HỌC VIỆN NGÂN HÀNG

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

**KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC**

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRÍCH CHỌN THÔNG TIN SỰ KIỆN Y TẾ TRÊN CÁC TRANG WEB TIẾNG VIỆT**

**TẠ QUỐC ANH**

**HÀ NỘI, NĂM 2019**

HỌC VIỆN NGÂN HÀNG

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

**KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC**

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRÍCH CHỌN THÔNG TIN SỰ KIỆN Y TẾ TRÊN CÁC TRANG WEB TIẾNG VIỆT**

**Giáo viên hướng dẫn: TS.Chu Thị Hồng Hải**

**Sinh viên thực hiện: Tạ Quốc Anh**

**Mã sinh viên: 18A4040015**

**Lớp: K18HTTTB**

**Khóa: K18**

**Hệ: Đại học chính quy**

***Hà Nội, tháng 5 / 2019***

**LỜI CẢM ƠN**

*Được lựa chọn là một trong các sinh viên của Khoa Hệ thống thông tin quản lý thực hiện viết khóa luận tốt nghiệp, cá nhân em cảm thấy thật vinh dự và tự hào. Việc thực hiện khóa luận tốt nghiệp đã tạo điều kiện cho em có thêm cơ hội tổng hợp, hệ thống hóa toàn bộ các kiến thức cùng những kỹ năng đã được trang bị trong suốt thời gian học tập tại Khoa Hệ thống thông tin quản lý, Học viện Ngân hàng. Cũng thông qua việc nghiên cứu, thực hiện khóa luận tốt nghiệp đã giúp em có cơ hội nâng cao khả năng tư duy, nghiên cứu, khả năng lập kế hoạch và giải quyết vấn đề,…Trong thời gian nghiên cứu, thực hiện hoàn thành khóa luận, em đã nhận được rất nhiều sự quan tâm, tạo điều kiện giúp đỡ của các thầy, cô giáo của Khoa Hệ thống thông tin quản lý, Học viện Ngân hàng. Đặc biệt là sự quan tâm sát sao, tạo điều kiện hướng dẫn tận tâm của Tiến Sĩ, giáo viên Chu Thị Hồng Hải để em hoàn thành khóa luận tốt nghiệp của mình.*

*Tuy nhiên vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên một số nội dung nghiên cứu và trình bày tại khóa luận không tránh khỏi những thiếu xót, chưa thực sự phù hợp với thực tế. Theo đó, em rất mong tiếp tục nhận sự góp ý, chỉ bảo thêm của các thầy cô để khóa luận của em được hoàn thiện hơn.*

*Qua đây, em xin trân trọng cảm ơn các thầy cô giáo đang giảng dạy tại Khoa Hệ thống thông tin quản lý, Học Học viện Ngân hàng đặc biệt là giáo viên hướng dẫn, Tiến sỹ Chu Thị Hồng Hải đã tận tình giúp đỡ, tạo điều kiện cho em trong suốt quá trình thực hiện và hoàn thành bài khóa luận tốt nghiệp này.*

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng 05 năm 2019*  *Sinh viên*  Tạ Quốc Anh |

**LỜI CAM KẾT**

Em xin cam đoan: Khóa luận tốt nghiệp với đề tài “Ứng dụng khai phá dữ liệu trích chọn sự kiện y tế từ các trang Web Tiếng Việt” là bài khóa luận được nghiên cứu thực hiện bởi cá nhân em và được sự hướng dẫn của giáo viên, Tiến Sĩ Chu Thị Hồng Hải. Các nội dung, kết quả trong bài khóa luận này là trung thực và chưa được công bố dưới bất kỳ hình thức nào không sao chép của bất cứ ai.

Em đã trích dẫn đầy đủ các loại tham khảo, công trình nghiên cứu liên quan ở trong nước và quốc tế. Tất cả những tham khảo đều được nêu nguồn gốc rõ ràng. Em xin chịu mọi trách nhiệm về nội dung trong bài khóa luận của mình.

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng 05 năm 2019*  *Sinh viên*  *Tạ Quốc Anh* |

**NHẬN XÉT**

**(Của giáo viên hướng dẫn)**

Về các mặt: Mục đích của đề tài; Tính thời sự và ứng dụng của đề tài; Bố cục và hình thức trình bầy đề tài; Kết quả thực hiện đề tài; Ý thức, thái độ của sinh viên trong quá trình thực hiện đề tài.

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

…………………………………………………………………………………………..

Kết luận : …………...…………………………………………………………………..

Hà Nội, ngày tháng năm 2019

Giáo viên hướng dẫn

*(Ký tên)*

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc9784416)

[CHƯƠNG 1. CƠ SỞ LÝ THUYẾT 4](#_Toc9784418)

[1.1. Tổng quan về khai phá dữ liệu 4](#_Toc9784419)

[1.1.1. Khái niệm khai phá dữ liệu 4](#_Toc9784420)

[1.1.2. Các kỹ thuật áp dụng trong khai phá dữ liệu 5](#_Toc9784422)

[1.1.3.Các nguồn dữ liệu có thế khai phá 7](#_Toc9784424)

[1.1.4. Ứng dụng của khai phá dữ liệu 7](#_Toc9784426)

[*1.1.4.1.Trí tuệ doanh nghiệp* 8](#_Toc9784427)

[*1.1.4.2.Công cụ tìm kiếm* 8](#_Toc9784428)

[1.2. Khai phá Web 10](#_Toc9784429)

[1.2.1. Khái niệm khai phá dữ liệu web 10](#_Toc9784430)

[1.2.2. Lợi ích của khai phá Web 12](#_Toc9784432)

[1.2.3. Các kiểu dữ liệu Web 13](#_Toc9784433)

[*1.2.4. Xử lý dữ liệu văn bản ứng dụng trong khai phá dữ liệu Web* 14](#_Toc9784435)

[1.2.4.1. Dữ liệu văn bản 14](#_Toc9784436)

[1.2.4.2. Một số vấn đề trong xử lý dữ liệu văn bản 14](#_Toc9784437)

[1.3. Tổng quan về sự kiện 15](#_Toc9784438)

[1.3.1. Định nghĩa sự kiện 16](#_Toc9784439)

[1.3.2. Trích chọn sự kiện 17](#_Toc9784440)

[1.3.3.Trích chọn sự kiện y tế từ các trang web 18](#_Toc9784441)

[*1.3.3.1. Bài toán trích chọn thông tin văn bản* 18](#_Toc9784442)

[*1.3.3.2.Phát hiện sự kiện* 19](#_Toc9784443)

[*1.3.3.3.Trích chọn sự kiện* 19](#_Toc9784444)

[1.4.Ý nghĩa của bài toán trích chọn sự kiện 20](#_Toc9784445)

[1.4.1. Ý nghĩa khoa học 20](#_Toc9784446)

[1.4.2. Ý nghĩa thực tiễn 20](#_Toc9784447)

[1.5. Kết luận chương 21](#_Toc9784448)

[CHƯƠNG 2: MỘT SỐ PHƯƠNG PHÁP TIẾP CẬN TRONG KHAI PHÁ DỮ LIỆU VÀ TRÍCH CHỌN THÔNG TIN 22](#_Toc9784449)

[2.1. Phương pháp tiếp cận dựa trên tập luật (Rule – Based) 22](#_Toc9784450)

[2.1.1. Luật cú pháp 22](#_Toc9784451)

[2.1.2. Luật ngữ nghĩa 23](#_Toc9784452)

[2.1.3. Hình dạng và biểu diễn của tập luật 24](#_Toc9784453)

[2.1.4. Các thuộc tính của các thẻ 24](#_Toc9784454)

[2.1.5. Các định luật đánh dấu ranh giới thực thể 25](#_Toc9784455)

[2.1.6. Các luật cho đa thực thể 26](#_Toc9784456)

[2.1.7. Chọn lựa hình dạng của các tập luật 26](#_Toc9784457)

[2.2 Phương pháp tiếp cận dựa trên học máy 28](#_Toc9784458)

[2.3 Phương pháp tiếp cận kết hợp luật và học máy 29](#_Toc9784459)

[2.4. Phương pháp Khai phá nội dung Web 30](#_Toc9784460)

[2.4.1. Khai phá kết quả tìm kiếm 30](#_Toc9784461)

[2.4.2. Khai phá văn bản Web 31](#_Toc9784462)

[*2.4.2.1. Nguồn dữ liệu* 31](#_Toc9784464)

[*2.4.2.2. Tiền xử lý dữ liệu* 32](#_Toc9784465)

[*2.4.2.3. Biểu diễn văn bản* 32](#_Toc9784466)

[*2.4.2.4. Trích rút ra các từ đặc trưng* 33](#_Toc9784467)

[*2.4.2.5. Khai phá dữ liệu văn bản* 33](#_Toc9784468)

[2.5. Phương pháp Khai phá cấu trúc Web 36](#_Toc9784469)

[2.5.1. Tiêu chuẩn đánh giá độ tương tự 37](#_Toc9784470)

[2.5.2. Khai phá và quản lý cộng đồng web 37](#_Toc9784471)

[2.6. Kết luận chương 37](#_Toc9784472)

[CHƯƠNG 3: ĐỀ XUẤT MÔ HÌNH TRÍCH CHỌN SỰ KIỆN Y TẾ 38](#_Toc9784473)

[3.1 Phát biểu bài toán 38](#_Toc9784474)

[3.2 Giải bài toán trích chọn sự kiện y tế 40](#_Toc9784475)

[3.2.1 Phương pháp đề xuất 40](#_Toc9784476)

[3.2.2.Thực hiện giải quyết bài toán phát hiện sự kiện và bài toán trích chọn sự kiện y tế 43](#_Toc9784478)

[*3.2.2.1.Bài toán thứ nhất (Pha 1) – Phát hiện sự kiện y tế* 43](#_Toc9784479)

[*3.2.2.2.Bài toán thứ hai (Pha 2) – Trích chọn sự kiện y tế* 50](#_Toc9784485)

[3.3. Kết luận chương 55](#_Toc9784490)

[CHƯƠNG 4: THỰC NGHIỆM 56](#_Toc9784491)

[4.1 Môi trường và công cụ để tiến hành thực nghiệm 56](#_Toc9784492)

[4.2 Bộ thu thập dữ liệu 57](#_Toc9784495)

[4.3.Tiền xử lý dữ liệu 59](#_Toc9784499)

[4.4.Bộ phát hiện sự kiện 61](#_Toc9784503)

[4.4.1.Mô tả thực nghiệm 61](#_Toc9784504)

[4.4.2.Đánh giá bộ lọc dữ liệu 62](#_Toc9784505)

[4.4.3 Đánh giá quá trình phân lớp 62](#_Toc9784507)

[4.5.Bộ trích chọn sự kiện 63](#_Toc9784509)

[4.5.1 Mô tả thực nghiệm 63](#_Toc9784510)

[4.5.2.Đánh giá quá trình trích chọn sự kiện 65](#_Toc9784511)

[4.6 Phân tích lỗi 66](#_Toc9784514)

[4.6.1 Phân tích lỗi quá trình phát hiện sự kiện 66](#_Toc9784515)

[4.6.2 Phân tích lỗi trong quá trình trích chọn sự kiện 67](#_Toc9784516)

[4.7. Một số kết quả phân tích các sự kiện 71](#_Toc9784518)

[4.7.1.Thống kê số ca mắc bệnh 71](#_Toc9784519)

[4.7.2.Thống kê số ca mắc bệnh 72](#_Toc9784521)

[4.7.3.Thống các dịch bệnh tại Hà Nội và Tp.Hồ Chí Minh 72](#_Toc9784523)

[4.8. Kết luận chương 69](#_Toc9784526)

[KẾT LUẬN 70](#_Toc9784527)

[DANH MỤC TÀI LIỆU THAM KHẢO 74](#_Toc9784528)

[PHỤ LỤC](#_Toc9784529) 75

**DANH MỤC CÁC CHỮ VIẾT TẮT**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Viết tắt** | **Cụm từ tiếng Anh** | **Cụm từ tiếng Việt** |
| 1 | CNTT | Information Technology | Công nghệ thông tin |
| 2 | CSDL | Database | Cơ sở dữ liệu |
| 3 | KDD | Knowledge Discovery in Database | Khám phá tri thức trong cơ sở dữ liệu |
| 4 | KPDL | Data mining | Khai phá dữ liệu |
| 5 | KPVB | Text Mining | Khai phá văn bản |
| 6 | PCDL | Data Clustering | Phân cụm dữ liệu |
| 7 | NER | Named Entity Recognition | Bộ nhận dạng thực thể |
| 8 | URL | Uniform Resource Locator | Định vị tài nguyên |

**DANH MỤC BẢNG BIỂU HÌNH VẼ**

[Hình 1.1. Quá trình khám phá tri thức 4](#_Toc9799977)

[Hình 1.2. Các lĩnh vực liên quan đến khám phá tri thức trong CSDL 5](#_Toc9799979)

[Hình 1.3: Các nguồn dữ liệu có thể khai phá 7](#_Toc9799981)

[**Hình 1.4. Các nội dung trong khai phá web 11**](#_Toc9799987)

[Hình 1.5. Phân loại dữ liệu web 13](#_Toc9799990)

[Hình 2.1: Quá trình khai phá văn bản Web 31](#_Toc9800019)

[Hình 3.1: Quá trình phát hiện và trích chọn sự kiện y tế 42](#_Toc9800033)

[Hình 3.2: Thành phần phát hiện sự kiện 43](#_Toc9800036)

[Hình 3.3: Tiêu đề bản tin có chứa tên các loại bệnh 45](#_Toc9800038)

[Hình 4.1 Quá trình thu thập danh sách địa chỉ URL các bài viết 57](#_Toc9800052)

[Hình 4.2 Kết quả quá trình thu thập URL bằng công cụ Octoparse 58](#_Toc9800053)

[Hình 4.3 Kết quả quá trình thu thập list URL 58](#_Toc9800054)

[Hình 4.4 Quá trình Import list URL của các trang web vào Octoparse 59](#_Toc9800057)

[Hình 4.5 Quá trình lựa chọn các thành phần trên trang web để thu thập 59](#_Toc9800057)

[Hình 4.6 Kết quả bộ dữ liệu sau khi đã được tiền xử lý 60](#_Toc9800058)

[Hình 4.7 Kết quả bộ dữ liệu sau quá trình Phát hiện sự kiện 61](#_Toc9800061)

[Hình 4.8 Dữ liệu đầu vào của Bộ trích chọn sự kiện 64](#_Toc9800068)

[Hình 4.9 Kết quả của Bộ trích chọn sự kiện 64](#_Toc9800069)

[Biểu đồ 1: Thống kê số ca mắc bệnh ở 4 tỉnh/thành phố lớn 71](#_Toc9800079)

[Biểu đồ 2: Thống kê số ca mắc bệnh theo các quý 72](#_Toc9800081)

[Biểu đồ 3: Thống kê các bệnh thường gặp tại Hà Nội 72](#_Toc9800083)

[Bảng 3.1 Bảng danh sách các loại bệnh truyền nhiễm 47](#_Toc9800039)

[Bảng 4.1: Cấu hình phần cứng 56](#_Toc9800049)

[Bảng 4.2: Công cụ phần mềm sử dụng 56](#_Toc9800050)

[Bảng 4.3: Các thành phần của một bảng tin 59](#_Toc9800056)

[Bảng 4.4: Tỷ lệ lỗi của quá trình lọc dữ liệu 62](#_Toc9800063)

[Bảng 4.5: Đánh giá kêt quả của quá trình phân lớp 63](#_Toc9800065)

[Bảng 4.6: Đánh giá quá trình trích chọn dữ liệu không qua bộ phân lớp 66](#_Toc9800071)

[Bảng 4.7: Đánh giá quá trình trích chọn dữ liệu thông qua bộ phân lớp 66](#_Toc9800072)

[Bảng 4.8: Một số lỗi trong quá trình trích chọn 68](#_Toc9800076)

**LỜI MỞ ĐẦU**

Chăm sóc sức khỏe là một trong những nhu cầu thiết yếu của cuộc sống đối với mỗi người. Chính vì vậy, việc tìm kiếm các thông tin liên quan tới lĩnh vực y tế lại càng trở nên cấp thiết và quan trọng hơn nhất là trong thời đại công nghệ thông tin đang phát triển một cách mạnh mẽ như hiện nay. Vấn đề này càng cần phải được quan tâm tới một cách sâu rộng khi chúng ta đang phải đối mặt với nhiều các loại dịch bệnh dịch nguy hiểm và có diễn biến khó lường. Một vài ví dụ điền hình có thể kể tới như bệnh Sốt xuất huyết, bệnh Sởi, Cúm… hiện đang bùng phát mạnh mẽ và có chiều hướng gia tăng trong thời gian gần đây. Những con số “biết nói” được thống kê trên các trang báo điện tử về lĩnh vực Y tế hằng năm đã phần nào phản ánh được hậu quả nghiêm trọng mà các dịch bệnh gây ra đối với sức khỏe. Bên cạnh việc phát triển mạnh mẽ của các tài nguyên trực tuyến thì việc tận dụng và khai thác một cách hợp lý nguồn tài nguyên này sẽ đem lại những hiệu quả tích cực trong công tác tuyên truyền và nâng cao sức khỏe cộng đồng.

Sự gia tăng nhanh chóng về số lượng các nguồn tài nguyên trực tuyến bên cạnh những mặt tích cực thì đi kèm với đó là một số những mặt còn hạn chế như việc xuất hiện những thông tin thừa, hay việc tổ chức thông tin tự do một cách không cấu trúc đã dẫn tới việc người sử dụng gặp rất nhiều khó khăn trong tìm kiếm theo dõi cũng như nắm bắt được các thông tin một cách nhanh và chính xác nhất. Hơn nữa công nghệ tìm kiếm thông tin theo hình thức truyền thống còn nhiều hạn chế, dễ thấy nhất là kết quả trả về không nhanh bằng hình thức trực tuyến, ít sự phong phú, thậm chí trong một số trường hợp có thể không đúng với những gì người sử dụng mong muốn, thêm vào đó là tính phức tạp trong việc triển khai diễn đạt các ngôn ngữ tự nhiên. Chính vì thế việc khai thác một cách tối ưu nguồn tài nguyên phong phú này trở thành một vấn đề vô cùng quan trọng, thu hút nhiều sự quan tâm của các nhà nghiên cứu.

Việc trích chọn sự kiện về vấn đề y tế để nhằm mục đích tập hợp, rút ngắn lượng dữ liệu lớn thành những dữ liệu ngắn gọn hơn nhưng vẫn đảm bảo tính chính xác, đầy đủ và súc tích từ đó giúp ích cho việc tìm kiếm, tổng hợp, phân tích xử lý dữ liệu trên các trang báo điện tử được tiến hành một cách hiệu quả hơn. Những thông tin thống kê này là nguồn dữ liệu vô cùng quý giá và hiệu quả để cung cấp cho các nhà quản lý và người dân nắm được tình hình diễn biến của dịch bệnh để từ đó có thể chủ động phòng trách và ngăn ngừa, giảm tối đa thiệt hại của bệnh dịch đối với sức khỏe cộng đồng. Đó cũng là lí do em chọn đề tài: “**Ứng dụng khai phá dữ liệu để trích chọn sự kiện y tế từ các trang Web Tiếng Việt**”.

**BỐ CỤC CÁC CHƯƠNG**

Chi tiết bài khóa luận được chia thành 4 chương như sau:

**Chương 1: Cơ sở lý thuyết**

Trong chương này cơ bản trình bày những khái niệm và các vấn đề liên quan của khai phá dữ liệu và khai phá dữ liệu web. Bên cạnh đó chương 1 còn trình bày các khái niệm cơ bản liên quan tới bài toán trích chọn sự kiện trong bối cảnh bùng nổ của công nghệ thông tin và mạng Internet. Cuối cùng đưa ra ý nghĩa khoa học và ý nghĩa thực tiễn của công tác trích chọn những sự kiện nói chung và sự kiện liên quan tới lĩnh vực y tế trên các trang web tiếng Việt nói riêng.

**Chương 2: Một số phương pháp tiếp cận trong khai phá dữ liệu và trích chọn thông tin sự kiện**

Chương 2 tập trung trình bày một số các phương pháp sử dụng để tiếp cận nhằm giải quyết được bài toán trích chọn sự kiện đó là phương pháp tiếp cận dựa trên luật, phương pháp tiếp cận dựa trên học máy và phương pháp kết hợp giữa luật và học máy. Ngoài ra trình bày thêm một số các phương án tiếp cận bổ sung như sử dụng mô hình để rút trích văn bản, phương pháp khai phá dữ liệu dựa trên nội dung web và cấu trúc web. Đây là các nền tảng kĩ thuật về công nghệ để thực hiện việc khai phá dữ liệu ở chương 3.

**Chương 3: Đề xuất mô hình trích chọn sự kiện y tế**

Chương này sẽ tiến hành mô tả bài toán cần giải quyết. Bên cạnh đó là tiến hành xây dựng các mô hình, tập luật để có thể giải quyết bài toán trên cơ sở lý thuyết

**Chương 4: Thực nghiệm và đánh giá**

Chương này khóa luận mô tả quá trình tiến hành thực nghiệm và đánh giá kết quả dựa trên các kết quả phân tích được ở chương 3. Ba bộ đo được sử dụng trong pha phát hiện sự kiện là độ đo chính xác, độ hồi tưởng và độ đo F1 sau đó tiến hành so sánh với kết quả thủ công cho pha trích chọn sự kiện. Thống kê và đánh giá biểu đồ các thuộc tính đã chọn.

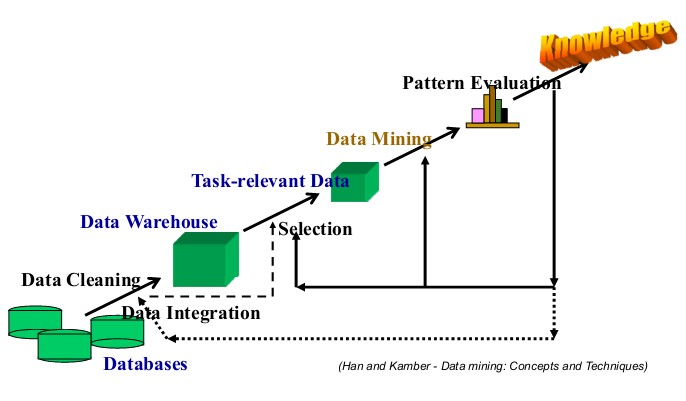
# CHƯƠNG 1. CƠ SỞ LÝ THUYẾT

# 1.1. Tổng quan về khai phá dữ liệu

## **1.1.1. Khái niệm khai phá dữ liệu**

Trong bối cảnh hiện nay, khai phá dữ liệu (KPDL) được xem như là một lĩnh vực nghiên cứu mới đầy hứa hẹn và triển vọng nhằm khai thác các thông tin, các tri thức mới từ các CSDL lớn tiềm ẩn một cách tự động nhằm phục vụ cho các tổ chức doanh nghiệp nâng cao hiệu quả sản xuất từ đó thúc đẩy các hoạt động phát triển kinh doanh và tạo ra những lợi thế cạnh tranh lớn cho những đơn vị tổ chức này. Dựa trên những kết quả nghiên cứu khoa học cùng những ứng dụng đang được triển khai một cách thành công đã cho thấy KPDL là một lĩnh vực có rất nhiều tiềm năng, mang lại lợi ích lớn và có thể phát triển một cách mạnh mẽ trong tương lai. Hiện nay KPDL đang được áp dụng ngày một rộng rãi trong nhiều các lĩnh vực đa dạng khác như thương mại, viễn thông, tài chính ngân hàng…

Một số các kỹ thuật phổ biến được tiến hành sử dụng trong lĩnh vực KPDL nhìn chung đều được kế thừa từ một số các lĩnh vực liên quan như trí tuệ nhân tạo, cơ sở dữ liệu…Chính vì vậy có thể khái quá hóa khái niệm khai phá dữ liệu như sau: **Khai phá dữ liệu là một quá trình tìm kiếm và phát hiện những tri thức mới và hữu ích đang tiềm ẩn trong cơ sở dữ liệu lớn.** Quá trình khai phá tri thức được thể hiện trong hình dưới đây:



### *Hình 1.1. Quá trình khám phá tri thức*

Khai phá dữ liệu còn là một trong các bước thuộc quy trình khám phá tri thức nhằm trích chọn ra các thông tin hữu ích có trong khối dữ liệu lớn. KDD là mục tiêu chính của KPDL, do vậy hai khái niệm KPDL và KDD được các nhà khoa học trên hai lĩnh vực xem là tương đương với nhau. Tuy nhiên nếu phân chia một cách chi tiết thì KPDL là một bước thuộc quá trình KDD.

## **1.1.2. Các kỹ thuật áp dụng trong khai phá dữ liệu**

KDD là một lĩnh vực liên ngành, bao gồm: Tổ chức dữ liệu, học máy, trí tuệ nhân tạo và các khoa học khác. Sự kết hợp này có thể được diễn tả như sau:

Các lĩnh vực

khoa học khác

Tổ chức dữ liệu



Học máy và

trí tuệ nhân tạo

### *Hình 1.2. Các lĩnh vực liên quan đến khám phá tri thức trong CSDL*

Đứng trên quan điểm của học máy, các kỹ thuật trong KPDL, bao gồm:

*Học có giám sát*: Là quá trình gán nhãn lớp cho các phần tử trong CSDL dựa trên một tập các ví dụ huấn luyện và các thông tin về nhãn lớp đã biết.

*Học không có giám sát:* Là quá trình phân chia một tập dữ liệu thành các lớp hay cụm dữ liệu tương tự nhau mà chưa biết trước các thông tin về lớp hay tập các ví dụ huấn luyện.

*Học nửa giám sát:* Là quá trình phân chia một tập dữ liệu thành các lớp dựa trên một tập nhỏ các ví dụ huấn luyện và các thông tin về một số nhãn lớp đã biết trước.

Nếu căn cứ vào lớp các bài toán cần giải quyết, thì KPDL bao gồm các kỹ thuật áp dụng sau:

*Phân lớp và dự báo*: Dựa vào các mẫu thu thập trong lịch sử giao dịch của đơn vị để đưa ra các dự báo các trường hợp có khả năng xảy ra tiếp theo. Dự báo với các giá trị liên tục sẽ sử dụng hồi quy. Phân tích hồi quy là kỹ thuật thống kê trong phân tích dữ liệu và xây dựng các mô hình từ thực nghiệm, cho phép mô hình hồi quy vừa được khám phá phục vụ cho mục đích dự báo (prediction), hoặc học cơ chế để tạo ra mẫu.

*Luật kết hợp*: Với một tập hợp dữ liệu các giao dịch lịch sử, cần xác định các luật dự đoán có khả năng suất hiện trong những giao dịch tiếp theo. Luật kết hợp với một tập hợp các giao dịch lịch sử, mục đích của bài toán là phát hiện tất cả các luật dựa vào độ hỗ trợ - minsup và độ tin cậy - minconf. Một trong số các giải thuật quan trọng của luật kết hợp là giải thuật Apriori.

*Phân tích chuỗi theo thời gian*: Tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

*Phân cụm*: Việc gom một tập các đối tượng có cùng đặc điểm giống nhau hay tương đồng nhau vào cùng một nhóm. Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng của cụm khác.

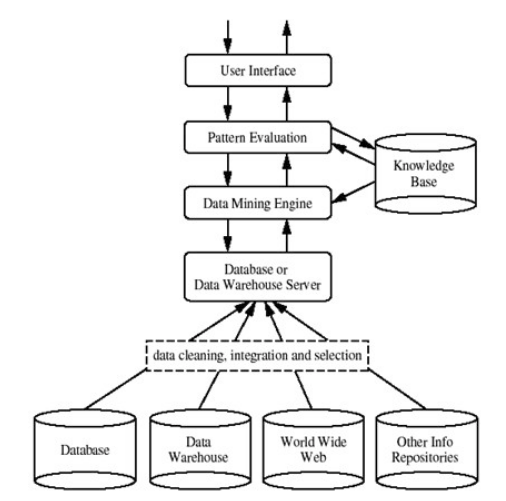
*Phân loại*: là dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu.

Quá trình gồm hai bước:

* Bước học (pha huấn luyện): xây dựng bộ phân loại (classifier) bằng việc phân tích/học tập huấn luyện.
* Bước phân loại (classification): phân loại dữ liệu trên đối tượng mới nếu độ chính xác của bộ phân loại được đánh giá là có thể chấp nhận được (acceptable).

Các giải thuật phân loại dữ liệu: cây quyết định (decision tree), mạng Bayesian, mạng neural, suy diễn dựa trên tình huống (case-based reasoning), phân loại dữ liệu dựa trên tiến hóa gen (genetic algorithms), phân loại dữ liệu với lý thuyết tập thô (rough sets), phân loại dữ liệu với lý thuyết tập mờ (fuzzy sets).

1.1.3.Các nguồn dữ liệu có thế khai phá



Hình 1.3: Các nguồn dữ liệu có thể khai phá

Việc KPDL được tiến hành trên tập dữ liệu lớn có trong các CSDL, các kho dữ liệu hoặc trên: Web, Other Info Repositories, trên các trang tính.... Các cơ sở dữ liệu lưu trữ tại các nguồn này có thể được bố trí đa dạng theo nhiều loại cấu trúc, rời rạc và đến từ nhiều nguồn khác nhau. Tất cả các dữ liệu phục vụ quá trình khai thác thông tin sẽ được làm sạch, tích hợp dữ liệu trước khi được đưa về các cơ sở dữ liệu chủ để phục vụ quá trình khai phá của người dùng.

## **1.1.4. Ứng dụng của khai phá dữ liệu**

Trong thời đại xã hội thông tin như hiện nay, có hàng ngàn thông tin được cập nhật mỗi ngày nhưng không phải thông tin nào cũng là cần thiết. Nhưng việc nắm bắt thông tin một cách kịp thời, chính xác lại tạo lợi thế cho con người đặc biệt là trong hoạt động sản xuất kinh doanh. Với lợi thế khai thác được thông tin ẩn trong kho dữ liệu khổng lồ, kĩ thuật khai phá dữ liệu được ứng dụng nhiều trong thực tế. Là một ứng dụng điều khiển có tính ổn định cao, khai phá dữ liệu đã chứng kiến những thành công lớn trong nhiều ứng dụng. Trí tuệ doanh nghiệp và các công cụ tìm kiếm là hai ứng dụng phổ biến và thành công trong khai phá dữ liệu.

**1.1.4.1.Trí tuệ doanh nghiệp**

Trí tuệ doanh nghiệp (BI) rất cần thiết và quan trọng cho các doanh nghiệp để có được một sự hiểu biết tốt hơn về bối cảnh thương mại, khách hàng, thị trường, nhà cung cấp, các nguồn lực đối thủ cạnh tranh. BI cung cấp quan điểm lịch sử, hiện tại và dự báo hoạt động kinh doanh, báo cáo, xử lý phân tích trực tuyến, quản lý hoạt động kinh doanh, tình báo cạnh tranh, điểm chuẩn, và các phân tích dự đoán.

"Trí tuệ doanh nghiệp có tầm quan trọng như thế nào?" Nếu không có khai phá dữ liệu, nhiều doanh nghiệp không thể thực hiện phân tích thị trường hiệu quả, so sánh thông tin phản hồi của khách hàng về các sản phẩm, khám phá những điểm mạnh và điểm yếu của đối thủ cạnh tranh của họ, giữ chân khách hàng có giá trị cao, và ra quyết định kinh doanh có hiệu quả hơn.

Rõ ràng, khai phá dữ liệu là cốt lõi của Trí tuệ doanh nghiệp. Công cụ phân tích xử lý trực tuyến trong Trí tuệ doanh nghiệp dựa trên kho dữ liệu và khai phá dữ liệu đa chiều. Phân loại và dự báo là kỹ thuật cốt lõi của phân tích dự đoán trong Trí tuệ doanh nghiệp, có nhiều ứng dụng trong việc phân tích thị trường, nguồn cung cấp và bán hàng. Hơn nữa, phân nhóm đóng một vai trò trung tâm trong quản lý quan hệ khách hàng, phân nhóm khách hàng dựa trên sự tương đồng của họ. Sử dụng kỹ thuật khai thác sự cá nhân hóa, chúng ta có thể hiểu rõ hơn về tính năng của từng nhóm khách hàng và phát triển các chương trình ưu đãi khách hàng cá nhân hoá.

**1.1.4.2.Công cụ tìm kiếm**

Một công cụ tìm kiếm Web là một máy chủ máy tính chuyên ngành để tìm kiếm thông tin trên Web. Các kết quả tìm kiếm của một truy vấn người dùng thường trở lại như một danh sách. Danh sách này có thể bao gồm các trang web, hình ảnh, và các loại tệp khác. Một số công cụ tìm kiếm cũng tìm kiếm và trả về dữ liệu có sẵn trong cơ sở dữ liệu công cộng hoặc thư mục mở. Công cụ tìm kiếm khác với thư mục web, trong đó các thư mục web được duy trì bởi các biên tập viên của con người trong khi công cụ tìm kiếm hoạt động thuật toán hoặc bởi một hỗn hợp của đầu vào thuật toán và con người.

Công cụ tìm kiếm web chủ yếu là các ứng dụng khai phá dữ liệu rất lớn. Kỹ thuật khai phá dữ liệu khác nhau được sử dụng trong tất cả các khía cạnh của công cụ tìm kiếm, thu thấp thông tin khác (ví dụ, quyết định các trang cần được thu thập thông tin và tần suất thu thâp), lập chỉ mục (ví dụ, lựa chọn các trang được lập chỉ mục và quyết định mức độ mà các chỉ số cần được xây dựng), và tìm kiếm (ví dụ như quyết định cách mà các trang nên được xếp hạng, trong đó quảng cáo cần phải được thêm vào, làm thế nào các kết quả tìm kiếm có thể được cá nhân hóa hoặc "nhận thức được bối cảnh").

Công cụ tìm kiếm là một thách thức lớn trong khai phá dữ liệu. Đầu tiên, họ phải xử lý một số lượng lớn và ngày càng tăng của dữ liệu. Thông thường, dữ liệu như vậy không thể được xử lý bằng cách sử dụng một hoặc một vài máy. Thay vào đó, công cụ tìm kiếm thường xuyên sử dụng mô hình điện toán đám mây, trong đó bao gồm hàng ngàn hoặc thậm chí hàng trăm ngàn máy tính hợp tác khai thác số lượng lớn dữ liệu. Nhân rộng các phương pháp khai phá dữ liệu trên mô hình điện toán đám mây và các bộ dữ liệu phân phối lớn là một lĩnh vực để nghiên cứu thêm.

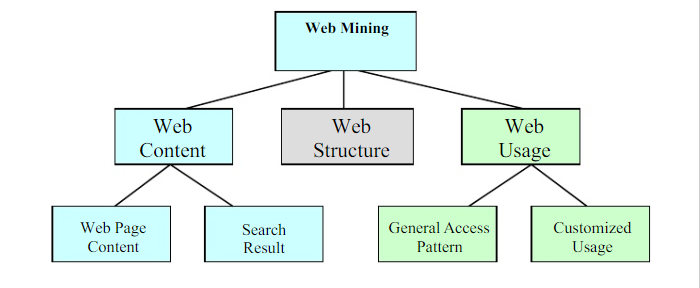
Thứ hai, công cụ tìm kiếm Web thường xuyên phải đối phó với dữ liệu trực tuyến. Một công cụ tìm kiếm có thể đủ khả năng xây dựng một mô hình ẩn trên bộ dữ liệu khổng lồ. Để làm điều này, nó có thể xây dựng một phân loại truy vấn mà chỉ định một truy vấn tìm kiếm để loại được xác định trước dựa trên chủ đề truy vấn (tức là, cho dù truy vấn tìm kiếm "Apple" có nghĩa là để lấy thông tin về một loại trái cây hoặc một thương hiệu của máy tính). Cho dù một mô hình được xây dựng ẩn, việc áp dụng các mô hình trực tuyến phải đủ nhanh để trả lời truy vấn người sử dụng trong thời gian thực. Một thách thức khác là duy trì và từng bước cập nhật một mô hình phát triển nhanh chóng trên luồng dữ liệu. Ví dụ, một phân loại truy vấn có thể cần phải được từng bước duy trì lien tục từ các truy vấn mới giữ loại đang nổi lên và được xác định trước và phân phối dữ liệu có thể thay đổi. Hầu hết các phương pháp đào tạo mô hình hiện đang offline và tĩnh và do đó không thể được sử dụng trong một kịch bản như vậy.

Công cụ tìm kiếm Web thường xuyên phải đối phó với các truy vấn được hỏi chỉ có một số rất ít lần. Giả sử một công cụ tìm kiếm muốn đưa ra khuyến nghị truy vấn nhận biết ngữ cảnh. Đó là, khi một người sử dụng đặt ra một truy vấn, công cụ tìm kiếm cố gắng để suy ra bối cảnh của các truy vấn bằng cách sử dụng của người sử dụng hồ sơ cá nhân và lịch sử truy vấn của mình để trở lại câu trả lời tùy biến hơn trong một phần nhỏ của một giây. Tuy nhiên, mặc dù tổng số các truy vấn yêu cầu có thể rất lớn, hầu hết các truy vấn có thể được yêu cầu chỉ một lần hoặc một vài lần. Dữ liệu sai lệch nghiêm trọng như đang thách thức đối với nhiều phương pháp khai phá dữ liệu và học máy.

# 1.2. Khai phá Web

## **1.2.1. Khái niệm khai phá dữ liệu web**

Trong những năm gần đây Internet đã đang và từng bước trở thành một trong những kênh thông tin đa dạng gồm rất nhiều các lĩnh vực khác nhau như khoa học, thông tin kinh tế, thương mại và quảng cáo. Chi phí thấp cho việc duy trì một trang web là lý do cho sự phát triển mạnh mẽ này. Khi so sánh với các công cụ truyền tải thông tin khác như báo giấy hay tạp chí thì chi phí cho một trang web rẻ hơn rất nhiều, hơn nữa thông tin được cung cấp tới cho người dùng trên các trang web còn được cập nhập một cách thường xuyên và nhanh chóng hơn tới với hàng triệu người dùng trên toàn thế giới. Có thể nói Internet là một môi trường khổng lồ được ví như Bách khoa toàn thư với nội dung và hình thức đa dạng, được ví như một xã hội ảo, bao gồm các thông tin về mọi mặt của đời sống và được trình bày dưới dạng văn bản, âm thanh…Tuy nhiên, Internet được nhìn nhận là một môi trường đa phương tiện và là sự kết hợp của tập hợp các cơ sở dữ liệu không đồng nhất, các chương trình và giao tiếp với người dùng. Vậy nên việc khai phá dữ liệu dưới dạng text chỉ là một phần rất nhỏ trong môi trường này. Việc khai phá dữ liệu web ngoài việc khai phá được nội dung các văn bản còn phải khai thác sâu hơn được các nguồn lực và mối tương quan liên kết giữa chúng. Về cơ bản các nội dung trong khai phá web bao gồm các thành phần sau:



### *Hình 1.4. Các nội dung trong khai phá web*

Có nhiều khái niệm khác nhau về khai phá Web, nhưng có thể tổng quát hóa như sau: **Khai phá Web là việc sử dụng các kỹ thuật KPDL để tự động hóa quá trình khám phá và trích rút những thông tin hữu ích từ các tài liệu, các dịch vụ và cấu trúc Web**. Hay nói cách khác khai phá Web là việc thăm dò những thông tin quan trọng và những mẫu tiềm năng từ nội dung Web, từ thông tin truy cập Web, từ liên kết trang và từ nguồn tài nguyên thương mại điện tử bằng việc sử dụng các kỹ thuật KPDL, nó có thể giúp con người rút ra những tri thức, cải tiến việc thiết kế các Web site và phát triển thương mại điện tử tốt hơn. Lĩnh vực này đã thu hút được nhiều nhà khoa học quan tâm. Quá trình khai phá Web có thể chia thành các công việc nhỏ, cụ thể như sau:

* Tìm kiếm nguồn tài nguyên: Thực hiện tìm kiếm và lấy các tài liệu Web phục vụ cho việc khai phá.
* Lựa chọn và tiền xử lý dữ liệu: Lựa chọn và tiền xử lý tự động các loại thông tin từ nguồn tài nguyên Web đã lấy về.
* Tổng hợp: Tự động khám phá các mẫu chung tại các Web site riêng lẻ cũng như nhiều Website với nhau.
* Phân tích: Đánh giá, giải thích, biểu diễn các mẫu khai phá được.

## **1.2.2. Lợi ích của khai phá Web**

Với sự phát triển nhanh chóng của thông tin trên www, KPDL Web đã từng bước trở nên quan trọng hơn trong lĩnh vực KPDL, người ta luôn hy vọng lấy được những tri thức hữu ích thông qua việc tìm kiếm, phân tích, tổng hợp, khai phá Web. Những tri thức hữu ích có thể giúp ta xây dựng nên những Web site hiệu quả để có thể phục vụ cho con người tốt hơn, đặc biệt trong lĩnh vực thương mại điện tử.

Khám phá và phân tích những thông tin hữu ích trên www bằng cách sử dụng kỹ thuật KPDL đã trở thành một hướng quan trọng trong lĩnh vực khám phá tri thức. Khai phá Web bao gồm khai phá cấu trúc Web, khai phá nội dung Web và khai phá các mẫu truy cập Web. Sự phức tạp trong nội dung của các trang Web khác với các tài liệu văn bản truyền thống. Chúng không đồng nhất về cấu trúc, hơn nữa nguồn thông tin Web thay đổi một cách nhanh chóng, không những về nội dung mà cả về cấu trúc trang. Chẳng hạn như tin tức, thị trường chứng khoán, thông tin quảng cáo, trung tâm dịch vụ mạng,...Tất cả thông tin được thay đổi trên Web theo từng giai đoạn. Các liên kết trang và đường dẫn truy cập cũng luôn thay đổi. Khả năng gia tăng liên tục về số lượng người dùng, sự quan tâm tới Web cũng khác nhau, động cơ người dùng rất đa dạng và phong phú. Vậy làm thế nào để có thể tìm kiếm được thông tin mà người dùng cần? Làm thế nào để có được những trang Web chất lượng cao?...

Những vấn đề này sẽ được thực hiện hiệu quả hơn bằng cách nghiên cứu các kỹ thuật KPDL áp dụng trong môi trường Web. Thứ nhất, ta sẽ quản lý các Website thật tốt; thứ hai, khai phá những nội dung mà người dùng quan tâm; thứ ba, sẽ thực hiện phân tích các mẫu sử dụng Web. Dựa vào những vấn đề cơ bản trên, ta có thể có những phương pháp hiệu quả cao để cung cấp những thông tin hữu ích đối với người dùng Web và giúp người dùng sử dụng nguồn tài nguyên Web một cách hiệu quả.

## **1.2.3. Các kiểu dữ liệu Web**

Ta có thể khái quát bằng sơ đồ sau:

Web data

Structure data

Usage data

User Profile data

Content data

Fr

ee Text

HTML file

XML file

Multimedia

Dynamic link

Static link

Dynamic content

### *Hình 1.5. Phân loại dữ liệu web*

Các đối tượng của khai phá Web bao gồm [21]: Server logs, Web pages, Web hyperlink structures, dữ liệu thị trường trực tuyến và các thông tin khác.

**Web logs:** Khi người dùng duyệt Web, dịch vụ sẽ phân ra 3 loại dữ liệu đăng nhập: sever logs, error logs, và cookie logs. Thông qua việc phân tích các tài liệu đăng nhập này ta có thể khám phá ra những thông tin truy cập.

**Web pages:** Hầu hết các phương pháp KPDL Web được sử dụng trong Web pages là theo chuẩn HTML.

**Web hyperlink structure:** Các trang Web được liên kết với nhau bằng các siêu liên kết, điều này rất quan trọng để khai phá thông tin. Do các siêu liên kết Web là nguồn tài nguyên rất xác thực.

**Dữ liệu thị trường trực tuyến:** Như lưu trữ thông tin thương mại điện tử trong các site thương mại điện tử. Các thông tin khác: Chủ yếu bao gồm các đăng ký người dùng, nó có thể giúp cho việc khai phá tốt hơn.

# 1.2.4. Xử lý dữ liệu văn bản ứng dụng trong khai phá dữ liệu Web

## **1.2.4.1. Dữ liệu văn bản**

Trong các loại dữ liệu hiện nay thì văn bản là loại dữ liệu phổ biến nhất và nó có mặt khắp mọi nơi, đặc biệt là đối với dữ liệu trên Web. Do vậy, các bài toán xử lý văn bản đã được đặt ra từ rất sớm và hiện nay nó vẫn là vấn đề rất được nhiều nhà nghiên cứu quan tâm, một trong những bài toán đó là tìm kiếm và trích dẫn văn bản, biểu diễn và phân loại văn bản,…

CSDL văn bản có thể chia làm 2 loại chính:

+ Dạng không có cấu trúc: Đây là những tài liệu văn bản thông thường mà ta đọc thường ngay trên các sách, báo, internet,…đây là dạng dữ liệu của ngôn ngữ tự nhiên của con người và nó không theo một khuôn mẫu định sẵn nào cả.

+ Dạng nửa cấu trúc: Đây là những văn bản được tổ chức dưới dạng cấu trúc lỏng, nhưng vẫn thể hiện nội dung chính của văn bản, như văn bản HTML, Email,…

## **1.2.4.2. Một số vấn đề trong xử lý dữ liệu văn bản**

Mỗi văn bản được biểu diễn bằng một vector Boolean hoặc vector số. Những vector này được xét trong một không gian đa chiều, trong đó mỗi chiều tương ứng với một từ mục riêng biệt trong tập văn bản. Mỗi thành phần của vector được gán một hàm giá trị *f*, nó là một số chỉ mật độ tương ứng của chiều đó trong văn bản. Nếu thay đổi giá trị hàm *f* ta có thể tạo ra nhiều trọng số khác nhau.

Một số vấn đề liên quan đến việc biểu diễn văn bản bằng mô hình không gian vector:

+ Không gian vector là một tập hợp bao gồm các từ.

+ Từ là một chuỗi các ký tự (chữ cái và chữ số); ngoại trừ các khoảng trống (space, tab), ký tự xuống dòng, dấu câu (như dấu chấm, phẩy, chấm phẩy, dấu cảm,…). Mặt khác, để đơn giản trong quá trình xử lý, ta không phân biệt chữ hoa và chữ thường (nếu chữ hoa thì chuyển về chữ thường).

+ Cắt bỏ từ: Trong nhiều ngôn ngữ, nhiều từ có cùng từ gốc hoặc là biến thể của từ gốc sang một từ khác. Việc sử dụng từ gốc làm giảm đáng kể số lượng các từ trong văn bản (giảm số chiều của không gian), nhưng việc cắt bỏ các từ lại rất khó trong việc hiểu văn bản.

Ngoài ra, để nâng cao chất lượng xử lý, một số công trình nghiên cứu đã đưa ra một số cải tiến thuật toán xem xét đến đặc tính ngữ cảnh của các từ bằng việc sử dụng các cụm từ/văn phạm chứ không chỉ xét các từ riêng lẽ. Những cụm từ này có thể được xác định bằng cách xem xét tần số xuất hiện của cả cụm từ đó trong tài liệu.

Bằng phương pháp biểu diễn không gian vector, có thể thấy rõ ràng là chiều của một vector sẽ rất lớn bởi số chiều của nó được xác định bằng số lượng các từ khác nhau trong tập hợp từ. Chẳng hạn, số lượng các từ có thể từ 103 đến 105 đối với các tập văn bản nhỏ. Vấn đề đặt ra là làm sao để giảm số chiều của vector mà vẫn đảm bảo việc xử lý văn bản đúng và chính xác, đặc biệt là trong môi trường www, sẽ xem xét đến một số phương pháp để giảm số chiều của vector.

# 1.3. Tổng quan về sự kiện

Trích chọn sự kiện với vai trò trích chọn ra các thông tin có ý nghĩa từ tập dữ liệu lớn và được cộng đồng khoa học rất quan tâm và đầu tư nghiên cứu.

Năm 1987, Message Understanding Conferences (MUC)quỹ nghiên cứu bộ quốc phòng Hoa Kỳ và lần đầu tiên khái niệm event (sự kiện) được đề cập. Sau đó, rất nhiều hội nghị được tổ chức tạo thành dãy hội nghị MUC. Với mỗi hội nghị, thông tin được quan tâm khác nhau nhưng đều có đặc điểm chung là chúng được trích xuất từ dữ liệu nói về khủng hoảng (crisis). 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 chọ thông tin dựa trên mẫu (scenariotemplate). Các mẫu được bam tổ chức quy định và các đội tham gia cần điền thông tin vào các mẫu này một cách 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 tuongwrv(recall) của các nghiên cứu tham dự MUC nằm trong khoảng 50% đến 60% [20].

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 hóa quốc gia Hoa Kỳ (NIST) và DAPRA nhằm giải quyết bài toán phát hiện, theo dõi và xâu chuỗi sự kiện. Một số nhóm nghiên cứu tham gia chương trình như sau: nhóm CMU của đại học Canegie Mellon, nhóm BBN từ công ty BBN Technologies, nhóm DRAGON của công ty Dragon, nhóm UPENN của trường đại học Pennsylvania (UPENN). Các bài toán quan trọng của TDT gồm: Story Segmentation, Topic Tracking, Topic Detection, First Story Detection, và Link Detection.

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 nầy 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.

## **1.3.1. Định nghĩa sự kiện**

Trích chọn sự kiện lần đầu tiên được giới thiệu như một chủ đề quan trọng trong Message Understanding Conference (MUC) năm 1987 [20]. Trong MUC, một sự kiện được định nghĩa như sau: “một sự kiện có tác nhân (actor), thời gian (time), địa điểm (place) và tác động tới môi trường xung quanh”.

Trong chương trình ACE, Đoington Deorge R và cộng sự đưa ra định nghĩa sự kiện như sau: “một sự kiện là một hàng động được tạo bởi những người tham gia” [20]. ACE chia sự kiện thành 8 loại khác nhau: LIFE (sự sống – chết), MOVEMENT (sự di chuyển), TRANSACTION (giao dịch), BUSINESS (kinh tế), CONFLICT (xung đột), CONTACT (giao thiệp), PERSONNEL (nhận – đổi việc), JUSTICE (pháp lý). Mỗi dạng sự kiện lại phân biệt từng dạng con. Ví dụ, LIFE có các dạng con như BE-BORN (chào đời), INJURE (bị thương), DIE (chết), hay PERSONAL có START- POSITION (vị trí khi nhận việc), END- POSITION (vị trí thôi việc), NOMINATE (bổ nhiệm), ELECT (bầu chọn)…

Có thể thấy rằng các nghiên cứu liệt kê ở trên đều đồng ý rằng sự kiện có thể coi như một mẫu (template) gồm nhiều các thuộc tính (elements). Quá trình trích chọn sự kiện quan tâm tới việc làm thế nào có thể điền các thông tin phù hợp từ các văn bản gốc tương ứng từng thuộc tính.

## **1.3.2. Trích chọn sự kiện**

Trích chọn sự kiện và trích chọn thông tin có điểm 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, “*Ngày 1/4 Trung tâm Kiểm soát bệnh tật Hà Nội cho biết trên địa bàn thành phố ghi nhận 79 ca mắc Sởi (tăng 19 trường hợp so với tuần liền trước*”. Từ ví dụ trên có thể thấy rằng những thông tin được trích chọn được đưa ra dưới dạng các kết của rời rạc như: 1/4, Hà Nội, Sởi, 79 ca mắc. Còn đối với trích chọn sự kiện thì quan tâm tới một bộ các thuộc tính biểu diễn cho sự kiện gồm {1/4, Hà Nội, Sởi, 79 ca mắc}. Rõ ràng, với tập dữ liệu trên, thông tin là hữu ích và đầy đủ hơn các thông tin rời rạc.

Một cách tổng quát, **có thể coi trích chọn sự kiện trong văn bản nhận đầu vào là các văn bản phi cấu trúc và đầu ra là tri thức được biểu diễn dưới dạng thông tin có cấu trúc.** Những thông tin này rất hữu ích cho việc khai thác dữ liệu như: thống kê, hệ thống giám sát, các hệ thống hỗ trợ ra quyết định. Trích chọn sự kiện có thể áp dụng cho một miền dữ liệu cụ thể như vụ tai nạn giao thông, thông tin các tour du lịch, bệnh dịch,…đồng thời đưa ra các thông tin xung quanh sự kiện đó thường bao gồm: thời gian, địa điểm, số lượng,…

Theo Grishman và cộng sự, trích chọn sự kiện là một bài toán khó do vấn đề xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) và đặc trưng dữ liệu [21]. Dễ ràng nhận thấy trích chọn sự kiện phụ thuộc nhiều vào NLP, cụ thể là bài toán nhận dạng thực thể (Named Enity Recognition – NER). Bên cạnh đó, dữ liệu đầu vào của V rất đa dạng nên sẽ ảnh hưởng tới tính hiệu quả của quá trình trích chọn.

## **1.3.3.Trích chọn sự kiện y tế từ các trang web**

### **1.3.3.1. Bài toán trích chọn thông tin văn bản**

Trích chọn thông tin (Information Extraction – IE), đặc biệt là trích chọn sự kiện (Event Extraction – EE) là một lĩnh vực con trong khai phá dữ liệu (Data Mining – DM). Những năm gần đây, trích chọn sự kiện đã thu hút nhiều sự quan tâm từ các nhà khoa học. Nó là bước đi tốt cho việc khai thác tri thức trên văn bản. Ngày nay thông tin đang phát triển một cách vô cùng mạnh mẽ và biến đổi một cách nhanh chóng. Việc trích chọn các thông tin theo chủ đề là rất cần thiết và mang tính ứng dụng thực tiễn cao. Cụ thể những thông tin được trích chọn liên quan tới lĩnh vực y tế sẽ đem lại những lợi ích rất lớn trong công tác phát hiện và phòng ngừa dịch bệnh, đảm bảo sức khỏe cộng đồng. Giúp ứng phó kịp thời và có hiệu quả trước các dịch bệnh nguy hiểm.

Trích chọn thông tin về sự kiện y tế như: thời gian (giờ trong ngày), thời gian (dd/mm/yyyy), thứ/tuần, tháng/năm, địa bàn nơi có dịch bệnh, số ca nhập viện hoặc số ca tử vong do dịch… Kết quả của quá trình trích chọn được làm đầu vào cho hệ thống khai thác như thống kê và trực quan hóa trên bản đồ những địa điểm nóng hay xảy ra dịch bệnh, nơi xuất hiện các ổ dịch mới, tháng nào hay mùa nào trong năm có nguy cơ xảy ra dịch bệnh nhiều hơn, độ tuổi có nguy cơ mắc bệnh. Những điều đó giúp ích cho các nhà quản lý có biện pháp khoanh vùng các ổ dịch, giúp khắc phục để giảm thiểu số số lượng người mắc bệnh, kịp thời đưa ra những khuyên cáo cho người dân phòng tránh bệnh lây lan trong cộng đồng.

Bài toán trích chọn sự kiện vụ tai nạn được phát biểu như sau:

**Đầu vào:** bản tin bất kỳ trên báo điện tử

**Đầu ra:** trích chọn những thông tin của sự kiện y tế (nếu có).

Bài toán trích chọn sự kiện y tế được chia thành hai bài toán. Bài toán thứ nhất, phát hiện sự kiện y tế, đầu vào là bản tin bất kỳ trên báo điện tử, bài toán phải chỉ ra đâu là sự kiện y tế. Kết quả của bài toán phát hiện sự kiện sẽ là dữ liệu đầu vào cho bài toán trích chọn; thông tin được trích chọn trong sự kiện y tế có thể là thời gian, địa điểm có dịch bệnh, số lượng các ca mắc, loại dịch bệnh, độ tuổi của những trường hợp mắc bệnh. Trong giới hạn đề tài, khóa luận sẽ tập trung vào việc trích chọn ra bộ các thuộc tính như: (thời gian, địa điểm có dịch bệnh, số lượng các ca nhập viện, loại dịch bệnh).

### **1.3.3.2.Phát hiện sự kiện**

Bài toán phát hiện sự kiện trả lời câu hỏi “*làm thế nào để phát hiện được một văn bản có chứa sự kiện y tế*”. Tức là, cho trước đầu vào là văn bản, làm thế nào để phát hiện văn bản đó có chưa sự kiện y tế? Theo Grishman và cộng sự [13], phát hiện sự kiện là quá trình học không giám sát, tác giả sử dụng các từ khóa để quyết định một văn bản có chứa sự kiện dịch bệnh hay không. Hai từ khóa được tác giả sử dụng là “outbreak of…” và “died from…”. Theo Doan và cộng sự [14], bài toán phát hiện sự kiện có thể coi như quá trình học có giám sát. Trong nghiên cứu của mình, tác giả sử dụng phương pháp để phân lớp các tài liệu. bộ phân lớp này dựa trên một tập các dữ liệu đã được gán nhãn. Qua quá trình huấn luyện, bộ phân lớp sẽ quyết định một văn bản đầu vào có chứa sự kiện dịch bệnh hay không.

Từ nghiên cứu của Grishman và cộng sự hoặc nghiên cứu của Doan và cộng sự, có các cách khác nhau để giải quyết bài toán phát hiện sự kiện dịch bệnh. Do đó, có thể vận dụng phương pháp này cho việc phát hiện sự kiện y tế cùng với việc xây dựng bộ từ khóa hoặc xây dựng một tập các dữ liệu đã được gán nhãn phù hợp cho sự kiện y tế.

### **1.3.3.3.Trích chọn sự kiện**

Nhiệm vụ của bài toán trích chọn sự kiện phải trả lời câu hỏi “làm thế nào để trích chọn các thuộc tính của một sự kiện”. Có nhiều phương pháp cho việc trích chọn sự kiện; trong đó phải kể đến phương pháp sử dụng luật (học không giám sát) được sử dụng từ rất sớm để giải quyết bài toán này [13]. Quá trình trích chọn bằng phương pháp này thường được sử dụng các luật dựa vào quá trình khảo sát dữ liệu để trích ra các thuộc tính của một sự kiện.

Phương pháp sử dụng học máy và các kỹ thuật NLP để giải quyết bài toán trích chọn sự kiện. quá trình này thường sử dụng Named Entity Recognition (NER) để lấy ra các thuộc tính cơ bản của sự kiện: thời gian, địa điểm, tên người… sau đó kết hợp các thuộc tính này thành một sự kiện.

Như vậy, bài toán trích chọn sự kiện nói chung hay bài toán trích chọn sự kiện y tế nói riêng có thể được chia thành hai bài toán con, đó là: phát hiện sự kiện và trích chọn sự kiện. trong khóa luận này, em sẽ mô tả chi tiết các kỹ thuật được áp dụng để giải quyết hai bài toán này ở chương 3.

# 1.4.Ý nghĩa của bài toán trích chọn sự kiện

## **1.4.1. Ý nghĩa khoa học**

Ý nghĩa khoa học của bài toán trích chọn sự kiện được rất nhiều các nhà nghiên cứu quan tâm. Kết quả của bài toán trích chọn sự kiện y tế có thể dùng làm tiền đề cho việc khai thác dữ liệu như thống kê, dự đoán xu hướng, hệ thống giám sát và hỗ trợ ra quyết định.

## **1.4.2. Ý nghĩa thực tiễn**

Kết của của việc trích chọn ra sự kiện y tế là dữ liệu đầu vào vô cùng quan trọng phục vụ cho những công việc khai thác và sử dụng dữ liệu cho các công việc tiếp theo: thống kê các con số có liên quan tới tình hịnh những dịch bệnh nguy hiểm đang lây lan trong cộng đồng, thời gian nào trong năm xảy ra những dịch bệnh truyền nhiễm, loại dịch bệnh đang hoành hành thuộc các nhóm bệnh nào (dịch sởi, tay chân miệng, sốt xuất huyết…), những vùng dịch, địa bàn để xảy ra dịch bệnh (các tỉnh, thành phố nơi xảy ra dịch), độ tuổi thường mắc bệnh (bệnh nhi, người trung niên, người cao tuổi…), số lượng thông kê các ca mắc bệnh phải nhập viện điều trị, số ca tử vong do dịch, ngoài ra có thể là những thông tin liên quan tới hướng dẫn phòng chống bệnh….Từ những con số thống kê trên những người làm trong bộ phận y tế có thể trực quan hóa trên bản đồ dịch bệnh để phân vùng dịch bệnh để thực hiện các công tác hỗ trợ triển khai đến với người dân để phòng tránh dịch. Người dân có thể tiếp cận được những thông tin vô cùng hữu ích này để có thể nắm được những thông tin về dịch bệnh như: Địa bàn dịch, loại dịch bệnh… Quan trọng hơn đó là mọi người sẽ chủ động có những biện pháp tự bảo vệ mình và những người thân trong gia đình trước nguy cơ dịch bệnh đang diễn ra. Từ đó xây dựng một cộng động khỏe mạnh, phòng chống nguy cơ lây lan rộng làm ảnh hưởng tới sức khỏe và tính mạng.

# 1.5. Kết luận chương

Trong chương này, khóa luận đã trình bày cơ bản các phần lý thuyết tổng quan của quá trình khai phá dữ liệu và khai phá dữ liệu web, bên cạnh đó là các vấn đề liên quan tới bài toán phát hiện sự kiện và bài toán trích chọn sự kiện. Ngoài ra, nêu lên được ý nghĩa khoa học, ý nghĩa thực tiễn và kèm với đó là những khó khăn và thách thức trong tiến hành quá trình giải quyết bài toán trích chọn sự kiện y tế. Trong chương tiếp theo, khóa luận sẽ trình bày các phương pháp tiếp cận để có thể giải quyết bài toán phát hiện sự kiện và trích chọn sự kiện y tế.

# CHƯƠNG 2: MỘT SỐ PHƯƠNG PHÁP TIẾP CẬN TRONG KHAI PHÁ DỮ LIỆU VÀ TRÍCH CHỌN THÔNG TIN

# 2.1. Phương pháp tiếp cận dựa trên tập luật (Rule – Based)

Phương pháp dựa trên tập luật hay còn được biết đến với một tên gọi khác là phương pháp dựa vào tri thức (knowledge – driven). Phương pháp này được dựa trên nền tảng tri thức. Thông thường phương pháp áp dụng những kiến thức của các chuyên gia miền để có thể sinh ra được tập luật, các chuyên gia này thuộc các lĩnh vực liên quan tới vấn đề ngôn ngữ và một số các chuyên gia miền dữ liệu). Phương pháp này đòi hỏi trước tiên là phải đọc và hiểu được bộ dữ liệu, tiếp theo đó mới là tiến hành sinh ra các tập luật.

## **2.1.1. Luật cú pháp**

Luật cú pháp được định nghĩa là mẫu cú pháp (lexico – syntactic patterns) có thể coi là một phương pháp sử dụng sớm có trong các bài toán liên quan tới chủ đề trích chọn các thông tin và sự kiện. Các chuyên gia miền (expert knowledge) sẽ sinh ra các mẫu này và các mẫu được biểu diễn thông qua các tập luật (rules) [4]. Một trong những điển hình khi nghiên cứu phương pháp này đó chính là việc các luật được biểu diễn thông qua hình thức dạng biểu thức chính quy (regular expression).

Các luật cú pháp được hình thành nhờ sự kết hợp của ba thành phần chính đó là sự kết hợp biểu diễn của các kí tự và các thông tin cú pháp, kết hợp cới đó là các biểu thức chính quy. Sau quá trình xây dựng các biểu mẫu chính quy các dữ liệu này sẽ được dùng để kết hợp với dữ liệu có trong các văn bản đầu vào để từ đó tiến hành công việc trích chọn ra được các thông tin tương tứng đối với mỗi thuộc tính. Tuy nhiên trong một số trường hợp thì luật cú pháp được hiểu diễn ở các dạng đơn giản hơn đó chính là các từ khóa. Tập luật cú pháp đã được sử dụng trong việc trích chọn các thông tin và sự kiện [5] [6] [7]. Trong nghiên cứu của mình thì tác giả Nishihara cùng các cộng sự đã đề cập và sử dụng ba từ khóa đó chính là: Địa điểm (place), đối tượng (Object) và hành vi (Action) để từ đó có thể biểu diễn được một sự kiện hay một thông tin có chứa trong các blogs. Đối với lĩnh vực y sinh, Yakushiji và các cộng sự đã sử dụng bộ phân tích và kết hợp với ngữ pháp nhằm xác định mối quan hệ và các sự kiện [16]. Còn đối với lĩnh vực tiền tệ và chính trị thì Aone cùng các cộng sự lại sử dụng và nghiên cứu hướng đi khác đó là việc dùng luật cú pháp để có thể trích chọn ra được những thông tin của sự kiện [24]. Luật cú pháp xác định các tham số bên trong văn bản tuy nhiên lại không tiến hành xác định ý nghĩa của văn bản.

Trong quá trình sử dụng luật để trích chọn các sự kiện thì đôi khi chúng ta phải trích chọn ra được các khai niệm mà chúng ẩn chứa những ý nghĩa đặc biệt hoặc mối quan hệ giữa các thành phần được trích chọn. Chính vì vậy, luật cú pháp không đủ để có thể đáp ứng được những nhu cầu này. Nhằm giải quyết vấn đề trên thì phương pháp thường được sử dụng ở trong (rule – based) là sử dụng luật ngữ nghĩa (lexico -semantic patterns).

## **2.1.2. Luật ngữ nghĩa**

Trong một số các trường hợp thì trích chọn thông tin sự kiện phải trích chọn các khái niệm có ý nghĩa đặc biệt hoặc mối quan hệ giữa các thành phần được tiến hành trích chọn. Chính vì vậy để có thể giải quyết được vấn đề này phương pháp được áp dụng trong (rule-based) là sử dụng luật ngữ nghĩa. Các luật ngữ nghĩa không chỉ đơn giản là các từ được biểu diễn dưới dạng biểu thức chính quy mà là các từ và mối liên hệ giữa chúng.

Trong thực tế, luật ngữ nghĩa được sử dụng với nhiều mục đích cũng như các lĩnh vực đa dạng khác nhau có thể kể tới như trong một vài các nghiên cứu sai: Li Fang và các cộng sự đã sử dụng luật ngữ nghĩa để trích chọn trông tin có ở các sàn chứng khoán (stock market) [25], Hay Cohen cùng các cộng sự [17] sử dụng khái niệm bộ nhận dạng (recognozer) trên miền dữ liệu y sinh để có thể trích chọn các thông tin liên quan tới y sinh có trong cập dữ liệu; Capet và các cộng sự lại sử dụng mẫu ngữ nghĩa để trích chọn sự kiện cho hệ thống cảnh báo sớm [27] còn Vargas-Vera và Celjuska đề xuất một bộ khung (framework) cho việc nhận diện các sự kiện tập trung trên báo Knowledge Media Institute (KMI)[26].

Có rất nhiền các ứng dụng đối với các lĩnh vực thông qua việc sử dụng trích chọn sự kiện phi cấu trúc vi dự như lĩnh vực tài chính, chứng khoán, y sinh, bản tin pháp luật…Có lẽ sẽ là chưa đầy đủ và chi tiết khi chưa đề tập tới các hình dạng và biểu diễn của tập luật trong trích chọn các thực thể.

## **2.1.3. Hình dạng và biểu diễn của tập luật**

Dựa vào tài liệu Information Extraction của Sunita Sarawagi [1], tài liệu này có đề cập tới một luật cơ bản có dạng: “Mẫu theo ngữ cảnh => hành động”. Một mẫu theo ngữ cảnh báo và trong mẫu này bao gồm một hoặc nhiều mẫu ghi lại tất cả các thuộc tính của một hoặc nhiều thực thể và bối cảnh xuất hiện trong văn bản. Mỗi một mẫu khi đã được gán nhãn đồng nghĩa với việc mẫu này là khớp so với một biểu thức chính quy và được xác định dựa vào các tính năng của thẻ trong băn bản và một nhãn tùy chọn. Các thuộc tính có thể được chỉ ra là thuộc tính của các thẻ trong văn bản và một nhãn tùy chọn. Các thuộc tính có thể được chỉ ra có thể là các thuộc tính của thẻ hoặc cũng có thể là ngữ cạnh hay thậm chí là các văn bản mà chúng có trong các thẻ xuất hiện.

Thông thường thì hầu hết các hệ thống hiện nay dựa trên luật được liên tầng, luật có thể được áp dụng trong nhiều giai đoạn khác nhau, mỗi một giai đoạn sử dụng các luật phù hợp để có thể liên kết được một dữ liệu đầu vào kèm theo đó có thể là một chú thích như là tính năng có các đầu vào để phục vụ cho các giai đoạn tiếp theo. Ví dụ như là việc trích chọn cho các địa điểm, địa chỉ liên lạc lạc của người được tạo ra thông qua hai giai đoạn của luật: giai đoạn đầu tiên là dùng nhãn thẻ cùng với nhãn thực thể chẳng hạn như: tên người, vị trí địa lý như tên đường hoặc tên các thành phố, và địa chỉ thư điện tử. Giai đoạn tiếp theo chính là xác định khối địa chỉ kèm theo đầu ra của giai đoạn thứ nhất được coi như một thuộc tính bổ sung.

## **2.1.4. Các thuộc tính của các thẻ**

Mỗi một thẻ có trong một câu thông thường được sử dụng kết hợp với cùng tập các thuộc tính thu được thông qua một số bộ các tiêu chí như sau:

* Các chuỗi mà nó đại diện cho các thẻ
* Các định dạng chính tả của thẻ có thể có dạng từ in hoa, từ in nhỏ, hỗn hợp, ký số, dấu cách và dấu chấm câu…
* Các phần phát biểu của thẻ (part of speech) có của thẻ
* Danh sách xuất hiện các thẻ của từ điển. Thông thường, điều này có thể được tiếp tục chắt lọc để có thể chỉ ra được, nếu các thẻ phù hợp với từ bắt đầu hoặc từ kết thúc, hoặc từ giữa của từ điển. Ví dụ như thẻ “New” phù hợp với từ đầu tiên của thành phố, tên sẽ được liên kết với các thuộc tính.

“Dictionary – Lookup = start of city”

* Chú thích kèm theo các bước xử lý trước đó

Luật để xác định một thực thể đơn (Rules to Indentify a Single Enity): Luật để nhận ra một thực thể đơn đầy đủ bao gồm ba loại mẫu.

* Một mẫu tùy chọn để ghi lại bối cảnh trước khi bắt đầu của một thực thể
* Một mẫu kết hợp các thẻ trong thực thể
* Một mẫu tùy chọn để ghi lại bối cảnh sau khi kết thúc thực thể

Ví dụ về một mẫu để xác định tên người có dạng “Dr. Yair Weiss” bao gồm một thẻ tiêu đề được liệt kê trong tập từ điển chức danh (có chưa các mục như: “Prof”, “Dr”, “Mr”) một dấu chấm, và hai từ viết hoa là

({Dictonary – Lookup = Titles} {String = “.”} {Orthography type = capitalization word} {2}) => Person Names

Mỗi điều kiện nằm bên trong dấu ngoặc nhọn là một điều kiện của một thẻ được theo sau cùng với số tùy chọn và chỉ ra số lần lặp này của các thẻ. Ví dụ về luật để đánh dấu tất cả các số đi sau các giới từ “by” và “in” là thực thể năm: (String = “by”|String = “in”})({Orthography type = Number}): y🡪Year=:y. Có thể trong hai mẫu trong luật này: mẫu đầu tiên để ghi lại ngữ cảnh xuất hiện các thực thể năm và mẫu thứ hai ghi lại các tính chất của thẻ tạo thành “year”. Một ví dụ khách cho việc tìm kiếm tên của công ty dạng “The XYZ Corp” or “ABC Ltd.” Được tạo bởi:

({String= ”The”}? {Orthography type = All Capitalization} {Orthography type = Capitalization word, Dictionary type = Company end}) 🡪 Company name.

## **2.1.5. Các định luật đánh dấu ranh giới thực thể**

Đối với một số các loại thực thể, trong các đơn vị dài một cách đặc biệt như tiêu đề cuốn sách, nó là một hình thức phân tích hiệu quả hơn nhằm mục tiêu xác định các luật đặc biệt để đánh dấu sự bắt đầu cũng như sự kết thúc ở một ranh giới cụ thể. Việc làm này được coi như là một cách để có thể loại bỏ độc lập và tất cả các thẻ ở giữa hai thẻ đánh dấu đầu cuối và được gọi là thực thể. Nhìn nhận các vấn đề dưới những khía cạnh khác nhau, mỗi luật cơ bản dẫn tới sự chèn của một đơn thẻ trong văn bản mà các đơn thẻ này có thể là một thẻ bắt đầu hoặc cũng có thể là một thẻ kết thúc. Để có thể giải quyết được vấn đề không nhất quán khi có hai thực thể bắt đầu đánh dấu trước và chỉ một thực thể đánh dấu kết thúc, điều này cần phải giải quyết bằng một cách đặc biệt. Ví dụ một quy tắc để có thẻ chèn một thẻ <journal> để đánh dấu sự bắt đầu của một tên tạp chí trong một bản trích dẫn:

({String=”to”} {String = “appear”} {String=”in”}): jstart

({Orthography type = Capitalization word {2,5})🡪 insert <journal> after: jstart

Nhiều hệ thống trích chọn dựa tên luật thành công đã dựa trên các luật tương tự như vậy

## **2.1.6. Các luật cho đa thực thể**

Một số các luật có cáu trúc dưới dạng các biểu thức chính quy với nhiều slot, mỗi một slot lại đại diện cho một thực thể khác nhau sao cho luật này dẫn đến sự công nhận của một lúc nhiều các đối tượng khác nhau. Ví dụ thống kê dựa vào luật WHISK [18] đã được chọn lựa để nhắm tới việc khai thác từ các hồ sơ có cấu trúc ví dụ như hồ sơ y tế, các bản ghi liên quan tới việc bảo trì các thiết bị và phân loại các quảng cáo. Các luật này được viết lại từ [18] để tiến hành trích chọn hai thực thể, số lượng các phòng ngủ hiện tại và các phòng đang được dùng để cho thuê từ một quảng cáo của một công ty cho thuê căn hộ

({Orthography type = Digit}) Bedrooms ({String = “BK”}) ({String = “$”}) type = Number): Price -🡪 Number of Bedroom , Rent =: Price

## **2.1.7. Chọn lựa hình dạng của các tập luật**

Hiện nay có rất nhiều các hệ thống dựa trên luật state-of-the-art điều này cho phép các chương trình có thể tùy ý được sử dụng các ngôn ngữ hàm thủ tục như Java hay C++ để thay thế cho cả hai thành phần mẫu và phần hành vi của các luật. Một ví dụ là GATE [9] hỗ trợ để có thể dùng được Java thay cho các ngôn ngữ hình thức các luật tùy chỉnh được gọi là JAPE có trong các hoạt động của một luật. Đây là một khả năng vô cùng mạnh mẽ vì nó có thể cho phép phần hành vi của các quy tắc để có thể tiến hành truy cập các thuộc tính khác nhau mà được sử dụng trong thành phần mẫu của các quy tắc và nhằm mục đích đó là để chèn các dạng chuẩn đối với một chuỗi chú thích. Ví dụ phần hoạt động có thể dẫn đến chèn các dạng chuẩn của một chuỗi có trong từ điển. Các trường mới được coi như một thuộc tính để bổ sung cho một luật trong các đường lấy tin riêng. Tương tự đối với các công thức trong Prolog-based từ [20] thì bất kì mã thủ tục nào đều hoàn toàn có thể được thay thế như là so khớp mẫu cho bất kì các tập con nào của các loại thực thể.

Nói chung, đối với các hệ thống tri thức (knowledge system) thì ban đầu phương pháp được sử dụng đó chính là tiếp cận dựa vào luật (rule – based). Những mặt tích cực của phương pháp này đó thứ nhất đó chính là không cần phải sử dụng quá nhiều dữ liệu huấn luyện nếu đem so sánh với phương pháp tiếp cận dựa trên dữ liệu, thứ hai phương pháp này dùng để xây xựng nên các biểu thức chính quy một cách tốt nhất để phù hợp cho việc trích chọn thông tin dựa trên cú pháp từ vựng và các thành phần ngữ nghĩa. Phương pháp tiếp cận trên phù hợp đối với các bài toán có liên quan tới việc chọn lựa các thông tin về thời gian (Ví dụ “Rạng sáng ngày hôm qua” “Giữa trưa nay”). Phương pháp này đưa lại kết quả có độ chính xác là rất cao (do nó được dùng để lấy ra những thông tin mang tính chất đặc biệt) và có độ hồi tưởng là tương đối nhỏ. Chính vì vậy phương pháp này rất phù hợp cho các bài toán chỉ có liên quan tới độ chính xác.

Tuy nhiên bên cạnh những ưu điểm vừa kể trên thì vẫn còn tồn tại một số những nhược điểm nhất định. Quá trình sử dụng phương pháp này đỏi hỏi người xây dựng phải đóng vai trò quan trọng được coi như là một chuyên gia miền dữ liệu, và đặc biệt cần am hiểu dữ liệu bên cạnh đó là khả năng nắm bắt được kiến thức về mặt ngôn ngữ, từ vựng cú pháp. Hơn thế nữa, thông thường luật được xây dựng để có thể lấy ra được những thông tin đặc biệt, chính vì vậy khi có sự thay đổi sang các miền dữ liệu khác thì chúng ta lại phải làm lại công đoạn xây dựng tập dữ liệu luật để phù hợp trong hoàn cảnh mới. Việc xây dựng tập luật này trong một số các trường hợp cụ thể đòi hỏi phải đầu tư thời gian và chi phí tương đối lớn.

# 2.2 Phương pháp tiếp cận dựa trên học máy

Phương pháp tiếp cận này thông thường được biết đến với cái tên gọi là tiếp cận dữ liệu (data -driven). Đây là phương pháp được sử dụng để nhằm phục vụ cho các ứng dụng để xử lý ngôn ngữ tự nhiên và có tập dữ liệu huấn luyện tương đối lớn để có thể huấn luyện một cách phù hợp nhất đối với các hiện tượng ngôn ngữ [9]. Phương pháp này thường được dựa trên thuật toán của các mô hình xác xuất (probabilistic – models) lý thuyết thông tin (information theory) và đại số tuyến tính (linear algeba). Một cách tiếp cận cơ bản thường được sử dụng trong trường hợp này đó dính là sử dụng Term Frequency – Inverse Document Frequency (TF – IDF) hay phân cụm.

Có rất nhiều các ví dụ điển hình có thể kể tới về việc áp dụng phương pháp tiếp cận dựa trên dữ liệu để tiến hành trích chọn các thông tin và sự kiện. Năm 2009 Okamoto và cộng sự [9] dựng một khung (framework) để phát hiện ra được những sự kiện cục bộ (local event). Trong nghiên cứu tác giả đã sử dụng các kĩ thuật phân cụm phân cấp. Trong khi đó phân cụm có thể sinh ra được những kết quả một cách thuận lợi nhất nhằm để phục vụ việc trích chọn các sự kiện. Liu M và các cộng sự [10] kết hợp các đồ thị trong số vô hướng chia đôi và phân cụm để có thể trích chọn các thực thể chính và các sự kiện có ý nghĩa từ các thông tin hằng ngày. Các kỹ thuật phân cụm cũng được Tanev và cộng sự [13] sử dụng để trích chọn các sự kiện bạo lực và thảm họa cho hệ thống giám sát.

Các tiếp cận dựa vào dữ liệu (data – driven) không yêu cầu người tạo dựng cần nắm được các kiến thức về ngôn ngữ và chuyên gia miền. Tuy nhiên phương pháp này đòi hỏi một lượng dữ liệu tương đối lớn để có thể dùng nó làm tập huấn luyện. Phương pháp tiếp cận dữ liệu cần xây dựng xác suất để xấp xỉ mô hình huấn luyện đối với dữ liệu. Phương pháp này có những ưu điểm như cách tiếp cận không nhất thiết phải có sự can thiệp của đội ngũ chuyên gia về ngôn ngữ cũng như các chuyên gia miền. Tiếp đó là các mô hình sau khi huấn luyện cũng có thể sử dụng với các miền dữ liệu khác nhau.

Tuy nhiên cũng còn tồn tại một số các nhược điểm cần phải khắc phục. Đầu tiên trong các bài toán trích chọn thông tin sự kiện phương pháp tiếp cận dựa trên dữ liệu không thể giải quyết vấn đề có liên quan tới ngữ nghĩa (ví dụ phương pháp này chỉ phát hiện các quan hệ có trong tập dữ liệu có sẵn mà không thể giải quyết được vấn đề ngữ nghĩa). Tiếp theo là phương pháp này cũng đòi hỏi yêu cầu phải có một bộ dữ liệu lớn để huấn luyện được mô hình. Đối với một số trường hợp chi phí và thời gian còn tốn kém. Bên cạnh đó do phương pháp tiếp cận dựa trên dữ liệu được xây dựng dựa trên các mô hình xác xuất thống kê do đó trong một số các trường hợp nếu quá trình xây dựng bộ huấn luyện không chính xác sẽ dẫn tới kết quả của quá trình trích chọn không được hiệu quả.

# 2.3 Phương pháp tiếp cận kết hợp luật và học máy

Phương pháp kết hợp hai yếu tố luật và học máy (lai – hybird) thường được sử dụng trong các bài toán trích chọn thông tin và sự kiện. Hầu hết các hệ thống dựa trên tri thức (knowledge – driven) được bổ sung bởi các phương thức dựa vào dữ liệu (data – driven) chính vì vậy có thể giải quyết được các khuyêt điểm của phương pháp dựa trên tri thức. Ví dụ Pisk và cộng sự [12] đã sử dụng các kỹ thuật bootstrapping trên hệ thống trích chọn sự kiện liên quan tới các hành vi bạo lực từ các bản tin trực tuyến đối với độ chính xác và độ hồi tưởng cao.

Morik [8] kết hợp các luật ngữ nghĩa với Conditional Random Fields được biểu diễn như đồ thị vô hướng để trích chọn các sự kiện từ phiên họp toàn thể của nghị viện Đức. Ở đây tác giả đã tiến hành giải quyết được những hạn chế của thuật toán học có giám sát đối với các cụm. Lee và cộng sự [8] sử dụng ontology mờ (ontology – based – fuzzy) để trích chọn sự kiện có trong các bảng tin tiếng Trung Quốc tác giả đã sử dụng thống kê trên ngữ pháp (gramma-based statistical) và gán nhãn từ loại (part of speech tagging). Chun và các cộng sự [3] trích chọn các sự kiện y sinh bằng các sử dụng các luật cú pháp để kết hợp với đồng tham chiếu. Như vậy có thể coi phương pháp này là một phương pháp lai, hội tụ những điểm mạnh đến từ hai phương pháp còn lại.

# 2.4. Phương pháp Khai phá nội dung Web

Khai phá nội dung của web chủ yếu tập trung vào nội dung chứa bên trong các trang web các nội dung này thông thường bao gồm các dạng như: hình ảnh, âm thanh, văn bản…Khi tiến hành khai phá web thì khai phá nội dung được coi như là kỹ thuật khai phá dữ liệu đối với CSDL quan hệ vì nó có thể nhận biết và phân loại các kiểu tương tự của tri thức có trong kho dữ liệu không cấu trúc trong các tài liệu Web. Đa phần các tài liệu Web là nửa cấu trúc hoặc là dữ liệu có cấu trúc giống như các thành phần dữ liệu trong CSDL dùng để tạo ra các trang HTML, nhưng nhìn chung đa phần dữ liệu văn bản là dạng dữ liệu không có cấu trúc. Điều này đặt ra thách thức cho công tác khai phá nội dung Web với những nhiệm vụ phức tạp hơn.

## **2.4.1. Khai phá kết quả tìm kiếm**

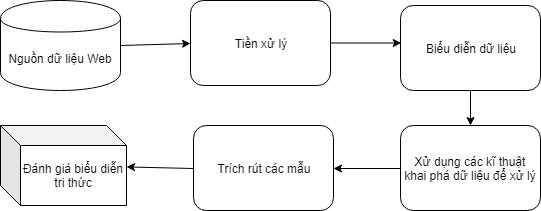
Hiện nay xu hướng đang là sử dụng một công cụ đó là Web Searching Engine để phục vụ công việc phân loại những tài liệu web một cách tự động. Công cụ này giúp chúng ta có thể đánh trọng số đối với các website. Việc đánh trọng số này được tiến hành qua các bước như sau:

* (1): Tải các dữ liệu web từ các website về;
* (2): Dùng công cụ Web Search Engine trích chọn ra những thông tin chỉ mục dùng để mô tả trang web đó đồng thời lưu trữ chúng cùng với URL của nó trong Web Search Engine;
* (3): Áp dụng những phương pháp KPDL để phân lớp một cách tự động và tiến hành phân loại các trang web một cách thuận lợi nhất, được tổ chức một cách chặt chẽ thông qua các cấu trúc siêu liên kết.

Trực quan hóa những kết quả tìm kiếm: Trong quá trình trích chọn và phân loại sẽ không tránh được các trường hợp lấy về cả các thông tin không liên quan tới nhau gây ảnh hưởng tới quá trình phân tích. Nếu chúng ta áp dụng kĩ thuật phân cụm để có thể phân tích và phân cụm được những kết quả tìm kiếm thì hiệu quả tìm kiếm sẽ được cải thiện một cách tốt và hiệu quả hơn rất nhiều. Nói đơn giản hơn, việc phân cụm tài liệu tìm kiếm được sẽ là nhóm các tài liệu “tương tự” nhau về nội dung sau đó sắp xếp chúng vào cùng một nhóm, đối với các tài liệu không tương tự nhau thì nhóm chúng lại vào một nhóm khác. Hoặc một cách khác nữa là chúng ta có thể tiến hành phân loại các tài liêu đã tìm kiếm được theo những tiêu chí nhất định để người sử dụng có thể tìm kiếm tài liệu một cách nhanh chóng và hiệu quả hơn do chúng được sắp xếp một cách khoa học và hợp lý.

## **2.4.2. Khai phá văn bản Web**

Khai phá văn bản chính là việc áp dụng những kĩ thuật KPDL trên các tập văn bản đã thu thập để phát hiện ra được những tri thức nằm bên trong đó. Chính vì vậy KPVB có liên quan mật thiết tới các kĩ thuật KPDL, tìm kiếm các thông tin và xử lý các ngôn ngữ tự nhiên. Nhìn chung các đối tượng của việc KPVB bao gồm: Các loại dữ liệu có cấu trúc, bán cấu trúc và dữ liệu không có cấu trúc. Kết quả của quá trình KPVB chính là việc phân loại các mục đích của tài liệu văn bản để phục vụ cho một nhu cầu cụ thể nào đó, tìm hiểu các trạng thái chung của mỗi tài liệu. Quy trình làm việc của KPVB được mô tả bằng hình như sau:



### *Hình 2.2: Quá trình khai phá văn bản Web*

### **2.4.2.1. Nguồn dữ liệu**

Nguồn dữ liệu web được định nghĩa là văn bản cục bộ có ở trên website đã được định dạng, nguồn dữ liệu này được tích hợp để tạo nên các tài liệu phù hợp với nhu cầu mong muốn để phục vụ việc khai phá và tiến hành phân phối trong nhiều dịch vụ Web thông qua việc ứng dụng các kỹ thuật truy xuất thông tin.

### **2.4.2.2. Tiền xử lý dữ liệu**

Tiền xử lý dữ liệu là công việc vô cùng quan trọng trước khi tiến hành các bước phân tích và khai phá. Đây là quá trình áp dụng những quy tắc hoặc kĩ thuật để thể hiện và làm rõ được dữ liệu. Cụ thể đây là công việc làm cho dữ liệu được rõ ràng và rành mạch hơn, loại bỏ đi những phần dữ liệu dư thừa, dữ liệu “rác” không có giá trị để quá trình khai phá được diễn ra một cách thuận tiện và dễ dàng hơn. Nhìn chung các bước cơ bản trong qua trình tiền xử lý dữ liệu là:

- (1): Tìm hiểu và xác định rõ được những nhu cầu của người dùng và xác định được những mối liên hệ giữa các tri thức đến từ các đối tượng dữ liệu;

- (2): Tiến hành chuẩn hóa và sắp xếp lại những tri thức này;

- (3): Kết quả đầu ra cuối cùng phải đảm bảo một số các tiêu chí như: Dữ liệu phải được thống nhất, giảm thiểu số chiều và tăng hiệu quả của việc phát hiện tri thức thông qua quá trình chuyển đổi, cưỡng bức dữ liệu, xóa bỏ những thuộc tính gây nhiễu và không liên quan để tránh việc có quá nhiều số chiều của dữ liệu còn tồn tại.

### **2.4.2.3. Biểu diễn văn bản**

Bản chất của KPVB Web là khai phá các tập tài liệu HTML. Chính vì vậy ta sẽ phải thực hiện công việc biến đổi và biểu diễn dữ liệu một cách hợp lý nhất để phục vụ cho quá trình phân tích và xử lý. Một cách mà chúng ta có thể sử dụng là lưu trữ chúng ở bên trong mảng 2 chiều nơi mà dữ liệu đó có thể phản ảnh được những nét đặc trưng nhất của tài liệu. Thông thường người ta sử dụng mô hình TF-IDF để nhằm vector hóa dữ liệu nhưng có một vấn đề xảy ra đó là việc sử dụng mô hình này sẽ làm cho số chiều của vector trở nên khá lớn. Vì thế chúng ta cần phải lựa chọn các đặc trưng một cách vô cùng cần thận và phải đảm bảo được việc nó sẽ trở thành khóa và nó có ảnh hưởng một cách trực tiếp tới quá trình KPVB.

Loại bỏ từ dừng và phân lớp: Chúng ta sử dụng định luật Zip để tiến hành xóa bỏ các từ có tần số xuất hiện và lặp lại cao (ví dụ như: và, hay, hoặc) những từ chứa đựng các hàm lượng ít thông tin và loại bỏ các từ ngữ có tần số xuất hiện thấp để giảm thiểu nhiễu. Bước tiếp theo là tiến hành chọn lọc trong các từ còn lại và thực hiện quá trình phân lớp chúng để mô tả các đặc trưng của tài liệu.

### **2.4.2.4. Trích rút ra các từ đặc trưng**

Việc rút ra đặc trưng là một phương pháp phổ biến. Trong phương pháp này có thể dùng để giải quyết được số chiều của vector đặc trưng lớn được mang lại thông qua kỹ thuật KPVB. Việc rút trích ra những đặc trưng dựa trên hàm trọng số.

Mỗi một từ ngữ đặc trưng sẽ nhận được một giá trị trọng số đáng tin cậy thông qua việc tính toán hàm trọng số tin cậy đó. Tần số xuất hiện cao của các từ mang đặc trưng chính là khả năng chắc chắn nó sẽ phản ánh được bao quát chủ đề của văn bản, tiếp theo ta sẽ gán cho nó một giá trị tin cậy lớn hơn. Bên cạnh đó nếu nó là tiêu đề, từ khóa hoặc phân cụm thì nó sẽ có giá trị tin cậy lớn hơn. Mỗi đặc trưng sẽ được lưu lại để có thể tiến hành xử lý một cách hiệu quả. Sau đó sẽ lựa chọn kích thước tương ứng của tập đặc trưng (kích thước phải nhận được từ thực nghiệm)

Việc rút ra các đặc trưng dựa trên việc phân tích các thành phần cốt lõi trong phân tích thống kê. Ý tưởng chính của phương pháp này là sử dụng thay thế từ đặc trưng bao hàm đối với một số ít các từ đặc trưng thể hiện nay trong phần mô tả để có thể giảm thiểu được số chiều. Bên cạnh đó ta cũng sử dụng phương pháp quy nạp thuộc tích dữ liệu để có thể giảm đi được số chiều của vector bằng cách tổng hợp nhiều dữ liệu thành một mức cao.

### **2.4.2.5. Khai phá dữ liệu văn bản**

Sau khi đã tiến hành các thao tác tập hợp, lựa chọn và rút trích ra được tập văn bản để hình thành nên các đặc trưng cơ bản, đây là cơ sở vững chắc để KPDL. Từ đây chúng ta có thể dễ dàng thực hiện thao tác phân loại, phân cụm cũng như nghiên cứu thêm về việc phân tích và dự đoán.

* *Trích rút văn bản:*

Việc trích rút văn bản có mục đích chính là để tìm ra được các từ ngữ quan trọng, có ý nghĩa chính bao quát tóm tắt tài liệu văn bản trong quá trình tổng hợp. Tiếp theo đó là việc những người sử dụng có thể nắm bắt được một cách nhanh chóng và hiệu quả những ý nghĩa chính của văn bản đang đọc mà không nhất thiết phải xem toàn bộ nội dung của văn bản. Phương pháp này thường được sử dụng bên trong searching engine dùng để thể hiện ra các văn bản trích dẫn. Công cụ này luôn chỉ ra được cho người dùng những câu mang tính chất dự báo trong khi tiến hành tìm kiếm và trả về kết quả. Một trong những cách hữu hiệu nhất để nắm bắt và thu thập được ý nghĩa của văn bản đó đính là thông qua việc sử dụng hệ thống các thuật toán khác nhau. Dựa vào đó thì quá trình tìm kiếm sẽ đem lại hiệu quả một cách tốt hơn và phù hợp với những mong muốn và sự lựa chọn của người tiêu dùng.

* *Phân lớp văn bản*

Đầu tiên rất nhiều các dạng tài liệu khách nhau được tiến hành phân lớp theo một cách vô cùng nhanh chóng mà mang lại hiệu quả cao. Thứ hai, đối với từng lớp văn bản sẽ được đưa vào một số các chủ đề khác nhau với mục đích để tạo sự phù hợp. Chính vì vậy nó rất thích hợp đối với công việc tìm kiếm và duyệt qua các tài liệu của Web thông qua người sử dụng. Thông thường có 2 phương pháp được sử dụng để phân lớp đó là thuật toán phân lớp Navie Bayesian và K-Nearest Neighbor để thực thi công việc khai phá thông tin văn bản. Đối với phân lớp văn bản nhiệm vụ đầu tiên là chúng ta phải phân loại được các tài liệu, công việc thứ hai là phải xác định được các đặc trưng thông qua số lượng các đặc trưng của tập tài liệu đã được phân lớp bằng cách sử dụng bất kì một thuật toán nào đó. Khi đó đối với các tài liệu có tính chất và độ tương tự cao sẽ nằm cùng với nhau trong một lớp. Độ tương tự sẽ được đo bằng các hàm đánh giá xác định trước, trong trường hợp số lượng tài liệu tương tự ít thì nó sẽ trở về số 0 trong trường hợp ngược lại nếu nó không giống như sự lựa chọn của phân lớp đã xác định trước thì trường hợp này được coi như là không phù hợp và chúng ta phải lựa chọn lại phân lớp. Đối với việc lựa chọn có hai giai đoạn đó là : Huấn luyện và phân lớp.

*Thuật toán phân cụm phân cấp*

* Trong tài liệu xác định;
* Chọn ngẫu nhiên 2 cụm;
* Lặp lại công việc.

Sau quá trình của phương pháp sắp xếp các liên kết thì sẽ có một cây được tạo thành, cây này phản ánh một cách rõ ràng và chi tiết mối quan hệ với nhau về độ tương tự của các tài liệu. Phương pháp này đem lại sự chính xác là khá cao tuy nhiên bất lợi của nó chính là tốc độ xử lý tương đối chậm nếu để đem ra so sánh độ tương tự có trong tất cả các cụm. Trong trường hợp mà tập tài liệu này có kích thước quá lớn thì chúng ta không thể sử dụng phương pháp này được

*Thuật toán phân cụm phân hoạch*

* Việc đầu tiên phải làm đó là tiến hành chia tập tài liệu của chúng ta thành các cụm khởi đầu dựa và việc tối ưu hóa đánh giá theo một nguyên tắc nhất định;
* Công việc trên được tiến hành liên tục và lặp lại đến khi nào tất cả các tài liệu đã được đưa vào bên trong các cụm xác định;

Đối với phương pháp này thì điểm mạnh là kết quả phân cụm sẽ được tiến hành một cách nhanh chóng với sự ổn định tương đối cao. Nhưng chúng ta phải xác định trước đó đâu là các phần tử khởi đầu và bên cạnh đó là số lượng các phần tử này, điều này sẽ ảnh hưởng một cách trực tiếp tới hiệu quả của công việc phân cụm.

* *Phân tích và dự đoán các xu hướng*

Việc phân tích các tài liệu web mang lại một lợi thế vô cùng lớn đó chính là việc chúng ta có thể nhìn nhận và phân tích được quan hệ phân phối đối với bộ dữ liệu thông qua mỗi giai đoạn khác nhau. Hơn nữa chúng ta cũng có thể tiến hành dự đoán được sự phát triển trong tương lai.

* *Đánh giá chất lượng mẫu*

KPDL Web được coi như là một quá trình nhỏ của học máy. Trong trường hợp này kết quả đầu ra của việc học máy chính là các biểu mẫu tri thức. Một trong những yếu tố quan trọng của học máy đó chính là đánh giá kết quả mẫu. Các tài liệu sẽ được tiến hành phân tích và áp dụng phương pháp phân lớp để đưa vào các tập huấn luyện, sau đó sẽ được kiểm tra. Việc học sẽ được lặp lại và tập huấn luyện, tập kiểm tra sẽ được kiểm thử. Bước cuối cùng là dùng chất lượng trung bình để đánh giá.

# 2.5. Phương pháp Khai phá cấu trúc Web

WWW được định nghĩa hệ thống thông tin toàn cầu, bên trong nó chứa đựng tất cả các website. Mỗi trang có thể được liên kết tới nhiều các trang khác nhau. Các siêu liên kết thay đổi chứa đựng ngữ nghĩa chung chủ đề của các trang. Mỗi khi có một siêu liên kết trỏ tới bất kì một trang web nào khác thì có thể được xem như là một chứng thực của trang web đó. Việc này tạo điều kiện thuận lợi và vô cùng hữu ích khi tiến hành thu thập những thông tin quan trọng dựa vào việc tìm ra và phân tích mối liên kết giữa các trang web. Sử dụng các thuật toán khai phá người sử dụng để phục vụ công việc trích xuất ra những thông tin hữu ích thông qua cấu trúc web, bên cạnh đó là tăng khả năng tìm ra những trang web có tính chất quan trọng và phát triển kế hoạch cụ thể để có thể tạo dựng nên một trang web phù hợp với nhu cầu của người sử dụng.

Mấu chốt của việc khai phá cấu trúc web là để có thể phát hiện thông tin về cấu trúc của web. Nếu như việc khai phá nội dung Web đi sâu vào phân tích và khám phá cấu trúc bên trong của tài liệu thì khai phá cấu trúc Web lại cố gắng để phát hiện cấu trúc liên kết của các siêu liên kết trong tài liệu. Các trang web sẽ được phân loại bằng hình thức khai phá cấu trúc web thông qua các mô hình hình học của siêu liên kết, tạo ra các thông tin như độ tương tự và các mối quan hệ giữa các website khác nhau. Trong trường hợp các trang web được liên kết một cách trực tiếp với nhau thì chúng ta sẽ tìm hiểu mối quan hệ giữa các trang web này. Những trang web này có thể thuộc dịch vụ web tương tự nhau do vậy chúng có có mối tương đồng với nhau về mặt nội dung bởi được tạo ra bởi cùng một người. Một số nhiệm vụ khác của công việc khai phá cấu trúc web đó là khám phá sự phân cấp tự nhiên hoặc mạng lưới của các siêu liên kết trong các website của một miền đặc biệt. Điều này có thể giúp tạo ra được những luồng thông tin trong website mà có thể đại diện cho nhiều miền đặc biệt, điều này giúp cho quá trình xử lý được dễ dàng và thuận tiện hơn.

## **2.5.1. Tiêu chuẩn đánh giá độ tương tự**

Khám phá ra một nhóm các trang web có cấu trúc giống nhau để khai phá, ta chỉ ra sự giống nhau của hai nút theo một tiêu chuẩn nào đó:

* Tiêu chuẩn 1: Đối với mỗi trang web d1 và d2. Ta nói d1 và d2 quan hệ với nhau nếu có một liên kết tới từ d1 đến d2 hoặc ngược lại;
* Tiêu chuẩn 2: Đồng trích dẫn: Độ tương tự giữa d1 và d1, được đo bởi số trang dẫn tới cả d1 và d2;
* Tương chỉ mục: Độ tương tự giữa d1 và d2, được đo bằng số trang mà cả d1 và d2 đều trở tới.

## **2.5.2. Khai phá và quản lý cộng đồng web**

Cộng đồng web được định nghĩa là một nhóm tập hợp các trang web cùng chia sẻ những mối quan tâm chung mà người sử dụng thường chú ý tới. Ở bên trong một cộng đồng web, các thành viên của cộng đồng có thể không hề biết được thực tế tình trạng đang tồn tại và diễn ra Nhận biết và hiểu được sự phát triển cũng như các đặc thù riêng biệt của cộng đồng web là một công việc vô cùng quan trọng. Nhìn chung cộng đồng Web thường có các đặc điểm sau:

* Các trang Web ở trong cùng một cồng đồng sẽ có những điểm tương tự nhau hơn khi so sánh với các trang web nằm bên ngoài cộng đồng;
* Các cụm trang web sẽ được tạo thành từ một cộng đồng web;
* Các cộng đồng web được xác định một các rõ ràng và cụ thể, được hầu hết tất cả mọi người biết đến.

# 2.6. Kết luận chương

Trong chương này đã trình bày một số các phương pháp tiếp cận khai phá dữ liệu và trích chọn thông tin. Mỗi phương pháp lại có những đặc trưng riêng biệt cũng như những lợi thế và hạn chế đối với từng phương pháp. Trong thực tế thì có thể sử dụng các phương pháp một cách linh hoạt để đưa ra được kết quả tốt nhất. Những phương pháp được sử dụng cho bài toán trích chọn sự kiện được trình bày ở chương tiếp theo

# CHƯƠNG 3: ĐỀ XUẤT MÔ HÌNH TRÍCH CHỌN SỰ KIỆN Y TẾ

Trong chương này, khóa luận sẽ tiến hành tập trung phân tích để có thể làm rõ được bài toán trích chọn những thông tin và sự kiện thuộc lĩnh vực y tế. Các công đoạn sẽ bao gồm việc tìm hiểu những thuộc tính đặc trưng của một sự kiện y tế, phát biểu bài toán, đề xuất mô hình để thực hiện bên cạnh đó là giải quyết chi tiết hai bài toán vô cùng quan trọng của khóa luận đó là phát hiện ra sự kiện y tế và bài toán trích chọn thông tin sự kiện y tế.

# 3.1 Phát biểu bài toán

Bài toán được đặt ra ở đây là việc trích chọn những thông tin và sự kiện y tế có trong các trang web Tiếng Việt được người sử dụng tiếp cận và truy cập thường xuyên để có thể tìm hiểu và nắm bắt được các tin tức mới nhất. Trong khuôn khổ nội dung khóa luận này, sẽ tiến hành tập trung vào việc giải quyết xử lý bài toán ***trích chọn sự kiện có ở trong các bản tin về y tế trên các website Tiếng Việt*.** Ở đây sẽ chỉ đề cập tới việc trích chọn sự kiện y tế. Chính vì vậy, cần làm rõ và nhấn mạnh sự phân biệt giữa thông tin y tế nói chung và thông tin sự kiện y tế để không bị nhầm lẫn giữa hai khái niệm này.

Cụ thể, bài toán trích chọn thông tin sự kiện y tế thông qua các trang web Tiếng Việt sẽ lấy các bản tin, bài viết từ các trang báo điện tử Việt Nam. Từ đó có thể trích chọn được ra những sự kiện y tế mà hiện tại đang được xã hội giành sự quan tâm. Những thông tin được trích chọn ra sẽ là thời gian ghi nhận xảy ra dịch bệnh, loại dịch bệnh, địa bàn nơi xảy ra dịch bệnh và số ca mắc…từ các văn bản phi cấu trúc. Bài toán được diễn đạt một cách cụ thể như sau:

**Đầu vào**: Bản tin, bài viết có trên các website báo điện tử

**Đầu ra:** Xác định, phân loại những bản tin có chứa sự kiện y tế và trích chọn ra những thông tin đặc trưng của sự kiện y tế đó.

Thông tin thu được nhờ khai phá dữ liệu sẽ được chứa trong một bản tin sự kiện y tế được định nghĩa là một bộ A bao gồm 4 thành phần chính đó là: Thời gian, địa bàn dịch, loại dịch bệnh, số lượng ca mắc. Có thể nói cách khác là tập A được định nghĩa như sau:

A = <thời gian, địa bàn, số trường hợp mắc, loại dịch>

Trong đó:

Thời gian: Là thời gian ghi nhận dịch bệnh bùng phát

Địa bản: Là khu vực (địa phương, tình/thành phố) có dịch bệnh bùng phát

Loại dịch bệnh**:** Thông tin được trích ra từ các bài viết là tên của bệnh dịch.

Số ca mắc: là thống kê số người mắc bệnh, số ca tử vong do bệnh dịch. Số lượng thiệt hại có thể là danh sách gồm hai trường số là số ca mắc bệnh và số lượng tử vong. Một ví dụ như có thông tin sau: “Sở Y tế Hà Nội vừa cho hay từ ngày 23 đến 30/10 trên địa bàn thành phố Hà Nội đã có thêm 11 ca sốt phát bang dạng sởi, nâng số ca mắc bệnh tính từ đầu năm lên con số 168 trường hợp trong đó có 4 trường hợp tử vong” Từ ví dụ trên, thông tin liên quan tới con số thương vong được thống kê ra dưới dạng một bảng danh sách như sau:

|  |  |
| --- | --- |
| **Số ca mắc** | **Số người tử vong** |
| 168 | 4 |

Xét một ví dụ khác với thông tin được trích ra từ một bài báo như sau: “Mới đây một ca mắc sốt xuất huyết đã tử vong tại phường Hiếp Thành quận 12. Đây là ca mắc đầu tiên trong năm nay tại thành phố Hồ Chí Minh. Trước đó con số mắc bệnh trên địa bàn thành phố vào năm ngoái là 20.000 ca mắc”.

|  |  |
| --- | --- |
| **Số ca mắc** | **Số người tử vong** |
| 20.000 | 1 |

Mặt khác, đối với một thông tin về sự kiện y tế của A như sau: A = <Ngày 24/4, 30 ca nhập viện do bị sởi trên địa bàn quận 1 thành phố Hồ Chí Minh>. Thông qua 4 thông tin cơ bản này, chúng ta có thể suy luận ra được thông tin rằng “Vào ngày 24/4 có 30 trường hợp phải nhập viện do mắc sởi trên địa bàn quận 1 thành phố Hồ Chí Minh”.

Đối với bài toán định nghĩa thì đầu vào của mô hình chính là các bản tin, những bài báo được có trên các website báo điện tử. Lý do chọn nguồn để lấy các thông tin ở đây chính là yếu tố an toàn và đảm bảo tính chính xác của thông tin. Thứ nhất các thông tin trên các trang web này là vô cùng đa dạng và phong phú với nhiều các khía cạnh có thể tiến hành phân tích. Lý do thứ hai chính là sự đảm bảo về độ tin cậy của những thông tin này là tương đối cao và kèm theo đó là tính cập nhập thường xuyên và liên tục, đáp ứng sự thay đổi một cách nhanh chóng mà mạnh mẽ hằng ngày. Hơn nữa lý do chọn lựa những báo điện từ này vì nhìn chung quá trình thu thập các thông tin ở đây cũng tương đối đơn giản và dễ dàng, ít gặp phải những khó khăn vướng mắc trong quá trình thu thập thông tin. Chính vì vậy nguồn dữ liệu luôn được đảm bảo tính đa dạng và tính cập nhập liên tục.

Đối với mô hình trong phần phát biểu bài toán lại được tiến hành chi nhỏ ra thành hai bài toán con như sau:

**- Bài toàn thứ nhất (Pha 1):** Phát hiện sự kiện y tế, bài toán ở pha này giải quyết vấn đề là tìm hiểu và phát hiện ra được một bài báo trên một website có chứa các thông tin về các sự kiện y tế hay không.

**- Bài toán thứ hai (Pha 2):** Trích chọn sự kiện y tế. Đối với bài toán ở Pha 2 sẽ đi giải quyết những vấn đề sau khi đã tiến hành việc kiểm tra xác thực đối với những thông tin có trong Pha 1. Pha 2 sẽ thực hiện công việc đó là trích chọn những thông tin về sự kiện y tế đó.

# 3.2 Giải bài toán trích chọn sự kiện y tế

## **3.2.1 Phương pháp đề xuất**

Ở chương 2, khóa luận đã giới thiệu tới việc sử dụng một số các phương pháp tiếp cận thông qua việc khai phá dữ liệu và các phương pháp trích chọn thông tin sự kiện để giải quyết bài toán đã được đặt ra là trích chọn sự kiện y tế. Bài toán sẽ được phân tích cụ thể hóa như sau:

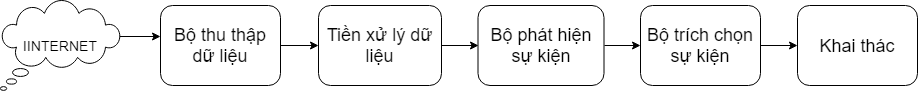
**Pha 1 - Phát hiện sự kiện y tế:** Ở pha đầu tiên này thì dữ liệu đầu vào là các bản tin có chủ đề liên quan tới lĩnh vực y khoa có ở trên các trang web báo điện tử. Số lượng của các bài báo này là vô cùng nhiều và đa dạng, thuộc nhiều khía cạnh khác nhau của lĩnh vực y tế. Chính vì vậy đối với pha 1 có thể được phân tích thành hai bước như sau:

* Bước 1: Sử dụng các luật để có thể lọc ra dữ liệu trong **miền sự kiện y tế**
* Bước 2: Dùng bộ lọc để nhận diện các bản tin có chứa những **sự kiện y tế**

**Pha 2 – Trích chọn sự kiện y tế:** Đối với pha này thì chúng ta phải tiến hành công việc trích chọn ra những thông tin về thời gian ghi nhận có xảy ra dịch bệnh, số lượng các ca mắc bệnh hoặc số lượng các ca phải nhập viện để tiến hành điều trị. Tiếp theo đó là loại dịch bệnh đang bùng phát và cuối cùng địa bàn nơi xảy ra dịch bệnh. Trích chọn thông tin được cụ thể hóa như sau:

* ***Trích chọn thông tin về thời gian:*** Thông tin có thể được ghi nhận ở dạng chuẩn (dd/mm/yyyy) hoặc ở dạng không chuẩn có thể ví dụ như: Trong 3 tháng đầu năm, nửa cuối năm,…nên ta dùng luật để có thể trích chọn ra các thông tin;
* ***Trích chọn thông tin về số ca mắc bệnh***: Sử dụng phương pháp nhận dạng thực thể và luật để có thể lọc ra thông tin;
* ***Trích chọn thông tin về loại dịch bệnh:*** Đối với các thông tin về mỗi loại dịch bệnh thì sẽ tiến hành dựng lên một bộ từ điển các loại dịch bệnh phổ biến hiện nay sau đó sử dụng luật để có thể khớp lại với bộ từ điển;
* ***Trích chọn thông tin về địa điểm xảy ra dịch bệnh:*** Dùng nhận dạng thực thể (NER) và sử dụng từ điển để có thể trích chọn được thông tin.

Nhằm giải quyết được một cách tốt nhất những vấn đề ở hai pha được nói tới ở phần trên, trong khóa luận này có đề xuất một mô hình để phát hiện và trích chọn sự kiện y tế. Mô hình này bao gồm 4 thành phần chính sau:



### *Hình 3.1: Quá trình phát hiện và trích chọn sự kiện y tế*

* (1) Thu thập dữ liệu: Thành phần này có nhiệm vụ tìm kiếm và thu thập dữ liệu một cách tự động từ các bài viết, bản tin có trên các báo điện tử trên môi trường Internet. Sau đó sẽ chuyển cho bước tiếp theo.
* (2) Tiền xử lý dữ liệu: Đây là công đoạn để xử lý dữ liệu đã được thu thập ở bước đầu tiên ở trên. Trong bước này chúng ta sẽ tiến hành loại bỏ đi những thẻ HTML, trích ra được những dữ liệu dạng thô (text) Sau đó chuyển sang bước phát hiện sự kiện.
* (3) Phát hiện sự kiện y tế: Là phát hiện ra được sự kiện y tế, dữ liệu được thu thập từ thành phần tiền xử lý dữ liệu, ta sử dụng các luật để có thể lấy ra các dữ liệu thuộc miền thông tin có liên quan tới lĩnh vực y tế, sau đó kiểm tra lại xem dữ liệu đó có phải là bản tin về các sự kiện y tế hay không. Trong trường hợp những thông tin này không thuộc vào thành phần những sự kiện y tế thì sẽ loại bỏ, nếu đúng thì lấy và chuyển dữ liệu đó sang cho bước trích chọn sự kiện y tế ở bước tiếp theo.
* (4) Trích chọn sự kiện: Là bước trích chọn ra sự kiện y tế. Đối với công đoạn cuối cùng trong mô hình này thì chúng ta phài làm công việc đó là trích chọn ra những thông tin đặc trưng thuộc vào các thành phần đã được thống nhất từ trước ví dụ như: Thời gian, số lượng, địa điểm, loại dịch bệnh, số người tử vong.

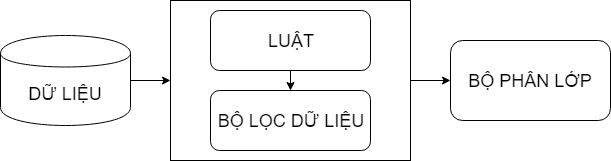
## **3.2.2.Thực hiện giải quyết bài toán phát hiện sự kiện và bài toán trích chọn sự kiện y tế**

Nhiệm vụ của bài toán thứ nhất. Từ bộ dữ liệu thô ban đầu đang ở dạng text ở công đoạn tiền xử lý dùng luật để lọc lấy dữ liệu trong miền thông tin liên quan tới y tế. Từ đó dùng bộ phân lớp để kiểm tra dữ liệu có xem đây có phải là dữ liệu thuộc các bản tin về y tế hay không. Trong trường hợp dữ liệu được tìm kiếm đúng là dữ liệu về bản tin về sự kiện y thế thì được chuyển sang bước 2 là Trích chọn sự kiện y tế.

# 3.2.2.1.Bài toán thứ nhất (Pha 1) – Phát hiện sự kiện y tế

Mục tiêu cốt lõi của pha 1 – Phát hiện sự kiện y tế, dữ liệu được sử dụng để phục vụ cho việc giải quyết được lấy từ bước tiền xử lý dữ liệu (Dữ liệu đang có dưới dạng thô – text). Dữ liệu đầu ra có chứa những sự kiện y tế hay không. Nói một cách hình thức thì bài toán được diễn giải như sau:

* **Đầu vào:** Một bản tin có trên các trang báo mạng và có dạng thô;
* **Đầu ra:** Xác định bản tin này có chứa sự kiện y tế hay không?



### *Hình 3.2: Thành phần phát hiện sự kiện*

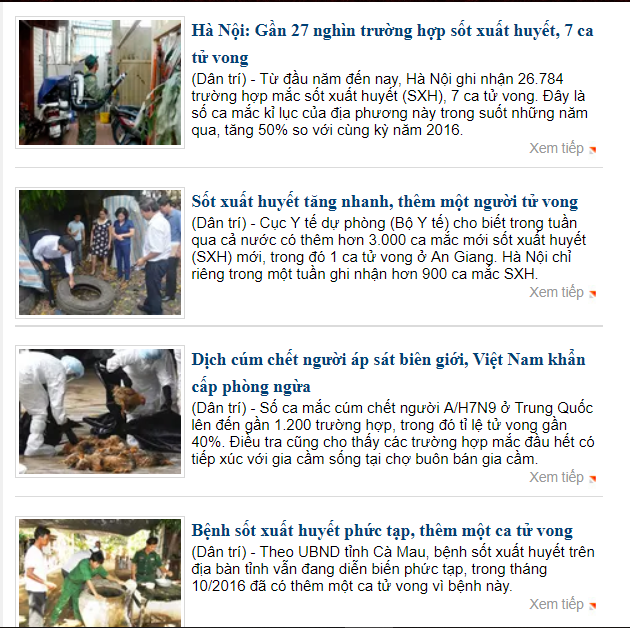
Trong Pha 1 này bao gồm hai thành phần chức năng: Thành phần thứ nhất chính là một bộ lọc dùng để lọc dữ liệu và thành phần thứ hai là một bộ phân lớp. Bộ lọc có chức năng chính là dùng để lọc dữ liệu từ bước tiền xử lý (dữ liệu dạng thô sau khi được lọc thẻ HTML từ bản tin thu thập được từ các trang báo điện tử), các bản tin có trong miền sự kiện y tế. Bên cạnh đó đối với chức năng phân lớp để kiểm tra xem bản tin có được có chứa sự kiện y tế hay không? Quá trình phát hiện sự kiện y tế được trình bày trong hình 3.2

## **Xây dựng tập luật cho bộ lọc dữ liệu**

Như các nội dung đã được trình bày ở phần trên pha phát hiện sự kiện bao gồm hai thành phần đó là một bộ lọc dữ liệu (sử dụng một tập luật để có thể lọc ra các bản tin nằm trong miền sự kiện y tế), tiếp theo đó là thành phần chức năng phân lớp, thành phần này có nhiệm vụ đó là sẽ kiểm tra các dữ liệu có chứa sự kiện liên quan tới y tế hay không

Thông qua việc khảo sát đối với dữ liệu thực tế, thông thường hầu hết các tiêu đề và phần tóm tắt của các bài viết nói lên một cách tương đối đầy đủ nội dung của bản tin đó. Chính vì vậy có thể lọc dữ liệu thông qua tiêu đề của các bản tin thay vì việc phải lọc thông qua toàn bộ nội dung của bài báo. Điều này giúp tiếp kiệm được thời gian cũng như công sức. Tuy nhiên, đối với một số trường hợp thì tiêu đề lại không phản ảnh được đầy đủ được hết những nội dung có chứa bên trong. Nếu muốn tìm được dữ liệu chính xác cho việc nghiên cứu thì người sử dụng bắt buộc phải lấy thông tin ở phần nội dung chính mà không lấy được các thông tin ở phần tiêu đề. Chính vì vậy, đối với các trang web khác nhau thì cần phải phải xác định được rõ xem phần thông tin cần thu thập nằm ở đâu? Có chính xác và đầy đủ thông tin cần thiết hay không? Từ đó việc thu thập thông tin có thể đạt được hiệu quả cao.

Qua việc tiến hành khảo sát thì hầu như các tiêu đề đối với các bài viết thuộc miền sự kiện y tế thường có các từ ngữ có liên quan tới tên một số các dịch bệnh hiện đang phổ biến, có khả năng nguy hiểm cao gây ảnh hưởng lớn tới sức khỏe của cộng đồng. Ví dụ như “Tại thành phố Hồ Chí Minh phát hiện thêm 5 trường hợp mắc bệnh Zika trong cùng một ngày” hoặc như “Tại Quảng Nam bùng phát dịch sốt xuất huyết khiến nhiều người phải nhập viện để điều trị. Tuy nhiên một số các bản tin lại có tiêu đề không đề cập tới cụ thể tên dịch bệnh ví dụ như “Tại Hà Nội, số trường hợp phải nhập viện để điều trị các bệnh truyền nhiễm trong mùa hè có dấu hiệu gia tăng. Đối với ví dụ trên tuy nó không chứa cụ thể các thông tin về thời gian, dịch bệnh nhưng lại chứa các từ ngữ liên quan như: dịch bệnh, truyền nhiễm. Ví dụ cụ thể như hình dưới đây:



### *Hình 3.3: Tiêu đề bản tin có chứa tên các loại bệnh*

Thông qua quá trình khảo sát dữ liệu thực tế, khóa luận đã đề xuất một tập danh sách tên các bệnh thường gặp, dễ lây lan trong cộng đồng gây nguy hiểm tới sức khỏe của mọi người. Tập này được gọi là từ điển các loại dịch bệnh. Chi tiết các loại dịch bệnh được liệt kê ở dưới bảng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên bệnh dịch** | **STT** | **Tên bệnh dịch** |
| 1 | Sốt thấp khớp cấp | 39 | Leptospirosis |
| 2 | Bệnh than | 40 | Nhiễm trùng xuất huyết |
| 3 | Lyssavirus Úc | 41 | Bệnh sốt rét |
| 4 | Dịch cúm gia cầm | 42 | Bệnh Sởi |
| 5 | Babesiosis | 43 | Bệnh não mô cầu |
| 6 | Virus rừng Barmah | 44 | MERS |
| 7 | Nhọt và nhiễm trùng da | 45 | Quai bị |
| 8 | Bệnh ngộ độc | 46 | Viêm não thung lũng |
| 9 | Brucellosis | 47 | Đại dịch cúm |
| 10 | Chlamydia | 48 | Ho gà |
| 11 | Thủy đậu | 49 | Bệnh phế cầu |
| 12 | Chikungunya | 50 | Viêm đa cơ |
| 13 | Chlamydia | 51 | Bệnh vẩy nến |
| 14 | Dịch tả | 52 | Bệnh dại |
| 15 | Bệnh CJD | 53 | Giun phổi |
| 16 | CMV | 54 | Virus hợp bào hô hấp RSV |
| 17 | Sốt xuất huyết | 55 | Virus SARS |
| 18 | Dientamoeba Fragilis | 56 | Shigellosis |
| 19 | Bạch hầu | 57 | Tấm lợp |
| 20 | Donovanosis | 58 | Bệnh giang mai |
| 21 | Bệnh do vi rút do Ebola | 59 | Thương hàn |
| 22 | Sốt ruột | 60 | Typhus |
| 23 | Viêm giác mạc do dịch | 61 | Số vàng da |
| 24 | Vi rút EBV | 62 | Virus Zika |
| 25 | Bệnh do thực phẩm | 63 | Uốn ván |
| 26 | Viêm dạ dày ruột |  |  |
| 27 | Sởi Đức (Rubella) |  |  |
| 28 | Viêm gan |  |  |
| 29 | Bệnh lậu |  |  |
| 30 | Sốt tuyến |  |  |
| 31 | Hội chứng tan máu bẩm sinh |  |  |
| 32 | Heamophilusenzea Loại b (Hib) |  |  |
| 33 | Bệnh tay chân miệng |  |  |
| 34 | Virus Hendra |  |  |
| 35 | Viêm gan A/B/C/D/E |  |  |
| 36 | Virus HIV |  |  |
| 37 | Cúm |  |  |
| 38 | Bệnh viêm não Nhật Bản |  |  |

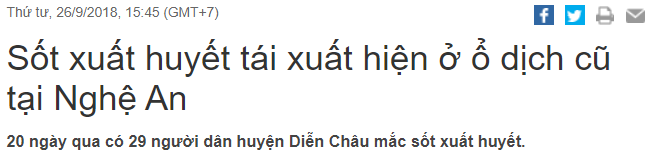
### *Bảng 3.1 Bảng danh sách các loại bệnh truyền nhiễm*

Từ đó tiến hành xây dựng luật cho hai trường hợp nói trên. Đối với trường hợp thứ nhất thì sử dụng mẫu 1, các tiêu đề bản tin so khớp với từ điển các bệnh dịch, nếu trùng khớp thì lọc ra. Trong trường hợp ngược lại thì sử dụng mẫu 2. Chi tiết về các mẫu như sau:

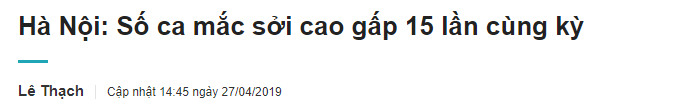
Mẫu 1 = “Tên loại dịch bệnh”

**Ví dụ minh họa cho mẫu 1:**

Từ tiêu đề “Sốt xuất huyết tái xuất hiện ở ổ dịch cũ tại Nghệ An”. Từ tiêu đề này chúng ta có thể tìm thấy từ “Sốt xuất huyết”



Hoặc một ví dụ khác như sau: “Hà Nội: Số ca mắc Sởi cao gấp 15 lần cùng kỳ”. Đối với tiêu đề này thì chúng ta có thể tìm thấy từ “Sởi”



Đối với một ví dụ khác như sau “Nắng nóng kéo dài: Dịch bệnh mùa hè gia tăng”. Đối với tiêu đề này thì nó không chứa các thông tin về dịch bệnh nên sẽ bỏ qua Mẫu 1 và chuyển sang sử dụng Mẫu 2 với cấu trúc như sau

Mẫu 2 = “Động từ” # “Danh từ”

Trong đó:

Động từ bao gồm các từ: Kéo dài, gia tăng…

Danh từ bao gồm các từ: Nắng nóng, dịch bệnh, mùa hè, cao…

## **Xây dựng mô hình phân lớp**

Đối với việc giải quyết bài toán thì bộ phân lớp có nhiệm vụ quan trọng đó là phát hiện ra một bài báo có chứa các sự kiện hay không. Bộ phân lớp sẽ được tiến hành phân ra làm hai lớp:

* Lớp chứa sự kiện y tế với nhãn của lớp là EVENT
* Lớp không chứa sự kiện y tế với nhãn của lớp là NOT\_EVENT

Dựa vào quá trình khảo sát có thể nhận thấy rằng hầu hết phần tiêu đề của bản tin gần như đã chứa đầy đủ nội dung chính của các bản tin. Chính vì vậy sẽ sử dụng các thông tin này để có thể xây dựng vector đặc trưng nhằm mục đích biểu diễn các đặc trưng có của văn bản. Các đặc trưng được sử dụng trong trình huấn luyện là 2-grams, 3-grams, 4-grams. Trong bài khóa luận này sẽ tiến hành xây dựng một tập để huấn luyện và dùng tập dữ liệu huấn luyện này nhằm mục đích xác định văn bản chứa sự kiện. Đối với công việc này thì chúng ta sẽ sử dụng một mô hình đó là mô hình Maximum Entropy (ME) bởi một số lý do sau:

(1) Dữ liệu được sử dụng trong quá trình huấn luyện là các dữ liệu ở dạng văn bản. Chính vì vậy khi biểu diễn thông qua dạng vector đặc trưng thì đây lại là những dữ liệu ở dạng dữ liệu thưa. Mà mô hình ME lại được sử dụng một cách tốt nhất khi dữ liệu ở dạng này

(2) Tốc độ huấn luyện của ME khá tốt và thực nghiệm cho thấy được rằng phương pháp sử dụng mô hình này đem lại kết quả khả quan đối với các dữ liệu dạng văn bản

(3) Do mã nguồn của ME là mã nguồn mở nên chúng ta có thể tùy biến đối với mã nguồn này

Mô hình ME dựa trên xác suất có điều kiện và nó cho phép việc tích hợp sự đa dạng của các đặc trung từ tập huấn luyện đối với một bài toán phân lớp. Ý tưởng của mô hình ME là mô hình phải xác định được một phân phối đều thỏa mãn các ràng buộc có trong tập dữ liệu huấn luyện mà không được phép tiến hành thêm bớt bất kì một giả định nào. Điều này đồng nghĩa với việc sự phân bố của mô hình phải thỏa mãn các ràng buộc của dữ liệu quan sát và càng gần với phân bố càng tốt.

Sau khi đã được tiến hành huấn luyện thì toàn bộ những dữ liệu của bộ lọc sẽ được đưa vào trong mô hình. Tại đây những văn bản có dán nhãn là EVENT sẽ được coi là đầu vào cho quá trình trích chọn. Trong trường hợp ngược lại thì những văn bản có nhãn NOT\_EVENT thì mô hình sẽ bỏ qua.

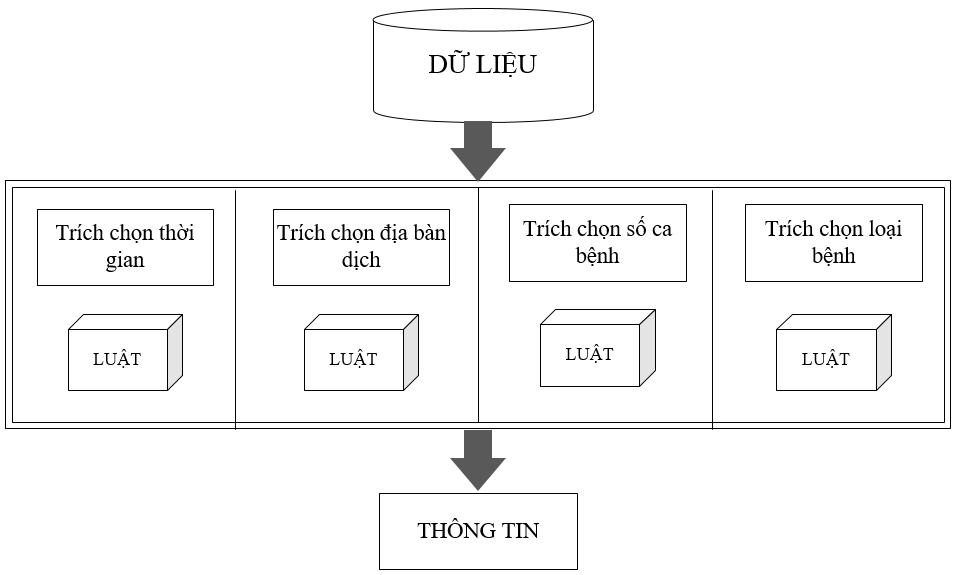
# 3.2.2.2. Bài toán thứ hai (Pha 2) – Trích chọn sự kiện y tế

Bộ trích chọn sự kiện được xem như là thành phần cốt lõi và quan trọng nhất của mô hình giải quyết bài toán. Nơi mà những thông tin của một sự kiện liên quan tới y tế được tiến hành trích chọn. Một cách hình thức thì bài toán trích chọn sự kiện có thể được phát biểu một cách như sau:

**Đầu vào:** Bản tin có chứa sự kiện y tế

**Đầu ra:** Các thông tin của một sự kiện y tế bao gồm: Thời gian ghi nhận, số ca bệnh, địa bàn dịch bệnh, loại dịch bệnh. Trong trường hợp này số ca mắc bệnh bao gồm số ca phải nhập viện điều trị và số ca tử vong do bệnh, vì vậy số ca mắc bệnh được liệt kê dưới dạng danh sách bao gồm hai trường đó là (số ca nhập viện, số ca tử vong) và kèm theo đó là một bản ghi tương ứng ghi nhận ra số ca nhập viện và số ca tử vong.

Bài toán trích chọn sự kiện được minh họa trong hình dưới đây



*Hình 3.2: Thành phần trích chọn sự kiện*

Bộ trích chọn bao gồm 4 đặc trưng chính: trích chọn thời gian, trích chọn số ca bệnh, trích chọn địa bàn có dịch bệnh, trích chọn loại bệnh. Đối với đặc trưng thứ nhất là thời gian thì chúng ta sử dụng luật để lấy các thông tin về thời gian công bố sự kiện y tế (Thông tin về thời gian ở đây là ngày tháng ghi nhận sự kiện xảy ra dịch bệnh). Hai đặc trưng về số ca bệnh và loại dịch bệnh được sử dụng các luật để có thể lấy ra thông tin có liên quan tới con số ca bệnh và trích ra được loại dịch bệnh. Còn đối với đặc trưng liên quan tới địa bàn nơi có ghi nhận dịch bệnh thì sẽ sử dụng một từ điển chứa các địa điểm để có thể tiến hành lấy ra được thông tin về địa điểm.

## **Trích chọn thời gian**

Thông qua quá trình khảo sát trên tập dữ liệu, kết quả cho thấy thông tin về thời gian được biểu diễn thông qua hai dạng chính đó là tương đối và tuyệt đối. Thời gian tuyệt đối thông thường có dạng DD/MM/YYYY hoặc dạng DD/MM trong đó: DD là chỉ ngày, MM là chỉ tháng và YYYY là chỉ năm. Ví dụ như sau:

Thời gian được biểu diễn thông qua hai cách nên sử dụng các luật đã được xây dựng từ sẵn để có thể lấy ra được các thông tin liên quan tới thời gian. Trong trường hợp thời gian được biểu diễn dưới dạng tuyệt đối thì có thể lấy ra các thông tin dựa vào biểu thức chính quy (Regular Expression – RE). Đối với trường hợp thứ hai khi thông tin thời gian được đưa ra dưới dạng tương đối, chúng ta có thể nhận ra nó chứa các thành phần: thành phần tiền tố và thời gian. Thành phần tiền tố là lập hợp các cụm từ chỉ thời gian một cách tương đối và thành phần thời gian được biểu diễn dưới dạng MM/YYYY. Ta có luật trích chọn thời gian được biểu diễn minh họa trong công thức : Thời gian = <Tiền tố> + <Ngày tháng> (3.2)

Ngày tháng có định dạng MM/YYYY hoặc DD/MM

Với những trường hợp bản tin không có đề tập tới ngày tháng thì thời gian sẽ được mặc định lấy là thời gian đăng bản tin. Dưới đây là một số ví dụ minh họa cho việc trích chọn thời gian thông qua biểu thức chính quy và thông qua luật:

Ví dụ 1: “Ngày 6/12 bác sĩ Nguyễn Đức Phước, Giám đốc Trung tâm Y tế huyện Trảng Bom cho biết, trên địa bàn huyện vừa ghi nhận 1 ca tử vong do sốt xuất huyết. Bệnh nhân ngụ tại xã Bình Minh huyện Trảng Bom. Trước đó bệnh nhân được người nhà đưa đến khám tại một phòng khám ở thành phố Biên Hòa. Sau đó do diễn biến bệnh nặng nên bệnh nhân được chuyển lên bệnh viện Nhi đồng thành phố Hồ Chí Minh để cấp cứu nhưng đã tử vong. Ngay sau khi ghi nhận trường hợp tử vong, trung tậm huyện Trảng Bom đã tiến hành điều tra xử lý, phun hóa chất để diệt lăng quăng tại nhà của bệnh nhân, hướng dẫn người nhà theo dõi tình hình dịch bệnh. Được biết, đây là trường hợp tử vong so sốt xuất huyết thứ 3 trên địa bàn toàn tỉnh. Đến nay toàn tỉnh cũng đã ghi nhận trên 5 ngàn ca mắc bệnh trong đó số trẻ em mắc bệnh chếm 50%.”

Ví dụ 2: “Theo ghi nhận của Trung tâm Y tế dự phòng tỉnh Đăk Lăk trong 2 tháng đầu năm trên địa bản tình đã ghi nhận 53 trường hợp mắc Sởi trong đó tại thành phố Buôn Ma Thuật có 18 ca, Krong Bông là 9 ca. Theo bác sĩ Phạm Văn Lào tình hình bệnh sởi đầu năm nay có những diễn biến phức tạp khi các ca bệnh đều được ghi nhận tại các vùng di dân ngoài kế hoạch, vùng sâu vùng xa, người dân ít được tiếp xúc với các dịch vụ chăm sóc y tế nên sức khỏe còn nhiều hạn chế. Theo như thống kê của Trung tâm y tế dự phòng tỉnh Đăk Lawk trong 2018 trên địa bàn tình có ít nhất 50 ca mắc sởi.”

## **Trích chọn địa điểm nơi có dịch bệnh**

Để có thể trích chọn ra được địa điểm chúng ta sử dụng bộ nhận dạng thực thể (NER) và từ điển địa điểm. Công việc được tiến hành thông qua các bước như sau:

Bước 1: Áp dụng NER

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

Bước 3: Kiểm tra ngược lại đối với từ điển về địa điểm để có thể tìm ra được chính xác các location

## **Trích chọn số ca bệnh**

Để có thể trích chọn ra thông tin về số ca mắc bệnh dịch thì sử dụng luật. Luật trích chọn số lượng ca bệnh được minh họa bằng công thức đưới đây:

Số ca mắc bệnh = <số> + <hậu tố>

Trong đó:

Số: là số bệnh nhân mắc phải bệnh dịch. Cũng có thể là dạng số hoặc chữ

Số = {“một”, “hai”, “ba”, “bốn”, “năm”, “sáu”, “bảy”, “tám”, “chín”, “mười”}; và các số [1…9].

Hậu tố: bao gồm một số các từ như: bệnh nhân, người, trường hợp, ca nhập viện….

Hậu tố = {“bệnh nhân”, “trường hợp”, “ca bệnh”, “người bệnh”, “ ca mắc bệnh”, “ trường hợp lây lan”};

Kết quả sau khi tiến hành thu thập thông tin được ghi dưới dạng một danh sách bao gồm hai trường dữ liệu và một bản ghi. Trường số ca mắc bệnh và trường số ca tử vong tương ứng với mỗi trường là các số liệu đã được thống kê dưới dạng bản ghi.

**Ví dụ 1:** Tại Hà Nội, từ đầu năm 2019 đến ngày 17/2/2019, thủ đô đã ghi nhận 192 trường hợp mắc bệnh sởi, tăng mạnh so với cùng kỳ năm 2018 (19 ca). Điều đáng nói là sau Tết, bệnh sởi đang có xu hướng gia tăng mạnh. Chỉ riêng tuần đầu sau Tết đã có tới 78 trường hợp mắc sởi. Tuy nhiên, số bệnh nhân phân bố rải rác khoảng 100/584 xã, phường của 25/30 quận huyện trên khắp thành phố; hiện tại chưa ghi nhận nơi tập trung ổ dịch. Độ tuổi mắc bệnh chủ yếu ở nhóm trẻ dưới 9 tháng tuổi và trên 5 tuổi (có 146 trường hợp, chiếm 76%). Hiện nay đã ghi nhận 30 trường hợp bệnh nặng. Có 5 ca bệnh đã tử vong.

Kết quả của ví dụ 1, số trường hợp tử vong là 5.

|  |  |
| --- | --- |
| **Số ca mắc bệnh** | **Số ca tử vong** |
| 192 | 5 |

**Ví dụ 2:** Trong giai đoạn từ ngày 1/1/2019 đến 14/2/2019, tại TP Hồ Chí Minh đã ghi nhận 184 ca nhập viện do bệnh sốt xuất huyết gây nên. Trung bình mỗi tuần trên địa bàn thành phố Hồ Chí Minh có gần 87 trường hợp nhập viện vì bệnh này, chưa kể còn có số lượng lớn người bị mắc sốt xuất huyết điều trị ngoại trú và tại các phòng khám tư. Theo thống kê, số ca mắc sốt xuất huyết tăng 49% so với cùng kỳ năm trước. Đặc biệt tính đến giữa tháng 2/2019 đã ghi nhận 7 ca tử vong.

Kết quả của ví dụ 2, số trường hợp tử vong là 7.

|  |  |
| --- | --- |
| **Số ca mắc bệnh** | **Số ca tử vong** |
| 184 | 7 |

**Ví dụ 3:** Từ tháng 1/2019 đến tháng 3/2019 tại Tỉnh Thanh Hóa đã ghi nhận hơn 42 trường hợp nhiễm bệnh thủy đậu. Dịch bệnh có tốc độ lây lan cao tại các huyện, xã. Theo thống kê thì có ít nhất 70% trẻ dưới 15 tuổi bị mắc thủy đậu và ít nhất 75% người lớn bị mắc bệnh. Đã có 2 trường hợp tử vong.

Kết quả ví dụ 3, số trương hợp tử vong là 2.

|  |  |
| --- | --- |
| **Số ca mắc bệnh** | **Số ca tử vong** |
| 42 | 2 |

## **Trích chọn loại dịch bệnh**

Sử dụng trích chọn thông tin sự kiện y tế thông qua luật để có thể trích chọn. Luật được áp dụng minh họa trong công thức như sau:

Loại dịch bệnh = <danh từ> + <động từ> (3.4)

Trong đó:

Danh từ: Bao gồm tên các loại dịch bệnh trong từ điển như: Sởi, xuất xuất huyết, thủy đậu, viêm gan B, cúm AH5N1, viêm gan, uống ván...Chi tiết của tập các loại dịch bệnh đã được liệt kê trong bảng 3.1

Động từ: Bao gồm các từ như: Lây lan, bùng phát, hoành hành, gây tử vong, diễn biến xấu…

Chi tiết của tập các động từ như sau, verbs = {“lây lan”, “bùng phát”, “báo động”, “diễn biến xấu”, “gia tăng”, “ảnh hưởng”, “tử vong”, “mắc”, “hoành hành”, “dịch bệnh”, “tấn công”, “đe dọa”, “tử vong”, “cảnh giác”, “phòng chống”…}

Ví dụ: Thống kê của Bộ Y tế trong năm 2016, chỉ trong 4 tháng đầu năm, nước ta đã xảy ra gần 30 vụ ngộ độc thực phẩm đặc biệt nghiêm trọng, có 1.386 người bị ngộ độc, có 2 trường hợp đã tử vong. Trong tháng 4/2016 đã xảy ra 9 vụ ngộ độc thực phẩm đã làm cho 375 người bị ngộ độc. Từ kết quả của ví dụ trên ta có thể rút ra được thông tin: Loại dịch bệnh là “Ngộ độc thực phẩm”

# 3.3 Kết luận chương

Trong chương này, khóa luận đã đề xuất ra được phương pháp và mô hình để nhằm giải quyết bài toán tổng quan trích chọn sự kiện y tế. Bên cạnh đó cũng trình bày chi tiết phương pháp và mô hình để có thể giải quyết hai bài toán phụ đó là: Bài toán phát hiện sự kiện y tế và bài toán trích chọn sự kiện y tế. Đối với bài toán thứ nhất sử dụng phương pháp kết hợp luật và học máy để phát hiện ra được sự kiện y tế, những dữ liệu này sau khi đã được phân loại sẽ trở thành nguồn dữ liệu đầu vào cho bài toán thứ hai là bài toán trích chọn sự kiện y tế. Những thông tin được trích chọn ra bao gồm: Thời gian, địa bàn có dịch, số ca bệnh, tên loại dịch bệnh.

# CHƯƠNG 4: THỰC NGHIỆM

Chương này khóa luận sẽ trình bày những thực nghiệm trên thực tế trong quá trình triển khai giải quyết bài toán. Bên cạnh đó là chứng minh tính hiệu quả của phương pháp thông qua hai bài toán quan trọng đó chính là bài toán phát hiện sự kiện và bài toán trích chọn sự kiện. Cuối cùng khóa luận trình bày một số những kiến nghị và bàn luận thêm về các kết quả thu được cũng như nêu một số những khó khăn vướng mắc trong quá trình triển khai bài toán trên thực tế và đề xuất một số các giải pháp để tối ưu hóa bài toán, mang lại hiệu quả cao hơn

# 4.1 Môi trường và công cụ để tiến hành thực nghiệm

Cấu hình phần cứng và các công cụ phần mềm để sử dung trong thực nghiệm của khóa luận được trình bày trong các bảng dưới đây:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành phần** | **Chỉ số** |
| 1 | CPU | 2.8GHz, Intel Core i5 |
| 2 | RAM | 8GB |
| 3 | Hệ điều hành | Win 10 |
| 4 | Bộ nhớ ngoài | 500GB |

### *Bảng 4.1: Cấu hình phần cứng*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên phần mềm** | **Chức năng** | **Nguồn** |
| 1 | Octoparse | Tải các bài viết từ các website | [https://www.octoparse.com/](https://www.octoparse.com/signup) |
| 2 | Python | Ngôn ngữ lập trình | <https://www.python.org/> |
| 3 | jvnTextPro | Phân tích câu Tiếng Việt | [https://github.com/JVnTextPro](https://github.com/garfieldnate/JVnTextPro) |

### *Bảng 4.2: Công cụ phần mềm sử dụng*

Trong bài khóa luận này, dữ liệu được lấy từ nhiều nguồn khác nhau. Một số các trang web báo điện tử uy tín và tin cậy được nhiều truy cập như (Báo Dân trí, VnExpress…) Bên cạnh đó là các website Trung tâm kiểm soát bệnh tật của các tỉnh và địa phương và những mục tin tức hằng ngày liên quan tới lĩnh vực y tế của các trang báo mạng uy tín khác. Lý do lựa chọn những website này là để đảm bảo tính cập nhật nhanh chóng và khá đầy đủ những thông tin y tế mới nhất, hơn nữa đây là những trang web chính thống, đảm bảo về tính chính xác của những thông tin được đăng tải. Quy trình tiến hành thực nghiệm được dựa trên mô hình giải quyết bài toán trích chọn sự kiện y tế đã được trình bày ở phần 3.2.2

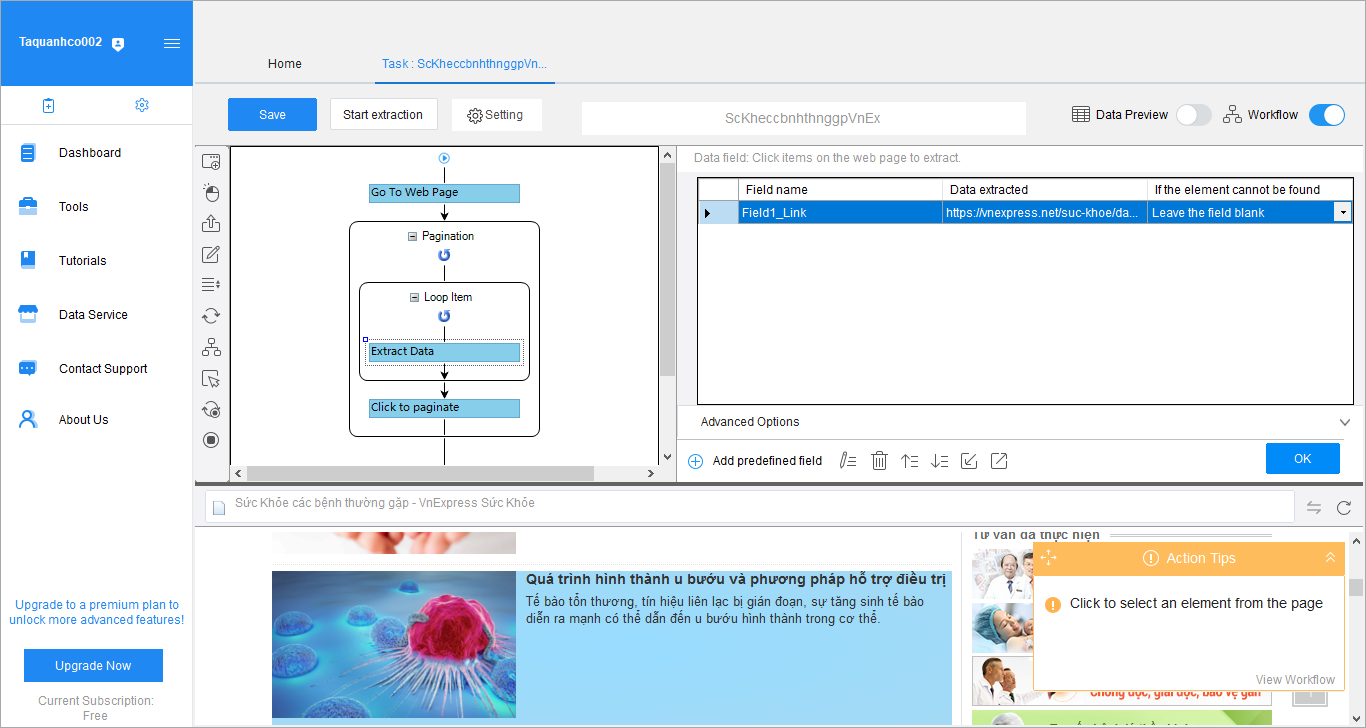
# 4.2 Bộ thu thập dữ liệu

**Mô tả thực nghiệm:** Việc tiến hành thu thập dữ liệu được thực hiện thông qua phần mềm Octoparse. **Octoparse là một công cụ phần mềm chuyên dùng để thu thập dữ liệu từ các website**. Phần mềm này thu thập toàn bộ các thông tin từ các website nói trên. Bên cạnh đó còn cho phép việc sử dụng một số lượng lớn điểm dữ liệu thuận tiện trong quá trình trích xuất và thống kê dữ liệu một cách hiệu quả. Đối với công đoạn đầu tiên sẽ tiến hành thu thập địa chỉ URL của các bài viết trên website.

**Các bước thực hiện**

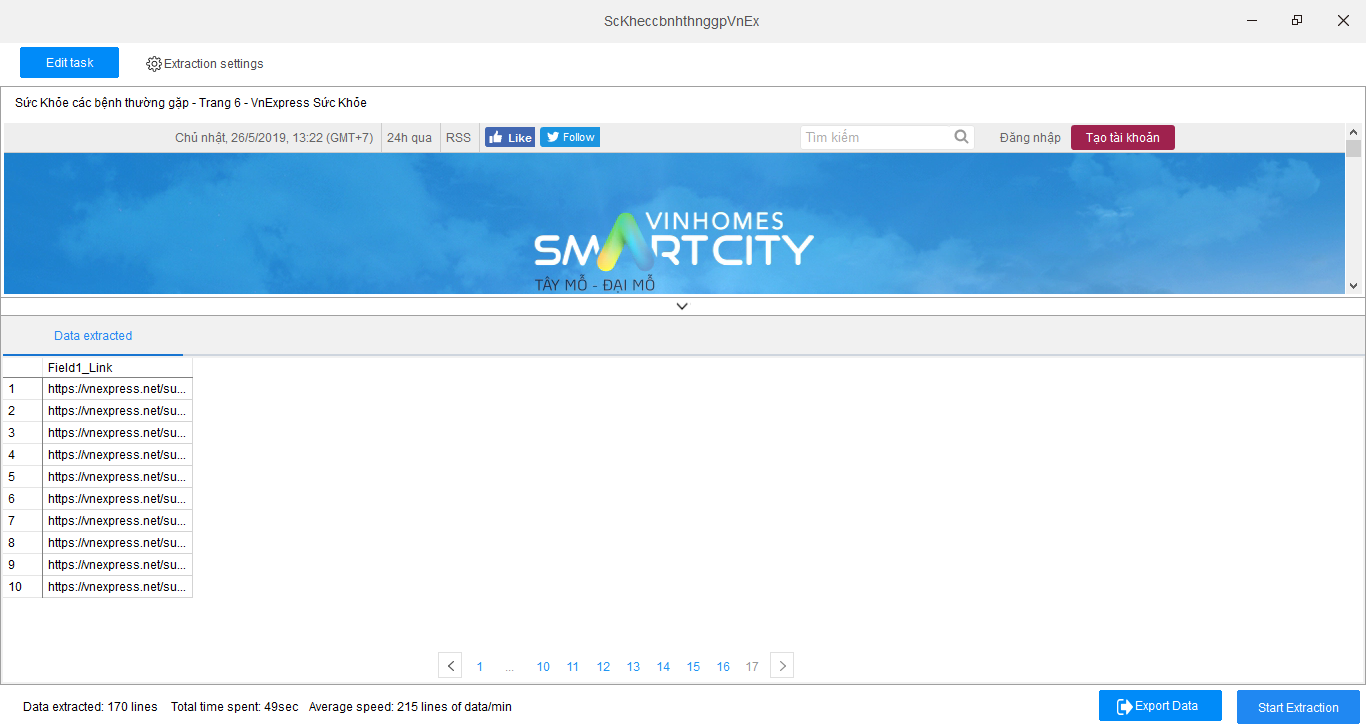
**Bước 1**: Nhập địa chỉ trang web muốn thu thập dữ liệu

**Bước 2:** Click chọn thành phần để lấy ra được địa chỉ URL tương ứng

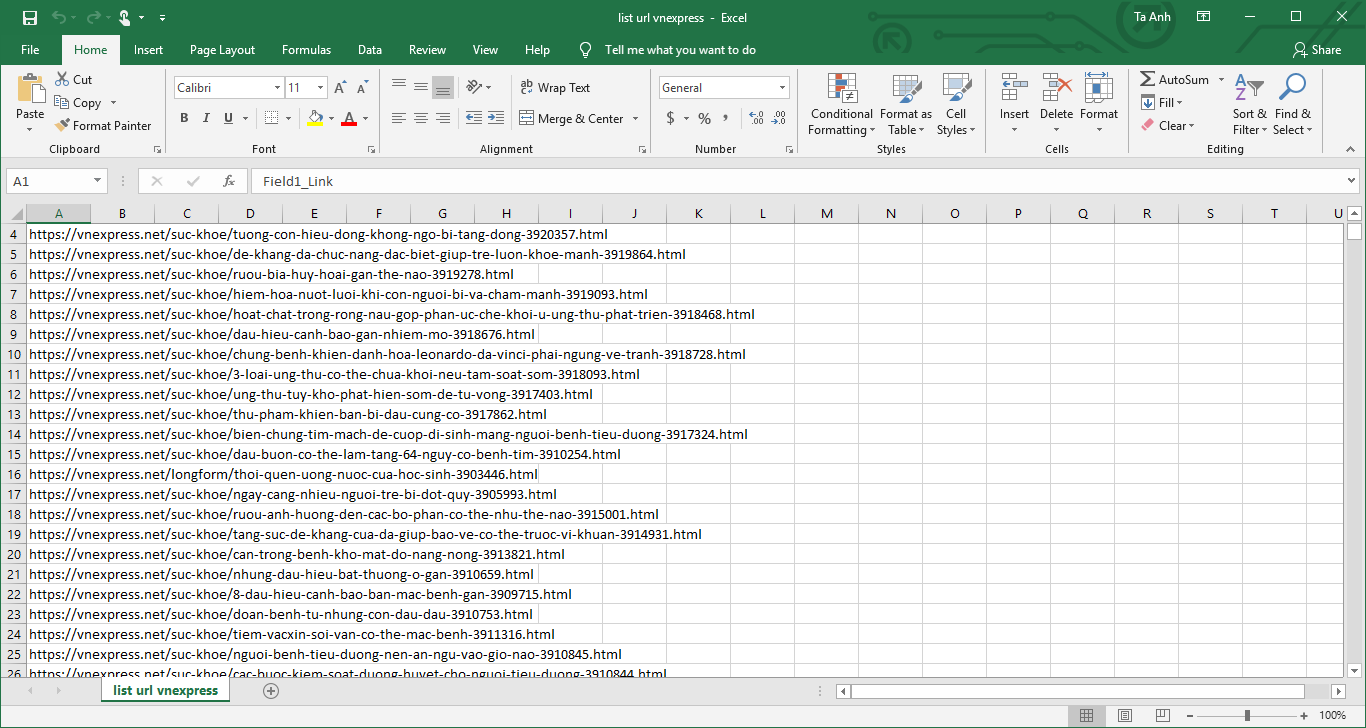


*Hình 4.1 Quá trình thu thập danh sách địa chỉ URL các bài viết*

**Kết quả:** Kết quả của quá trình thu thập dữ liệu sẽ được một bộ dữ liệu bao gồm danh sách địa chỉ các URL của trang web**.** Kết quả quả quá trình được minh họa thông qua hình dưới đây:



### *Hình 4.2 Kết quả quá trình thu thập URL bằng công cụ Octoparse*



### *Hình 4.3 Kết quả quá trình thu thập list URL*

# 4.3.Tiền xử lý dữ liệu

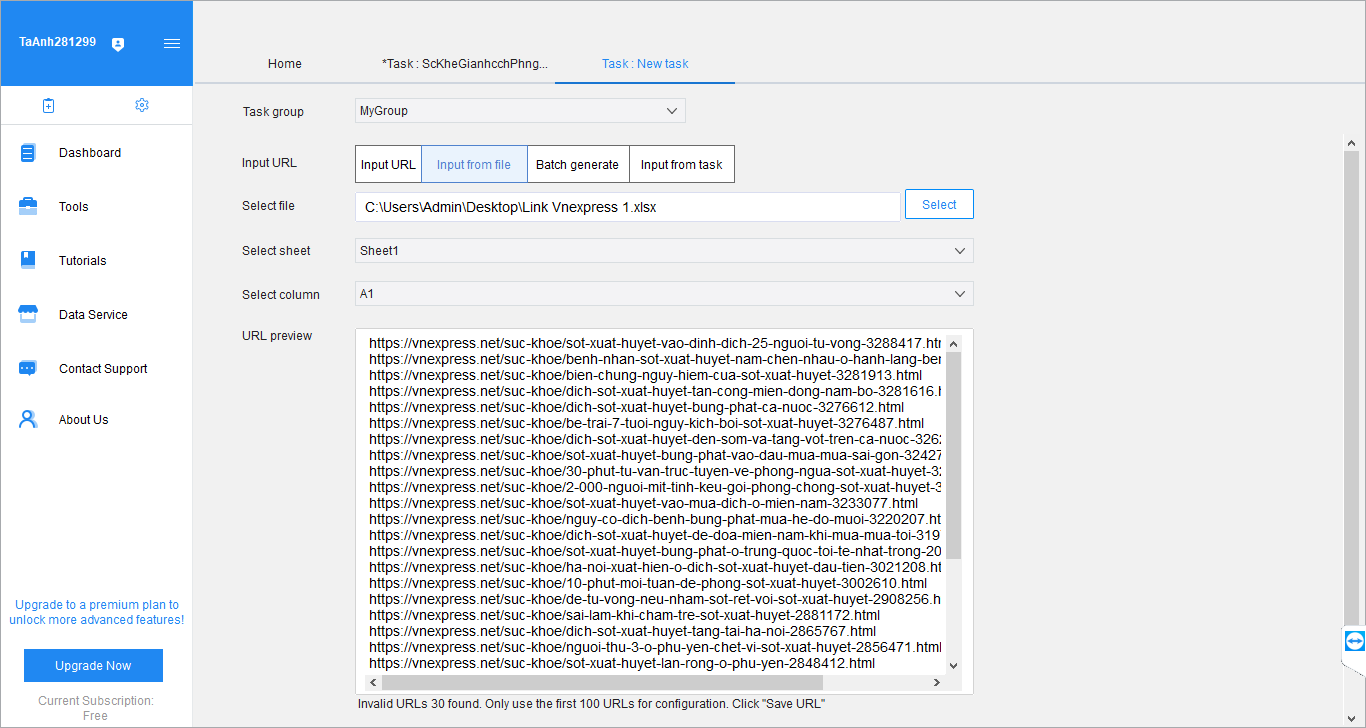
**Mô tả thực nghiệm:** Sau khi đã có được danh sách các URL của mỗi bài viết thì tiếp tục sử dụng công cụ Octoparse để thu thập các thành phần của bài viết như: Thời gian đăng bài, tiêu đề, tóm tắt…Dữ liệu lấy là dữ liệu dạng văn bản thô (text), Cụ thể các thành phần trong một bản tin được minh họa trong bảng dưới đây:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thành phần** | **Mô tả** |
| 1 | Tiêu đề bài viết | Tên tiêu đề của bản tin |
| 2 | Tóm tắt | Phần tóm tắt các ý chính |
| 3 | Ngày đăng | Ngày mà bài viết được đăng tải |
| 4 | Nội dung | Nội dung của bản tin |

### *Bảng 4.3: Các thành phần của một bảng tin*

**Các bước thực hiện:**

**Bước 1:** Import bảng danh sách các URL đã thu thập được ở Bộ thu thập dữ liệu vào phần mềm Octoparse

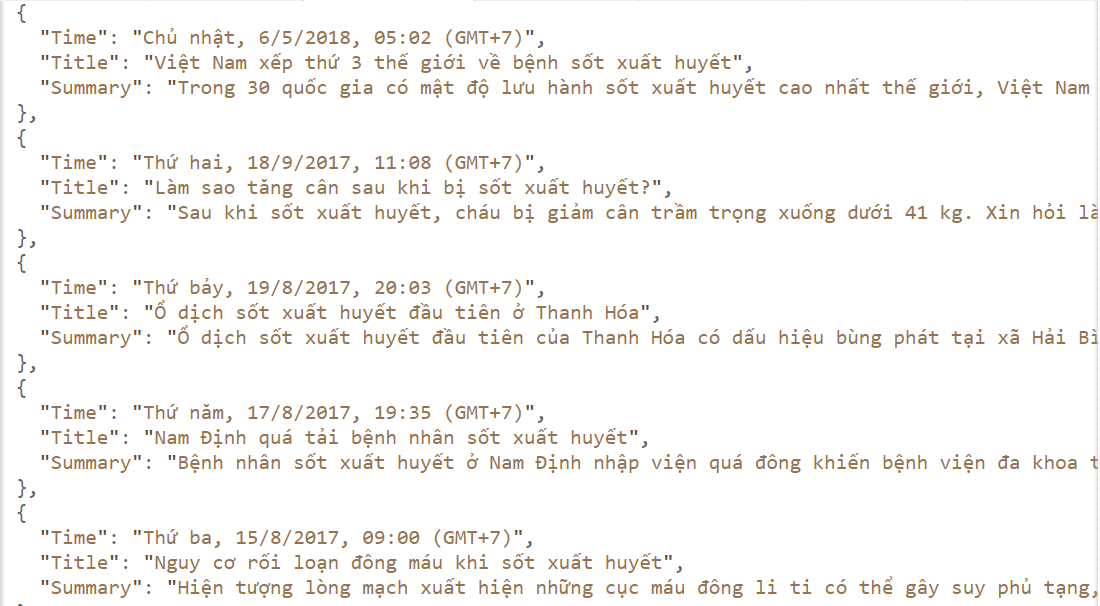


### *Hình 4.4 Quá trình Import list URL của các trang web vào Octoparse*

**Bước 2:** Sau khi đã có được danh sách URL của các bài viết trên website thì sẽ tiến hành chọn lựa các thành phần của bài viết mà chúng ta muốn thu thập về. sẽ tạo một vòng lặp sau đó mới lựa chọn các thành phần trong bài viết để phần mềm có thể thu thập được toàn bộ những bài viết trên site với cấu trúc tương tự. Sau đó phần mềm sẽ tự động trích xuất dữ liệu theo các thành phần đã chọn.

### *Hình 4.5 Quá trình lựa chọn các thành phần trên trang web để thu thập*

**Kết quả:** Kết quả của quá trình tiền xử lý dữ liệu sẽ là thu được một bảng dữ liệu lưu trữ dưới dạng Excel, JSON… bao gồm (Thời gian đăng tải bài viết, tiêu đề bài viết, tóm tắt nội dung bài viết). Bộ dữ liệu được mô tả bằng hình dưới đây:



### *Hình 4.6 Kết quả bộ dữ liệu sau khi đã được tiền xử lý*

# 4.4. Bộ phát hiện sự kiện

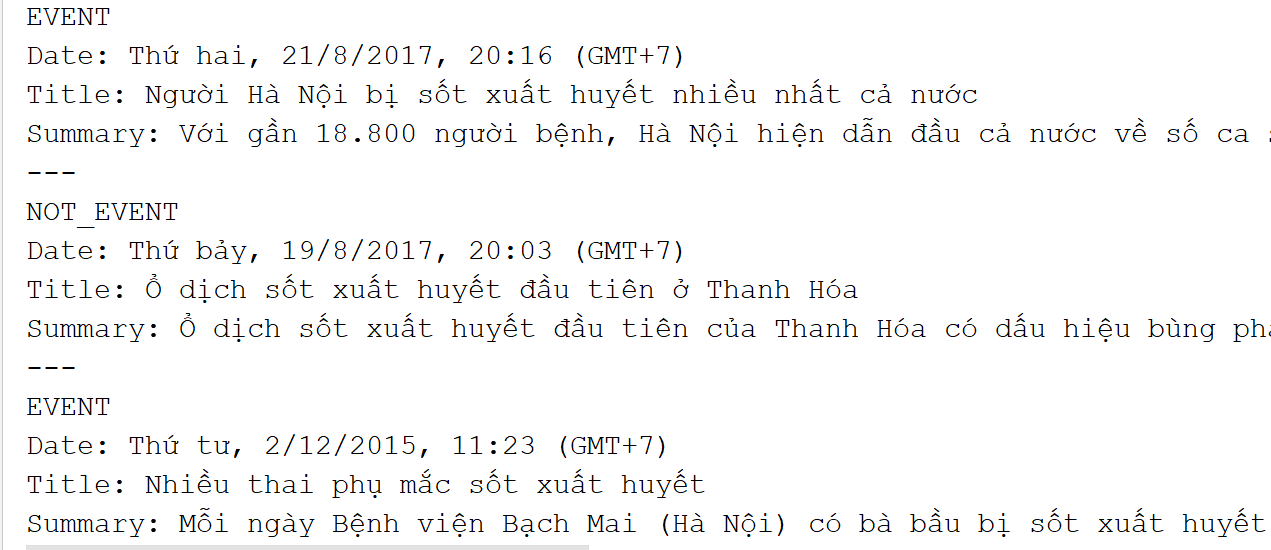
## **4.4.1.Mô tả thực nghiệm**

Để có thể phân loại các bản tin được thu thập được từ thành phần Tiền xử lý dữ liệu ở trên thì khóa luận đã tiến hành xây dựng một công cụ dùng để phát hiện những thông tin trong sự kiện y tế. Công cụ được xây dựng bao gồm:

* Ngôn ngữ sử dụng: Python ( Phần code được để ở phần phụ lục)
* Nguyên tắc xây dựng: Công cụ được thiết kế dựa trên việc sử dụng các tập luật đã được phân tích và xây dựng trong phần giải quyết bài toán Phát hiện sự kiện (Pha 1) đã được trình bày ở mục 3.2.2.1

**Đầu vào:** Danh sách các bài viết đã thông qua quá trình Tiền xử lý dữ liệu

**Đầu ra:** Kết quả đầu ra của Bộ phát hiện sự kiện sẽ là một bộ dữ liệu bao gồm danh sách các bài viết, bản tin có chứa các sự kiện y tế. Những bản tin này sau đó sẽ được chuyển qua Bộ trích chọn sự kiện để tiến hành trích chọn những thông tin của sự kiện. Kết quả của Bộ phát hiện sự kiện được miêu tả bằng hình dưới đây:



### *Hình 4.7 Kết quả bộ dữ liệu sau quá trình Phát hiện sự kiện*

## **4.4.2.Đánh giá bộ lọc dữ liệu**

**Dữ liệu thực nghiệm**: là 500 bản tin

Sau khi tiến hành lọc dữ liệu thu được tổng số là 514 bản tin thuộc miền dữ liệu có liên quan tới lĩnh vực y tế. Trong số những dữ liệu được thu thập về thì những bản tin không liên quan chiếm tỉ lệ phần trăm tương đối ít và không ảnh hưởng quá nhiều tới quá trình thu thập thông tin, lý dó là hầu hết các bài viết đều được lấy trên những website về y tế nên tỉ lệ những bài viết được thu thập thuộc đúng miền thông tin là tương đối cao. Công thức tính tỉ lệ lỗi được tỉnh bày theo công thức dưới đây.

Tỷ lệ phần trăm lỗi của quá trình lọc dữ liệu được tính thông qua công thức sau:

Tỷ lệ lỗi =

Trong đó:

* Tổng số bài viết: Là tổng số bản tin thu thập được sau khi tiến hành lọc;
* Số bài không chính xác: Là những bài viết không liên quan với chủ đề y tế và không thuộc miền dữ liệu y tế.

|  |  |  |
| --- | --- | --- |
| **Tổng số bản tin** | **Số bản tin không liên quan** | **Tỷ lệ lỗi** |
| 312 | 12 | 3.8% |

### *Bảng 4.4: Tỷ lệ lỗi của quá trình lọc dữ liệu*

## **4.4.3 Đánh giá quá trình phân lớp**

**Dữ liệu thực nghiệm**: Dữ liệu của mỗi một lần tiến hành đánh giá là 50 bản tin được lấy ra một cách ngẫu nhiên từ các bản tin được lọc bởi bộ dữ liệu. Kết quả được trình bày trong bảng 4.5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Số bản tin chính xác** | **Số bản tin không chính xác** | **Số bản tin không tìm thấy** | **Precision**  **%** | **Recall**  **%** | **Độ đo F1**  **%** |
| 1 | 80 | 18 | 2 | 81 | 97 | 88 |
| 2 | 85 | 12 | 3 | 87 | 96 | 91 |
| 3 | 79 | 10 | 3 | 88 | 96 | 92 |
| 4 | 86 | 17 | 4 | 83 | 95 | 89 |

### *Bảng 4.5: Đánh giá kêt quả của quá trình phân lớp*

Thông qua kết quả thực nghiệm trong bảng 6 chúng ta thấy được quá trinhg phân lớp có độ chính xác (P – Precision) đạt, dộ đo hồi tưởng (R-Recall) đạt, độ đo F-1 đạt 91%.

# 4.5.Bộ trích chọn sự kiện

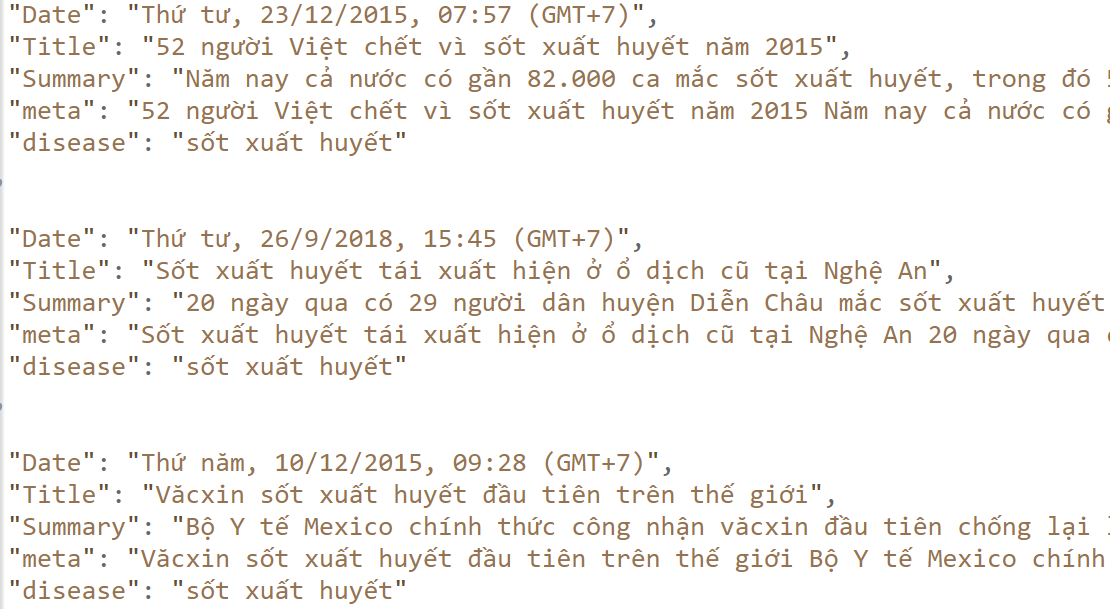
## **4.5.1 Mô tả thực nghiệm**

Trong phần này khóa luận tiếp tục tiến hành chỉnh sửa phát triển công cụ được xây dựng từ Bộ phát hiện sự kiện để có thể thực hiện được mục tiêu là trích chọn ra được các thành phần đặc trưng của sự kiện bao gồm 4 thành phần chính là: Thời gian, địa điểm nơi xảy ra dịch bệnh, số ca mắc bệnh và tên loại bệnh. Quá trình xây dựng tập luật cho từng đặc trưng của sự kiện đã được trình bày trong phần giải quyết bài toán Trích chọn sự kiện (Pha 2) được trình bày ở mục 3.2.2.2

**Đầu vào:** Một bản tin trong miền sự kiện y tế.

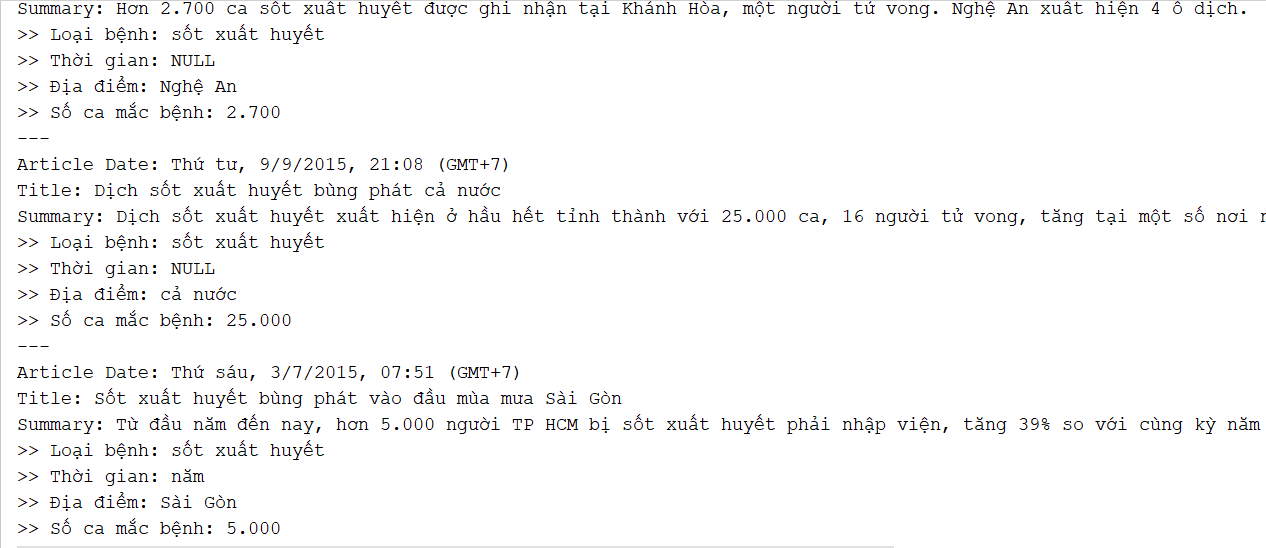
**Đầu ra:** Các thông tin về sự kiện y tế được công bố bao gồm: Thời gian ghi nhận diễn biễn của dịch bệnh, địa bản nơi có dịch bệnh, số ca bệnh (Số ca nhập viện, số ca tử vong) và cuối cùng là loại dịch bệnh. Kết quả của Bộ trích chọn sự kiện được thể hiện qua hình sau:

Dữ liệu đầu vào của Bộ trích chọn sự kiện chính là những bản tin ở đầu ra của Bộ phát hiện sự kiện. Những bản tin này sau khi được tiến hành qua bộ lọc sẽ nhận biết được bản tin đó có chứa các sự kiện y tế hay không.



### *Hình 4.8 Dữ liệu đầu vào của Bộ trích chọn sự kiện*

Đối với những bản tin này thì kết quả đầu ra sẽ là: Bộ trích chọn sự kiện sẽ xác định và tìm được 4 đặc trưng chính của sự kiện bao gồm: Thời gian ghi nhận xảy ra dịch bệnh, địa điểm, số trường hợp nhập viện hoặc tử vong, tên loại bệnh.



### *Hình 4.9 Kết quả của Bộ trích chọn sự kiện*

## **4.5.2. Đánh giá quá trình trích chọn sự kiện**

**Thực nghiệm không thông qua bộ phân lớp**

Dữ liệu thực nghiêm: Lấy ngẫu nhiên một số bản tin từ các bản tin trong miền sự kiện y tế chưa được đưa qua bộ phân lớp.

Ta có 1 sự kiện A được định nghĩa là một bộ bao gồm thời gian, địa điểm, số ca mắc bệnh và tên loại bệnh. Điều này đã được trình bày ở trong công thức 3.1. Như vậy nếu một sự kiện được coi là sự kiện đúng nếu chứa đủ cả bốn thành phần kể trên. Trong trường hợp ngược lại nếu sự kiện không có những yếu tố trên được xem như là một sự kiện sai và chúng ta sẽ tiến hành loại bỏ sự kiện này.

Để có thể đánh giá khả năng trích chọn của sự kiện, em đã sử dụng ba bộ đo lần lượt là: Độ chính xác (P-Precision), độ hồi tưởng (R -Recall), và độ đo F1 (F -Score). Các độ đo này được biểu diễn thông qua các công thức dưới đây.

Độ chính xác (P) =

Trong đó:

Số sự kiện đúng: Số sự kiện được mô hình trích chọn một cách chính xác

Số sự kiện sai: Là số sự kiện mà hô hình trích chọn sai

Độ hồi tưởng (R) =

Trong đó:

Số sự kiện đúng: Số sự kiện được mô hình trích chọn một cách chính xác

Số sự kiện không được trích chọn: Số sự kiện mà mô hình không trích chọn ra

F1 = 2 x

Dựa vào các công thức trên chúng ta áp dựng vào để đưa ra được bảng đánh giá mô hình trích chọn. Chi tiết được trình bày trong bảng dưới đây.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tên website** | **Số sự kiện đúng** | **Số sự kiện sai** | **Số sự kiện không tìm thấy** | **P%** | **R%** | **F1%** |
| [vnexpress.net](https://vnexpress.net/suc-khoe/cac-benh) | 234 | 30 | 10 | 88 | 95 | 91 |
| [soyte.hanoi.gov.vn](http://soyte.hanoi.gov.vn/vi/news/benh-truyen-nhiem/) | 112 | 25 | 9 | 81 | 92 | 86 |
| dantri.com.vn | 135 | 14 | 5 | 90 | 96 | 92 |

### *Bảng 4.6: Đánh giá quá trình trích chọn dữ liệu không qua bộ phân lớp*

**Thực nghiệm thông qua bộ phân lớp**

Dữ liệu thực nghiệm: Dữ liệu là 100 bản tin được lấy từ các bản tin chứa sự kiện y tế (gán nhãn EVENT). Kết quả của quá trình trích chọn sự kiện cũng sử dụng các công thức tính ở trên để có thể đánh giá được thực nghiệm. Kêt quả được mô tả trong bảng dưới đây:

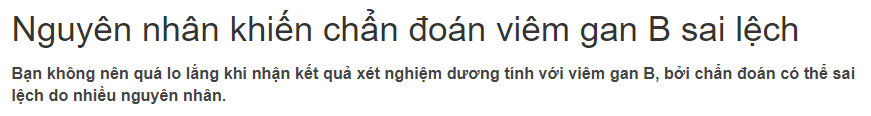
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tên website** | **Số sự kiện đúng** | **Số sự kiện sai** | **Số sự kiện không tìm thấy** | **P%** | **R%** | **F1%** |
| [vnexpress.net](https://vnexpress.net/suc-khoe/cac-benh) | 92 | 4 | 3 | 95 | 96 | 95 |
| dantri.com.vn | 90 | 5 | 2 | 94 | 97 | 96 |
| [soyte.hanoi.gov.vn](http://soyte.hanoi.gov.vn/vi/news/benh-truyen-nhiem/) | 89 | 2 | 4 | 97 | 95 | 96 |

### *Bảng 4.7: Đánh giá quá trình trích chọn dữ liệu thông qua bộ phân lớp*

# 4.6 Phân tích lỗi

## **4.6.1 Phân tích lỗi quá trình phát hiện sự kiện**

Qua quá trình tiến hành khảo sát và thống kê dữ liệu sau khi đã thực nghiệm nhận thấy có một vài lỗi liên quan tới quá trình phát hiện sự kiện như sau sau. Một số các bản tin được thu thập về mặc dù có chứa từ khóa nằm trong từ điển về các loại dịch bệnh nhưng nội dung bài viết lại không thuộc miền dữ liệu sự kiện y tế. Những bài viết này trong quá trình lọc có thể làm ảnh hưởng tới kết quả tìm kiếm và thu thập thông tin.



Một số những khó khăn trong quá trình thu thập và phát hiện sự kiện: Đối với lĩnh vực y tế thì thông tin thu thập được tuy nhiều nhưng những bài viết có nội dung liên quan tới những sự kiện y tế thì tương đối ít và cập nhập không được thường xuyên. Chủ yếu các bài viết trên các website đều là những thông tin y tế tổng quát hoặc liên quan tới cách phòng chống dịch bệnh. Điều này khiến cho số lượng những bài viết thu thập về đáp ứng đủ nhu cầu để có thề tiến hành trích chọn sự kiện là tương đối ít.

## **4.6.2 Phân tích lỗi trong quá trình trích chọn sự kiện**

Đối với pha trích chọn thông tin thì khả năng trích chọn các thông tin còn tương đối thấp. Nguyên nhân của việc này có thể được phân tích thông qua một số những lỗi như sau: Đối với việc trích chọn thời gian, thông thường những thống kê liên quan tới sự kiện y tế được thu thập và công bố theo các thời kì và giai đoạn (Ví dụ: Trong quý 4 năm 2018, trong 3 tháng đầu năm, 2 tuần đầu tiên của kì nghỉ lễ,…). Những trường hợp này không thể xác định được một giai đoạn cụ thể nào ghi nhận các trường hợp mắc bệnh. Đối với việc trích chọn địa điểm nơi ghi nhận bùng phát dịch bệnh, đôi khi trong các bản tin thường chỉ nhắc tới tên của các quận/huyện chứ không đề cập tới tên thành phố nên những trường hợp này không cho ra được những kêt quả địa điểm chính xác. Đối với việc trích chọn thông tin về số lượng các ca mắc bệnh hoặc số lượng các ca tử vòng thì một số bài viết không đề cập tới con số cụ thể các ca mắc mà nói chung ví dụ như: “Dịch thủy đậu bùng phát ở nhiều tỉnh thành trên cả nước làm cho số lượng trẻ nhập viện tăng lên một cách đột biến”, trong trường hợp này không thể tìm ra được thông tin về số lượng ca bệnh để có thể trích chọn ra được thông tin. Chi tiết hơn về một số ví dụ các lỗi được trình bày ở bảng dưới đây:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thông tin chung** | **Thông tin trích chọn** |
| 1 | Trên địa bàn thành phố Hà Nội | Hà Nội |
| 2 | 2 ngày nghỉ tết, quý 1 năm 2019 | Null |
| 3 | Nhiều trường hợp phải nhập viện | Null |
| 4 | Khống chế dịch Sởi | Sởi |

### *Bảng 4.8: Một số lỗi trong quá trình trích chọn*

**4.6. Các để xuất sử dụng kết quả phân tích**

Kết quả của quá trình trích chọn có thể được sử dụng trong lĩnh vực y tế để đem lại hiệu quả cao trong các công tác phòng tránh và ngăn ngừa dịch bệnh:

* Từ những số liệu thống kê liên quan tới thời gian có thể sử dụng để biết được trong giai đoạn nào thì dịch bệnh bùng phát và khiến cho nhiều người mắc bệnh. Những mùa nào trong năm thì có những loại bệnh nào nguy hiểm cần phải hết sức lưu ý
* Đối với thông tin về số lượng các ca mắc bệnh thì những nhà phân tích có thể dựa vào đó để biết dịch bệnh này đang ảnh hưởng tới bao nhiều người. Con số mắc bệnh và tử vong cụ thể là bao nhiêu.
* Những địa bàn nơi ghi nhận vùng dịch bệnh có chứa các thông tin về vùng dịch bệnh đang diễn ra. Dữ liệu này để các cơ quan phòng chống dịch bệnh phân vùng được ổ dịch. Nắm bắt được xem địa phương nào đang để xảy ra dịch bệnh nguy hiểm.

Nhìn chung các thông tin được trích chọn có thể dùng để tổng hợp và thống kê lại thành các báo cáo nhằm phục vụ công tác ngăn ngừa và phòng chống dịch bệnh nguy hiểm đang lây lan trong cộng đồng. Những nhà quản lý, các đơn vị liên quan có thể sử dụng những nguồn dữ liệu báo cáo này để nắm bắt được tình hình và diễn biến của dịch bệnh từ đó đưa ra những phản ứng và thông cáo hướng dẫn đến với người dân một cách nhanh chóng nhất, đề phòng dịch bệnh diễn biến phức tạp gây ảnh hưởng tới sức khỏe của người dân.

# 4.8. Kết luận chương

Trong chương này khóa luận đã trình bày quá trình tiến hành thực nghiệm bên cạnh đó là xem xét và đánh giá kết quả của mô hình trích chọn thông tin trong văn bản liên quan tới lĩnh vực y tế và sức khỏe cộng đồng. Thông qua kết quả thu được từ thực nghiệm đã cho thấy được tính khả thi của mô hình giải quyết bài toán trích chọn sự kiện y tế.

# KẾT LUẬN

**Kết quả đạt được của khóa luận**

Trong thời gian thực hiện bài khóa luận này em đã tìm hiểu được các phương pháp liên quan tới việc trích chọn sự kiện. Cùng với đó là phương pháp sử dụng kết hợp các luật và ứng dụng khai phá dữ liệu để áp dụng cho bài toán phát hiện sự kiện và bài toán trích chọn thông tin sự kiện. Dựa trên cơ sở đó, đề xuất mô hình và phương pháp giải quyết một cách chi tiết đối với bài toán phát hiện sự kiện y tế và bài toán trích chọn sự kiện y tế. Đối với mỗi bài viết được xác định là thuộc miền dữ liệu liên quan tới y tế và có chứa sự kiện y tế thì các sự kiện được tiến hành trích chọn ra bao gồm bốn thành phần chính là: Thời gian, địa điểm, số ca mắc bệnh (tử vong), tên loại bệnh dịch. Thông qua quá trình thực nghiệm nhận thấy còn một số vấn đề khó khăn khi triển khai thực tế. Bên cạnh đó là đề xuất thêm một số các phương án để phát triển đề tài trong tương lai.

**Khó khăn và hạn chế**

* Kết qủa của bộ phân lớp chưa có độ chính xác cao. Nguyên nhân là do vẫn còn có sự chưa rõ ràng để phân biệt được giữa những bản tin có chứa sự kiện y tế và các bản tin có liên quan tới y tế nói chung.
* Việc xây dựng tập luật hiện tại đang được tiến hành bằng công tác thủ công. Đây là lý do khiến cho bộ dữ liệu khó có thể được bao phủ một cách hoàn toàn. Điều này dẫn tới việc tập luật chưa thực sự chặt chẽ, trong một số các trường hợp có thể bị bỏ sót những dữ liệu có liên quan tới miền dữ liệu;
* Trích chọn thông tin thời gian còn chưa rõ ràng do đặc thù những báo cáo thống kê về dịch bệnh thường được cập nhập theo kì (nhóm số liệu của nhiều tháng) thay vì được cập nhập hằng ngày;
* Trong một vài trường hợp viết tắt, khi trích chọn thông tin còn chưa chính xác.

**Định hướng trong tương lai**

Các kết quả của quá trình trích chọn sự kiện sẽ được sử dụng để tiến hành lập nên các biểu đồ dùng cho việc phân tích số liệu. Một số các loại biểu đồ có thể hiện mối tương quan giữa các thành phần trích chọn của sự kiện bao gồm: số ca, thời gian, địa điểm và tên loại bệnh dịch. Trong tương lai, những biểu đồ và các báo cáo thống kê sẽ được đa dạng hóa để mang lại hiệu quả cao. Ví dụ về một số các bảng thống kê dữ liệu được trình bày dưới đây:

## **4.7.1.Thống kê số ca mắc bệnh**

Biểu đồ dưới đây mô tả số lượng các trường hợp nhập viện do mắc một số các bệnh truyền nhiễm đang có xu hướng bùng phát mạnh trong thời gian gần đây như: Sốt xuất huyết, Sởi, Cúm… Số liệu được ghi nhận tại một số tỉnh/thành phố lớn trên cả nước. Dựa vào thống kê có thể thấy số lượng ca mắc bệnh đối với các bệnh nguy hiểm như sốt xuất huyết hay sởi là tương đối lớn trên các thành phố.

### *Biểu đồ 1: Thống kê số ca mắc bệnh ở 4 tỉnh/thành phố lớn*

## **4.7.2.Thống kê số ca mắc bệnh**

Biều đồ dưới đây thông kê số ca mắc bệnh được ghi nhận trung bình qua các quý trong năm. Có thể nhận thấy rằng thông thường thì trong các quý 2 và quý 3 thì các dịnh bệnh có xu hướng tăng mạnh.

### *Biểu đồ 2: Thống kê số ca mắc bệnh theo các quý*

## **4.7.3.Thống các dịch bệnh tại Hà Nội và Tp.Hồ Chí Minh**

### *Biểu đồ 3: Thống kê các bệnh thường gặp tại Hà Nội*

Định hướng nghiên cứu tiếp theo của khóa luận là tiếp tục cải tiến và hoàn thiện mô hình trích chọn sự kiện y tế từ các trang web Tiếng Việt. Nghiên cứu và phát triển thêm để có thể trích chọn được nhiều những thông tin quan trọng khác như: Độ tuổi trung bình của các ca mắc bệnh; dấu hiệu nhận biết của bệnh; diễn biến tình trạng bệnh của các bệnh nhân sau khi được nhập viện để điều trị và một số thông tin liên quan khác. Những dữ liệu được rút trích có thể là một nguồn thông tin tham khảo hữu ích cho các nhà nghiên cứu, đội ngũ y bác sĩ, các cơ quan chuyên trách về y tế để có thể nắm bắt được diễn biến và tình trạng của dịch bệnh đang xảy ra. Từ đó đưa ra được những phương án kịp thời để cảnh báo và ngăn ngừa nguy cơ lây lan của dịch bệnh và bảo vệ sức khỏe của cộng đồng. Ngoài ra có thể áp dụng những lý thuyết và mô hình của bài toán trích chọn này cho rất nhiều những lĩnh vực khác nhau trong cuộc sống. Một số những lĩnh vực khác có thể áp dụng bài toán đó là trích chọn thông tin sự kiện vụ tai nạn giao thông hoặc trích chọn thông tin liên quan tới dịch vụ du lịch. Tùy vào mục đích và đối tượng nghiên cứu cụ thể.

# DANH MỤC TÀI LIỆU THAM KHẢO

[1] Sunita Sarawagi (2008), *Information Extraction*, Indian Institute of Technology, CSE, Mumbai 400076, India,

[2] Douglas E. Appelt. Introduction to information extraction technology. *In Tutorial held at IJCAI-99, Stockholm, Swenden*,1999.

[3] Young-Sook Hwang Chun Hong-Woo and Hae-Chang Rim. Unsupervised event extraction from biomedical literature using co-occurrence information and basic pattems*. In: Ist International Joint Conference on Natural Language Processing (*IJCNLP 2004). Lecture Notes in Computer Science. Springer- Verlag Berlin Heidelberg, vol.3248:7772786,2004.

[4] Uzay Kaymak Frederik Hogenboom, Flavius Frasincar and Franciska de Jong. An overview of event extratiom from text. *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011),* 779:pp.48257, 2011.10

[5] M.A Heast. Automatic acquisition of hyponyms from large text corpora. *In: 14th Conference on Computational Linguistics (COLING 1992),* vol. 2;5392 545, 1992.

[6] M.A Hearst. Wordnet: An electronic lexical database and come of its applications. *In Automated Discovery of WordNet Relations, pp, 1312 151. MIT Press, 1998.*

[7] Frederik Hogenboom jethro Borsje and Flavius Frasonacar. Semi-automatic financial events discovery based on lexico-semantic patterms. *International journal of Web Engineering and Technology*, 6(2): 1152 140, 2010.

[8] Yea-Juan Chen Lee Chang-Shing and Zhi-Wei Jian. Ontology-besed fuzzy event extraction agent for chinese e-news summarization. *In Expert Systems with Applications* 25(3),4312 447,2003.

[9] Okamoto Masayyki and Masaaki Kikuchi. Discovering volatile events in your neightborhood: Local-area topic extraction from blog entries. *In: 5th Asia Information Retrieval Symposium (AIRS 2009). Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg*, vol , 5839:1812 192, 2009.

[10] Liang Xiang Xing Chen Mingrong Liu, Yicen Liu and Qing Yang. Extracting key entities and significant events from online daily news*.In :9th International Conference on Intel- ligent Data Engineering and Automated Learning Heidelberg*, vol.5326:2012 209,2008.

[11] L.Peshkin and A.Pfeffer. Bayesian information extraction network. *In Proc.of the 18th International Joint Conference on Artificial Intelligence (IJCAI)* 2003.

[12] Hristo Tenev Piskorski Jakub and Pinar Oezden Wennerberg. Extracting violent events from on-line news for ontology population. *In: 10th International Conference on Business Information Systems (BIS 2007). Lecture Notes in Computer Science. Springer- Verlag Berlin Heidelberg,* vol. 4439:2872 300, 2007.

[13] Silja Huttunen Ralph Grishman and Roman Yangaber. Information extraction for endenced access to disease outbreak reports. Journal of Biomerdical Informastic, 35(4):pp.2362 246,2002.

[14] Ai kawazoe Son Doan and Nigel Collier. Global health monitoer- a web- based system for detecting and mapping infectious discases. Proc*. International Joint Conference on Natural Language Processing ( IJCNLP), Companion Voulume, Hyderabad, India*: pp, 9512 956, 2008.

[15] William H. Hsu Svitlana Volkova, Doina Caragea and Swathi Bujuru.

*Animal disease event recognition and classification.*2010

[16] Yusuke Miyao Akane Yakushiji, Yuka Tateisi and Jun ichi Tsujii.

Event extraction from biomedical papers using a full parser. In *In: 6th Pacific Symposium on Biocomputing (PSB 2001)* :pp, 4082 419, 2001.

[17] Helen L.Johnson Chris Rocder Philip V. Ogren- William A.Baumgartner Jr.Elizabeth White Hannah Tipney K. Bretonnel Cohen, Karin Verspoor and Lawrence Hunter. High-precision biological event extraction with a concept recognizwer. In *In: Workshop on Bio NLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. pp. 502 58. Association for Computational Linguistics, 2009*.

[18] S.Soderland, “Learning information extraction rules for semi – structured and free text,” *Machine Learning,* vol. 34, 1999.

[19] H Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “Gate: A framework and graphical development environment for robust nlp tools and applications,” in *Proceedings of the 40th Anniversary Meeting of the Association for Comutational Linguistics*, 2002.

[20] W. Shen, A . Doan, J, F. Naughton, and R. Ramakrishnan, “Declarative information extraction using datalog with embedded extraction predicates,” in VLDB, pp. 1033-1044, 2007.

[21] Bing Liu, Web mining Springer 2007.

# 

# PHỤ LỤC

**Xây dựng code cho thành phần Bộ phát hiện sự kiện:**

*# Change execution scope to this script's directory*os.chdir(os.sys.path[0])  
  
*# Get the disease keywords***with** open(**r'keywords\dise\_keywords.txt'**, encoding=**'utf-8'**) **as** f:  
 dise\_keywords = f.read().splitlines()  
  
*# Get the list of crawled articles***with** open(**r'articles\crawled\_pages.json'**, encoding=**'utf-8'**) **as** f:  
 raw\_articles = json.loads(f.read().strip(**'\ufeff'**))  
  
*# Make up a 'meta' key for later look up***for** article **in** raw\_articles:  
 article.update({**'meta'**: article[**'Title'**]+**' '**+article[**'Summary'**]})  
  
*# Mark articles True if keywords can be found, False otherwise*marked\_articles = []  
**for** article **in** raw\_articles:  
 passed = **False  
 for** kw **in** dise\_keywords:  
 **if** kw.lower() **in** article[**'meta'**].lower():  
 passed = **True  
 break** article.update({**'disease'**: kw.lower()})  
 marked\_articles.append((article, passed))  
  
*# Store accepted articles for later analysis***with** open(**r'data\bai1.json'**, **'w'**, encoding=**'utf-8'**) **as** f:  
 accepted = [article **for** article,passed **in** marked\_articles **if** passed]  
 marked\_json = json.dumps(accepted, ensure\_ascii=**False**, indent=2)  
 f.write(marked\_json)  
  
**Xây dựng code cho thành phần phân lớp bản tin sự kiện:**

*# Change directory to this script's directory*os.chdir(sys.path[0])  
  
*# Load accepted articles from the previous step's result***with** open(**r'data\bai2.json'**, encoding=**'utf-8'**) **as** f:  
 accepted = json.loads(f.read())  
  
*# Display results*screen\_width = os.get\_terminal\_size().columns - 1  
print(**'-'** \* screen\_width)  
**for** article **in** accepted:  
 print(**'Article Date: {}'**.format(article[**'Date'**]))  
 print(**'Title: {}'**.format(article[**'Title'**]))  
 print(**'Summary: {}'**.format(article[**'Summary'**]))  
 print(**'>> Loại bệnh: {}'**.format(article[**'disease'**]))  
 print(**'>> Thời gian: {}'**.format(article[**'time'**] **if 'time' in** article **else 'NULL'**))  
 print(**'>> Địa điểm: {}'**.format(article[**'place'**] **if 'place' in** article **else 'NULL'**))  
 print(**'>> Số ca mắc bệnh: {}'**.format(article[**'case'**] **if 'case' in** article **else 'NULL'**))  
 print(**'-'** \* screen\_width)

**Xây dựng code cho thành phần Bộ trích chọn sự kiện:**

*# Change directory to this script's directory*os.chdir(os.sys.path[0])  
  
*# Load keywords into a dict, each key represents a keyword type*keywords\_dict = {}  
**for** file\_name **in** [**'time\_keywords.txt'**,  
 **'place\_keywords.txt'**,  
 **'case\_keywords.txt'**,  
 **'number\_keywords.txt'**]:  
 f = open(**r'keywords\{}'**.format(file\_name), encoding=**'utf-8'**)  
 key = file\_name.replace(**'\_keywords.txt'**, **''**)  
 keywords\_dict.update({key: f.read().splitlines()})  
 f.close()  
  
*# Load accepted articles from previous step***with** open(**r'data\bai1.json'**, encoding=**'utf-8'**) **as** f:  
 articles = json.loads(f.read())  
  
*# A function to update the dict, article, only when necessary***def** update(dict\_, key, value):  
 **if** key **not in** dict\_:  
 dict\_.update({key: value})  
  
*# Search for data fields*accepted = []  
**for** article **in** articles:  
 string = article[**'meta'**]  
  
 *# Search for occurrences of places* **for** kw **in** keywords\_dict[**'place'**]:  
 **if** kw.lower() **in** string.lower():  
 update(article, **'place'**, kw)  
 **break** *# Search for data that is composed with a number* numbers\_pos = []  
 words = string.split(**' '**)  
 **for** i **in** range(len(words)):  
 **if** re.match(**r'((\d)+|(\d)+(\.){1}(\d)+)$'**, words[i]):  
 numbers\_pos.append((words[i], i))  
  
 **for** i **in** range(len(numbers\_pos)):  
 num\_str, num\_str\_pos = numbers\_pos[i]  
 **if** i < len(numbers\_pos) - 1:  
 next\_num\_str\_pos = numbers\_pos[i+1][1]  
 **else**:  
 next\_num\_str\_pos = -1  
  
 **if** num\_str\_pos == len(words) - 1:  
 neighbor = words[num\_str\_pos - 1]  
 **if** neighbor **in** keywords\_dict[**'time'**]:  
 update(article, **'time'**, neighbor+**' '**+num\_str)  
 **break  
  
 if** next\_num\_str\_pos == -1:  
 examinables = words[num\_str\_pos+1:]  
 **else**:  
 examinables = words[num\_str\_pos+1 : next\_num\_str\_pos]