BỘ QUỐC PHÒNG

**HỌC VIỆN KỸ THUẬT QUÂN SỰ**

**NGUYỄN HỒNG SƠN**

**NGHIÊN CỨU PHƯƠNG PHÁP TÌM KIẾM NGỮ NGHĨA**

**SỬ DỤNG ONTOLOGY VÀ ỨNG DỤNG XÂY DỰNG**

**HỆ THỐNG TRA CỨU, TÌM KIẾM VĂN BẢN MẪU BỆNH**

 Chuyên ngành: Cơ sở toán học cho tin học

 Mã số: 9 46 01 10

**TÓM TẮT LUẬN ÁN TIẾN SĨ**

**HÀ NỘI – 2020**

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI**

**HỌC VIỆN KỸ THUẬT QUÂN SỰ - BỘ QUỐC PHÒNG**

**Người hướng dẫn khoa học:**

  **1. TS. Dương Trọng Hải**

 **2. TS. Hoa Tất Thắng**

**Phản biện 1: PGS. TS Lê Trọng Vĩnh, Trường Đại học KHTN, Đại học Quốc gia Hà Nội**

**Phản biện 2: PGS. TS Ngô Thành Long, Học viện Kỹ thuật Quân sự**

**Phản biện 3: PGS. TS Cao Tuấn Dũng, Trường Đại học Bách khoa Hà Nội**

Luận án được bảo vệ tại Hội đồng đánh giá luận án cấp Học viện theo quyết định số 1109/QĐ-HV, ngày 15 tháng 04 năm năm 2020 của Giám đốc Học viện Kỹ thuật Quân sự, họp tại Học viện Kỹ thuật Quân sự vào hồi …… giờ … ngày …. tháng …. năm 2020.

**Có thể tìm hiểu luận án tại:**

- Thư viện Học viện Kỹ thuật Quân sự

- Thư viện Quốc gia 0...

MỞ ĐẦU

1. Tính cấp thiết của đề tài luận án

 Hiện nay, các hệ thống tìm kiếm văn bản phần lớn vẫn dựa trên cách sử dụng từ khóa, người dùng phải diễn đạt nội dung mình cần tìm kiếm dưới dạng các từ khoá sao cho đảm bảo mối quan hệ ngữ nghĩa giữa các từ khóa (cụm từ khoá) với nội dung cần tìm. Đây là vấn đề khó khăn, đặc biệt khi người dùng không hiểu về lĩnh vực cần tìm kiếm, điều này được thể hiện rất rõ khi người dùng tìm kiếm bệnh dựa trên các triệu chứng ban đầu chưa được biểu đạt rõ ràng.

Đề tài “Nghiên cứu phương pháp tìm kiếm ngữ nghĩa sử dụng Ontology và ứng dụng xây dựng hệ thống tra cứu, tìm kiếm văn bản mẫu bệnh” nghiên cứu các phương pháp tìm kiếm, các phương pháp xây dựng, trích rút thông tin từ văn bản làm giàu ontology bệnh, trích rút thông tin từ văn bản, khai phá các luật kết hợp trong ontology bệnh và ứng dụng xây dựng Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh có hỗ trợ tương tác với người sử dụng bằng các gợi ý dựa trên tập luật kết hợp giữa các triệu chứng và luật kết hợp ngữ nghĩa từ các mối quan hệ trên Ontology bệnh.

2. Đối tượng và phạm vi nghiên cứu của đề tài luận án

- Phương pháp thu thập thông tin bằng trích rút đặc trưng văn bản;

- Phương pháp xây dựng, tổ chức lưu trữ, quản lý và khai thác sử dụng tri thức dưới dạng ontology;

- Phương pháp khai phá luật kết hợp trong Ontology bệnh;

- Các mô hình tìm kiếm;

- Phân tích thiết kế và xây dựng Hệ thống tìm kiếm ngữ nghĩa có tương tác thông tin bệnh.

3. Nội dung nghiên cứu đề tài luận án

* Thu thập thông tin nhờ trích rút đặc trưng văn bản;
* Xây dựng, tổ chức lưu trữ, quản lý và khai thác sử dụng tri thức Ontology bệnh *DO* (Disease Ontology);
* Xác định luật kết hợp trong Ontology bệnh;
* Giải pháp tìm kiếm ngữ nghĩa có tương tác thông tin bệnh gợi ý người sử dụng xác định câu truy vấn cho phép nhận được kết quả tìm kiếm đúng ý định;
* Hệ thống tìm kiếm ngữ nghĩa có tương tác thông tin bệnh.

4. Phương pháp nghiên cứu đề tài luận án

* Cơ sở lý thuyết về các mô hình tìm kiếm và phương pháp sử dụng ontology trong các hệ thống tìm kiếm; Cơ sở lý thuyết xây dựng ontology;
* Phương pháp xây dựng và làm giàu Ontology bệnh tiếng Việt;
* Phương pháp xây dựng hệ thống kiếm ngữ nghĩa có tương tác và áp dụng xây dựng hệ thống tìm kiếm ngữ nghĩa có tương tác thông tin bệnh.

5. Những điểm mới của đề tài luận án

 (1) Đề xuất phương pháp trích rút bộ ba (triple) dựa trên mô hình cú pháp, trích rút các bộ ba từ dữ liệu văn bản phục vụ cho việc xây dựng đặc trưng của văn bản [CTLA3];

 (2) Đề xuất phương pháp tìm kiếm đa diện dữ liệu văn bản và cá nhân hoá tìm kiếm đa diện sử dụng dữ liệu định hướng xử lý nhập nhằng của *Wikipedia Disambiguation* [CTLA3], [CTLA4];

 (3) Đề xuất phương pháp khai phá luật kết hợp trong Ontology bệnh (ASO-Apriori) [CTLA2] dựa trên 02 độ đo mới: độ hỗ trợ mở rộng và độ tin cậy mở rộng và khai phá luật kết hợp ngữ nghĩa giữa các mối quan hệ trong Ontology bệnh;

 (4) Xây dựng Ontology bệnh tiếng Việt [CTLA1];

 (5) Đề xuất phương pháp và xây dựng Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh có hỗ trợ tương tác với người sử dụng bằng các gợi ý dựa trên tập luật kết hợp giữa các triệu chứng và luật kết hợp ngữ nghĩa từ các mối quan hệ trên Ontology bệnh [CTLA1].

6. Ý nghĩa khoa học

Luận án mở ra hướng nghiên cứu mới về tích hợp ontology và sử dụng luật kết hợp trong hệ thống tìm kiếm thông minh có tương tác. Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh thực sự hữu ích trong thực tế, giúp người sử dụng dễ dàng lựa chọn các truy vấn phù hợp với ý định của họ mà không cần nhớ toàn bộ truy vấn.

Các đề xuất của luận án còn có thể ứng dụng trong các hệ thống khác nhau, như: Đề xuất phương pháp trích rút bộ ba (triple) dựa trên mô hình cú pháp, trích rút các bộ ba từ dữ liệu văn bản phục vụ cho việc xây dựng đặc trưng của văn bản có thể sử dụng trong các hệ thống phân loại, phận cụm và tóm tắt văn bản; Đề xuất phương pháp tìm kiếm đa diện dữ liệu văn bản và cá nhân hoá tìm kiếm đa diện sử dụng dữ liệu định hướng xử lý nhập nhằng của *Wikipedia Disambiguation* có thể áp dụng cho các hệ thống khuyến nghị, các hệ chuyên gia.

7. Cấu trúc của luận án

 Luận án gồm 4 chương như sau:

Chương 1. Kiến thức cơ sở

Chương 2. Tìm kiếm ngữ nghĩa dựa trên nội dung văn bản

Chương 3.Tìm kiếm ngữ nghĩa có tương tác

Chương 4. Xây dựng Hệ thống tra cứu, tìm kiếm ngữ nghĩa thông tin bệnh

CHƯƠNG 1. KIẾN THỨC CƠ SỞ

1.1. Ontology

 **Khái niệm:**Theo [36], trong triết học Ontology được định nghĩa là “*Triết lý về sự tồn tại*”. Trong lĩnh vực tin học, Tom Gruber[36] định nghĩa Ontology là “*một biểu diễn tường minh các khái niệm*”, *“một đặc tả tường minh, hình thức và chia sẻ về các khái niệm dùng chung*”.

 **Ứng dụng Ontology:** Ontology cho phép cấu trúc hoá các mối quan hệ giữa các đối tượng, thuộc tính, sự kiện, quá trình trong thế giới thực [71].

 **Các ngôn ngữ Ontology:**RDF,DAML+OIL,OWL

 **Tiến trình xây dựng ontology:** gồm 04 giai đoạn

1.2. Trích rút thông tin từ văn bản

1.2.1. Trích rút đặc trưng phổ biến

1.2.1.1. TF-IDF trích rút thông tin văn bản

 TF-IDF là mô hình được sử dụng rộng rãi trong việc trích rút thông tin văn bản. Mô hình trích rút đặc trưng văn bản TF-IDF gắn với mỗi từ một con số thống kê thể hiện mức độ quan trọng của từ này trong văn bản, trong ngữ cảnh văn bản nằm trong một tập hợp nhiều văn bản. TF tính tần số xuất hiện của từ trong văn bản. Các tài liệu dài ngắn khác nhau sẽ dẫn đến số lần xuất hiện của một từ khác nhau. Do đó, tần số xuất hiện của một từ phải được chuẩn hóa bằng cách chia cho độ dài tài liệu [64].

1.2.1.2. Sử dụng NER trong trích rút thông tin văn bản

 NER cho phép xác định danh từ riêng trong văn bản và phân loại chúng vào các lớp thực thể có tên tương ứng [64]. Xác định thực thể có tên cũng là một nhiệm vụ quan trọng trong các nhiệm vụ xử lý ngôn ngữ tự nhiên.

1.2.2. Trích rút quan hệ ngữ nghĩa trong văn bản

Hệ thống *Open Information Extraction*(OpenIE) trích rútcác bộ ba (arg1, rel, arg2) từ văn bản dựa trên các quan hệ với động từ, ở đây arg1 và arg2 là các đối số của quan hệ và rel là quan hệ ngữ nghĩa. Trong khi các hệ thống IE khác chỉ tập trung vào tập các quan hệ đã được định nghĩa trước, hệ thống OpenIE dựa trên các phương pháp khai thác không có giám sát. Do đó số lượng các quan hệ sẽ linh động. Điểm mạnh của OpenIE là không yêu cầu phải cung cấp dữ liệu đã dán nhãn

1.3. Các mô hình tìm kiếm

1.3.1. Tìm kiếm tương tác

 Tìm kiếm tương tác (*Interactive search*)[4], [101], [105] là phương pháp tìm kiếm dựa vào sự tương tác của người dùng với hệ thống, từ đó hệ thống đưa ra những kết quả tương ứng cho phù hợp với ý định tìm kiếm của người dùng.

1.3.2. Tìm kiếm ngữ nghĩa

1.3.2.1. Giới thiệu về tìm kiếm ngữ nghĩa

 Tìm kiếm ngữ nghĩa *(Semantic search)* [36], [38], [75] hướng tới tăng độ chính xác của việc trích xuất thông tin bằng cách nắm được ý định tìm kiếm của người dùng, cũng như là ngữ cảnh trong không gian tìm kiếm. Hệ thống tìm kiếm ngữ nghĩa xem xét rất nhiều nội dung gồm có: ngữ cảnh tìm kiếm, địa điểm, ý định người dùng, quan hệ ngữ nghĩa giữa các khái niệm.

*1.3.2.2. Các công trình nghiên cứu về tìm kiếm ngữ nghĩa*

Nhìn chung, các nghiên cứu về tìm kiếm dựa trên ngữ nghĩa hiện nay chủ yếu tập trung cải thiện hiệu quả tìm kiếm theo bốn hướng chính: Khai thác những nguồn tri thức như WordNet, UMLS, Sensus; Trích rút thông tin từ tài liệu và mở rộng câu truy vấn; Sử dụng các kỹ thuật khác để hỗ trợ quá trình tìm kiếm như xử lý ngôn ngữ tự nhiên, logic mờ (*fuzzy*), khử nhập nhằng, phân loại (*classification*); Xây dựng, biểu diễn và so khớp các cấu trúc khái niệm.

*1.3.2.3. Tìm kiếm đa diện (faceted search)*

Tìm kiếm đa diện là một kỹ thuật trong tìm kiếm ngữ nghĩa cho phép truy xuất thông tin đã được đánh chỉ mục dưới dạng hệ thống phân lớp đa diện (facet).

1.4. Luật kết hợp

Luật kết hợp (*Association Rules*) được sử dụng rộng rãi để biểu thị sự kết hợp trong các dòng dữ liệu [97]. Quan sát một lượng lớn dữ liệu các phiên giao dịch, người ta muốn tìm ra được các quy luật, phục vụ cho các dự đoán [54]. Nghiên cứu về luật kết hợp đóng vai trò quan trọng trong số các phương pháp khai phá dữ liệu [97].

1.5. Kết chương

Chương 1 đã trình bày các kiến thức cơ bản về ontology; các phương pháp trích rút thông tin từ văn bản; các mô hình tìm kiếm và phương pháp khai phá luật kết hợp làm cơ sở cho đề xuất tích hợp ontology và hỗ trợ tương tác với người sử dụng trong hệ thống tìm kiếm ngữ nghĩa.

CHƯƠNG 2. TÌM KIẾM NGỮ NGHĨA DỰA TRÊN NỘI DUNG VĂN BẢN

2.1. Trích rút triple dựa trên mô hình ngữ pháp và từ vựng

2.1.1.Trích rút triple

Trong [CTLA3], luận án đề xuất phương pháp phân tích cấu trúc ngữ pháp tiếng Anh dựa vào loại động từ, cụm động từ để phân tích một mệnh đề thành <subject, predicate>, sau đó sử dụng mô hình cú pháp (Syntax Model) để phân tích 02 thành phần này một cách chi tiết hơn. Mô hình cú pháp không những thể hiện được quan hệ giữa chủ ngữ và tân ngữ trong câu mà còn thể hiện được quan hệ giữa các từ khác trong câu, vì vậy sẽ tăng khả năng trích rút được các bộ ba phù hợp trong một mệnh đề.

2.1.2. Quy trình trích rút thông tin dữ liệu văn bản

1. NER nhận diện các thực thể có tên.
2. Tính tần suất xuất hiện và độ quan trọng của các từ xuất hiện trong toàn bộ văn bản TF-IDF.
3. Trích rút *triple* sử dụng kết quả của NER và xác lập quan hệ ngữ nghĩa giữa các đối tượng trong câu, các câu trong văn bản.
4. Biểu diễn văn bản dưới dạng cấu trúc các đối tượng có mối quan hệ ngữ nghĩa.

***2.1.3. Đánh giá***

Phương pháp đề xuất trích rút chính xác gấp 1.7 lần so với ClausIE. So với hệ thống của Ollie, phương pháp đề xuất đạt gấp 2.7–2.8 lần

2.2. Tìm kiếm đa diện dữ liệu văn bản

 Trong [CTLA5], luận án đã đề xuất một phương pháp để xây dựng công cụ tìm kiếm đa diện dữ liệu văn bản gồm các bước sau:

1. Xây dựng tham chiếu định hướng xử lý nhập nhằng sử dụng Wikipedia (*Wikipedia Disambiguation*).
2. Xây dựng không gian tìm kiếm ngữ nghĩa sử dụng tham chiếu định hướng xử lý nhập nhằng.

2.2.1. Xây dựng tham chiếu định hướng xử lý nhập nhằng sử dụng Wikipedia

Luận án sử dụng các trang dữ liệu định hướng xử lý nhập nhằng để xây dựng các facet tìm kiếm định hướng xử lý nhập nhằng.

2.2.2. Xây dựng không gian tìm kiếm ngữ nghĩa sử dụng tham chiếu định hướng xử lý nhập nhằng

Không gian tìm kiếm ngữ nghĩa được xây dựng bằng cách lưu trữ các tài liệu theo tham chiếu định hướng xử lý nhập nhằng. Để tính toán độ tương tự giữa tài liệu và khái niệm (từ, cụm từ) tham chiếu đến, luận án sử dụng phép đo độ tương tự cosin:

là đo độ tương tự giữa văn bản *dj* và khái niệm cụ thể *q, t*rong đó, *Wj* là vector đặc trưng của văn bản *dj*, *Wq* là vector đặc trưng của *q*.

2.2.3. Đánh giá kết quả trích rút dữ liệu định hướng lý nhập nhằng từ Wikipedia Disambiguation

Luận án đã tiến hành đánh giá kết quả trích rút dữ liệu định hướng xử lý nhập nhằng từ Wikipedia Disambiguation bằng cách so sánh dữ liệu thô với kết quả được trích rút. Kết quả độ chính xác là trên 93%.

2.3. Phương pháp cá nhân hóa facet trong tìm kiếm đa diện

Trong [CTLA4], luận án đề xuất phương pháp cá nhân hóa tìm kiếm đa diện sử dụng dữ liệu định hướng xử lý nhập nhằng (*Wikipedia Disambiguation)* và mạng xã hội nhằm đưa ra kết quả tìm kiếm phù hợp với profile của người sử dụng gồm các bước sau:

• Bước 1: Chuẩn bị dữ liệu, thực hiện các bước lấy và xử lý dữ liệu định hướng xử lý nhập nhằng (*Wikipedia Disambiguation)*.

• Bước 2: Chuẩn bị hồ sơ người dùng, tạo hồ sơ người dùng lấy từ hồ sơ người dùng Facebook.

Kết quả cho thấy việc sử dụng profile của người dùng đã giúp xác định được các kết quả tìm kiếm phù hợp với người sử dụng.

2.4. Kết chương

 Chương 2 đã đề xuất phương pháp rút trích triple dựa trên mô hình cú pháp dùng để thu thập thông tin phục vụ xây dựng cơ sở tri thức; đề xuất phương pháp tìm kiếm đa diện dữ liệu văn bản dựa trên dữ liệu định hướng xử lý nhập nhằng của Wikipedia Disambiguationvà đề xuất phương pháp cá nhân hoá tìm kiếm đa diện. Tìm kiếm đa diện cho phép phân loại kết quả theo các facet rất phù hợp với việc sử dụng ontology trong các hệ thống tìm kiếm ngữ nghĩa có tương tác.

CHƯƠNG 3.TÌM KIẾM NGỮ NGHĨA CÓ TƯƠNG TÁC

3.1. Sử dụng luật kết hợp trong tìm kiếm ngữ nghĩa thông tin bệnh

Để xác định mối tương quan giữa các triệu chứng bệnh với nhau, luận án sử dụng luật kết hợp để khai thác mối quan hệ trong dữ liệu, cụ thể: *Một bệnh có nhiều triệu chứng, một triệu chứng có thể xuất hiện ở nhiều bệnh khác nhau. Một số triệu chứng có mối quan hệ kết hợp, cùng xuất hiện trong một bệnh.* Từ đó luận án phát biểu bài toán [CTLA2] như sau:

* *D* là tập các căn bệnh, *D={d1, ..., dn},* với *di, i=1,2,…,n* là các căn bệnh.
* *T* là một tập triệu chứng bao gồm các triệu chứng khác nhau *T={t1, ..., tm}*, với *tj, j = 1,2,...,m* là các triệu chứng.

Mỗi căn bệnh *di* ứng với một tập các triệu chứng *tj (tj ⊂ T)*

Như vậy vấn đề cần giải quyết là:*Tìm tập các triệu chứng thường xuyên xuất hiện cùng nhau trong một bệnh, tính tương quan, tính kết hợp giữa các tập triệu chứng.*

3.2. Luật kết hợp trong Ontology bệnh (ASO-Association rule in DiseaseOntology)

Luận án đề xuất một phương pháp khai phá luật kết hợp trong Ontology bệnh, hỗ trợ người dùng lựa chọn các triệu chứng, lựa chọn facet trong tìm kiếm. Cụ thể:

- Luật kết hợp giữa các triệu chứng được sử dụng để gợi ý các triệu chứng tiếp theo cho người sử dụng.

- Luật kết hợp ngữ nghĩa giữa các quan hệ (subclasses, properties) trong Ontology bệnh. Các quan hệ subclasses được sử dụng trong tìm kiếm đa diện, mỗi lớp con là một facet. Các quan hệ properties được sử dụng để gợi ý người dùng lựa chọn thuộc tính (property) tiếp theo.

Để khai phá các luật kết hợp trên, luận án sử dụng dữ liệu đầu vào là: Ontology bệnh, mỗi bệnh sẽ có một danh sách các triệu chứng tương ứng và Ontology triệu chứng [113], trong đó giữa các triệu chứng có 02 mối quan hệ tổng quát hơn (“*more\_generation*”), cụ thể hơn (“*more\_specification*”.

**\* Thuật toán ASO-Apriori dựa trên 02 nguyên tắc sau:**

**+ Nguyên tắc 1:** Nếu một tập phần tử A là tập phổ biến (*frequent itemset*), thì không tồn tại tập con phần tử thuộc tập phần tử A có mối quan hệ “*more\_specification*” hoặc “*a\_part\_of*” với một tập phần tử không phổ biến khác.

**Định nghĩa 3.1:** Một phần tử (triệu chứng) *y* được gọi là **InstanceOf** của phần tử *x* nếu như phần tử *y* có mối quan hệ “*more\_specification*” hoặc “*a\_part\_of*” với phần tử *x*.

Ký hiệu là: *y IsOf x.*

Ví dụ, triệu chứng *severe abdominal cramp* là *IsOf* của triệu chứng *abdominal cramp* do triệu chứng *severe abdominal cramp* có quan hệ *more\_specification* với triệu chứng *abdominal cramp.*

**Định nghĩa 3.2:** Một tập phần tử (tập triệu chứng) *Y=y1y2…yk* được gọi là InstanceOf của tập phần tử *X=x1x2…xk* nếu như *yi=xi* hoặc *yi* có mối quan hệ “*more\_specification*” hoặc “*a\_part\_of*” với phần tử *xi*với mọi *i=1..k*.

Ký hiệu là: *Y IsOf X.*

Ví dụ, tập triệu chứng {*severe abdominal cramp, abdominal discomfort*} là *IsOf* của tập {*abdominal cramp, abdominal discomfort*} vì triệu chứng *severe abdominal cramp* có quan hệ *more\_specification* với triệu chứng *abdominal cramp.*

**Định nghĩa 3.3:** Cho một tập phần tử *X*, giao dịch *Ti*. Giao dịch *Ti*được gọi là hỗ trợ mở rộng đối với tập *X* nếu như tồn tại một tập *Y* là *IsOf* của *X* thoả mãn *Y⊆Ti*

**Định nghĩa 3.4:** Độ hỗ trợ mở rộng (*Supportext*)

Độ hỗ trợ mở rộng (*Supportext*) của tập phần tử *X* là tỷ số giữa số lượng giao dịch trong CSDL *D* hỗ trợ mở rộng *X* và tổng số giao dịch của D (gọi là N).

trong đó:

- *δ(X\*)* : số lượng giao dịch trong *D* có chứa *X\**.

**Định nghĩa 3.5:** Độ tin cậy mở rộng (*Confidenceext*)

Độ tin cậy mở rộng (*Confidenceext*) của *XY* là tỷ số giữa số lượng giao dịch trong CSDL *D* hỗ trợ mở rộng tập *{X∪Y}* và số lượng giao dịch trong CSDL *D* hỗ trợ mở rộng tập *X*.

Trong đó:

* *δ()*: số lượng giao dịch trong *D* có chứa .
* *δ()* : số lượng giao dịch trong *D* có chứa .

**Mệnh đề 3.1:** Tính chất Apriori mở rộng.

Cho hai tập triệu chứng *X, Y*. Nếu *X⊆Y* thì ≤

Chứng minh:

Xét giao dịch *Ti* bất kỳ trong CSDL *D*, nếu *Ti* hỗ trợ mở rộng tập *Y* thì *Ti* hỗ trợ mở rộng tập *Y*. Thật vậy:

*Ti* hỗ trợ *Y*, tức là tồn tại một *Y\** là *IsOf Y* mà *Y\*⊆Ti*

Từ *Y\** là *IsOf Y* và *X⊆Y* suy ra tồn tại *X\**thoả mãn *X\*⊆Y\**và*X\** là *IsOf X*.

Như vậy, theo Định nghĩa 3.3 ta có *Ti* hỗ trợ mở rộng *X* (đpcm).

**Định nghĩa 3.6:** Tập triệu chứng phổ biến.

Một tập triệu chứng *X* được gọi là tập triệu chứng phổ biến nếu như ≥ *min\_sup.*

**Định nghĩa 3.7:** Luật kết hợp trên tập triệu chứng

 được gọi là luật kết hợp trên tập triệu chứng nếu như ≥ *minSup* và ≥*min\_conf*

**+ Nguyên tắc 2:** Ontology hàm chứa các luật kết hợp ngữ nghĩa thể hiện qua các mối quan hệ.

**\* Tư tưởng chính của thuật toán ASO-Apriori**

- Tìm tất cả các tập triệu chứng phổ biến từ CSDL *D* và tập *I,*trong đó CSDL *D* chứa các giao dịch tương ứng với các bệnh, mỗi giao dịch chứa một tập triệu chứng. Tập *I* chứa tất cả các triệu chứng và mối quan hệ: “*more\_specification*”, “*a\_part\_of*” giữa các triệu chứng.

- Sinh luật kết hợp từ tập triệu chứng phổ biến.

- Sinh ra luật kết hợp ngữ nghĩa dựa trên các mối quan hệ trong Ontology bệnh.

**Thuật toán ASO-Apriori có 02 điểm mới sau đây:**

- *ASO-Apriori sử dụng* ***độ hỗ trợ mở rộng và độ tin cậy mở rộng*** *để xác định tập triệu chứng phổ biến và luật kết hợp giữa các triệu chứng.*

*- Khai phá luật kết hợp ngữ nghĩa dựa trên các mối quan hệ trong Ontology bệnh.*

 **\* Thuật toán ASO-Apriori**

**Giai đoạn 1: Sinh tập triệu chứng phổ biến - ASO-Apriori-GenFrequentSymptom**

1. Duyệt *CSDL,* tính độ hỗ trợ mở rộng của tập chỉ gồm 1 triệu chứng phổ biến (frequent 1-symptom), so sánh với *min\_sup*, để có được tập tất cả các tập 1 triệu chứng phổ biến (gọi là *L1*).

2.1. Sử dụng *Lk-1*: Nối (join) các cặp *Lk-1* để sinh ra các tập ứng viên *k*-triệu chứng, gọi là *Ck*.

2.2. Loại bỏ các ứng viên *k*-triệu chứng dựa trên **mệnh đề 3.1.**

2.3. Duyệt CSDL để xác định độ hỗ trợ mở rộng của mỗi candidate k-symptom, so sánh với min\_sup để nhận được tập *k*-triệu chứng phổ biến frequent *k*–symptom (gọi là *Lk*).

3. Lặp lại từ bước 2 cho đến khi tập ứng viên *k*-triệu chứng *Ck* trống.

4. Trả về tập tất cả các tập triệu chứng phổ biến.

**Giai đoạn 2**:

Đầu vào: Tập tất cả các tập triệu chứng phổ biến *Lk*, ngưỡng minconf.

Đầu ra: Tập luật

**ASO-Apriori\_GenRules (*Lk*)** {

**for** (frequent symptomset *lk ,k*≥ 2) {

 call **Genrules** (*lk, lk*);

}

**Genrules**(lk:frequent *k*-symptomset; *am*: frequent *m*-symptomset) {

*A={(m-l)-symptomset am-1 | am-1⊂ am};*

***for****( am-1∈ A ) {*

 *conf =supportexxt (lk)/supportexxt (am-1);*

***if*** *(conf ≥ minconf) {*

 *output the rule am-1⇒ (lk –am-1),*

***With*** *confidence =conf*

*and support=supportext (lk);*

***if*** *(m-1> l) {*

 *call* ***Genrules****(lk, am-1);*

 *}*

*}*

**- Sinh các luật kết hợp ngữ nghĩa theo Nguyên tắc 2*:***

Đầu vào: Ontology *O*, bệnh *d*;

Đầu ra: Tập các quan hệ *R* (subclasses, properties) của bệnh *d* trong *O*.

***ASO-Apriori\_GenRulesFromOntology (O, d) {***

*R=;*

 *for (quan hệ r ∈ (O,d) ) { // xét từng mối quan hệ của bệnh d trong O;*

 *R=R∪ r;// luật kết hợp dạng: d r;*

 *}*

 *return R;*

 ***}***

3.3. Đánh giá

3.3.1. Phương pháp đánh giá độ chính xác và độ bao phủ

Sử dụng phương pháp đánh giá trong truy hồi thông tin với tiêu chí đánh giá *F* là sự kết hợp của hai tiêu chí đánh giá *Precision* (độ chính xác) và *Recall* (độ bao phủ):

+ Độ chính xác (*Precision*): Trong tập hợp tìm được thì có bao nhiêu phần tìm được là đúng.

+ Độ bao phủ (*Recall*): Trong số phần tử tồn tại thì tìm ra được bao nhiêu phần tử.

3.3.2. Kết quả

- Tổng số khái niệm bệnh: 9.801

- Tổng số triệu chứng: 3.098

- Tổng số *Transaction*: 1.798

Nếu xem mỗi khái niệm bệnh là một transaction thì ta có đến 9.801 transaction, tuy nhiên ở đây NCS chỉ chọn lọc 1.798 bởi vì đa phần các khái niệm bệnh khác không bao gồm các triệu chứng bệnh. Với min\_support = 2 và Confidence = 1, thì tổng số luật kết hợp (Association rule) là 692. NCS sử dụng 300 bệnh thường gặp đã được rút trích ở trên để thực nghiệm hệ thống với hai giải thuật khác nhau: Apriori và ASO-Apriori.

3.4. Kết chương

Chương 3 đã trình bày phương pháp sử dụng luật kết hợp trong tìm kiếm ngữ nghĩa thông tin bệnh và đề xuất phương pháp mới khai phá luật kết hợp trong Ontology bệnh được gọi là ASO-Apriori.

CHƯƠNG 4. XÂY DỰNG HỆ THỐNG TRA CỨU, TÌM KIẾM NGỮ NGHĨA THÔNG TIN BỆNH

4.1. Tổng quan hệ thống tìm kiếm ngữ nghĩa thông tin bệnh

 Hình 4.1 dưới đây mô tả các thành phần của Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh. Các nội dung tiếp theo sẽ chỉ ra các thành phần của hệ thống.



Hình 4.1. Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh

4.2. Xây dựng Ontology bệnh

4.2.1. Lựa chọn disease ontology

Ontology bệnh (*Disease Ontology* – DO) cho phép tích hợp ngữ nghĩa các loại bệnh và từ vựng y tế trên cơ sở kết hợp các kho dữ liệu bệnh như: DO (*http://disease-ontology.org*), MeSH, ICD, từ điển đồng nghĩa NCI, SNOMED và OMIM [10-13]. Với nhiều ưu điểm và sự tiện lợi của CSDL này, CSDL DO được Việt hoá để dùng làm thông tin. Mục tiêu là cung cấp cho cộng đồng y sinh các đặc điểm miêu tả bệnh một cách nhất quán, có khả năng tái sử dụng và bền vững.

4.2.2. Dịch tự động Google Translation

 Dịch tự động Google Translation cho phép cập nhật tự động do các phiên bản ontology luôn phát triển mới.

4.2.3. Tinh chỉnh bán tự động

 Mục tiêu của giai đoạn này là tạo được cấu trúc cơ bản ban đầu cho Ontology bệnh, dựa trên các tài liệu chuẩn như ICD,… Ontology này sau đó sẽ được dùng để so khớp với DO thông qua mã ICD hoặc tên tiếng Anh nhằm chỉnh sửa lại các lỗi dịch tự động.

4.2.4. Chỉnh sửa cộng tác Ontology bệnh

 Luận án dùng phương pháp thống kê với hai cách tiếp cận: tiếp cận dữ liệu và tiếp cận đồng thuận của Bác sĩ (phương pháp chuyên gia) trong xây dựng Ontology bệnh. Đầu tiên, tên của 6.000 bệnh sẽ được kiểm tra để xem có tồn tại tên bệnh qua các hệ thống tìm kiếm. Các tên bệnh không tồn tại trong tìm kiếm sẽ được các bác sĩ phiên dịch dựa trên đồng thuận.

a. Tiếp cận theo dữ liệu

Theo tiếp cận dữ liệu, có hai giả thuyết như sau:

1. Tên bệnh được dịch chính xác là tên bệnh tồn tại, được người dùng thường hỏi đáp, chia sẻ trên các hệ thống Internet.
2. Tên bệnh được dịch chính xác là tên bệnh tồn tại, được các chuyên gia hay tổ chức y tế đề cập trong các tài liệu trên mạng Internet.

b. Tiếp cận đồng thuận của các Bác sĩ

1. Xây dựng nhóm cộng tác, gồm nhiều Bác sĩ ở Bệnh viện Trung Ương Huế.
2. Ban đầu, các bác sĩ làm việc độc lập, bằng phương pháp thống kê hoặc kinh nghiệm có được, các bác sĩ đề xuất ra 300 bệnh thường gặp độc lập.
3. Người điều phối sẽ tổng hợp lại theo một thống kê để tính toán sự đồng thuận của nhóm.
4. Sau đó, trả kết quả về cho từng người, mỗi người tiếp tục chỉnh sửa kết quả của mình cho phù hợp với kết quả đạt được. Kết quả đạt được xem như một tham khảo nhất định.
5. Quay lại bước 2, cho đến khi có sự đồng thuận giữa các bác sĩ tham gia dịch các bệnh chưa tồn tại trên các hệ thống tìm kiếm.
6. Kết thúc quá trình cộng tác.

4.3. Thu thập dữ liệu

Hơn 3.000 tài liệu về các bệnh được thu thập tại các thư viện uy tín và các trường đại học y danh tiếng trên địa bàn thành phố Hồ Chí Minh như Thư viện tổng hợp thành phố Hồ Chí Minh, Thư viện Đại học Y Dược, Thư viện Đại học Y Phạm Ngọc Thạch, ...Trong CSDL hiện có hơn 6.000 khái niệm bệnh đã được Việt hóa.Đây là một CSDL rất quý giá cho việc xử lý các truy vấn và tìm kiếm các thông tin về các căn bệnh đang có. Các file mềm liệt kê danh sách các bệnh tật đã được nhóm nghiên cứu cập nhật và miêu tả trong CSDL có trong hồ sơ minh chứng.

4.4. Xây dựng phương pháp tìm kiếm ngữ nghĩa có tương tác điều hướng luồng tìm kiếm

Khi hệ thống chưa có nhiều người dùng, các tương tác giữa người dùng và hệ thống chưa nhiều, sử dụng luật kết hợp giữa các triệu chứng sẽ cho phép điều hướng luồng tìm kiếm, đưa ra gợi ý cho người dùng các triệu chứng bệnh tiếp theo dựa vào những triệu chứng ban đầu mà người dùng đã cung cấp.

4.5. Xây dựng hệ thống

4.5.1. Cấu trúc cơ sở dữ l iệu

Để sử dụng ontology *DO*, thì *DO* cần phải được lưu trữ trong một hệ quản trị CSDL, ở đây luận án thiết kế lược đồ CSDL bệnh, và sử dụng MySQL cho việc quản trị dữ liệu bệnh.

4.5.2. Khung tìm kiếm

a. Gợi ý từ khóa tìm kiếm

Chức năng *Gợi ý từ khóa tìm kiếm* cung cấp tính năng gợi ý đề nghị để tăng tính tương tác giữa người sử dụng và hệ thống.

b. Gợi ý từ khóa liên quan

Chức năng *Gợi ý từ khóa liên quan* nhằm giới thiệu các triệu chứng có liên quan hoặc thuộc bệnh khác, ví dụ khi người sử dụng lựa chọn một số từ khóa tìm kiếm “sốt cao” (*fever*), “đau đầu” (*headache*) và họ dừng lại để suy nghĩ, tìm từ khóa tiếp theo; lúc này hệ thống sẽ giới thiệu các triệu chứng liên quan với các triệu chứng gợi ý là “co giật” (*convulsions*).

c. Tìm kiếm kết quả bệnh

Khi người dùng thực hiện tìm kiếm, hệ thống sẽ tìm bệnh tương ứng với những triệu chứng bệnh được người dùng đã nhập vào: Với dữ liệu bệnh có trong hệ thống, hệ thống sẽ chọn ra dữ liệu bệnh tương ứng được sắp xếp theo số lượng đã xem và được đánh giá; Với dữ liệu bệnh không có trong hệ thống (chưa được cập nhật CSDL bệnh), thì hệ thống sẽ tìm kiếm trên Google.

*d. Hiển thị kết quả tìm kiếm*

 Kết quả tìm kiếm sẽ hiển thị theo từng khối cho mỗi bệnh, và được sắp xếp từ trái sang phải, từ trên xuống dưới.

* Phần trên hiển thị ảnh đại diện của bệnh.
* Phần giữa hiển thị tên bệnh, và số lượng người xem bệnh này.
* Phần cuối hiển thị đánh giá của người dùng về nội dung bệnh.

4.6. Kết chương

 Chương 4 trình bày Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh có hỗ trợ tương tác với người sử dụng bằng các gợi ý dựa trên tập luật khai thác được từ Ontology triệu chứng và Ontology bệnh [CTLA1].

KẾT LUẬN VÀ CÁC NGHIÊN CỨU TIẾP THEO

1. Kết luận

 Luận án đã nghiên cứu các phương pháp xây dựng một hệ thống tìm kiếm ngữ nghĩa và ứng dụng xây dựng Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh. Hệ thống này mang lại hiệu quả cho việc tìm kiếm thông tin về bệnh khi ban đầu chưa có đầy đủ thông tin đầu vào. Quá trình tương tác giữa người dùng và Hệ thống dựa trên luật kết hợp mang lại hiệu quả cao trong quá trình suy diễn hơn so với việc tìm kiếm dựa trên luật kết hợp truyền thống. Điều này cũng dễ dàng lý giải bởi bản thân Ontology bệnh đã hàm chứa các luật kết hợp bên trong nó – quan hệ giữa các khái niệm và các thuộc tính. Ngoài ra, Ontology bệnh giúp việc phân loại các tài liệu bệnh theo một phân cấp ngữ nghĩa, điều này dễ dàng cho việc tìm kiếm. Các kết quả chính của luận án gồm:

(1) Đề xuất phương pháp trích rút bộ ba (triple) dựa trên mô hình cú pháp, trích rút các bộ ba từ dữ liệu văn bản phục vụ cho việc xây dựng đặc trưng của văn bản;

(2) Đề xuất phương pháp tìm kiếm đa diện dữ liệu văn bản và cá nhân hoá tìm kiếm đa diện sử dụng dữ liệu định hướng xử lý nhập nhằng của *Wikipedia Disambiguation*;

(3) Đề xuất phương pháp khai phá luật kết hợp trong Ontology bệnh (ASO-Apriori) dựa trên 02 độ đo mới: độ hỗ trợ mở rộng và độ tin cậy mở rộng và khai phá luật kết hợp ngữ nghĩa giữa các mối quan hệ trong Ontology bệnh;

(4) Xây dựng Ontology bệnh tiếng Việt;

(5) Đề xuất phương pháp và xây dựng Hệ thống tìm kiếm ngữ nghĩa thông tin bệnh có hỗ trợ tương tác với người sử dụng bằng các gợi ý dựa trên tập luật kết hợp giữa các triệu chứng và luật kết hợp ngữ nghĩa từ các mối quan hệ trên Ontology bệnh.

2. Các nghiên cứu tiếp theo

Tiếp tục nghiên các phương pháp trích rút thông tin văn bản, các phương pháp khai phá luật kết hợp trong ontology, xử lý nhập nhằng và điều hướng luồng tìm kiếm sử dụng ontology.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ CỦA TÁC GIẢ CÓ LIÊN QUAN ĐẾN LUẬN ÁN

1. **Nguyễn Hồng Sơn**, Dương Trọng Hải, Hoa Tất Thắng, *Hệ thống tương tác thông minh trong tra cứu, tìm kiếm thông tin bệnh*, Tạp chí Nghiên cứu Khoa học và Công nghệ Quân sự, Số 53, 02/2018, trang 160-169, Viện Khoa học và Công nghệ Quân sự, Hà Nội.
2. **Hong Son Nguyen**, Minh Hieu Le, Chan Quan Loi Lam, Trong Hai Duong, *Smart interactive search for Vietnamese disease by using data mining-based ontology*, Journal of Information and Telecommunication, volume 1 - issue 2, 2017, pages 176-191.
3. Huỳnh Ái Loan, **Nguyễn Hồng Sơn**, Dương Trọng Hải, *Triple Extraction Using Lexical Pattern-based Syntax Model*, Advanced Computational Methods for Proceedings of the 4th International Conference on Computer Science, Applied Mathematics and Applications, ICCSAMA 2016, 2-3 May, 2016, Vienna, Austria, 2016, vol 453, pages 265-279.
4. **Nguyễn Hồng Sơn**, Phạm Hồng Phúc, Dương Trọng Hải, Nguyễn Thị Phương Trang, *Personalized Facets for Faceted Search Using Wikipedia Disambiguation and Social Network*, Advanced Computational Methods for Knowledge Engineering: Proceedings of the 4th International Conference on Computer Science, Applied Mathematics and Applications, ICCSAMA 2016, 2-3 May, 2016, Vienna, Austria, 2016, vol 453, pages 229-241.
5. Đặng Đức Bình, **Nguyễn Hồng Sơn**, Nguyễn Thanh Bình, Dương Trọng Hải, *A framework of faceted search for unstructured documents using wiki disambiguation*, Computational Collective Intelligence, ICCCI 2015, Springer International Publishing, 2015, pages 502-511.