cho mô hình là hàm lỗi hỗn loạn chéo, có dạng như công thức (8) dưới đây:

$$f = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{8}$$

Bước 4: phân tích và trích xuất các thông tin chi tiết về đối tượng liên quan đến hoạt động đã được phân loại;

Sau khi thực hiện phân loại hoạt động dựa trên một từ trọng tâm (hay còn gọi là từ kích hoạt), sử dụng chính từ đó cũng như loại hoạt động làm đầu vào tiếp theo để trích xuất các thông tin chi tiết liên quan đến loại hoạt động đó. Hình 4 mô tả quy trình phân tích và trích xuất thông tin từ văn bản.

Mô hình ở phần này cũng sử dụng cơ chế tập trung qua kiến trúc máy biến hình, quá trình huấn luyện cũng được thực hiện tương tự như bước trước. Tuy nhiên, mô hình được sử dụng là dạng kiến trúc nhiều-nhiều, bài toán thường sử dụng là gán nhãn chuỗi, nghĩa là mỗi một từ thành phần sẽ có nhãn riêng của nó.

Mục tiêu của mô hình là xác định được vị trí của các thông tin liên quan đến hoạt động đã phân loại, cụ thể là xác định vị trí bắt đầu và kết thúc của một thành phần thông tin. Các thông tin này được định nghĩa dựa trên một lược đồ từ trước, ứng với từng loại hoạt động sẽ có các loại thông tin, thành phần cụ thể. Một số thông tin chung giữa các hoạt động là thời gian, địa điểm và đối tượng tham gia hoạt động đó.

Đầu vào của mô hình trích xuất thông tin sẽ bao gồm hai thành phần: một là các từ thành phần đã được tách thông qua bước tách từ kèm theo loại hoạt động, hai là một chuỗi phân khúc có độ dài bằng với số lượng các từ thành phần nhằm xác định vị trí của từ kích hoạt đã phát hiện ở bước trước. Cụ thể, thực hiện chèn thêm loại hoạt động vào trước vị trí của từ kích hoạt. Đối với chuỗi phân khúc, vị trí của từ kích hoạt và loại hoạt động được gán là 1, còn lại là 0:

$$S = w_1 < action > w_2\ w_3 w_4 w_5 \dots \tag{9}$$

$$segment = 011000 \tag{10}$$

Mô hình cũng sử dụng kiến trúc biểu diễn bộ mã hóa hai chiều từ máy biến hình như ở bước trước, tuy nhiên kiến trúc đầu ra lại có phần khác biệt. Sau khi đi qua bộ mã hóa có ngữ cảnh và có được các véc-tơ mã hóa cho từng thành phần như (11). Mô hình sau đó đi qua một lớp mạng nơ-ron tuyến tính và trả ra kết quả đầu ra là xác suất của từng loại thành phần liên quan đến hoạt động tại từng vị trí.

$$\{h_1, \dots, h_s, h_t, h_n\} = encoder(w_1, \dots, s, t, \dots, w_n) \tag{11}$$

Hàm lỗi được sử dụng cho qua trình tối ưu là hàm lỗi hỗn loạn chéo có dạng như (8) nhưng áp dụng cho đầu ra của từng từ.

Về loại thành phần liên quan đến hoạt động, mỗi loại hoạt động sẽ có tập các loại thành phần riêng biệt, và mỗi loại thành phần đó lại được chia làm hai loại: $B-$ và $I-$. Loại $B-$ tượng trưng cho vị trí của từ bắt đầu của một thành phần, loại $I-$ tương trưng cho những từ nằm trong thành phần. Ngoài ra còn có loại O, nghĩa là không nằm trong thành phần nào.

Ví dụ: "*A Romanian and a U.K citizen died on the 600-foot carrier Mercer Street when the vessel came under attack on Thursday while sailing from Tanzania's Dar es Salam to Fujairah in the United Arab Emirates*". Hoạt động ở đây là hoạt động tấn công, trong đó:

- *the*: $\boldsymbol{B - Target}$
- *600-foot*: $\boldsymbol{I - Target}$
- *carrier*: $\boldsymbol{I - Target}$
- *Mercer*: $\boldsymbol{I - Target}$
- *Street*: $\boldsymbol{I - Target}$

- *Thursday*: $\boldsymbol{B-Time}$

- Các từ còn lại có nhãn là $\boldsymbol{O}$.

Bước 5: hợp nhất, theo dõi các mục tiêu được trích xuất theo thời gian, địa điểm;

Bước cuối cùng là tổng hợp thông tin và thực hiện theo dõi đối tượng. Việc theo dõi sẽ được xem xét trên tiêu chí về mức độ ưu tiên của hoạt động đã định nghĩa, các bản tin đề cập các hoạt động có mức độ ưu tiên cao cần được xếp trên, từ đó thực hiện theo dõi các đối tượng liên quan đến các hoạt động đó. Ngoài ra, việc theo dõi còn được căn cứ trên thời gian diễn ra hoạt động và địa điểm diễn ra hoạt động. Các bản tin mô tả hoạt động với thời gian mới trong một ngưỡng nhất định, có thể là một đến hai ngày, hay các hoạt động diễn ra ở vùng trọng điểm sẽ cần mức ưu tiên cao hơn. Ngoài ra, chỉ các bản tin chứa đầy đủ các thông tin liên quan đến hoạt động cụ thể mới được xem xét.

Cụ thể, các bước thực hiện bao gồm:

- Sắp xếp và lọc các bản tin dựa trên mức độ ưu tiên của các loại hoạt động
- Xét trên các bản tin đã được lọc, thực hiện gom nhóm các bản tin miêu tả cùng đối tượng.
- Thực hiện sắp xếp các bản tin có cùng các đối tượng theo trình tự thời gian.
- Đưa ra các thông tin khác liên quan đến đối tượng trong các bản tin và loại hoạt động khác nhau.

Ví dụ ta xem xét hai bản tin đề cập đến cùng một đối tượng như sau:

Bản tin 1: "*The drone attack which targeted at the 600-foot carrier Mercer Street happened surprisingly without being noticed on Thursday*"

Bản tin 2: "*The carrier Mercer Street successfully avoided the Chinese drone attack at South China Sea and sailed to its homeland on Friday morning*"

Ở đây cả hai bản tin mô tả hai loại hoạt động khác nhau (*Attack* và *Travel*) nhưng có mô tả đến cùng một đối tượng *Mercer Street*, từ đó sắp xếp theo trình tự thời gian.

Yêu cầu bảo hộ

1. Phương pháp theo dõi hoạt động của mục tiêu trên biển dựa trên phân tích dữ liệu văn bản bao gồm:

bước 1: gom nhóm các bản tin cần phân tích theo các nhóm chủ đề định nghĩa trước; các chủ đề được định nghĩa trước, bao gồm các chủ đề cần quan tâm và chủ đề khác (không cần quan tâm), thông qua các từ khóa tương ứng với các chủ đề, các bản tin được phân nhóm theo các chủ đề được định nghĩa; theo đó, đầu vào là một tập dữ liệu văn bản và một tập các chủ đề được định nghĩa trước, đầu ra là các từ khóa tìm được ứng với từng loại chủ đề, và dựa trên đó thực hiện phân nhóm chủ đề; cụ thể:

với một tập dữ liệu văn bản D và tập các chủ đề $C = \{c_1, \dots, c_n\}$, mô hình thực hiện trích xuất một tập từ khóa $S_i = \{w_1, \dots, w_m\}$ từ tập dữ liệu văn bản D tương ứng với từng loại chủ đề c_i;

quá trình mã hóa thông tin chủ đề riêng biệt được tối ưu thông qua hàm lỗi (1), và tín hiệu giám sát yếu được truyền qua các từ khác thông qua công thức (2) và (3); hàm lỗi (1) là hàm lỗi phân kỳ Kullback-Leibler được sử dụng để tính toán độ hỗn loạn tương đối giữa hai phân bố xác suất; theo góc nhìn về quá trình học bộ nhúng, hàm lỗi (1) thúc đẩy các bộ nhúng của từng chủ đề trở thành các điểm riêng biệt trong không gian nhúng sao cho chúng nằm xa nhau, đồng thời được vây quanh bởi các từ tượng trưng cho chủ đề đó; công thức (2) và (3) mô tả sự kết hợp của một hệ số học k_w cho mỗi từ w nhằm tối ưu quá trình huấn luyện đồng thời của bộ từ nhúng và xác định tính cụ thể của phân phối xác suất các từ một cách không giám sát;

$$\mathcal{L}_{topic} = \sum_{c_i \in C} \sum_{w \in S_i} KL(l_w || p_w) \quad (1)$$

trong đó $\boldsymbol{l_w}$ là một véc-tơ với một giá trị là 1, còn lại là 0, $\boldsymbol{p_w}$ là phân bố xác suất của một từ w trên tất cả các lớp chủ đề, $\boldsymbol{p_w}$ có dạng $[p(c_1|w) \dots p(c_n|w)]^T$ với $p(c_i|w)$ là xác suất từ w thuộc vào lớp c_i;

$$p(w_i|d) = \frac{\exp\left(k_{w_i} u_{w_i}^T d\right)}{\sum_{d' \in D} \exp\left(k_{w_i} u_{w_i}^T d'\right)} \quad (2)$$

$$p(w_{i+j}|w_i) = \frac{\exp\left(k_{w_i} u_{w_i}^T v_{w_{i+j}}\right)}{\sum_{w' \in V} \exp\left(k_{w_i} u_{w_i}^T v_{w'}\right)} \quad (3)$$

trong đó $\boldsymbol{u}_w$ là véc-tơ biểu diễn đầu vào của một từ khóa w (thông thường là bộ nhúng từ), $\boldsymbol{v}_w$ là véc-tơ biểu diễn đầu ra của từ khóa w có xét yếu tố ngữ cảnh, $\boldsymbol{d}$ là bộ nhúng tài liệu (một tài liệu bao gồm nhiều từ khóa), $\boldsymbol{c}_i$ là bộ nhúng của chủ đề;

bước 2: thực hiện tiền xử lý văn bản tùy theo loại dữ liệu; ở bước này, các kỹ thuật tiền xử lý được áp dụng bao gồm:

chuyển từ viết hoa thành viết thường;

tách từ thực hiện phân tách văn bản thành các đơn vị nhỏ hơn là mức từ hoặc từ con;

đối với dữ liệu mạng xã hội, cần thực hiện thêm các kỹ thuật chuẩn hóa như xóa bỏ các kí tự đặc biệt khác trong văn bản, xóa bỏ các đường liên kết trang mạng (website);

đối với dữ liệu báo chí, để giảm bớt khối lượng tính toán cho các bước sau, cần sử dụng thêm mô hình nhận dạng thực thể có tên, mô hình gán nhãn phần hội thoại để xác định các loại từ quan trọng; sử dụng mô hình nhận dạng thực thể để loại bỏ đi các từ thuộc vào dạng thực thể có tên; sử dụng mô hình gán nhãn phần hội thoại để xác định các loại từ như danh từ, động từ trong văn bản;

bước 3: phát hiện và phân loại các hoạt động được đề cập trong văn bản; sau khi qua bước tiền xử lý dữ liệu, văn bản được xử lý tiếp tục được đưa qua mô hình phát hiện và phân loại hoạt động; mô hình này hoạt động theo kiến trúc nhiều-một, nghĩa là đầu vào là số nhiều và đầu ra chỉ có một; ngoài ra, mô hình còn sử dụng cơ chế tập trung thông qua kiến trúc máy biến hình nhằm tăng độ chính xác cho quá trình phân loại khi kết hợp yếu tố ngữ nghĩa và yếu tố ngữ cảnh; giống như với chủ đề, các hoạt động cần phân loại cũng được định nghĩa trước, và phải phù hợp với các chủ đề quan tâm; trong quá trình dự đoán cũng như huấn luyện của mô hình, thực hiện chèn thêm một số kí tự đặc biệt để đánh dấu vị trí của từ cần xét; việc làm này giúp xác định được đâu là từ cần xác định để phân loại, đồng thời kết hợp được thông tin ngữ cảnh của cả câu văn bản; cụ thể:

cho một câu S bao gồm 1 chuỗi các từ $w_1 w_2 \dots w_i \dots w_n$ với $i \in \{1, \dots, n\}$ ứng với các thành phần của câu đã được tách, thực hiện thêm hai kí tự đặc biệt bao quanh từ mục tiêu

$$S = w_1 < special > w_2 < special > w_3 w_4 w_5 \dots \tag{4}$$

mô hình sử dụng kiến trúc máy biến hình, cụ thể là kiến trúc biểu diễn bộ mã hóa hai chiều từ máy biến hình, để mã hóa và ánh xạ thông tin đầu vào là các từ đã tách qua vùng không gian mới mang nhiều ý nghĩa hơn; công thức (5) biểu diễn quá trình mã hóa, với h_i là giá trị mã hóa ứng với từ thứ i, $encoder$ là bộ mã hóa:

$$\{h_1, \dots, h_n\} = encoder(w_1, \dots, t, \dots, w_n) \tag{5}$$

sau khi có được bộ mã hóa, thực hiện cơ chế đa tổng hợp động trên các bộ mã hóa, từ đó có được véc-tơ tổng hợp x với $[.]_j$ là thành phần thứ j của một véc-tơ và i là vị trí của từ kích hoạt;

$$\begin{gathered}
[\overleftarrow{x}]_j = \max\{[h_1]_j, \dots, [h_i]_j\} \\
[\overrightarrow{x}]_j = \max\{[h_{i+1}]_j, \dots, [h_n]_j\} \\
x = [\overleftarrow{x}; \overrightarrow{x}]
\end{gathered} \tag{6}$$

đầu ra của mô hình là phân bố xác suất của các loại hoạt động, chọn loại hoạt động có xác suất cao nhất; quá trình dự đoán bắt ta phải thực hiện chạy qua từng từ thành phần của câu, tuy nhiên số lượng đã được giảm thiểu qua bước tiền xử lý ở bước trước;

với véc-tơ x tổng hợp được, thực hiện tính toán xác suất cho từng loại hoạt động sử dụng công thức SOFTMAX với $\sigma(\vec{z})_i$ là xác suất hoạt động thứ i xảy ra:

$$z = W.x$$

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{7}$$

hàm lỗi cần tối ưu để cập nhật trọng số cho mô hình là hàm lỗi hỗn loạn chéo, có dạng như công thức (8) dưới đây:

$$f = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{8}$$

bước 4: phân tích và trích xuất các thông tin chi tiết về đối tượng liên quan đến hoạt động đã được phân loại;

sau khi thực hiện phân loại hoạt động dựa trên một từ trọng tâm (hay còn gọi là từ kích hoạt), sử dụng chính từ đó cũng như loại hoạt động làm đầu vào tiếp theo để trích xuất các thông tin chi tiết liên quan đến loại hoạt động đó; mô hình ở phần này cũng sử dụng cơ chế tập trung qua kiến trúc máy biến hình, quá trình huấn luyện cũng được thực hiện tương tự như bước trước; tuy nhiên, mô hình được sử dụng là dạng kiến trúc nhiều-nhiều, bài toán thường sử dụng là gán nhãn chuỗi, nghĩa là mỗi một từ thành phần sẽ có nhãn riêng của nó;

đầu vào của mô hình trích xuất thông tin sẽ bao gồm hai thành phần: một là các từ thành phần đã được tách thông qua bước tách từ kèm theo loại hoạt động, hai là một chuỗi phân khúc có độ dài bằng với số lượng các từ thành phần nhằm xác định vị trí của từ kích hoạt đã phát hiện ở bước trước; cụ thể, thực hiện chèn thêm loại hoạt động vào trước vị trí của từ kích hoạt; đối với chuỗi phân khúc, vị trí của từ kích hoạt và loại hoạt động được gán là 1, còn lại là 0:

$$S = w_1 < action > \boldsymbol{w}_2\ w_3 w_4 w_5 \ldots \tag{9}$$

$$segment = 011000 \tag{10}$$

mô hình cũng sử dụng kiến trúc biểu diễn bộ mã hóa hai chiều từ máy biến hình như ở bước trước, tuy nhiên kiến trúc đầu ra lại có phần khác biệt; sau khi đi qua bộ mã hóa có ngữ cảnh và có được các véc-tơ mã hóa cho từng thành phần như (11); mô hình sau đó đi qua một lớp mạng nơ-ron tuyến tính và trả ra kết quả đầu ra là xác suất của từng loại thành phần liên quan đến hoạt động tại từng vị trí:

$$\{h_1, \ldots, h_s, h_t, h_n\} = encoder(w_1, \ldots, s, t, \ldots, w_n) \tag{11}$$

hàm lỗi được sử dụng cho qua trình tối ưu là hàm lỗi hỗn loạn chéo có dạng như (8) nhưng áp dụng cho đầu ra của từng từ;

về loại thành phần liên quan đến hoạt động, mỗi loại hoạt động sẽ có tập các loại thành phần riêng biệt, và mỗi loại thành phần đó lại được chia làm hai loại: $B-$ và $I-$; loại $B-$ tượng trưng cho vị trí của từ bắt đầu của một thành phần, loại $I-$ tương trưng cho những từ nằm trong thành phần; ngoài ra còn có loại O, nghĩa là không nằm trong thành phần nào;

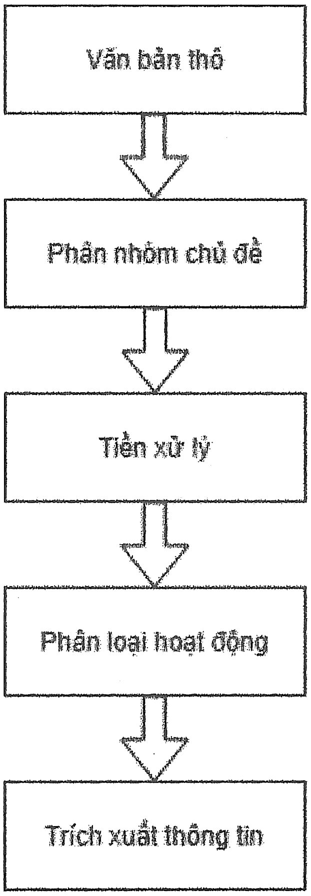
bước 5: hợp nhất, theo dõi các mục tiêu được trích xuất theo thời gian, địa điểm; cụ thể, các bước thực hiện bao gồm:

sắp xếp và lọc các bản tin dựa trên mức độ ưu tiên của các loại hoạt động;

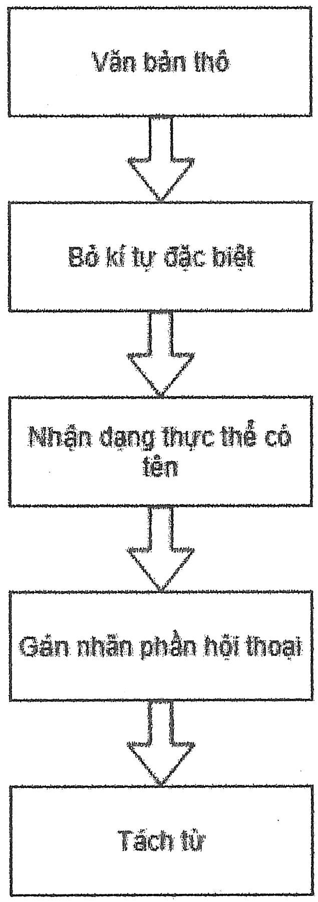
xét trên các bản tin đã được lọc, thực hiện gom nhóm các bản tin miêu tả cùng đối tượng;

thực hiện sắp xếp các bản tin có cùng các đối tượng theo trình tự thời gian;

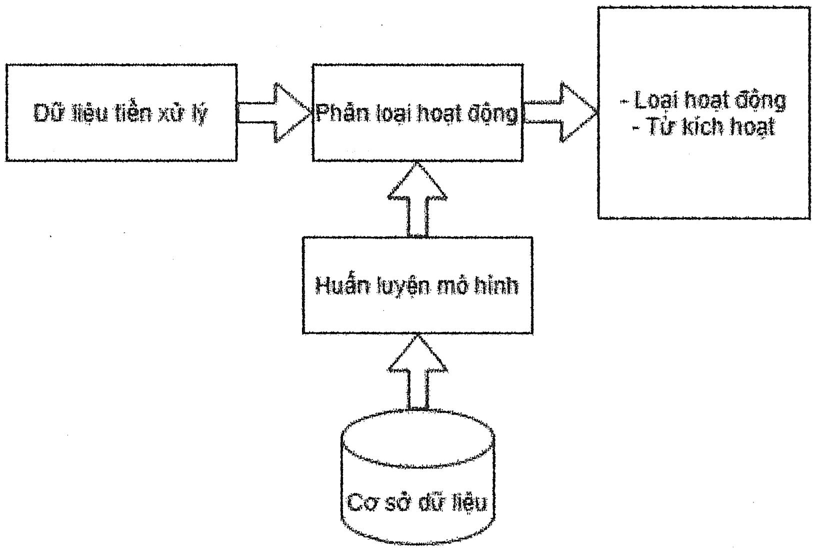
đưa ra các thông tin khác liên quan đến đối tượng trong các bản tin và loại hoạt động khác nhau.


Văn bản thô
Phân nhóm chủ đề
Tiền xử lý
Phân loại hoạt động
Trích xuất thông tin

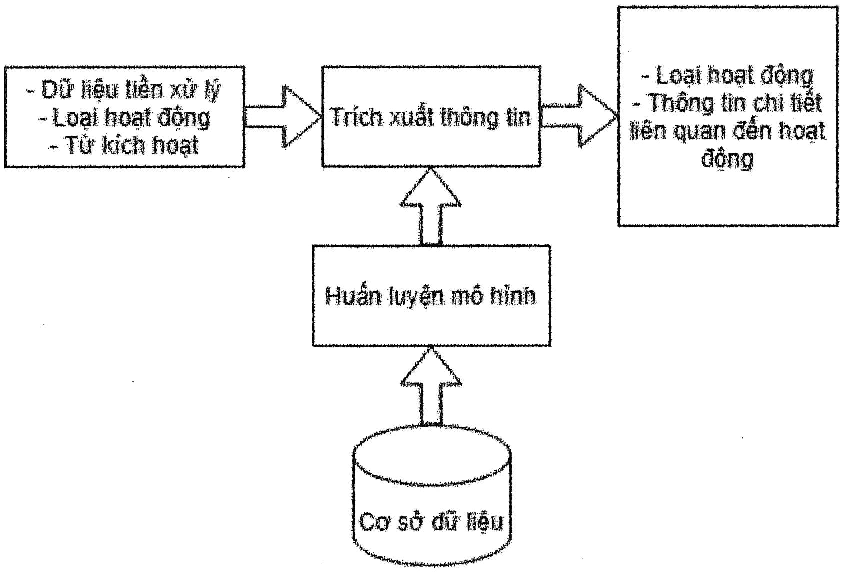
Hình 1


Văn bản thô
Bỏ kí tự đặc biệt
Nhận dạng thực thể có tên
Gán nhãn phần hội thoại
Tách từ

Hình 2


Dữ liệu tiền xử lý
Phân loại hoạt động
- Loại hoạt động
- Từ kích hoạt
Huấn luyện mô hình
Cơ sở dữ liệu

Hình 3


- Dữ liệu tiền xử lý
- Loại hoạt động
- Từ kích hoạt
Trích xuất thông tin
- Loại hoạt động
- Thông tin chi tiết liên quan đến hoạt động
Huấn luyện mô hình
Cơ sở dữ liệu

Hình 4