(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) **1-0053648**

(51)[7] **G10L 15/00**

(13) **B**

(21) 1-2018-05992

(22) 27/12/2018

(45) 25/11/2025 452

(43) 27/05/2019 374A

(73) Tập đoàn Công nghiệp - Viễn thông Quân đội (VN)
Số 1 đường Trần Hữu Dực, phường Mỹ Đình 2, quận Nam Từ Liêm, thành phố Hà Nội.

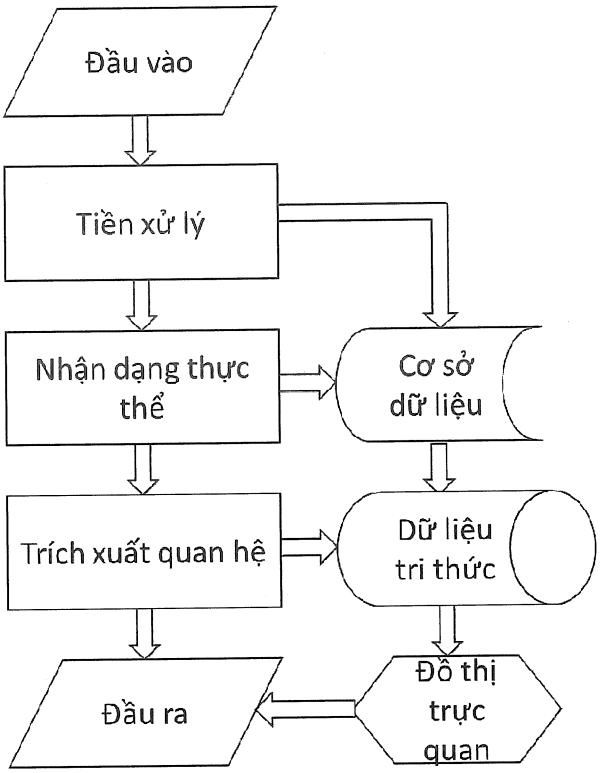
(72) Bùi Tăng Bảo Ngọc (VN); Chu Văn Tạo (VN).

(74) Công ty TNHH Tư vấn Quốc Dân (NACILAW)

(54) HỆ THỐNG TỰ ĐỘNG TRÍCH XUẤT QUAN HỆ GIỮA CÁC THỰC THỂ XUẤT HIỆN TRONG VĂN BẢN TIẾNG VIỆT

(21) 1-2018-05992

(57) Sáng chế hệ thống trích xuất tự động quan hệ giữa các thực thể xuất hiện trong văn bản tiếng việt đã xử lý tự động trên tập dữ liệu rất lớn thu thập được từ những nguồn khác nhau trên mạng internet hay các nguồn văn bản có sẵn. Việc sử dụng phương pháp tiếp cận bán giám sát giúp giảm thiểu thời gian và công sức cần thiết để gán nhãn dữ liệu trên một khối lượng dữ liệu quá lớn vượt quá khả năng của con người. Kết quả của sáng chế cũng có thể coi là tiêu chuẩn giúp các hệ thống trích xuất thông tin dựa trên những kết quả này tinh chỉnh kết quả tốt hơn.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến một hệ thống tự động trích xuất thông tin từ dữ liệu thu thập được dựa trên các tin tức, báo chí, mạng xã hội từ internet hay nguồn dữ liệu định dạng văn bản từ các nguồn khác nhằm mục đích tự động phát hiện những mỗi quan hệ giữa các thực thể xuất hiện trong đó.

Tình trạng kỹ thuật của sáng chế

Phần lớn giao tiếp giữa con người, cho dù đó là trong văn bản, lời nói hoặc ngôn ngữ tự nhiên thường không có cấu trúc. Các ngữ nghĩa cần thiết để giải thích thông tin phi cấu trúc này thường tiềm ẩn đằng sau và được bắt nguồn bằng cách sử dụng thông tin cơ bản và suy luận. Dữ liệu phi cấu trúc tương phản với dữ liệu có cấu trúc, chẳng hạn như dữ liệu trong các bảng cơ sở dữ liệu truyền thống, nơi dữ liệu được xác định rõ và ngữ nghĩa rõ ràng. Khi dữ liệu có cấu trúc được sử dụng, các truy vấn được chuẩn bị để trả lời các câu hỏi được xác định trước trên cơ sở kiến thức cần thiết và đầy đủ về ý nghĩa của các tiêu đề bảng (ví dụ: Tên, Địa chỉ, Mục, Giá và Ngày). Với sự phổ biến rộng rãi của nội dung điện tử trên web hiện nay và trong các doanh nghiệp, thông tin phi cấu trúc (ví dụ: văn bản, hình ảnh và lời nói) đang phát triển nhanh hơn nhiều so với thông tin có cấu trúc. Cho dù đó là tài liệu tham khảo chung, sách giáo khoa, tạp chí, hướng dẫn kỹ thuật, tiểu sử hoặc trang cá nhân, nội dung này chứa kiến thức có giá trị cao thường rất quan trọng đối với việc ra quyết định sáng suốt. Thông tin từ các trang web, mạng xã hội đã trở thành một kho dữ liệu khổng lồ, theo như thống kê thì trong vòng 5 năm gần đấy đã có gần 60 tỷ trang web được tạo ra, hay như trang web live-counter.com chỉ ra rằng dữ liệu hiện nay có trên internet là hơn 14000 petabytes tương đương 14.000.000.000.000.000.000 bít tức là nếu coi một từ tương đương với 1 bít thì chúng ta hiện có xấp xỉ 10^{18} từ hay tương đương 10^{15} bài báo, trong đó chúng chứa nhiều loại thông tin ngữ nghĩa có giá trị về các thực thể trong thế giới thực, chẳng hạn như về con người, tổ chức hay địa điểm, thời gian,... Thật không may, những mối quan hệ này có thể không phải lúc nào cũng có thể tự động phát hiện, tự động xác định

hoặc thậm chí có thể tìm kiếm được. Trong nhiều trường hợp, các mối quan hệ này chỉ có thể được phát hiện bằng con người và mất rất nhiều thời gian công sức. Tuy nhiên, do lượng dữ liệu hiện có trên Web là quá lớn, việc nhập thủ công nhận dạng mối quan hệ như vậy sẽ quá tốn thời gian và kém hiệu quả để có thể phát hiện ra những thông tin như vậy.

Hiện tại trên thế giới chưa có các công cụ hoàn thiện để phát hiện và trích xuất hiệu quả thông tin mối quan hệ thực thể một cách đầy đủ trọn vẹn. Các công cụ trích xuất hiện tại chỉ đơn thuần xác định và trích xuất thông tin dựa trên các mối quan hệ được chỉ định trước. Một số giải pháp đã được đề xuất trên thực tế được nêu sau đây có một số nhược điểm nhất định ảnh hưởng đến khả năng ứng dụng trên thực tế.

Để giải quyết bài toán xác định mối quan hệ giữa các thực thể, ta thường chuyển về trực tiếp bài toán phân loại. Cho trước cặp thực thể cùng xuất hiện trong câu, ta cần phân loại có mối quan hệ giữa hai thực thể này hay không và nếu có thì mối quan hệ giữa hai thực thể này được phân loại vào một trong các quan hệ được định nghĩa trước. Để đơn giản, ta chỉ quan tâm đến hai thực thể trong một câu, bài toán có thể mở rộng thành quan hệ giữa hai thực thể nằm ở các câu khác nhau. Các phương pháp phân loại dựa trên lựa chọn đặc trưng (feature-based) đặc biệt quan trọng trong bài toán này. Từ đây, ta có thể áp dụng phương pháp Nhân (kernel-based) sử dụng các thuật toán Support-Vector-Machine (SVM) để thực hiện phân loại. Kernel ở đây có thể xây dựng dựa trên chuỗi các từ loại, cây phụ thuộc (dependency trees), cây phân tích cú pháp (syntactic parse trees). Cả hai phương pháp feature-based và kernel-based đều cần một tập ngữ liệu rất lớn để huấn luyện. Ngoài ra có một hướng tiếp cận sử dụng mạng học sâu (Deep learning) một nhánh của phương pháp học có giám sát cũng đang trở thành xu hướng gần đây khi mà tốc độ tính toán xử lý của các máy tình ngày càng cao. Các mạng tích chập nơ ron (Convolutional Neural Networks-CNNs) thường được sử dụng trong việc học có giám sát nhằm tìm kiếm các mối quan hệ giữa các thực thể. Các kết quả tốt nhất hiện nay theo hướng tiếp cận này có thể kể đến như CNNs của Thien Nguyen và Grishman 2016, Piecewise CNNs (PCNNs) của Zeng và các cộng sự vào năm 2015, PCNN + Attention của Lin 2016 hay Multi-instance Multi-label CNNs của Jiang 2016. Kỹ thuật chính cho việc học tập được giám sát có tên là Mồi

(bootstrapping), bằng cách boostrapping đối với tập dữ liệu nhỏ trước, sau đó dần dần tìm ra các quan hệ mới cũng như các mẫu (pattern) tìm được trong corpus. Tuy nhiên việc học có giám sát đòi hỏi rất nhiều tài liệu được dán nhãn để đào tạo mô hình, cái này đòi hỏi rất nhiều công sức và tiền bạc không phải doanh nghiệp nào cũng có điều kiện đề đầu tư. Do đó, việc học bán giám sát chỉ cần một lượng rất nhỏ dữ liệu được làm mồi ban đầu để có thể rút trích quan hệ từ tập dữ liệu văn bản (corpus) lớn mà không đòi hỏi nhiều quá trình gán nhãn ban đầu là một phương pháp giúp chúng ta có thể giải quyết vấn đề này.

Đặc biệt, đối với tiếng việt chưa có một hệ thống trọn vẹn nào được đề xuất và xây dựng riêng nhằm mục đích trích xuất thông tin. Do đó, cần có các hệ thống và phương pháp trích xuất mối quan hệ đủ mạnh để xác định các mối quan hệ mới và xử lý lượng dữ liệu trên quy mô dữ liệu lớn từ các thông tin đa dạng trên internet.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng Việt có khả năng xác định các mối quan hệ trong ngữ nghĩa câu tiếng Việt trên một lượng lớn dữ liệu quy mô lớn từ các thông tin đa dạng trên internet.

Để đạt được mục đích trên, sáng chế đề xuất hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt bao gồm khối tiền xử lý, khối nhận dạng thực thể, khối trích xuất quan hệ, khối dữ liệu và hiển thị, trong đó

khối tiền xử lý thực hiện tiền xử lý dữ liệu văn bản thu được, tốt nhất nếu sử dụng mô đun học máy huấn luyện CBOW;

khối nhận dạng thực thể thực hiện tìm kiếm và gán nhãn tất cả các thực thể (đối tượng) trong văn bản; sử dụng mô hình học Mạng bộ nhớ dài ngắn hai chiều (Bidirectional Long Short-Term Memory) và Trường điều kiện ngẫu nhiên (Conditional Random Field);

khối trích xuất các mối quan hệ giữa các thực thể bao gồm mô đun trích xuất các mối quan hệ giữa các thực thể (RE), thực hiện phát hiện và trích xuất được mối quan hệ giữa các thực thể có trong cơ sở dữ liệu;

khối dữ liệu và hiển thị, sử dụng nền tảng hệ quản trị cơ sở dữ liệu đồ thị neo4j, bao gồm giao diện hiển thị và hệ thống xây dựng dữ liệu tri thức.

Hơn nữa, theo phương án thực hiện sáng chế, hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt trong đó, khối tiền xử lý bao gồm các công cụ trọn vẹn cho đối tượng là văn bản tiếng việt: lọc các thẻ html, tách từ, tách câu, biến đổi từ, cụm từ thành véc tơ, phân tích cú pháp phụ thuộc trong câu.

Hơn nữa, theo phương án thực hiện sáng chế, hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt trong đó, tại khối trích xuất các mối quan hệ giữa các thực thể, việc trích xuất tự động các mối quan hệ giữa các thực thể được thực hiện theo số vòng lặp định nghĩa trước, bao gồm các bước:

bước 1: cung cấp tập các hạt mẫu để làm đầu vào cho phương pháp;

bước 2: khởi động vòng lặp với điều kiện dừng là khi không có tập hạt mẫu mới được sinh ra;

bước 3: tìm tập các cặp tương tự hạt mẫu xuất hiện trong văn bản bằng cách tính *ĐộTươngTự* của một bộ quan hệ xuất hiện trong văn bản;

bước 4: trích xuất tất cả các nhóm chứa cặp quan hệ có nghĩa bằng cách chỉ chọn những bộ mẫu mới có *ĐộTươngTự* lớn hơn một ngưỡng đặt trước τ_{nguong} thì cặp thực thể sẽ được phân nhóm với nhau;

bước 5: sinh ra các cặp quan hệ mới bằng cách hệ thống sẽ quét lại lần nữa tất cả bộ dữ liệu có cùng định dạng với bộ hạt mẫu, đồng thời đi tìm những bộ quan hệ nào có độ tương tự lớn nhất với các nhóm đã tìm được ở trên sẽ phân loại chúng vào các nhóm đó;

bước 6: sinh ra các hạt mẫu mới bằng cách đi tính độ $Tincay_p$ trong mỗi nhóm ở trên, sau đó loại bỏ những bộ quan hệ nào có độ $Tincay_p$ nhỏ hơn ngưỡng. Những bộ quan hệ nào được chọn tiếp sẽ là các hạt mẫu mới

bước 7: tiếp tục vòng lặp với đầu vào là tập các hạt mẫu mới ở bước 6.

Hơn nữa, theo phương án thực hiện sáng chế, hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt trong đó, tại khối trích xuất các

mối quan hệ giữa các thực thể sẽ được sử dụng cơ sở dữ liệu đồ thị xây dựng thành cơ sở dữ liệu tri thức gồm các bước;

bước 1: so sánh giá trị các thực thể mới trích xuất được với các giá trị đã có trong các bảng dữ liệu về thực thể NGƯỜI, CƠ QUAN/TỔ CHỨC, ĐỊA ĐIỂM, THỜI GIAN, nếu đã có thực hiện đồng nhất dữ liệu đã có, nếu chưa tạo dữ liệu thực thể mới;

bước 2: với trường hợp cả hai thực thể trích xuất được giống với thực thể đã tồn tại thực hiện so sánh quan hệ mới trích xuất được với tập quan hệ đã có nếu giống không tạo mới quan hệ, nếu quan hệ mới trích xuất được chưa xuất hiện thì thực hiện tao mới quan hệ, trong trường hợp một trong hai thực thực thể khác với thực thể đã có thực hiện tạo mới mối quan hệ giữa hai thực thể.

Hơn nữa, theo phương án thực hiện sáng chế, hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt trong đó, để tìm những quan hệ giống nhau giữa các câu khác nhau chứ các cặp thực thể quan hệ với nhau chúng ta sử dụng công thức:

$$\begin{aligned} ĐộTươngTự(S_i, S_j) = {} & \alpha \cdot \cos(TRUOC_i, TRUOC_j) \\ & + \beta \cdot \cos(GIUA_i, GIUA_j) \\ & + \gamma \cdot \cos(SAU_i, SAU_j) \end{aligned} \quad (5)$$

trong đó, S_i, S_j là hai câu chứa cặp thực thể như miêu tả ở (1), α, β, γ là các hệ số trọng số để tính toán độ tương tự giữa hai câu, TRUOC, GIUA, SAU là véc tơ có được sau khi sử dụng bộ biến đổi từ thành véc tơ ở bước 1 của hệ thống.

Mô tả vắn tắt hình vẽ

Hình 1 là hình vẽ thể hiện lược đồ tổng quát hệ thống tự động trích xuất quan hệ giữa các thực thể;

Hình 2 là hình vẽ thể hiện nội dung chi tiết của lược đồ trong Hình 1;

Hình 3 là hình vẽ thể hiện chi tiết module trích xuất các mối quan hệ giữa các thực thể trong Hình 2;

Hình 4 là hình vẽ thể hiện cơ sở dữ liệu đồ thị ở mức tổng quát trong việc xây dựng mô đun dữ liệu tri thức ở Hình 3;

Hình 5 là hình vẽ thể hiện ở mức chi tiết cơ sở dữ liệu đồ thị của Hình 4;

Hình 6 là hình vẽ thể hiện một ví dụ minh họa bằng mạng lưới biểu diễn quan hệ biểu diễn các mối quan hệ giữa hai thực thể người tự động trích xuất được từ hệ thống.

Mô tả chi tiết sáng chế

Mội mối quan hệ thường hiểu là mối quan hệ giữa hai thực thể (đối tượng) xuất hiện trong một câu hay giữa các câu trong văn bản. Ví dụ, với câu *"Anh Ngọc sống ở Hà Nội"*, có một quan hệ giữa thực thể *Người (Anh Ngọc)* và thực thể *Vị trí (Hà Nội)*. Tổng quát hóa, giả sử một câu *S* được biểu diễn dưới dạng một chuỗi của *n* từ (được ký hiệu bởi w_i với $i \in \{1, \dots, n\}$), tồn tại thực thể e_1 (là một từ hay cụm từ) có quan hệ *r* với thực thể e_2 (là một từ hay cụm từ)

$$S = w_1 w_2 \dots e_1 \dots e_2 \dots w_{n-1} w_n \quad (1)$$

Quan hệ *r* giữa hai thực thể e_1 và e_2 được biểu diễn dưới dạng bộ dữ liệu như sau:

$$(e_1, e_2, r) \quad (2)$$

Trong sáng chế vì trong quá trình phân tích tìm hiểu các lĩnh vực tiếng việt khác nhau, tác giả nhận thấy có những mối quan hệ xảy ra phụ thuộc thời gian ví dụ như chức danh, hay quan hệ giữa người với người,...vì thế chúng tôi đưa vào một định nghĩa mới với những loại quan hệ dạng này được mô tả như sau

$$(e_1, e_2, r, t) \quad (3)$$

trong đó t chỉ thời gian. Và mục đích của một hệ thống trích xuất quan hệ giữa các thực thể thực chất là một hệ thống có thể học được hàm phân loại F_r có biểu thức như sau:

$$F_r = f(x) = \begin{cases} 1, \text{ nếu } e_1 \text{ có quan hệ với } e_2 \text{ theo quan hệ } r \\ -1, \text{ ngược lại} \end{cases} \quad (4)$$

Tham chiếu hình 1, hệ thống tự động trích xuất quan hệ giữa các thực thể xuất hiện trong văn bản tiếng việt bao gồm các khối thành phần: khối tiền xử lý, khối nhận dạng thực thể, khối trích xuất quan hệ, Khối dữ liệu và hiển thị. Trong đó, tại mỗi khối thành phần, khác biệt ở chỗ:

Khối tiền xử lý thực hiện tiền xử lý dữ liệu văn bản thu được, cụ thể là thực hiện tách câu, tách từ, tách cụm từ, word2vec, nhãn cú pháp cho các từ, cây phân tích cú pháp phụ thuộc,..

Đối tượng xử lý đầu vào tại khối này là các dữ liệu văn bản từ nhiều nguồn trên internet như: báo chí, tin tức, trang cá nhân, mạng xã hội facebook, twitter hay cơ sở dữ liệu mở như wikipedia, freebase, dbpedia,…

Tiếp đến các dữ liệu thô thu thập được sẽ được xử lý làm sạch.

Phần sau đây đề cập đến ví dụ làm sạch tập dữ liệu thô như sau:

Phần sau đây là một đoạn trong tập dữ liệu thô sau khi tải về từ trang vnexpress.net *"<link rel="dns-prefetch" href= "//www.google-analytics.com"/> <link rel="dns-prefetch" href="//www.googleta gmanager.com"/> <meta name="apple-mobile-web-app-capable" content="yes"/> <meta name="apple-mobile-web-app-title" content="Vnexpress.net"/> <meta name="description" content="Thông tin nhanh & mới nhất được cập nhật hàng giờ. Tin tức Việt Nam & thế giới về xã hội, kinh doanh, pháp luật, khoa học, công nghệ, sức khoẻ, đời sống, văn hóa, rao vặt, tâm sự..."> "*.

Dữ liệu này bằng cách sử dụng các công cụ biểu thức chính quy (regular express viết tắt là regexp, regex) sẽ được loại bỏ các thẻ html, để chỉ lấy các thông tin nội dung quan trọng trong văn bản, như ví dụ trên thông tin sau khi làm sạch sẽ có dạng *content = "Thông tin nhanh & mới nhất được cập nhật hàng giờ. Tin tức Việt Nam & thế giới về xã hội, kinh doanh, pháp luật, khoa học, công nghệ, sức khoẻ, đời sống, văn hóa, rao vặt, tâm sự..."* .

Các dữ liệu này sau đó sẽ được một tác vụ khác của hệ thống xử lý bằng cách tách văn bản thành các câu, tách câu thành các phần nhỏ hơn như cụm từ hay từ, ngoài ra sẽ có một mô đun khác trong hệ thống sử dụng mô hình học máy phù hợp để thực

hiện việc huấn luyện trên tập mẫu lớn để chuyển các từ, cụm từ thành các véc tơ với số chiều 300 chiều.

Theo phương án thực hiện sáng chế, có thể sử dụng các mô hình nơ ron học máy khác như Skip-gram, hay các mô hình học sâu,….

Hơn nữa, theo phương án thực hiện sáng chế, tốt nhất nếu hệ thống sử dụng mô hình học máy CBOW (Continuous Bag-of-Words) để thực hiện việc huấn luyện.

Ngoài ra cần thực hiện thêm việc phân tích cây cú pháp phụ thuộc vì đối với một câu dài hay với một đoạn văn bản việc loại bỏ bớt các từ không cần thiết trong việc tìm hiểu ngữ nghĩa của các từ trong câu tạo nên quan hệ giữa các thực thể giúp loại bỏ bớt nhiễu ảnh hưởng đến kết quả phân tích việc sử dụng cây cú pháp phụ thuộc giúp chúng ta chỉ giữ lại những phần quan trọng trong câu hay trong đoạn văn bản do đó loại bỏ được những thành phần không có nghĩa.

Tất cả dữ liệu thu được sau khi đi ra khối tiền xử lý được lưu trữ vào cơ sở dữ liệu để phục vụ cho việc xử lý dữ liệu tiếp theo.

Khối nhận dạng thực thể thực hiện tìm kiếm và gán nhãn tất cả các thực thể (đối tượng) trong văn bản. Khối nhận dạng thực thể sử dụng mô hình học Mạng bộ nhớ dài ngắn hai chiều (Bidirectional Long Short-Term Memory) và Trường điều kiện ngẫu nhiên (Conditional Random Field) là mô hình tốt nhất hiện nay đối với bộ dữ liệu lớn chúng tôi thu thập bao gồm 1.575.000 bài báo tiếng việt thu thập trên 35 website phổ biến nhất hiện này trong nước và thời gian lấy trong vòng 2 năm trở lại.

Trong sáng chế này chúng tôi đề cập đến trong phươn án thực hiện sang chế một ứng dụng của hệ thống liên quan đến tự động trích xuất các mối quan hệ của con người trong xã hội.

Theo một phương án thực hiện sáng chế, để thực hiện việc theo dõi các mối quan hệ của một người dựa trên những tiến hành phân tích ngôn ngữ trên tập bộ dữ liệu thu được, bộ các thực thể đề xuất cần trích xuất bao gồm:

- NGƯỜI, ví dụ như: Nguyễn Văn Anh, ông Hùng, bà Mai, …
- CƠ QUAN/TỔ CHỨC, ví dụ như: Tập đoàn công nghiệp viễn thông Viettel, Đảng cộng sản Việt Nam, Trường Đại học Bách khoa Hà Nội, …

- ĐỊA ĐIỂM, ví dụ như: Số 1, Đại Cồ Việt, Hà nội; thành phố Đà nẵng; sông Hồng; Việt nam; tòa nhà landmark 72; …
- THỜI GIAN, ví dụ như: 11:00 ngày 24 tháng 9 năm 2013; buổi sáng ngày 15 tháng 9 năm 2008; mùa xuân năm 68; …

Theo phương án thực hiện sáng chế, tương ứng với những ứng dụng trong lĩnh vực khác, hoàn toàn có thể sử dụng cấu trúc này của hệ thống để trích xuất thông tin tuy nhiên cần phải có chuyên gia phân tích lại các thực thể để phù hợp với lĩnh vực mới.

Thông tin thực thể được trích xuất như miêu tả ở trên sẽ được chuyển hóa theo định dạng nhất định và lưu trữ vào cơ sở dữ liệu. Cụ thể, tương ứng với từng bộ thực thể chính được gắn với từ/ cụm từ thể hiện nội dung đó tại đoạn văn bản đang xem xét. Theo đó, thu được dữ liệu trích xuất về thông tin thực thể trong cơ sở dữ liệu.

Ví dụ sau đây mô tả kết quả sau khi đã được chuyển hóa thành định dạng thích hợp và lưu trữ vào cơ sở dữ liệu:

"<NGƯỜI>Nguyễn Phú Trọng</NGƯỜI> sinh ngày <THỜI GIAN>14 tháng 4 năm 1944 </THỜI GIAN> tại <ĐỊA ĐIỂM> thôn Lại Đà, xã Hội Phụ, tổng Hội Phụ, huyện Đông Anh, phủ Từ Sơn, tỉnh Bắc Ninh </ĐỊA ĐIỂM> (sau là <ĐỊA ĐIỂM> thôn Lại Đà, xã Đông Hội, huyện Đông Anh, ngoại thành Hà Nội </ĐỊA ĐIỂM>). <NGƯỜI> Ông </NGƯỜI> hiện cư trú nhà công vụ ở <ĐỊA ĐIỂM> Số 5, Thiền Quang, phường Nguyễn Du, quận Hai Bà Trưng, thành phố Hà Nội </ĐỊA ĐIỂM>".

Khối trích xuất các mối quan hệ giữa các thực thể bao gồm mô đun trích xuất các mối quan hệ giữa các thực thể (RE), thực hiện phát hiện và trích xuất được mối quan hệ giữa các thực thể có trong cơ sở dữ liệu. Các cặp thực thể cùng với mối quan hệ giữa chúng sẽ được lưu trữ theo định dạng như đã mô tả ở (2) hoặc (3). Cụ thể, tập mẫu có thể lựa chọn biểu diễn (2) nếu là mối quan hệ thông thường A có quan hệ r với B, hoặc (3) A có quan hệ r với B tại thời gian t.

Với ví dụ trên chúng ta có mô tả các mối quan hệ giữa các thực thể như sau:

<Nguyễn Phú Trọng, 14 tháng 4 năm 1944, sinh ngày>; <Nguyễn Phú Trọng, thôn Lại Đà_xã Hội Phụ_tổng Hội Phụ_huyện Đông Anh_phủ Từ Sơn_tỉnh Bắc Ninh, sinh tại>; <Nguyễn Phú Trọng, Số 5_Thiền Quang_phường Nguyễn Du_quận Hai Bà Trưng_thành phố Hà Nội, hiện cư trú>.

Việc trích xuất mối quan hệ thực thể ở quy mô dữ liệu lớn có thể được thực hiện bằng cách nhận các hạt quan hệ mẫu làm đầu vào cho một quy trình lặp. Trong ngữ cảnh này, các hạt mẫu mối quan hệ có thể là các bộ dữ liệu quan hệ ban đầu (nghĩa là, danh sách dữ liệu mối quan hệ được sắp xếp) có chứa nhận dạng của các thực thể nhất định (như người, công ty, thời gian hoặc địa điểm) và mối quan hệ của chúng. Các hạt quan hệ có thể được tạo thành từ các thực thể được tìm thấy trong cơ sở dữ liệu hoặc một hoặc nhiều từ khóa quan hệ. Trong quá trình lặp lại, các bộ dữ liệu mới có thể được trích xuất, các mẫu mới có thể được tạo và chọn từ các bộ dữ liệu được trích xuất và các mẫu được chọn sau đó có thể được sử dụng làm đầu vào để tiếp tục trích xuất các mối quan hệ mới. Quá trình lặp có thể xác định và trích xuất các bộ dữ liệu mối quan hệ mới từ cơ sở dữ liệu cho đến khi không có bộ dữ liệu mối quan hệ mới nào được trích xuất. Ngoài ra, tác vụ trích xuất có thể được xác định ở cấp độ thực thể, cấp độ câu, cấp độ trang hoặc cấp độ văn bản. Cuối cùng, các bộ dữ liệu mối quan hệ được trích xuất có thể được nhóm lại để kết nối các bộ dữ liệu cùng loại và dữ liệu mối quan hệ có thể được xuất ra cho các mục đích khác nhau. Quá trình lặp này được mô tả như trong Hình 3.

Phương pháp tự động trích xuất quan hệ ở khối trích xuất các mối quan hệ giữa các thực thể thông qua một quá trình lặp. Tuy nhiên để hiểu rõ hơn quá trình lặp này ở đây chúng tôi giới thiệu một số khác niệm và thuật toán chính để giải thích cho quá trình này.

Đối với ngôn ngữ tiếng Việt các mối quan hệ giữa hai thực thể xuất hiện trong các trường hợp sau:

Trường hợp 1: Các từ giữa hai thực thể trong một câu báo hiệu một mối quan hệ. Mẫu của trường hợp này thường được mô tả e_1 <*words*> e_2.

Ví dụ 1: Ông Nguyễn Anh Tuấn <đang làm việc cho> công ty Viettel.

Ví dụ 2: Chị Đào Thị Hạnh <đang sống ở> Hà Nội.

Ví dụ 3: Hội nghị bảo tồn văn hóa dân tộc <tổ chức tại > thành phố Huế vào ngày 21/3/2001. Trong các ví dụ trên, các cụm từ *"làm việc cho", "sống ở", "tổ chức tại"* là dấu hiệu cho mối quan hệ giữa "NGƯỜI-CƠ QUAN/TỔ CHỨC", "NGƯỜI-ĐỊA ĐIỂM", "CƠ QUAN/TỔ CHỨC – ĐỊA ĐIỂM".

Trường hợp 2: Các từ ở phía trước và giữa hai thực thể trong một câu báo hiệu một mối quan hệ. Mẫu của trường hợp này thường được mô tả <*words*> e_1 <*words*> e_2.

Ví dụ 4: <Đám cưới giữa> ông Hà Văn Đông <và> cô Vũ Thị Lê Trang diễn ra tại khách sạn Nha Trang.

Ví dụ 5: <Mùa hè vừa rồi> đội bóng Nghệ An đã <chuyển sân nhà về sân vận động mới> tại phường Lê Mao.

Trong hai ví dụ trên, các cụm từ *"Đám cưới giữa", "và"* là dấu hiệu cho mối quan hệ giữa *"NGƯỜI-NGƯỜI" trong quan hệ vợ chồng, "CƠ QUAN/TỔ CHỨC-ĐỊA ĐIỂM"*.

Trường hợp 3: Các từ ở giữa và phía sau hai thực thể trong một câu báo hiệu một mối quan hệ. Mẫu của trường hợp này thường được mô tả e_1 <*words*> e_2 <*words*>.

Ví dụ 6: Anh Nam <và> cô Mai <đã ly hôn>.

Ví dụ 7: Anh Hiếu <với> anh Thiện <cùng theo đuổi cô gái ấy>.

Trong hai ví dụ trên, các cụm từ *"và" "đã ly hôn", "với", "cùng theo đuổi cô gái ấy"* là dấu hiệu cho mối quan hệ giữa *"NGƯỜI-NGƯỜI" trong quan hệ vợ chồng, yêu đương.*

Dựa trên những phân tích này để tìm những quan hệ giống nhau giữa các câu khác nhau chứ các cặp thực thể quan hệ với nhau chúng ta sử dụng công thức:

$$\begin{aligned} \text{ĐộTươngTự}(S_i, S_j) = {} & \alpha \cdot \cos(TRUOC_i, TRUOC_j) \\ & + \beta \cdot \cos(GIUA_i, GIUA_j) \\ & + \gamma \cdot \cos(SAU_i, SAU_j) \end{aligned} \quad (5)$$

Trong đó, S_i, S_j là hai câu chứa cặp thực thể như miêu tả ở (1), α, β, γ là các hệ số trọng số để tính toán độ tương tự giữa hai câu, TRUOC, GIUA, SAU là véc tơ có được sau khi sử dụng bộ biến đổi từ thành véc tơ ở khối tiền xử lý của hệ thống. Hệ thống sẽ quét toàn bộ tập dữ liệu và tìm kiếm nếu có cả hai thực thể của một hạt mẫu cùng xuất hiện trong một câu hay trong một câu trước hoặc sau trong văn bản, sau đó câu hay đoạn đó sẽ được xem xét và hệ thống sẽ trích xuất ba bối cảnh văn bản TRUOC, GIUA, SAU. Nếu như *ĐộTươngTự* của một bộ mẫu mới lớn hơn một ngưỡng đặt trước τ_{nguong} thì cặp thực thể sẽ được phân nhóm.

Ngoài ra chúng tôi cũng sử dụng độ tin cậy để loại bỏ bớt những thực thể nào được trích xuất mà không có nghĩa tức là những mối quan hệ giữa hai thực thể nếu có độ tin cậy trên ngưỡng thì sẽ được sử dụng làm hạt mẫu cho vòng lặp tiếp theo. Công thức tính độ tin cây như sau:

$$Tincay_P = \frac{|P|}{|P|+W_{sai}\cdot|S|+W_{kb}\cdot|K|} \quad (6)$$

Trong đó, P, S, K là tập các trích xuất đúng, sai và không biết được chọn lựa tương ứng với thực thể đúng hoàn toàn với tập hạt mẫu, đúng một phần với tập hạt mẫu và không có thực thế nào giống tập mẫu, W_{sai} và W_{kb} là các trọng số tương ứng với các trích xuất sai và không biết. Ví dụ, nếu chúng ta cần tìm các mối quan hệ trụ sở chính giữa CƠ QUAN và ĐỊA ĐIỂM, ta có thể sử dụng hạt mẫu như *<tập đoàn Viettel, Trần Hữu Dực_Nam Từ Liêm_Hà Nội >*, *<Fpt, phố Duy Tân_phường Dịch Vọng Hậu_quận Cầu Giấy_Hà Nội >*, *<Vinamilk, phố Tân Trào_Phường Tân Phú_Quận 7_Hồ Chí Minh>*. Dựa trên bộ dữ liệu lớn từ internet chúng ta thu thập được, chúng ta đi tìm các cặp thực thể này đồng thời xuất hiện trong văn bản. Ta giả định rằng nếu hai thực thể này đồng thời xuất hiện gần với quan hệ trong hạt mẫu thì nhiều khả năng đây chính là quan hệ mà ta đang tìm. Ví dụ, ta có thể tìm thấy các mẫu câu như "Trụ sở chính – Tập đoàn công nghiệp Viettel : Số 1, Trần Hữu Dực, Nam Từ Liêm, Hà Nội, Việt Nam", "Tòa nhà tập đoàn Fpt tọa lạc tại số 17 phố Duy tân phường Dịch Vọng Hậu, quận Cầu Giấy, Hà Nội, Việt Nam" và trích xuất ra được pattern "Trụ sở chính – CƠ QUAN, ĐỊA ĐIỂM", "CƠ QUAN tọa lạc ĐỊA ĐIỂM". Với mẫu tìm được, ta lại

có thể tìm kiếm tiếp <*CƠ QUAN, ĐỊA ĐIỂM*> có điểm tương đồng với quan hệ trụ sở chính. Ta thêm cặp thực thể này vào bộ hạt mẫu và lặp lại quá trình. Điểm quan trọng là làm sao ta lọc ra các quan hệ không liên quan.

Phương pháp tự động trích xuất các mối quan hệ giữa hai thực thể xuất hiện trong văn bản được mô tả trong hình 3 theo các bước sau đây:

Bước 1. Cung cấp tập các hạt mẫu để làm đầu vào cho phương pháp;

Bước 2. Khởi động vòng lặp với điều kiện dừng là khi không có tập hạt mẫu mới được sinh ra;

Bước 3. Tìm tập các cặp tương tự hạt mẫu xuất hiện trong văn bản bằng cách tính *ĐộTươngTự* của một bộ quan hệ xuất hiện trong văn bản;

Bước 4. Trích xuất tất cả các nhóm chứa cặp quan hệ có nghĩa bằng cách chỉ chọn những bộ mẫu mới có *ĐộTươngTự* lớn hơn một ngưỡng đặt trước τ_{nguong} thì cặp thực thể sẽ được phân nhóm với nhau;

Bước 5. Sinh ra các cặp quan hệ mới bằng cách hệ thống sẽ quét lại lần nữa tất cả bộ dữ liệu có cùng định dạng với bộ hạt mẫu, đồng thời đi tìm những bộ quan hệ nào có độ tương tự lớn nhất với các nhóm đã tìm được ở trên sẽ phân loại chúng vào các nhóm đó;

Bước 6. Sinh ra các hạt mẫu mới bằng cách đi tính độ $Tincay_p$ trong mỗi nhóm ở trên, sau đó loại bỏ những bộ quan hệ nào có độ $Tincay_p$ nhỏ hơn ngưỡng. Những bộ quan hệ nào được chọn tiếp sẽ là các hạt mẫu mới;

Bước 7. Tiếp tục vòng lặp với đầu vào là tập các hạt mẫu mới ở bước 6.

Theo phương án đề xuất, các quan hệ không liên quan được loại bỏ trong quá trình lặp bằng hai điều kiện loại bỏ những quan hệ dưới ngưỡng cho phép ở bước 4 và bước 6 nêu trên.

Khối dữ liệu và hiển thị, sử dụng hệ quản trị cơ sở dữ liệu đồ thị neo4j, bao gồm giao diện hiển thị và hệ thống xây dựng dữ liệu tri thức.

Các bộ mô tả mối quan hệ giữa các cặp thực thể trích xuất này sẽ được lưu trữ vào bộ cơ sở dữ liệu tri thức dưới dạng cơ sở dữ liệu đồ thị (Graph database) chúng tôi đề xuất giải pháp sử dụng hệ quản trị cơ sở dữ liệu đồ thị neo4j để lưu trữ cơ sở dữ liệu tri thức này. Việc thiết lập hệ thống cơ sở dữ liệu đồ thị được mô tả ở mức kiến trúc tổng quát như trong Hình 4 hay ở mức độ chi tiết như trong Hình 5.

Các dữ liệu lưu trữ dưới dạng đồ thị này trong hệ quản trị cơ sở dữ liệu đồ thị neo4j sẽ được hiển thị trực tiếp ở đầu ra hệ thống thông qua một giao diện xây dựng sẵn hoặc qua bước trung chuyển trung gian bằng các cho người dùng truy vấn dựa trên bộ cơ sở dữ liệu tri thức và hiển thị qua các mạng lưới quan hệ giữa các thực thể như minh họa ở Hình 6.

Yêu cầu bảo hộ

1. Hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt bao gồm khối tiền xử lý, khối nhận dạng thực thể, khối trích xuất quan hệ, khối dữ liệu và hiển thị, trong đó

khối tiền xử lý thực hiện tiền xử lý dữ liệu văn bản thu được, tốt nhất nếu sử dụng mô đun học máy huấn luyện CBOW;

khối nhận dạng thực thể thực hiện tìm kiếm và gán nhãn tất cả các thực thể (đối tượng) trong văn bản; sử dụng mô hình học mạng bộ nhớ dài ngắn hai chiều (Bidirectional Long Short-Term Memory) và trường điều kiện ngẫu nhiên (Conditional Random Field);

khối trích xuất các mối quan hệ giữa các thực thể bao gồm mô đun trích xuất các mối quan hệ giữa các thực thể (RE), thực hiện phát hiện và trích xuất được mối quan hệ giữa các thực thể có trong cơ sở dữ liệu;

khối dữ liệu và hiển thị, sử dụng nền tảng hệ quản trị cơ sở dữ liệu đồ thị neo4j, bao gồm giao diện hiển thị và hệ thống xây dựng dữ liệu tri thức.

2. Hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt theo điểm 1, trong đó,

khối tiền xử lý bao gồm các công cụ trọn vẹn cho đối tượng là văn bản tiếng việt: lọc các thẻ html, tách từ, tách câu, biến đổi từ, cụm từ thành véc tơ, phân tích cú pháp phụ thuộc trong câu.

3. Hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt theo các điểm bất kỳ từ điểm 1 đến điểm 2 trong đó tại khối trích xuất các mối quan hệ giữa các thực thể,

việc trích xuất tự động các mối quan hệ giữa các thực thể được thực hiện theo số vòng lặp định nghĩa trước, bao gồm các bước:

bước 1: cung cấp tập các hạt mẫu để làm đầu vào cho phương pháp;

bước 2: khởi động vòng lặp với điều kiện dừng là khi không có tập hạt mẫu mới được sinh ra;

bước 3: tìm tập các cặp tương tự hạt mẫu xuất hiện trong văn bản bằng cách tính *ĐộTươngTự* của một bộ quan hệ xuất hiện trong văn bản;

bước 4: trích xuất tất cả các nhóm chứa cặp quan hệ có nghĩa bằng cách chỉ chọn những bộ mẫu mới có *ĐộTươngTự* lớn hơn một ngưỡng đặt trước τ_{nguong} thì cặp thực thể sẽ được phân nhóm với nhau;

bước 5: sinh ra các cặp quan hệ mới bằng cách hệ thống sẽ quét lại lần nữa tất cả bộ dữ liệu có cùng định dạng với bộ hạt mẫu, đồng thời đi tìm những bộ quan hệ nào có độ tương tự lớn nhất với các nhóm đã tìm được ở trên sẽ phân loại chúng vào các nhóm đó;

bước 6: sinh ra các hạt mẫu mới bằng cách đi tính độ $Tincay_p$ trong mỗi nhóm ở trên, sau đó loại bỏ những bộ quan hệ nào có độ $Tincay_p$ nhỏ hơn ngưỡng, những bộ quan hệ nào được chọn tiếp sẽ là các hạt mẫu mới;

bước 7: tiếp tục vòng lặp với đầu vào là tập các hạt mẫu mới ở bước 6.

4. Hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt theo các điểm bất kỳ từ điểm 1 đến điểm 3 trong đó tại khối trích xuất các mối quan hệ giữa các thực thể sẽ được sử dụng cơ sở dữ liệu đồ thị xây dựng thành cơ sở dữ liệu tri thức gồm các bước;

bước 1: so sánh giá trị các thực thể mới trích xuất được với các giá trị đã có trong các bảng dữ liệu về thực thể NGƯỜI, CƠ QUAN/TỔ CHỨC, ĐỊA ĐIỂM, THỜI GIAN, nếu đã có thực hiện đồng nhất dữ liệu đã có, nếu chưa tạo dữ liệu thực thể mới;

bước 2: với trường hợp cả hai thực thể trích xuất được giống với thực thể đã tồn tại thực hiện so sánh quan hệ mới trích xuất được với tập quan hệ đã có nếu giống không tạo mới quan hệ, nếu quan hệ mới trích xuất được chưa xuất hiện thì thực hiện tao mới quan hệ, trong trường hợp một trong hai thực thực thể khác với thực thể đã có thực hiện tạo mới mối quan hệ giữa hai thực thể.

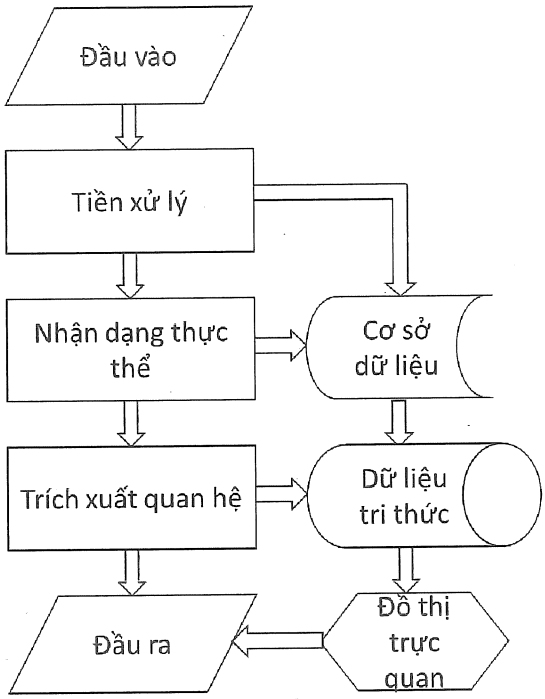
5. Hệ thống trích xuất tự động các mối quan hệ giữa các thực thể trong văn bản tiếng việt theo các điểm bất kỳ từ điểm 1 đến điểm 4, theo đó để phù hợp với đặc trưng

ngôn ngữ tiếng việt, cụ thể, để tìm những quan hệ giống nhau giữa các câu khác nhau chứ các cặp thực thể quan hệ với nhau chúng ta sử dụng công thức:

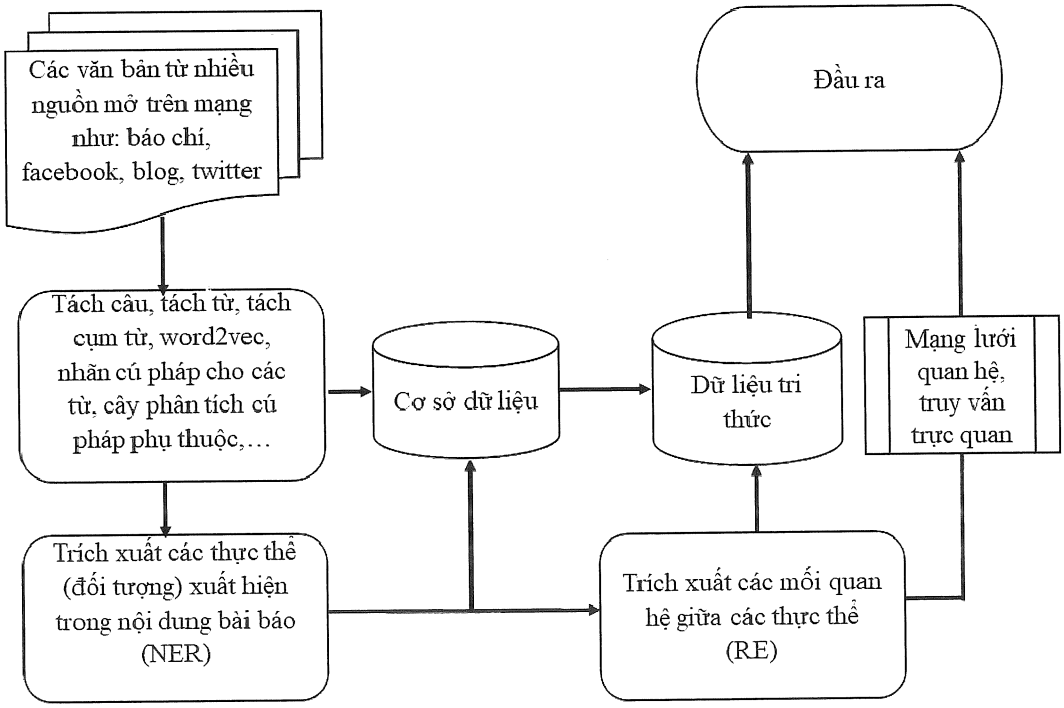
$$\begin{aligned} \text{ĐộTươngTự}(S_i, S_j) = {} & \alpha \cdot \cos(TRUOC_i, TRUOC_j) \\ & + \beta \cdot \cos(GIUA_i, GIUA_j) \\ & + \gamma \cdot \cos(SAU_i, SAU_j) \end{aligned} \quad (5)$$

trong đó, S_i, S_j là hai câu chứa cặp thực thể như miêu tả ở (1), α, β, γ là các hệ số trọng số để tính toán độ tương tự giữa hai câu, TRUOC, GIUA, SAU là véc tơ có được sau khi sử dụng bộ biến đổi từ thành véc tơ ở khối tiền xử lý của hệ thống.

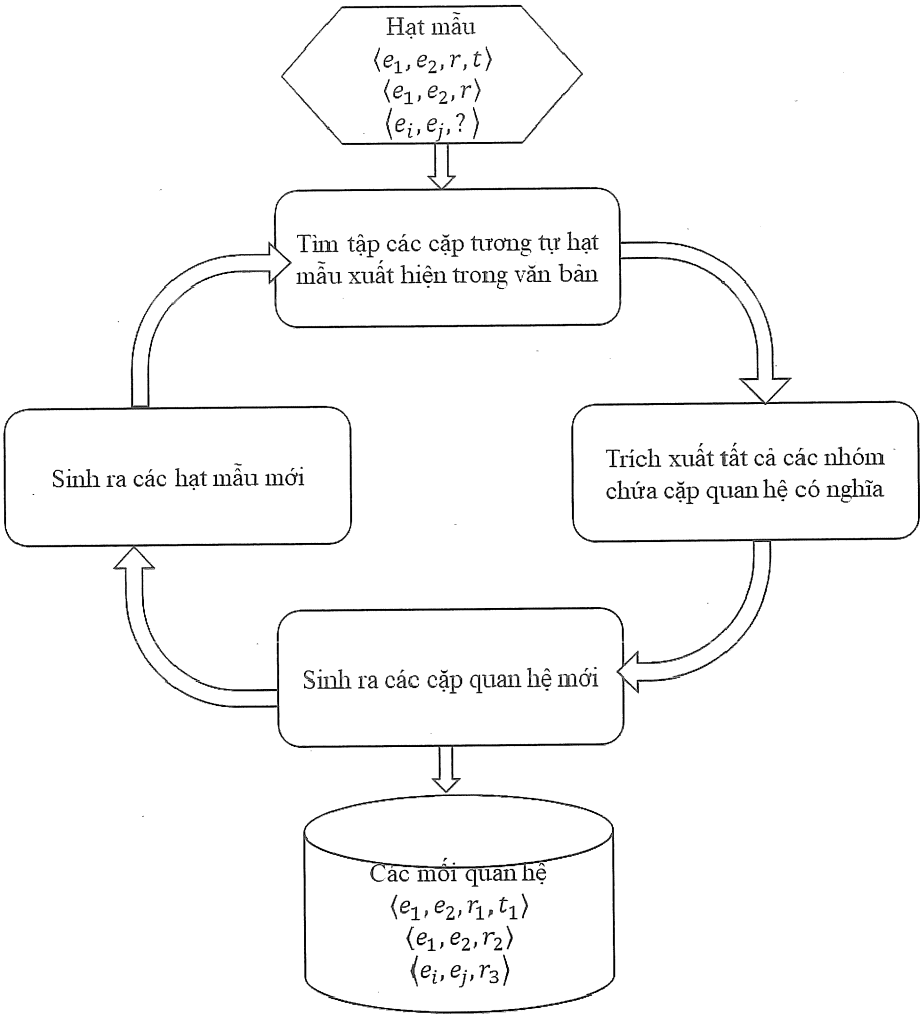
Hình vẽ



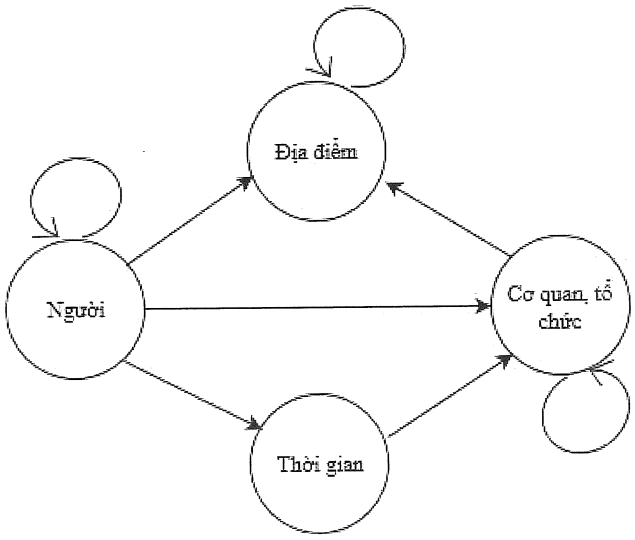
Hình 1



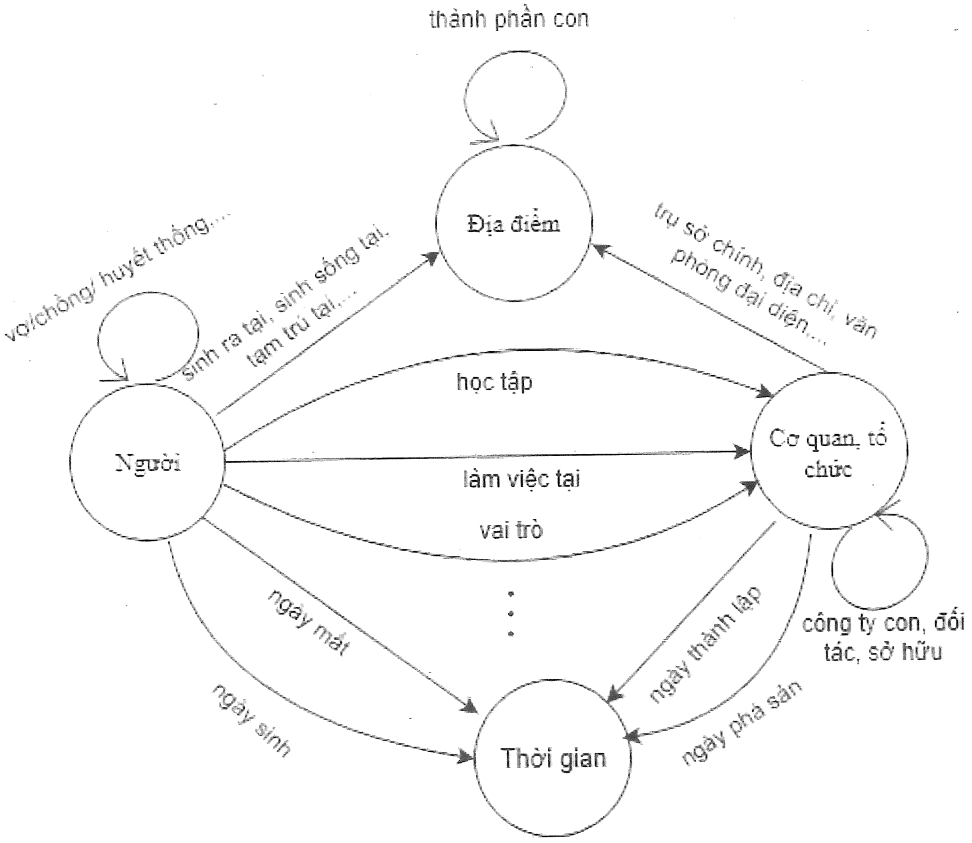
Hình 2


Hạt mẫu
$\langle e_1, e_2, r, t \rangle$
$\langle e_1, e_2, r \rangle$
$\langle e_i, e_j, ? \rangle$
Tìm tập các cặp tương tự hạt mẫu xuất hiện trong văn bản
Trích xuất tất cả các nhóm chứa cặp quan hệ có nghĩa
Sinh ra các cặp quan hệ mới
Sinh ra các hạt mẫu mới
Các mối quan hệ
$\langle e_1, e_2, r_1, t_1 \rangle$
$\langle e_1, e_2, r_2 \rangle$
$\langle e_i, e_j, r_3 \rangle$

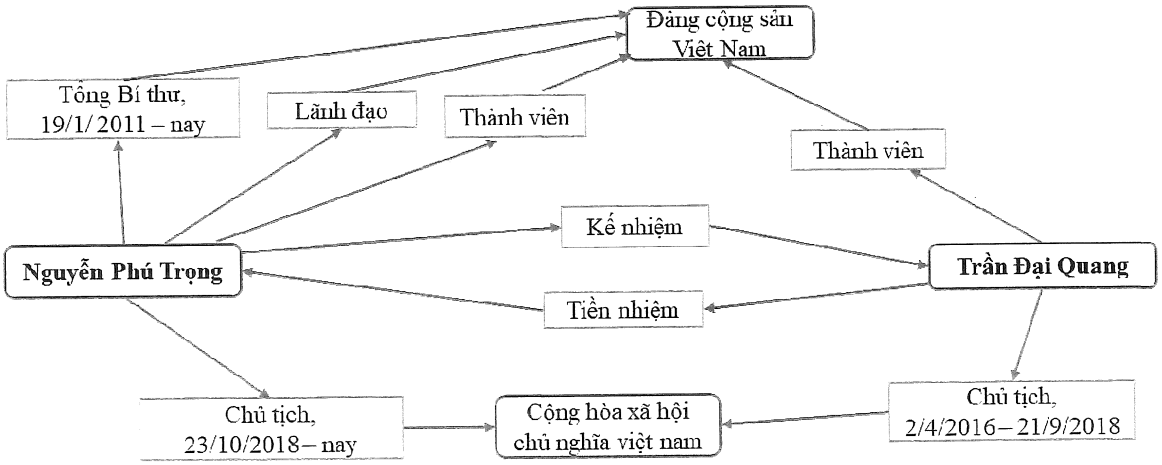
Hình 3


Địa điểm
Người
Cơ quan, tổ chức
Thời gian

Hình 4


thành phần con
Địa điểm
vợ/chồng/ huyết thống....
sinh ra tại, sinh sống tại, tạm trú tại....
trụ sở chính, địa chỉ, văn phòng đại diện....
học tập
Người
Cơ quan, tổ chức
làm việc tại
vai trò
ngày mất
ngày thành lập
công ty con, đối tác, sở hữu
ngày sinh
ngày phá sản
Thời gian

Hình 5


Đảng cộng sản Việt Nam
Tổng Bí thư, 19/1/ 2011 – nay
Lãnh đạo
Thành viên
Thành viên
Kế nhiệm
Nguyễn Phú Trọng
Trần Đại Quang
Tiền nhiệm
Chủ tịch, 23/10/2018– nay
Cộng hòa xã hội chủ nghĩa việt nam
Chủ tịch, 2/4/2016– 21/9/2018

Hình 6