Journal of educational equipment: Applied research, Volume 1, Issue 308 (March 2024)
ISSN 1859 - 0810

Trích chọn sự kiện trong văn bản tiếng Việt

Đào Ngọc Tú*

**Khoa Công nghệ thông tin, Trường Đại học Hải Phòng*
Received: 15/2/2024; Accepted: 23/2/2024; Published: 26/2/2024

Abstract: *Event extraction plays the role of extracting meaningful information from large data sets and is of great interest and research investment in the scientific community. Information Extraction (IE), especially Event Extraction (EE), is a subfield of data mining (Data Mining - DM). In recent years, event extraction has attracted much attention from scientists. It is a good step for exploiting knowledge on text*

Keywords*: Information Extraction, Data Mining, Event Extraction*

1.Đặt vấn đề

Trích chọn sự kiện và trích chọn thông tin có điểm gì chung? Có thể nói rằng trích chọn sự kiện là một lĩnh vực con của trích chọn thông tin. Nếu như trích chọn thông tin chỉ quan tâm các dữ liệu rời rạc (tên người, địa điểm, các con số,…) thì trích chọn sự kiện quan tâm nhiều hơn tới tính cấu trúc và mức độ liên quan của thông tin trong một sự kiện. Từ đó, người đọc có thể dễ ràng suy luận ra các thông tin có ý nghĩa. Ví dụ, *"ngay sáng ngày 30/4, trên đường Xuân Thuỷ, thủ đô Hà Nội đã xảy ra vụ tai nạn nghiêm trong làm 2 người trên xe máy bị thương nặng. Nguyên nhân bước đầu được cho là do tài xế tắc-xi đã tăng tốc khi nhận điểm nên đã xô thẳng vào xe máy đi cùng chiều."* Trong ví dụ này, trích chọn thông tin đưa ra các kết quả rời rạc như: 30/4, Hà Nội, 2 hoặc tắc xi; trong khi đó trích chọn sự kiện thì quan tâm tới một bộ các thuộc tính biểu diễn cho sự kiện gồm {30/4, Hà Nội, 2 người bị thương, tắc-xi}. Rõ ràng, với tập dữ liệu trên, thông tin là hữu ích và đầy đủ hơn các thông tin rời rạc.

2.Nội dung nghiên cứu

2.1. Các nghiên cứu liên quan

Năm 1987, Message Understanding Conferences (MUC)[3] được tổ chức với sự hỗ trợ của Quỹ nghiên cứu Bộ quốc phòng Hoa Kỳ[4] và lần đầu tiên khái niệm event (sự kiện) được đề cập. Các chủ đề trong dữ liệu thường là tội phạm, khủng bố, đánh bom… một trong những đóng góp lớn của MUC là đưa ra việc trích chọn thông tin dựa trên mẫu (scenariotemplate). Các mẫu được ban tổ chức quy định và các đội tham gia cần điền thông tin vào các mẫu này một các tự động. Cuối cùng, các sự kiện được trích chọn gồm các thông tin: tổ chức, đối tượng tham gia (người, sự vật, sự việc), thời gian, địa điểm, số lượng…Độ chính xác (precision) và hồi tưởng (recall) của các nghiên cứu tham dự MUC nằm trong khoảng 50% đến 60%[2].

Chương trình Phát hiện và theo dõi chủ đề (Topic Detection and Tracking, TDT)[8] được tổ chức từ năm 1997 thu hút nhiều nhóm nghiên cứu từ các trường đại học tham gia. Chương trình này được phối hợp bởi Viện Công nghệ và Chuẩn hoá quốc gia Hoa Kỳ (NIST) và DAPRA nhằm giải quyết bài toán phát hiện, theo dõi và xâu chuỗi sự kiện.

Chương trình Trích chọn nội dung tự động (Automatic Content Extraction, ACE) của đại học Pennsylvania cũng thu hút được nhiều quan tâm từ các cộng đồng nghiên cứu và trích chọn thông tin cũng như trích chọn sự kiện. Chương trình này tập trung vào các ngôn ngữ như tiếng Anh, Trung Quốc và Ả rập. Các thông tin được trích chọn gồm các thực thể, quan hệ giữa các thực thể, và các sự kiện chúng tham gia vào.

Như vậy, có thể thấy rằng trích chọn thông tin nói chung và trích chọn sự kiện nói riêng là một vấn đề quan trọng và thời đại, nhận được rất nhiều quan tâm từ cộng đồng khoa học.

2.2. Trích chọn sự kiện trong văn bản tiếng Việt

**CÁC ĐẶC TÍNH CỦA SỰ KIỆN VỤ TAI NẠN*

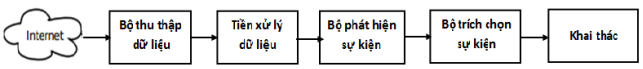
Quá trình khảo sát trên miền dữ liệu là thông tin vụ tai nạn chỉ ra rằng trong quá trình phát hiện sự kiện vụ tai nạn cần phải phân biệt rõ đâu là thông tin vụ tai nạn giao thông, đâu là thông tin tai nạn giao thông. Thông tin vụ tai nạn giao thông là cái mà luận văn quan tâm trong bài toán trích chọn sự kiện vụ tai nạn, ví dụ như *"sáng ngày 25/5 một vụ tai nạn thảm khốc đã xảy ra trên quốc lộ 1A"*; còn thông tin tai nạn giao thông như tiêu đề bài báo *"làm thế nào để giảm thiểu số vụ tai nạn giao thông"*, hay *"sốc về con số thiệt mạng do tai nạn trong nửa đầu năm 2022"* thì đây không phải thông tin vụ tai nạn giao thông mà chỉ là thông tin tai nạn giao thông.

**PHÁT BIỂU BÀI TOÁN*

Đầu vào: một bản tin trên báo điện tử

Đầu ra: bản tin ở đầu vào có phải sự kiện vụ tai nạn giao thông không, nếu có thì trích chọn ra thông tin về vụ tai nạn giao thông.

Mô hình phát hiện và trích chọn sự kiện vụ tai nạn



**GIẢI QUYẾT BÀI TOÁN PHÁT HIỆN SỰ KIỆN VÀ BÀI TOÁN TRÍCH CHỌN SỰ KIỆN VỤ TAI NẠN*

+Bài toán 1- Pháp hiện sự kiện vụ tai nạn (pha 1)

-Phát biểu bài toán

Đầu vào: một bản tin trên các trang báo có dạng thô.

Đầu ra: bản tin đó có chứa sự kiện tai nạn hay không?

+Xây dựng tập luật

Mẫu 1 =*" phương tiện giao thông"; M*ẫu 2= *"động từ" # "danh từ"*

Trong đó:

Động từ gồm các từ: Tai nạn, TNGT,…; Danh từ gồm các từ: giao thông, thương tâm, …

Ví dụ minh hoạ cho mẫu 2:

"tai nạn" # "thương tâm" ; "tai nạn" # "giao thông"

+Xây dựng mô hình phân lớp

Bộ phân lớp có nhiệm vụ phát hiện một bài báo có chứa sự kiện hay không. Bộ phân lớp sẽ phân ra thành hai lớp: lớp có chứa sự kiện vụ tai nạn nhãn là EVENT và lớp không chứa sự kiện vụ tai nạn nhãn là NOT_EVENT. Quá trình khảo sát cho thấy rằng phần tiêu đề và tóm tắt của bản tin đã chứa đầy đủ nội dung chính của cả bản tin. Nên, tác giả dùng thông tin này để xây dựng véctơ đặc trưng biểu diễn văn bản. Các đặc trưng được sử dụng trong quá trình huấn luyện là 2-grams, 3-grams, 4-grams. Tác giả xây dựng một tập huấn luyện và dùng tập dữ liệu huấn luyện này để xác định văn bản chứa sự kiện.

+Bài toán 2- Trích chọn sự kiện vụ tai nạn (pha 2)

Đầu vào: bản tin chứa sự kiện vụ tai nạn

Đầu ra: các thông tin của một vụ tai nạn gồm: thời gian, địa điểm, số thương vong, phương tiện gây tai nạn. Ở đây số thương vong bao gồm số nạn nhân tử vong và số nạn nhân bị thương. Số thương vong được liệt kê dưới dạng danh sách gồm hai trường (số tử vong, số bị thương), và một bản ghi tương ứng ghi ra số tử vong và số bị thương

+Trích chọn thời gian: Thời gian = <Tiền tố> + <Ngày tháng> Trong đó, Tiền tố gồm các từ: vào, ngày, sáng, trưa, chiều, tối, nửa đêm, trưa nay, sáng nay, chiều nay, vào giờ tan tầm, hôm qua, hôm nay, tối qua, đêm qua, rạng sáng nay, tháng.

+Trích chọn địa điểm: Trong trích chọn địa điểm, sử dụng NER và từ điển địa điểm.

Bước 1: áp dụng NER

Bước 2: lấy về các thực thể được gán thẻ <loc>

Bước 3: kiểm tra ngược lại với từ điển địa điểm để tìm các location chính xác

+Trích chọn số thương vong

Số nạn nhân = <số> + <hậu tố>; Số: chính là số nạn nhân. Có thể là số hoặc chữ; số={"một", "hai", "ba", "bốn", "năm", "sáu", "bảy", "tám", "chín", "mười"}; và các số [1..9]

Hậu tố: là các từ tử vong, bị thương, thiệt mạng, chết, nhập viện....; hậu tố={"bị thương", "chết", "tử vong", "thiệt mạng", "chết thảm", "thương nặng", "thương nhẹ", "cấp cứu","bệnh viện"};

Trong phần này, tác giả đã đề xuất PP và mô hình giải quyết bài toán tổng quan trích chọn sự kiện vụ tai nạn. Đồng thời tác giả cũng trình bày chi tiết PP và mô hình giải quyết hai bài toán: bài toán phát hiện sự kiện vụ tai nạn và bài toán trích chọn sự kiện vụ tai nạn; bài toán thứ nhất tác giả đã dùng PP kết hợp luật và học máy để phát hiện sự kiện vụ tai nạn giao thông và dữ liệu này được làm đầu vào cho bài toán thứ hai. Ở bài toán thứ hai, các thông tin được trích chọn là: thời gian, địa điểm, số thương vong, và phương tiện gây tai nạn. Trong cả hai bài toán đều dùng kết hợp giữa luật và học máy.

2.3. Thực nghiệm và đánh giá

****Xây dựng tập dữ liệu***

- Thu thập dữ liệu

Dữ liệu được thu thập trên trang http://vovgiaothong.vn/giao-thong-trong-nuoc/ (kênh VOV Giao thông Quốc gia – Đài Tiếng nói VN) và trang http://antoangiaothong.gov.vn/tai-nan-giao-thong/ (của Uỷ ban An toàn giao thông Quốc gia).

- Tiền xử lý dữ liệu

***Đánh giá quá trình phát hiện sự kiện**

+Đánh giá bộ lọc dữ liệu

Mô tả thực nghiệm: mục đích của thực nghiệm này đánh giá khả năng của bộ lọc dữ liệu.

Phát biểu thực nghiệm

- Đầu vào: một tập các bản tin được thu thập từ trang http://vovgiaothong.vn/giao-thong-trong-nuoc/ và trang http://antoangiaothong.gov.vn/tai-nan-giao-thong/

- Đầu ra: các bài báo liên quan tới miền dữ liệu tai nạn giao thông

Dữ liệu thực nghiệm: là 3.000 bản tin

Sau quá trình lọc dữ liệu thu được tổng sô 919 bản tin thuộc miền tai nạn giao thông, trong đó số bản tin không liên quan đến tai nạn giao thông rất ít, và có thể tính tỷ lệ lỗi. Chi tiết được trình này trong bảng 4.4.

Bảng 4.4. Tỷ lệ lỗi của quá trình lọc dữ liệu

Tổng số bản tin	Số bản tin không liên quan	Tỷ lệ lỗi
919	19	3.9%

Công thức tính tỷ lệ lỗi của quá trình lọc dữ liệu:

$$\text{Tỷ lệ lỗi} = \frac{\text{Số bài không chính xác}}{\text{Tổng số}} \quad (4.1)$$

Trong đó: Tổng số: là tổng số bản tin thu được sau quá trình lọc; Số bài không chính xác: là số bản tin không thuộc miền tai nạn giao thông.

Kết quả của quá trình này, được trình bày trong bảng 4.4, thu được kết quả độ chính xác khá cao.

+Đánh giá quá trình phân lớp: Dữ liệu thực nghiệm: dữ liệu của mỗi lần đánh giá là 100 bản tin được lấy ngẫu nhiên từ các bản tin được lọc bởi bộ dữ liệu. Kết quả của các quá trình đánh giá được trình bày trong bảng 4.5.

Bảng 4.5. Đánh giá kết quả phân lớp

TT	Số bản tin chính xác	Số bản tin Không chính xác	Số bản tin không tìm thấy	Preci-sion	Recall	độ đo F-1
1	85	12	3	88%	97%	92%
2	81	16	3	84%	96%	90%
3	83	15	2	85%	98%	91%
4	85	11	4	89%	96%	92%
5	80	17	3	82%	96%	89%
Trung bình	82.8	14.2	3	85%	97%	91%

Kết quả thực nghiệm trong bảng 4.5, cho thấy quá trình phân lớp cho thấy độ chính xác (P-Precision) đạt 85%, độ đo hồi tưởng (R-Recall) đạt 97%, độ đo F-1 đạt 91%.

*Đánh giá quá trình trích chọn sự kiện

+Thực nghiệm không qua bộ phân lớp

Mô tả thực nghiệm: mục đích của phần này là đánh giá khả năng trích chọn.

Phát biểu thực nghiệm

Đầu vào: một bản tin trong miền tai nạn giao thông

Đầu ra: thông tin về sự kiện vụ tai nạn gồm: thời gian xảy ra vụ tai nạn, địa điểm xảy ra vụ tai nạn, số thương vong (số tử vong, số bị thương), và phương tiện gây tai nạn.

Dữ liệu thực nghiệm: dữ liệu là 200 bản tin lấy ngẫu nhiên từ các bản tin.

Tên website	Số sự kiện đúng	Số sự kiện sai	Số sự kiện không tìm thấy	P	R	F1
antoangiaothong.gov.vn	160	34	6	82%	96%	89%
vovgiaothong.vn	154	37	9	81%	94%	87%
Trung bình	314	71	15	82%	95%	88%

+Thực nghiệm qua bộ phân lớp

Dữ liệu thực nghiệm: dữ liệu là 100 bản tin được lấy từ các bản tin chứa sự kiện vụ tai nạn (gán nhãn EVENT). Kết quả được mô tả chi tiết trong bảng 4.7.

Bảng 4.7. Đánh giá quá trình trích chọn - dữ liệu qua bộ phân lớp.

Tên website	Số sự kiện đúng	Số sự kiện sai	Số sự kiện không tìm thấy	P	R	F1
antoangiaothong.gov.vn	91	5	4	95%	96%	95%
vovgiaothong.vn	93	4	2	96%	98%	97%
Trung bình	184	9	6	95%	97%	96%

+ Nhận xét

Từ thực nghiệm được chi tiết trong bảng 4.6 (dữ liệu không qua bộ phân lớp) và bảng 4.7 (dữ liệu được xử lý qua bộ phân lớp). Kết quả cho thấy dữ liệu được xử lý qua bộ phân lớp cho kết quả cao hơn. Điều đó chứng tỏ tầm quan trọng của bộ phân lớp trong mô hình.

3.Kết luận

Trong bài báo này, tác giả đã tìm hiểu các PP trích chọn sự kiện, PP kết hợp luật và học máy được sử dụng cho bài toán phát hiện sự kiện và bài toán trích chọn sự kiện. Trên cơ sở đó, xây dựng mô hình và PP giải quyết chi tiết cho bài toán phát hiện sự kiện vụ tan nạn và bài toán trích chọn sự kiện vụ tai nạn. Kết quả thực nghiệm của quá trình trích chọn sự kiện trên miền dữ liệu vụ tai nạn với độ đo P đạt 95%, độ đo R đạt 97 %, và độ đo F1 đạt 96%, điều đó chứng tỏ tính khả thi của mô hình.

Tài liệu tham khảo

[1] Uzay Kaymak Frederik Hogenboom, Flavius Frasincar and Franciska de Jong (2011). *An overview of event extraction from text. Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference*

[2] M.A Hearst (1992). Automatic acquisition of hyponyms from large text corpora. *In: 14th Conference on Computational Linguistics*

[3] M.A Hearst (1998). Wordnet: An electronic lexical database and some of its applications. *In Automated Discovery of WordNet Relations*