Chương 2 : CÁC NGHIÊN CỨU VỀ LẬP CHỈ MỤC TRÊN KHÁI NIỆM

2.1 Tổng quan

Việc lập chỉ mục theo khái niệm là rút trích các khái niệm có trong nội dung văn bản để làm chỉ mục biểu diễn cho nội dung văn bản. Việc rút trích khái niệm có thể được thực hiện theo nhiều phương pháp mà một trong những phương pháp đó là sử dụng một Ontology cho trước. Tuy nhiên không phải khi nào cũng có sẵn Ontology phù hợp và việc xây dựng một Ontology phù hợp đòi hỏi nhiều thời gian, công sức cũng như kinh phí. Vì đó nhiều công trình lập chỉ mục trên khái niệm đã tìm các giải pháp sao cho không cần so khớp tài liệu với Ontology. Từ đó việc lập chỉ mục trên khái niệm chia ra 2 hướng tiếp cận lớn :

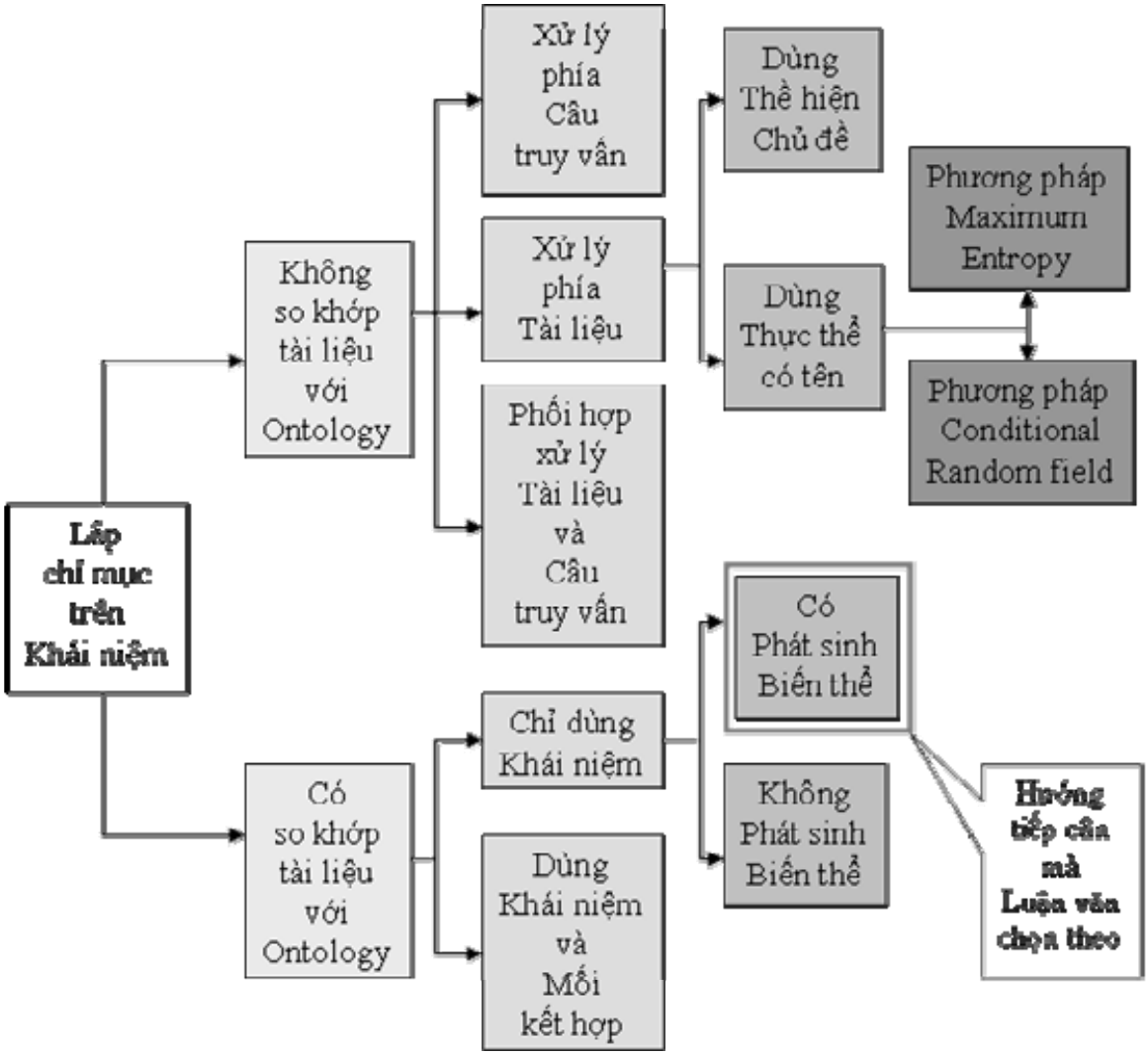
i. Lập chỉ mục trên khái niệm không so khớp tài liệu với Ontology

ii. Lập chỉ mục trên khái niệm có so khớp tài liệu với Ontology

Trong hướng tiếp cận không so khớp tài liệu với Ontology, các giải pháp có thể tập trung xử lý trên tài liệu như xu hướng thông thường. Tuy nhiên cũng có giải pháp tập trung xử lý câu truy vấn như [41] hoặc phối hợp xử lý trên cả tài liệu lẫn câu truy vấn để tăng hiệu quả lập chỉ mục như [39]. Vì câu truy vấn rất thiếu thông tin ngữ cảnh nên nó cần được viết theo khuôn mẫu cho trước thì quá trình xử lý mới khả thi. Ngược lại, tài liệu rất giàu ngữ cảnh và có thể được xử lý bằng nhiều kỹ thuật khác nhau : Dùng Thể hiện chủ đề (Thematic Representation) như [4] hoặc dùng Thực thể có tên (Named Entity) như [8, 12, 43].

Trong hướng tiếp cận có so khớp tài liệu với Ontology, các giải pháp có thể chỉ dùng cụm từ gốc (cụm từ thực sự hiện diện trong tài liệu) như [11] hoặc dùng cả những biến thể của chúng như [1, 2, 9, 23, 35, 42]. Bên cạnh đó cũng có những công trình tận dụng cả mối kết hợp giữa các khái niệm trong Ontology nhằm giúp cho việc truy vấn được chi tiết hơn như [26, 28, 29, 36].

Tổng quan về các hướng tiếp cận và mối quan hệ phân cấp giữa chúng được mô tả trong hình 2-1 sau đây :



Hình 2-1 : Lược đồ tổng quan các hướng tiếp cận lập chỉ mục trên khái niệm

2.2 Lập chỉ mục trên khái niệm không so khớp tài liệu với Ontology

2.2.1 Hướng tiếp cận xử lý phía câu truy vấn

Hướng tiếp cận này chủ trương rằng về phía tài liệu chỉ cần lập chỉ mục bằng từ khóa như cách truyền thống. Việc rút trích khái niệm được xử lý hoàn toàn ở phía câu truy vấn.

Trong [41], các tác giả khảo sát nhu cầu truy vấn tài liệu Y khoa của các bác sĩ và thiết kế các mẫu câu truy vấn. Người dùng phải truy vấn theo đúng các mẫu như hình 2-2 sau đây :

Hình 2-2 : Cấu trúc mẫu câu truy vấn

Vì các mẫu câu truy vấn là có cấu trúc nên dễ dàng trích ra các khái niệm trong câu truy vấn mà người dùng đưa ra. Các khái niệm trong nhóm A và C nằm trong số 190,000 khái niệm của danh mục MeSH (Medical Subject Heading) hay một trong số 1,700,000 gien trong CSDL gien Entreze.

Các biến thể sẽ được [41] phát sinh cho khái niệm trong nhóm A và nhóm C bằng những kỹ thuật khác nhau tương ứng từng loại biến thể (thông tin chi tiết về các loại biến thể và các kỹ thuật phát sinh tương ứng sẽ được trình bày trong chương 4 của luận văn) :

- Biến thể ngữ nghĩa (semantic variant) của khái niệm gốc (biến thể đồng nghĩa, biến thể tổng quát hóa, biến thể chuyên biệt hóa…) được tra ra từ MeSH và Entreze nhờ mạng ngữ nghĩa (Semantic network) của ULMS.
- Biến thể từ điển (Lexical variant) tra từ CSDL viết tắt ADAM, kết hợp với một số heuristic cho tên gien (như Số Latin và số La Mã có thể được dùng thay nhau trong tên gien, phần chữ và phần số trong tên gien có thể viết liền hoặc cách nhau khoảng trắng hoặc liên kết nhau bằng gạch nối…)
- Biến thể hình thái cho khái niệm trong MeSH được chọn là những khái niệm trong MeSH mà khác biệt không quá 2 ký tự so với khái niệm gốc.

Sau đó, các biến thể của nhóm A được giao với các biến thể của nhóm C tạo thành nhóm B. Các tác giả của [41] tính độ liên kết giữa mỗi khái niệm b trong B với tập A (gọi là I(b,A)) và với tập C (gọi là I(b,C)) theo công thức (2.1) và (2.2) sau :

$$I(b,A) = \log \frac{P(b,A)}{P(b).P(A)} \quad (2.1) \qquad I(b,C) = \log \frac{P(b,C)}{P(b).P(C)} \quad (2.2)$$

$$\text{Trong đó} \quad P(x) = \frac{\text{Số tài liệu có x}}{\text{Tổng số tài liệu}} \quad \text{với } X \subset \{A, b, C\} \quad (2.3)$$

Từ I(b,A) và I(b,C), nhóm tác giả tính điểm cho mỗi khái niệm b ứng với câu truy vấn Q theo công thức (2.4) sau :

$$Score(b,Q) = \frac{|\{x : x \in B \wedge I(x,A) \leq I(b,A) \wedge I(x,C) \leq I(b,C)\}|}{|\{x : x \in B \wedge I(x,A) \geq I(b,A) \wedge I(x,C) \geq I(b,C)\}|} \quad (2.4)$$

k khái niệm có điểm cao nhất trong B được chọn giữ lại để mở rộng câu truy vấn : Các khái niệm trong nhóm A, nhóm C cùng với k khái niệm giữ lại trong nhóm B tạo thành thể hiện hoàn chỉnh cho một câu truy vấn Q và được dùng để tìm kiếm các tài liệu lên quan Q theo kỹ thuật sau :

- Độ liên quan giữa câu truy vấn Q và tài liệu D được chia ra 2 mức : Mức khái niệm và Mức từ vựng
- Độ liên quan khái niệm ConceptSim(Q,D) được tính theo công thức (2.5) sau :

$$\frac{\sum_{c \in d \text{ và } c \in A} idf_c}{\sum_{c \in A} idf_c} * \frac{\text{Tổng số tài liệu}}{\text{Số tài liệu chứa đủ khái niệm trong A}} + \frac{\sum_{c \in d \text{ và } c \in C} idf_c}{\sum_{c \in C} idf_c} * \frac{\text{Tổng số tài liệu}}{\text{Số tài liệu chứa đủ khái niệm trong C}} \quad (2.5)$$

- Độ liên quan từ vựng WordSim(Q,D) được tính theo công thức (2.6) sau :

$$\sum_{w \in q} \log \frac{\text{Số tài liệu - Số tài liệu có w + 0.5}}{\text{Số tài liệu có w + 0.5}} * \frac{2.2 * tf_w}{1.2(0.25 - 0.75 \frac{\text{Độ dài d}}{\text{Độ dài trung bình các tài liệu}}) + tf_w} \quad (2.6)$$

- Cho 2 văn bản D1 và D2, D1 được xem là liên quan Q nhiều hơn D2 khi thỏa 1 trong 2 điều kiện sau :
  - ConceptSim(Q,D1) > ConceptSim(Q,D2)
  - ConceptSim(Q,D1) = ConceptSim(Q,D2) ∧ WordSim(Q,D1) > WordSim(Q,D2)

Theo [41], phương pháp này cho kết quả ánh xạ tài liệu đạt độ bao phủ 54%.

2.2.2 Hướng tiếp cận xử lý phía tài liệu

2.2.2.1 Phương pháp dùng Thể hiện Chủ đề (Thematic Representation)

Theo hướng tiếp cận này, mỗi tài liệu đều thuộc về một hoặc một số chủ đề nhất định, trong đó các chủ đề chính gọi là chủ đề trung tâm, các chủ đề còn lại liên quan nhiều hay ít đến tài liệu tùy mức độ. Từ đó, tài liệu được biểu diễn thành một cây phân cấp có các nút là các chủ đề liên quan đến nó (từ chủ đề tổng quát nhất đến chủ đề chuyên biệt nhất). Cây phân cấp ấy gọi là Thể hiện Chủ đề, trong đó có đánh dấu những chủ đề trung tâm của tài liệu.

Mỗi nút chủ đề gồm một Trung tâm Chủ đề (Thematic center) và các cụm từ liên quan. Trung tâm Chủ đề là cụm từ chứa tên của chủ đề, cụm từ này có xuất hiện tường minh trong tài liệu. Các cụm từ liên quan là những biến thể đồng nghĩa, tổng quát hóa, chuyên biệt hóa của Trung tâm Chủ đề (được tra ra từ Wordnet, không xuất hiện tường minh trong tài liệu)

Công trình [4] đề xuất một kỹ thuật xây dựng Thể hiện Chủ đề như sau :

- Nhóm tác giả sử dụng một tài nguyên do chính họ tự xây dựng : Một từ điển chuyên môn về Chính trị Xã hội. Trong đó, mỗi thuật ngữ có một hoặc một số diễn giải. Mỗi diễn giải có các tham chiếu đến các diễn giải liên quan.
- Từ điển này được dùng để tra các thuật ngữ xuất hiện trong tài liệu nhằm lấy ra diễn giải cho mỗi thuật ngữ. Trong trường hợp thuật ngữ có nhiều diễn giải, diễn giải liên quan đến nhiều diễn giải đã tra được nhất sẽ được chọn.

- Kết quả là tài liệu được biểu diễn bằng một loạt diễn giải thay vì các thuật ngữ. Với mỗi diễn giải được tra ra, các diễn giải liên quan mà nó tham chiếu trong từ điển cũng được lấy ra và chúng hình thành một nút Chủ đề (với diễn giải gốc đóng vai trò Trung tâm Chủ đề)
- Tiếp theo, hệ thống [4] lựa chọn nút chủ đề chính (chủ đề trung tâm) cho tài liệu bằng heuristic : Đó là những nút chủ đề có Trung tâm Chủ đề nằm trong các tiêu đề và các câu đầu của mỗi đoạn văn.
- Tuy nhiên, các tác giả nhận xét rằng như vậy vẫn có thể bỏ xót chủ đề quan trọng. Những diễn giải thuộc về các nút chủ đề chính thì thường xuất hiện cùng nhau xuyên suốt trong tài liệu. Càng nhiều diễn giải của 2 nút chủ đề đứng gần nhau trong tài liệu thì độ liên kết giữa hai nút chủ đề đó xem như càng cao. Do vậy, các tác giả chọn thêm một tập các nút chủ đề mà độ liên kết giữa chúng cao hơn hẳn giữa những nút chủ đề còn lại.

Mỗi chủ đề là một khái niệm trong lĩnh vực Chính trị Xã hội. Sau khi xác định được các nút chủ đề chính yếu, các tác giả dùng chúng để xác định một tài liệu có liên quan những khái niệm nào. Hiện chưa có công trình nào thực hiện phương pháp này cho việc lập chỉ mục tài liệu tiếng Việt.

2.2.2.2 Hướng tiếp cận dùng thực thể có tên (Named Entity)

Hướng nghiên cứu Nhận dạng thực thể có tên (Named entity recognization - NER) nguyên thủy chỉ quan tâm nhận dạng các thực thể *con người*, *địa danh* và *tổ chức* xuất hiện trong văn bản. Gần đây người ta đã quan tâm nhận dạng các thực thể Y khoa (như tên bệnh, tên gien, tên tế bào, tên protein …). Các nghiên cứu này hứa hẹn khả năng đóng góp cao cho việc lập chỉ mục khái niệm trên các tài liệu Y khoa : Áp dụng mô hình vector nhưng chỉ mục được lập trên các thực thể có tên (Named Entity - NE) thay vì từ khóa thông thường. Có hai phương pháp chính dùng để nhận dạng thực thể Y khoa : Phương pháp Maximum Entropy và Phương pháp Conditional Random Field

Cả hai phương pháp này đều sử dụng những hàm đặc trưng nhị phân cho quá trình huấn luyện. Tuy nhiên các hàm đặc trưng này có độ quan trọng không bằng nhau nên mang những trọng số khác nhau, những trọng số này được tự động xác định trong quá trình huấn luyện.

- Phương pháp Maximum Entropy ([43]) xác định các trọng số bằng thuật toán Generalized Iterative Scaling. Đồng thời, phương pháp này cũng xây dựng hai danh sách từ ngữ cảnh : các danh từ phía bên phải NE và các bổ từ phía bên trái các NE để mở rộng phạm vi các NE khi có một NE nhỏ nằm trong một NE lớn (Nested NE). Kết quả thử nghiệm của [43] cho thấy phương pháp Maximum Entropy có thể nhận diện NE với độ chính xác 72.7% và độ bao phủ 71.5%.
- Phương pháp Conditional Random Field thì xác định các trọng số bằng thuật toán Modified Iterative Scaling ([8]) hoặc thuật toán Numerical Optimization ([12]). Kết quả thử nghiệm của [8] cho độ chính xác 69.3% và độ bao phủ 70.3% trong khi của [12] cho độ chính xác 70.16% và độ bao phủ 72.27%.

Hướng tiếp cận này chủ yếu dựa trên thống kê nên không phụ thuộc nhiều vào ngôn ngữ. Tuy nhiên lại cần một corpus huấn luyện khá lớn đã gắn nhãn sẵn và hiện nay phương pháp này chỉ mới xác định được các thực thể Y khoa là : Tên gene, tên tế bào, loại tế bào, tên protein, tên virus và tên một số bệnh. Hiện chưa có công trình nào thực hiện phương pháp này cho việc lập chỉ mục tài liệu tiếng Việt.

2.2.3 Hướng tiếp cận phối hợp xử lý cả câu truy vấn và tài liệu

Trong [38], tác giả nhận định rằng sự không tương xứng giữa tài liệu và câu truy vấn dẫn đến việc độ chính xác thấp trong tìm kiếm thông tin. Do vậy đã có nhiều nỗ lực nghiên cứu nhằm mở rộng câu truy vấn và mở rộng tài liệu sao cho vector biểu diễn tài liệu và vector biểu diễn câu truy vấn được tiến gần nhau hơn. Phương pháp dùng trong mở rộng câu truy vấn hoặc mở rộng tài liệu là tương tự nhau, chia làm 3 hướng chính :

- Phương pháp dựa trên tập dữ liệu (collection based - [37]) còn được gọi là phương pháp Phân tích Toàn cục (global analysis). Phương pháp này sử dụng một tập tài liệu lớn và phân tích ngữ cảnh toàn cục của các thuật ngữ trong toàn bộ tập tài liệu (chứ không phải trong tài liệu đơn) nhằm tìm ra những thuật ngữ tương tự như thuật ngữ trong câu truy vấn (hay trong tài liệu) để mở rộng câu truy vấn (hay tài liệu).
- Phương pháp Phân tích Cục bộ (local analysis) giới hạn ngữ cảnh của thuật ngữ trong một tập thông tin nhỏ hơn. Tập thông tin này có được từ những kỹ thuật như relevance feedback, pseudo feedback [3], [18] hoặc có được từ những thông tin cộng tác (collaboration information) như hồ sơ người dùng (user profile)... [22]
- Phương pháp Cơ sở Tri thức (Knowledge based) sử dụng nguồn tri thức bên ngoài. Chẳng hạn như [20] và [33] sử dụng một từ điển đại trà là Wordnet nhằm tra ra mối liên hệ ngữ nghĩa giữa từ với từ, nhờ đó tìm ra những thuật ngữ liên quan với thuật ngữ của câu truy vấn (hay tài liệu). Tuy nhiên những nhập nhằng về phương diện từ vựng làm cho kết quả còn hạn chế.

Công trình [38] đi theo hướng Cơ sở Tri thức với nguồn tri thức là UMLS vì sự nhập nhằng thuật ngữ trong UMLS chỉ xuất hiện ở 0.25% tổng số thuật ngữ. Trong [38] một cải tiến được đóng góp : Kết hợp xử lý khái niệm trên câu truy vấn với xử lý khái niệm trên tài liệu theo hai hướng ngược nhau : Câu truy vấn được mở rộng bằng các khái niệm chuyên biệt hơn trong khi tài liệu được mở rộng bằng các khái niệm tổng quát hơn. Điều này dựa trên ý tưởng là người truy vấn thường đưa ra các khái niệm tổng quát nhưng tài liệu thường mô tả cặn kẽ vào các khái niệm chi tiết. Kỹ thuật phối hợp mở rộng truy vấn và mở rộng tài liệu của [38] giúp cải thiện hiệu quả truy vấn, độ đo MAP(DFR) tăng 66% so với giải pháp chỉ mở rộng tài liệu. Hiện chưa có công trình nào thực hiện phương pháp này cho việc lập chỉ mục tài liệu tiếng Việt.

2.3 Lập chỉ mục trên khái niệm có so khớp tài liệu với Ontology

Nếu chưa có sẵn Ontology, trước tiên phải xây dựng Ontology để sử dụng cho việc lập chỉ mục trên khái niệm. Việc xây dựng Ontology đòi hỏi rất nhiều kiến thức chuyên môn trong từng lĩnh vực. Vì đó hầu hết Ontology (như UMLS, SKOS, Wordnet…) được xây dựng thủ công bởi các chuyên gia. Bên cạnh đó cũng có những nỗ lực xây dựng Ontology một cách tự động. Sau khi đã có Ontology, công việc tiếp theo là ánh các tài liệu vào các khái niệm trong Ontology.

2.3.1 Xây dựng Ontology

Công trình [34] giới thiệu phương pháp xây dựng Ontology tự động bằng corpus đa ngữ. Nhóm tác giả sử dụng sự dóng hàng giữa các bản dịch (ở những ngôn ngữ khác nhau) của cùng một bản gốc để gom cụm các bản dịch của cùng một từ thành một cụm, nhờ đó khái niệm tạo nên bởi pha sau là đa ngữ. Pha tiếp theo dùng thuật toán Fuzzy C-mean. Với thuật toán này, một bộ gồm C khái niệm được lập sẵn bởi chuyên gia, mỗi khái niệm có một thuật ngữ làm định danh khái niệm. Các tên gọi này phải có hiện diện trong corpus đa ngữ. Thuật toán Fuzzy C-mean dùng corpus đa ngữ và gom cụm các cụm từ trong corpus thành C cụm thuật ngữ có trung tâm là C định danh khái niệm ban đầu. Với C cụm tìm được, thuật toán tính lại trung tâm của mỗi cụm. Với trung tâm mới, thuật toán tính lại ranh giới cụm. Quá trình lặp dừng khi kết quả lần lặp i+1 không khác lần lặp i, khi đó mỗi cụm là một khái niệm trong một Ontology có C khái niệm. Khoảng cách giữa mỗi thuật ngữ với trung tâm cụm được dùng làm trọng số thành viên của thuật ngữ ấy đối với cụm của nó. Chi tiết thuật toán được mô tả trong [34].

Công trình [35] vận dụng ý tưởng trên và xây dựng một Ontology đa ngữ có tên gọi Balkanat. Trong đó, các tên gọi của một khái niệm trong cùng một ngôn ngữ tạo thành một tập đồng nghĩa. Các tập đồng nghĩa khác nhau thuộc những ngôn ngữ khác nhau của cùng một khái niệm được ánh xạ về tập đồng nghĩa tương ứng trong tiếng Anh thông qua một chỉ mục liên ngữ. Một cụm như vậy, với ngôn ngữ tiếng Anh ở trung tâm và các ngôn ngữ khác ở xung quanh, tạo thành một khái niệm.

2.3.2 Lập chỉ mục – Chỉ sử dụng khái niệm

2.3.2.1 Có phát sinh biến thể cụm từ

Các công trình thuộc hướng tiếp cận này đều thực hiện ánh xạ tài liệu vào một Ontology. Một số Ontology như ULMS được sử dụng bởi [1, 2, 9, 26, 29]; Balkanat được dùng bởi [35]; SKOS được dùng bởi [23]…

Quá trình xử lý được chia làm 3 tác vụ lớn :

- Phân tích cú pháp
- Phát sinh biến thể
- Ánh xạ tài liệu vào danh mục khái niệm.

2.3.2.1.1 Phân tích cú pháp

Mục đích của tác vụ này là tiền xử lý trên văn bản thô, sao cho rút trích ra được những cụm danh từ (vì định danh khái niệm trong Ontology cũng là các cụm danh từ). Cụm danh từ được rút trích là những cụm danh từ đơn giản, nghĩa là không có cụm danh từ con và cũng không có mệnh đề tính từ (Relative Clause).

Để làm việc này, công trình [1] và [2] sử dụng một từ điển 60,000 từ và bộ gán nhãn từ loại Xerox Stochastic để phân tích cú pháp câu trong văn bản, từ đó rút trích được các cụm danh từ đơn giản. Mỗi cụm danh từ được phân ra danh từ trung tâm (head-noun) và phần bổ nghĩa (modifier). Tuy nhiên một thách thức là từ điển không phủ hết mọi từ có trong văn bản, do đó trong [42] nhóm tác giả xây dựng thêm một bộ luật gồm 600 luật nhằm xử lý những từ không có trong từ điển (bằng cách xác định mối liên hệ giữa từ chưa biết với một từ gần nhất có trong từ điển). Đồng thời, nhóm tác giả nhận xét rằng phân tích cú pháp để tách cụm danh từ có chi phí khá cao nên đã đề xuất một bảng từ tách (table of break words). Từ tách là những từ thường đóng vai trò phân cách các cụm từ. Hệ thống sẽ nhận diện những từ tách này trong văn bản để tách ra các cụm danh từ tương ứng.

2.3.2.1.2 Phát sinh biến thể

Mỗi khái niệm có thể xuất hiện trong tài liệu bằng nhiều cụm từ khác nhau (ví dụ như *ung thư dạ dày* và *ung thư bao tử* là cùng một khái niệm). Không những vậy, cụm từ không phải lúc nào cũng xuất hiện tường minh trong tài liệu (ví dụ cụm từ *rối loạn tiêu hóa* không xuất hiện tường minh trong *rối loạn tuần hoàn và tiêu hóa*). Do đó để hạn chế sự bỏ sót khái niệm trong tài liệu, cần thực hiện phát sinh các biến thể cho mỗi cụm danh từ rút trích được.

Có nhiều loại biến thể (như biến thể hình thái, biến thể từ điển, biến thể ngữ nghĩa, biến thể cú pháp…) nhưng không phải mọi công trình đều phát sinh đầy đủ các loại biến thể. Chẳng hạn như [2] chỉ phát sinh biến thể từ điển (dùng một từ điển đồng nghĩa là Illustrated Medical Dictionary), biến thể ngữ nghĩa (dùng Cơ sở tri thức Specialist) và biến thể hình thái (dùng bộ luật Derivational Morphological Rules). Biến thể và các kỹ thuật phát sinh biến thể được mô tả chi tiết trong chương 4.

2.3.2.1.3 Ánh xạ tài liệu vào danh mục khái niệm

Mục đích của tác vụ này là chọn ra từ Ontology những khái niệm thực sự liên quan đến tài liệu. Trước tiên, các khái niệm trong Ontology có định danh khái niệm giống với một cụm danh từ gốc hoặc biến thể nào đó trong tài liệu (giống toàn bộ hoặc giống một phần) thì đều được lấy ra làm khái niệm ứng viên. Công trình [35] đề xuất một cải tiến khi rút trích khái niệm ứng viên : Ngay từ đầu ta chỉ chọn ra các từ loại quan trọng (n, v, adj, adv) và tính trọng số tf.idf cho chúng. Sau đó chỉ giữ lại các từ có trọng số vượt một ngưỡng cho trước (được xem là những từ quan trọng) để xử lý rút trích khái niệm, nhờ vậy giảm nhiễu đáng kể.

Sau đó mỗi khái niệm ứng viên sẽ được chấm điểm thông qua một hoặc một số độ đo so khớp chuỗi. Tùy công trình mà những độ đo khác nhau được sử dụng. Cuối cùng, những ứng viên có điểm vượt một ngưỡng cho trước sẽ thực sự được chọn.

Công trình [1] và [2] sử dụng 4 độ đo sau đây với những trọng số khác nhau :

- Độ trọng tâm (Centrality) mang trọng số bằng 1.
- Độ biến động (Variation) mang trọng số bằng 1.

- Độ phủ lấp (Coverage) mang trọng số bằng 2.
- Độ cố kết (Cohensiveness) mang trọng số bằng 2.

Sau khi hệ thống Metamap [1] đã được xây dựng và đạt khả năng truy vấn với độ chính xác trung bình 55.2% và độ bao phủ trung bình 93.3%, Lo¨ıc Maisonnasse và đồng sự thực hiện một cải tiến bằng cách phối hợp Metamap với 2 công cụ tuyển chọn khái niệm khác là MiniPar và TreeTagger (cùng chạy trên UMLS) để cải tiến độ chính xác tăng 3%. Chi tiết hệ thống được mô tả trong [28]. Chi tiết về một số độ đo so khớp nêu trên và cách phối hợp giữa chúng để cho ra một độ liên quan duy nhất giữa tài liệu và khái niệm được trình bày chi tiết trong chương 5 của luận văn.

Công trình [26] cũng sử dụng ULMS nhưng đề xuất một giải pháp so khớp hoàn toàn khác với Meta-map và phát triển một hệ thống lập chỉ mục gọi là Conann (Concept Annotation). Ý tưởng của [26] trước tiên xuất phát từ nhận xét rằng có những từ hiếm, xuất hiện trong rất ít định danh khái niệm, những từ như vậy là dấu hiệu rất đặc trưng để nhận biết các khái niệm ấy. Ngược lại, có những từ phổ biến, xuất hiện trong rất nhiều khái niệm. Những từ như vậy không chuyển tải thông tin đặc trưng của bất kỳ khái niệm nào nên khi tham gia so khớp sẽ gây nhiễu. Do đó [26] tính độ đo IPF (Inverse Phrase Frequency) cho mỗi từ phân biệt trong Ontology. Từ càng xuất hiện trong nhiều định danh khái niệm thì IPF càng nhỏ và càng ít có trọng lượng khi tham gia so khớp. Mục tiêu của giai đoạn so khớp là tìm ra những khái niệm ứng viên cho mỗi cụm từ trong tài liệu. Các ứng viên được tuyển chọn qua 3 lần sàng lọc. Mỗi lần sàng lọc sử dụng một tập độ đo riêng và chỉ giữ lại những ứng viên có độ đo thỏa một ngưỡng cho trước. Sau lần sàng lọc thứ 3, các ứng viên còn trụ lại được xem là thực sự liên quan đến cụm từ đang xét. Mô tả chi tiết các độ đo, cách phối hợp chúng và cách tính ngưỡng trong mỗi lần sàng lọc được trình bày chi tiết trong chương 5 của luận văn.

Việc tuyển chọn khái niệm ứng viên và tính toán độ liên quan (giữa cụm từ và khái niệm ứng viên) được thực hiện cho từng cụm danh từ trong tài liệu. Sau cùng, độ liên quan của mỗi khái niệm ứng viên được cộng dồn trên đầu cụm từ mà nó làm

ứng viên, rồi chia trung bình cho tổng số cụm từ. Do vậy khái niệm nào làm ứng viên cho càng nhiều cụm từ thì được xem là càng liên quan đến tài liệu.

Hướng tiếp cận này còn gặp một số thách thức như việc phát sinh biến thể có thể cho ra các biến thể khôn lường, đồng thời chưa hoàn toàn chọn được ứng viên tốt nhất khi nhiều khái niệm ứng viên có cùng một độ so khớp. Về vấn đề nhập nhằng khi một thuật ngữ có thể ám chỉ nhiều hơn một khái niệm, theo [38], nếu dùng UMLS, không cần lo lắng vì 99.75% thuật ngữ trong UMLS chỉ liên quan 1 khái niệm.

2.3.2.2 Không phát sinh biến thể cụm từ

Công trình [11] thực hiện Việt hóa một phần Ontology Y khoa UMLS để phục vụ việc lập chỉ mục trên khái niệm cho các tài liệu Y khoa tiếng Việt. Tuy nhiên việc ánh xạ khái niệm vào Ontology chỉ mới dừng lại ở sự so khớp các cụm danh từ, sử dụng các hệ số so khớp chuỗi, tác giả chưa thực hiện phát sinh biến thể cho cụm danh từ, do đó còn bỏ qua các khái niệm không xuất hiện tường minh trong tài liệu.

Để rút trích từ chỉ mục trong văn bản tiếng Việt, tác giả của [11] đề nghị sử dụng kết hợp phương pháp ngôn ngữ với phương pháp thống kê để rút trích cụm từ trong tài liệu. Cụ thể là tác giả sử dụng công cụ Wordseg để trích từ có trong từ điển, sau đó sử dụng N-gram với đơn vị là từ rồi áp dụng các hệ số thống kê để trích ra cụm từ. Về thống kê, tác giả sử dụng hệ số Dice để lọc ra các cụm từ. Cụ thể là thử nghiệm sẽ lần lượt lấy ra các cụm từ có 2 từ (2-gram), 3 từ (3-gram), 4 từ (4-gram) từ để tính hệ số Dice. Việc trích cụm từ theo N-gram tổng quát được [11] tính như sau: giả sử cho một chuỗi N-gram được biểu diễn là S = w1 w2… wN (N từ 2 đến 4), với mỗi N-gram, tính hệ số Dice của tất cả các tổ hợp từng hai phần tử có thể có của nó, nếu kết quả tính hệ số Dice của bất kỳ tổ hợp nào lớn hơn một ngưỡng cho trước thì cụm từ tương ứng được đưa vào danh sách cụm từ kết quả.

Ngoài ra, thử nghiệm còn sử dụng một số heuristic của tri thức ngôn ngữ như: một cụm từ có nghĩa thì không thể bắt đầu hay kết thúc bằng các hư từ (stopword)

(stopword là các từ như và, là, cái, bị,...), một cụm từ có nghĩa thì không thể bắt đầu bằng một con số.

Để so khớp cụm từ vào danh mục khái niệm, tác giả lấy từng cụm từ rút trích được trong mỗi tài liệu để so khớp với từng khái niệm có trong danh mục khái niệm. Trong so khớp tác giả chọn cách so khớp dựa trên các độ đo tương tự giữa hai chuỗi: Hệ số Overlap, hệ số Dice, hệ số Cosine, hệ số Jaccard, hệ số R_Over. Chi tiết về từng hệ số và cách phối hợp chúng được trình bày trong [11]. Kết quả thử nghiệm của [11] đạt độ chính xác trung bình 58.5% và độ bao phủ trung bình 74.2%.

2.3.3 Lập chỉ mục – Sử dụng khái niệm và Mối kết hợp giữa chúng

UMLS bao gồm 3 thành phần chính. Bên cạnh Bộ từ vựng Chuyên gia (Specialist Lexicon – chứa tập biến thể từ vựng của các thuật ngữ Y khoa) và Siêu từ điển Chuyên môn (MetaThesaurus – chứa 1,700,000 thuật ngữ thuộc 797,359 khái niệm Y khoa trong 9 ngôn ngữ khác nhau trên thế giới), còn có một Mạng ngữ nghĩa (Semantic Network) chứa mối kết hợp giữa tất cả các khái niệm Y khoa trong UMLS. Do vậy một số công trình đề nghị sử dụng các mối kết hợp này để giúp việc lập chỉ mục trên khái niệm được chi tiết và hiệu quả hơn.

2.3.3.1 Tổng quan về cách tổ chức của Mạng ngữ nghĩa

Mạng ngữ nghĩa phân nhóm 797,359 khái niệm của UMLS thành 134 loại ngữ nghĩa (Semantic type). Mỗi loại ngữ nghĩa có một định danh duy nhất (Type Unique Identifier – TUI). Từ đó mối kết hợp giữa các khái niệm được tổng quát hóa thành mối kết hợp giữa các loại ngữ nghĩa và phân thành 54 nhóm cho 54 lĩnh vực con (SubDomain) khác nhau. Mỗi mối kết hợp đều là một bộ ba (TUI – TUI – TUI)

Hình 2-3 : Cấu trúc mối kết hợp trong mạng ngữ nghĩa

Trong đó A và B là những loại ngữ nghĩa của các đối tượng hoặc quy trình Y khoa, còn B là loại ngữ nghĩa của những mối quan hệ trong các vấn đề Y khoa.

2.3.3.2 Phương pháp thực hiện

Ý tưởng là khi lập chỉ mục, không chỉ trả lời câu hỏi "Một tài liệu D có liên quan những khái niệm nào ?" mà còn trả lời câu hỏi "Trong tài liệu D, các khái niệm tìm được có quan hệ gì với nhau ?". Để làm được điều đó, cần rút trích từ mạng ngữ nghĩa những mối kết hợp có hiện diện trong tài liệu.

Công trình [28] đề xuất giải pháp là nếu hai khái niệm a (thuộc loại ngữ nghĩa A) và b (thuộc loại ngữ nghĩa B) cùng xuất hiện trong một câu, và nếu Mạng ngữ nghĩa có định nghĩa một (hay một số) mối kết hợp giữa A và B thì (các) mối kết hợp ấy được xem là có hiện diện trong tài liệu và được rút trích ra.

Nhưng theo [36] thống kê, các loại ngữ nghĩa là quá tổng quát nên trung bình các mối kết hợp định nghĩa giữa các loại ngữ nghĩa chỉ đúng cho 17% cặp khái niệm thành viên. Trong số 17% ấy chỉ có 34% là những mối kết hợp quan trọng. Từ đó cho thấy mối kết hợp thừa được rút trích rất nhiều. Ngược lại, Mạng ngữ nghĩa cũng không phủ hết được mọi quan hệ có trong đời thực, nên trong tài liệu sẽ tồn tại những mối kết hợp có ý nghĩa mà lại không được rút trích. Do vậy [36] đề nghị hai bước tinh chỉnh :

- Lọc bỏ mối kết hợp thừa
- Bổ sung mối kết hợp thiếu

2.3.3.2.1 Lọc mối kết hợp thừa

Lọc bằng IDF

Kỹ thuật này dựa trên nhận định rằng mối kết hợp quan trọng là mối kết hợp liên kết các khái niệm quan trọng trong tài liệu. Khái niệm quan trọng trong một tài liệu là những khái niệm xuất hiện rất nhiều lần trong tài liệu ấy nhưng không (hoặc rất ít) xuất hiện trong hầu hết các tài liệu còn lại.

Từ đó độ đo phù hợp được chọn là IDF. Những mối kết hợp rút trích ra mà liên kết các khái niệm có IDF thấp hơn một ngưỡng θ cho trước thì bị lọc bỏ (từ thực nghiệm θ chọn bằng 2.7)

Lọc bằng Chỉ thị Động từ (Verbal marker)

Kỹ thuật này dựa trên nhận định rằng mối kết hợp đúng thì thường được thể hiện trong tài liệu bằng một động từ phù hợp với nó.

Từ đó [36] lập nên một ma trận tương thích, trong đó 1 chiều là các động từ trong từ điển và chiều còn lại là các mối kết hợp. Ma trận này cho biết 1 mối kết hợp ở dòng i có tương thích với động từ ở cột j hay không. Khi các mối kết hợp trong Mạng ngữ nghĩa được rút trích ra giữa hai khái niệm trong một câu, mà động từ tương ứng trong câu này không tương thích thì mối kết hợp ấy bị lọc bỏ.

2.3.3.2.2 Bổ sung mối kết hợp thiếu

Mối kết hợp mới được tìm bằng cách Sử dụng định danh của các thuật ngữ trong MeSH của UMLS. Tổng quát về cách tổ chức của MeSH như sau : UMLS tổ chức các thuật ngữ trong MeSH thành cây quan hệ với 15 Node cấp cao nhất (ký hiệu là A, B, C …, M, N, Z). Dưới đó là 114 node ở cấp thứ 2 (ký hiệu như 267, C23, E7, C2…). [36] chỉ sử dụng đến cấp thứ 2 chứ không đi sâu thêm xuống các cấp dưới.

Với mỗi mối kết hợp trong Semantic Network, các tác giả xây dựng một danh sách các mẫu (patterns) chỉ rõ các Node cấp 2 có thể quan hệ với nhau bằng mối kết hợp đang xét (Ví dụ : Quan hệ chữa trị : 267|C23, D3|C23, …). Nhờ đó hệ thống phát hiện sự hiện diện của mối kết hợp trong tài liệu dựa vào sự xuất hiện của cặp node cấp 2 của mẫu nào đó trong cùng một câu.

Công trình [29] đề nghị cách bổ sung mối kết hợp tổng quát hóa – chuyên biệt hóa bằng thống kê thay vì dùng Semantic Network dựa trên nhận định rằng : Y là chuyên biệt của X nếu $P(X|Y)=1$ và $P(Y|X)<1$. Về sau, từ thực nghiệm, các tác giả đã điều chỉnh các hằng số và điều kiện trên trở thành $P(X|Y)\geq 0.8$ và $P(Y|X)<1$

Nhận xét :

Kết quả thử nghiệm của [26] cho thấy :

- Dùng phương pháp lọc bớt mối kết hợp thừa, độ chính xác giảm 1.6%
- Dùng phương pháp phát hiện mối kết hợp mới, độ chính xác tăng 6.4%

Như vậy phương pháp lọc bớt kết hợp thừa không cho kết quả khả quan (vì chỉ chấp nhận các mối kết hợp chỉ thị bằng động từ, bỏ quên vai trò của các từ loại khác). Phương pháp phát hiện mối kết hợp mới cho cải thiện rõ rệt.

2.4 Lựa chọn của đề tài

Hiện có một Ontology tiếng Việt thuộc lĩnh vực Y khoa được xây dựng bởi [11]. Ontology này được [11] Việt hóa từ một phần của UMLS với sự trợ giúp của PGS. TS. BS Nguyễn Đỗ Nguyên và tập thể giảng viên bộ môn Dịch Tễ - Khoa Y Tế Công cộng – Đại Học Y Dược TP. HCM. Do vậy luận văn chọn đi theo hướng tiếp cận có so khớp tài liệu vào Ontology.

Tuy nhiên, Ontology tiếng Việt này hiện chưa có thành phần mạng ngữ nghĩa chứa mối kết hợp giữa các khái niệm. Do vậy luận văn chỉ sử dụng khái niệm trong xử lý so khớp, tạm thời chưa quan tâm mối kết hợp giữa các khái niệm.

Như đã trình bày trên đây, xử lý của hướng tiếp cận này chia làm ba tác vụ chính, trong đó hai tác vụ đầu tiên là Phân tích cú pháp và Phát sinh biến thể.

Vì bài toán của luận văn là Lập chỉ mục trên khái niệm, và hầu hết khái niệm đều xuất hiện trong các tài liệu dưới dạng cụm danh từ nên cần phải có tác vụ Phân tích cú pháp để rút trích cụm danh từ trong tài liệu. Những cụm danh từ này cần được chia ra thành phần trung tâm và các thành phần bổ ngữ nhằm tạo cơ sở cho việc phát sinh biến thể và việc ánh xạ tài liệu vào Ontology (vì những bộ phận ngữ pháp khác nhau trong cụm danh từ có thể có độ quan trọng khác nhau khi so khớp vào các khái niệm trong Ontology). Chương 3 sau đây sẽ trình bày giải pháp được vận dụng trong luận văn để rút trích và cấu trúc hóa cụm danh từ. Sau đó, chương 4 sẽ trình bày giải pháp mà luận văn lựa chọn để thực hiện tác vụ Phát sinh biến thể nhằm phục vụ cho việc lập chỉ mục trên khái niệm.