(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **2-0001862**

(51)[7] **G06F 17**

(13) **Y**

(21) 2-2010-00144
(22) 06.07.2010
(45) 25.10.2018 367
(43) 25.06.2012 291
(73) ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH (VN)
Phường Linh Trung, quận Thủ Đức, thành phố Hồ Chí Minh
(72) Phan Thị Tươi (VN), Nguyễn Chánh Thành (VN)

(54) PHƯƠNG PHÁP PHỤC VỤ HỎI ĐÁP VÀ TRUY XUẤT THÔNG TIN DẠNG VĂN BẢN CÓ HỖ TRỢ TIẾNG VIỆT

(57) Giải pháp hữu ích (GPHI) đề cập tới việc nghiên cứu giải pháp chương trình máy tính phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt. Với ưu điểm là sẽ hỗ trợ người dùng truy vấn thông tin một cách thông minh hơn và uyển chuyển hơn, cho kết quả mang độ chính xác cao hơn. Hệ thống sẽ giúp các thư viện của các cơ quan trường học, viện nghiên cứu các tòa soạn báo, đài phát thanh/truyền hình triển khai phục vụ người dùng trong việc khai thác thông tin được hiệu quả hơn góp phần xây dựng và củng cố uy thế cạnh tranh cho các sản phẩm và công nghệ nội địa về Web có ngữ nghĩa, truy vấn thông tin đa phương tiện hướng đến ngữ nghĩa có hỗ trợ tiếng Việt trong tương lai.

Lĩnh vực kỹ thuật được đề cập

Phương pháp thuộc lĩnh vực công nghệ thông tin, phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt.

Tình trạng kỹ thuật của sáng chế

Liên quan đến phương pháp khai thác thông tin dạng văn bản (bao gồm cả truy xuất thông tin và hỏi đáp thông tin), hiện nay đã có các phương pháp nghiên cứu và công trình khoa học được công bố trên thế giới được tóm lược như sau:

Trên thế giới:

- ***Các phương pháp Khai thác thông tin trên Web:***

- Các phương pháp rất hiệu quả đã được dùng để phát triển những động cơ tìm kiếm thông tin, như: PageRank và Hilltop cho Google, TrustRank cho Yahoo, …

- Các phương pháp rút trích thông tin (Information Extraction), như học máy thống kê (với HMM, MaxEnt, MEMM, CRF,…), máy vector hỗ trợ (SVM). Những phương pháp này được ứng dụng để phát triển phương pháp của đề tài.

- Các phương pháp tóm lược văn bản (Text Summarization). Các phương pháp này dựa trên học thống kê trên kho ngữ liệu, được sử dụng để phát triển cho việc khai thác thông tin có hỗ trợ tiếng Việt của đề tài.

- ***Các phương pháp Truy vấn thông tin hướng ngữ nghĩa:***

- Các dự án nghiên cứu Neurocommons, FOAF, SIOC, SIMILE, Linking Open Data hướng đến phát triển Web ngữ nghĩa để triển khai các ứng dụng.

- Dự án SIR của nhóm Iryna Gurevych ở Darmstadt hướng đến việc gắn kết những yếu tố ngữ nghĩa vào mô hình IR, kết hợp các tri thức ngôn ngữ như German WordNet, Web 2.0, Wikipedia, … để cải thiện chất lượng kết quả của truy vấn thông tin.

- ***Các phương pháp Hỏi đáp hướng ngữ nghĩa:***

- Hiện nay có một số hệ thống hỏi đáp có hỗ trợ tiếng Anh như *Ask.com*hay *Lexxe.com* đã áp dụng các kết quả xử lý ngôn ngữ tự nhiên như: mẫu (pattern) hỏi đáp; phân tích câu hỏi; phân nhóm câu hỏi và các câu trả lời liên quan; rút trích và tóm tắt nội dung trả lời. Hiệu quả của phương pháp này phụ thuộc vào dạng câu hỏi, số lượng và chất lượng mẫu câu hỏi, mức độ phong phú của kho ngữ liệu. Hiện nay, các hệ thống này được triển khai dạng thương phẩm ***tuy nhiên chưa có hệ thống hỏi đáp nào có sự hỗ trợ tiếng Việt.***

- Một số hệ thống khác trong lĩnh vực này (tiếng Anh, và một phần tiếng Hoa) được công bố trong *NIST Special Publication 500-274: The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)* như:

+ Hệ thống LCC's CHAUCER-2 của nhóm tác giả Andrew Hickl phát triển phương thức chỉ mục và động cơ truy xuất dữ liệu để tìm các tài liệu hoặc đoạn văn trong các trường hợp như (a) thực thể có tên và kiểu, (b) quan hệ phụ thuộc ngữ nghĩa, (c) khung ngữ nghĩa.

+ Hệ thống Lymba's PowerAnswer4 của nhóm tác giả Dan Moldovan cung cấp giải pháp phân lớp truy vấn và câu hỏi, đề xuất chiến lược và kỹ thuật xử lý riêng biệt cho từng lớp như Question Processing (QP), Passage Retrieval (PR), và Answer Processing (AP) dựa trên việc áp dụng các kết quả trong lĩnh vực xử lý ngôn ngữ tự nhiên.

+ Hệ thống CSAIL của nhóm Boris Katz đề xuất phương pháp: phân tích câu hỏi; truy cập tài liệu; trả lời dạng câu hỏi trung tâm hoặc danh mục; xác định mục tiêu của câu hỏi và dạng câu trả lời trong bước phân tích đầu tiên. Phương pháp này sử dụng Wikipedia để hỗ trợ huấn luyện nhằm tăng hiệu quả xử lý.

+ Hệ thống OHSU Biomedical Question Answering - của nhóm tác giả A. M. Cohen đã xây dựng hạ tầng kiến trúc hệ thống hỏi đáp với giải pháp ống xử lý (pipeline) cho đối tượng Blob và tập các kết quả dạng [tên: trị] tích lũy được. Hệ thống này đang trong giai đoạn phát triển.

Trong nước:

- ***Phương pháp Khai phá Web:***

- Đề tài NCKH cấp ĐHQG, "Phát triển một Hệ thống Search Engine (S.E) Hỗ trợ tìm kiếm thông tin, thuộc lãnh vực Công nghệ Thông tin trên Internet qua từ khoá bằng tiếng Việt.", PGS.TS. Trương Mỹ Dung, ĐHKHTN, 2004.

- Đề tài NCKH trọng điểm ĐHQG Tp.HCM "Xây dựng hệ thống truy xuất thông tin hỗ trợ tiếng Việt", PGS.TS. Phan Thị Tươi, ĐHBK.

- "Hệ thống tìm kiếm thông tin xuyên ngữ CLIRS" của TS. Hồ Bảo Quốc và PGS.TS. Đồng Thị Bích Thuỷ, thuộc Bộ môn Hệ thống Thông tin, ĐHKHTN Tp.HCM.

- ***Phương pháp Web ngữ nghĩa:***

- Đề tài trọng điểm quốc gia KC.01.02, của nhóm nghiên cứu PGS.TS. Cao Hoàng Trụ, ĐHBK Tp.HCM, 2004.

Bản chất kỹ thuật của sáng chế

Trong những năm gần đây, nhu cầu truy cập internet và truy cập tin tức của người dân Việt nam rất cao (khoảng 20 triệu lượt) xuất phát từ các nhu cầu chủ yếu như xem tin tức, trao đổi thông tin, và đặc biệt bao gồm số lượng rất lớn tác vụ tìm kiếm thông tin cần thiết. Tuy nhiên, những tác vụ này, thông qua các động cơ tìm kiếm thông tin sẵn có như Google – Yahoo – Alta Vista..., thường không nhận được kết quả cần thiết như mong muốn.

Nguyên nhân chủ yếu ảnh hưởng đến vấn đề này là: Người dùng không cung cấp đủ thông tin truy vấn cần thiết, động cơ tìm kiếm thông tin không nhận biết được ngữ cảnh của truy vấn, động cơ tìm kiếm thông tin hoạt động dựa trên phương thức so trùng từ khóa và chưa quan tâm yếu tố ngữ nghĩa, các động cơ tìm kiếm hiện có thường hỗ trợ chính cho tiếng Anh, và thiếu tiếng bản xứ.

Điều này dẫn đến tình trạng người sử dụng phải dành thời gian khá lớn để chọn lọc các kết quả thích hợp. Người sử dụng không nhận được một kết quả trả lời trọn vẹn hoàn chỉnh về một vấn đề cần tìm vì thiếu một hệ thống tìm kiếm thông tin nhanh, linh hoạt để tìm các thông tin trong tài liệu tiếng Anh và tiếng Việt theo phương thức so trùng từ khóa. Chưa có hệ thống truy vấn thông tin có hỗ trợ tiếng Việt (trong phạm vi xác định trước) ...

Do đó, ***phương pháp phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt*** được đề xuất nhằm giải quyết các vấn đề khó khăn nêu trên. Phương pháp này gồm các bước:

– Phân tích nội dung câu hỏi tiếng Việt trong một miền/lĩnh vực thông tin xác định trước.

– Truy vấn thông tin (có hỗ trợ tiếng Việt) dạng văn bản trong hệ thống truy xuất thông tin xuyên ngôn ngữ.

– Suy diễn tìm kiếm nội dung trả lời cho câu hỏi (có hỗ trợ tiếng Việt) dạng văn bản

trong một miền/lĩnh vực thông tin xác định trước.

Sau đây là phần mô tả chi tiết các đặc điểm kỹ thuật của phương pháp đề xuất.

1. Phân tích nội dung câu hỏi tiếng Việt trong một miền/lĩnh vực thông tin xác định trước.

Nội dung phần này đóng vai trò vô cùng quan trọng cho hệ thống hỏi đáp có hỗ trợ tiếng Việt (**VQAS**) do nhóm tác giả xây dựng.

● ***Cách thức thực hiện***

a) Xác định miền/lĩnh vực thông tin để xác định các khái niệm, các đối tượng trừu tượng, vật lý và các mối quan hệ chức năng ngữ nghĩa giữa các đối tượng trong miền xác định đó.

b) Xây dựng tập mẫu các dạng câu hỏi, các câu truy vấn tiếng Việt trên miền thông tin được xác định. Các dạng câu truy vấn gồm có: câu hỏi tổng quát (dạng yes/no question); câu hỏi xác định thành phần (dạng Wh-question); câu hỏi lựa chọn (dạng OR/AND)

Một số dạng câu hỏi trong tập mẫu

Câu hỏi tổng quát (Yes/No questions)

- Dạng [Tác nhân–*Hành vi?*–Đối tượng]
- Dạng [Đối tượng–*Hành vi?*–Tác nhân]
- Dạng [Đối tượng–Tác nhân–*Hành vi?*]

Câu hỏi có từ nghi vấn (Wh–questions)

- Dạng [Tác nhân–Hành vi–*Đối tượng*?]
- Dạng [*Tác nhân*?–Hành vi–Đối tượng]
- Dạng [*Đối tượng*?–Tác nhân–Hành vi
- Dạng [Tác nhân–Hành vi–*Đối tượng*?–Đối tượng không trực tiếp]
- Dạng [Tác nhân–Hành vi–Đối tượng–*Đối tượng không trực tiếp*?]
- Dạng [*Đối tượng*?–Tác nhân–Hành vi–Đối tượng không trực tiếp]
- Dạng [Đối tượng–Đối tượng không trực tiếp–Hành vi–*Tác nhân*?]

c) Xây dựng từ điển các từ trong phạm vi miền thông tin; các từ/các cụm từ đồng nghĩa của câu truy vấn để phục vụ cho việc mở rộng truy vấn hướng đến ngữ nghĩa.

d) Tiền xử lý câu truy vấn: phân đoạn từ (Word Segmentation), gán nhãn từ loại (Part Of Speech Tagging), xác định cấu trúc cụm danh từ, động từ, trạng từ (Phrase Chunker) .

e) Xác định các cụm từ chức năng ngữ nghĩa trong câu tiếng Việt, như: *tác nhân*

(AGENT), *hành vi* (ACTION), *đối tượng* chịu tác động hành vi (THEME), *công cụ* thực hiện hành vi(INSTR), *vị trí* xảy ra hành vi (LOC), *đối tượng không trực tiếp* (TO-POS, BENECIFIARY) … bằng bộ phân tích cú pháp và xử lý ngữ nghĩa.

f) Xây dựng tập luật phân tích cú pháp và ngữ nghĩa trên cơ sở văn phạm phụ thuộc và ràng buộc ngữ nghĩa giữa các đối tượng trong câu truy vấn

◦ ***Ví dụ*** *minh họa việc xây dựng luật cho câu hỏi tổng quát (Y/N question)*

- Câu truy vấn: "***Bài báo*** ***ABC*** ***thuộc*** ***lĩnh vực*** ***IR*** ***đã*** ***tham khảo***
N_baibao *Ne* *Tsh_thuoc* *N_linhvuc* *Ne* □ *V_thamkhao*

công trình ***XYZ*** ***phải không?***
N_tacpham *Ne* *Thđs_pk*

Xử lý cú pháp

- 1. N_baibao Ne → <C_baibao : Ne>
- 2. N_linhvuc Ne → <C_linhvuc : Ne>
- 3. V_thamkhao N_tacpham Ne → (Ref: (<C_sach : Ne>)) | (Ref: (<C_baibao : Ne>))

g) Xác định dạng diễn dịch ngữ nghĩa cho các câu truy vấn

Bộ ba xuất ra đã được phân tích cú pháp dựa trên văn phạm phụ thuộc là dạng diễn dịch ngữ nghĩa của câu truy vấn. Trong trường hợp này hệ thống xuất ra hai dạng diễn dịch ngữ nghĩa:

- (<C_baibao : Ne>, <C_linhvuc : Ne>, (Ref: (<C_sach: Ne>)))
- | (<C_baibao : Ne>, <C_linhvuc : Ne>, (Ref: (<C_baibao: Ne>)))

h) Xây dựng giải thuật giải quyết nhập cho diễn dịch ngữ nghĩa các câu truy vấn tiếng Việt.

Như trường hợp trên, hệ thống sẽ giải quyết nhập nhằng để tìm ra dạng diễn dịch phù hợp nhất bằng sự ràng buộc lựa chọn hoặc bằng suy diễn tìm nội dung câu trả lời.

• ***Phương tiện thực hiện***

- Tham khảo các ontology WordNet tiếng Anh để phát triển WordNet tiếng Việt phục vụ cho việc xác định nghĩa từ, từ đồng nghĩa để thực hiện ***bước c)***

- Nhóm tác giả phát triển mã nguồn mở xử lý ngôn ngữ tự nhiên (GATE) cho tiếng Việt và công cụ vnSegmentation, vnPosTagger của nhóm nghiên cứu xử lý ngôn ngữ tự nhiên của khoa Khoa học và Kỹ thuật máy tính của trường Đại học Bách khoa-ĐHQG TP.HCM để thực hiện tiền xử lý, như phân đoạn từ, gán nhãn từ loại, xác định các cụm từ, phân tích nội dung câu hỏi tiếng Việt- để thực hiện ***bước d).***

- Nghiên cứu một số ứng dụng mã nguồn mở GATE (mô đun Jape) để sinh tự động bộ phân tích cú pháp và xử lý ngữ nghĩa cho câu tiếng Việt trên cơ sở ***bước b), c) và f)*** để hiện *thực* ***bước e) và g)***

- Sử dụng WordNet như là cơ sở tri thức phân cấp nghĩa từ để xác định ràng buộc ngữ nghĩa. Từ đó làm cơ sở xây dựng giải thuật giải quyết nhập nhằng dạng diễn dịch ngữ nghĩa của câu truy vấn- để thực hiện ***bước h)***

• ***Điều kiện thực hiện***

- Nghiên cứu lý thuyết ngôn ngữ: nghiên cứu văn phạm phụ thuộc có ràng buộc ngữ nghĩa (Constraint Dependency Grammar-CDG của Maruyama,1990 và Schr"oder, 2002) để áp dụng phân tích cú pháp và ngữ nghĩa cho các câu tiếng Việt;

- Nghiên cứu văn phạm tiếng Việt để xác định các dạng câu hỏi, bao gồm câu hỏi tổng quát, cấu hỏi thành phần, câu hỏi lựa chọn …; nghiên cứu cấu trúc cụm từ: danh từ, động từ để xác định các cụm từ chức năng ngữ nghĩa tiếng Việt. Cơ sở lý thuyết ngôn ngữ tiếng Việt nhóm tác giả dựa vào tài liệu của "*Diệp Quang Ban, 2004. Ngữ pháp tiếng Việt , tập 1,2. NXB Giáo dục"* và "Hồ Lê, 1993. *Cú pháp tiếng Việt – Cú pháp tình huống.* NXB Khoa học Xã hội".

Sau đây là các cấu trúc cụm danh từ tiếng Việt mà chúng tôi đã nghiên cứu, với T1 và T2 là danh từ trung tâm; t1 và t2 các từ phụ trước cho danh từ trung tâm; s1 và s2 là các từ phụ sau

t1 t2 T1 T2 s1 s2	t1 t2 T1 s1 s2	t1t2T2s1	t1T2
t1 t2 T1 T2 s1	t1T1s1s2	t2 T1 s1	
t1 T1 T2 s1	s1 s2	t2 T2 s1	
t2 T1 T2 s1	t1 t2 T2 s1 s2	t1 t2 T1 T2 s2	
t1 t2 T1 s1	t1 T2 s1 s2	t1 T1 T2 s2	
t1 T1 s1	t2 T2 s1 s2	t2 T1 T2 s2	
t2 T1 s1	t1 t2 T1 T2 s1	t1 t2 T1 s2	
t1 t2 T2 s1	t1 T1 T2 s1	t1 T1 s2	
t1 T2 s1	t2 T1 T2 s1	t2 T1 s2	
t2 T2 s1	t1 t2 T1 s1	t1 t2 T2 s2	
t1 T1 T2 s1 s2	t1 T1 s1	t1 T2 s2	
t2 T1 T2 s1 s2	t2 T1 s1	t2 T2 s2	

- Phát triển các kết quả nghiên cứu của đề tài Vietnamese Information Retrieval (của

nhóm tác giả) vào phân tích câu truy vấn tiếng Việt.

- Kết hợp với công tác nghiên cứu, hướng dẫn nghiên cứu sinh, đây là nguồn lực để thực hiện các công việc đã nêu trên.

2. Truy vấn thông tin (có hỗ trợ tiếng Việt) dạng văn bản trong hệ thống truy xuất thông tin xuyên ngôn ngữ

Hiện nay nhu cầu truy xuất thông tin giữa các ngôn ngữ ngày một cao, để có thể đáp ứng yêu cầu này, thì hệ thống truy xuất thông tin (Information Retieval- IR) phải được phát triển theo hướng kết hợp với xử lý ngôn ngữ tự nhiên nhằm thực hiện việc hỗ trợ chuyển ngữ cho truy vấn, đó chính là lĩnh vực CLIR (Cross Language Information Retrieval). Thực tế, khi nói đến CLIR người ta thường hướng đến truy xuất thông tin xuyên ngôn ngữ dạng văn bản (Cross-Language Text Retrieval - CLTR) vì hầu hết các nguồn thông tin thường được thể hiện ở dạng văn bản. Chúng tôi đã thực hiện xây dựng hệ thống truy xuất thông tin xuyên ngôn ngữ Việt-Anh **(VIRS)** như sau.

• ***Cách thức thực hiện***

Hầu hết các hệ thống CLIR/CLTR đều được phân làm ba nhóm chính, đó là xử lý theo hướng chuyển ngữ cho truy vấn ở ngôn ngữ nguồn cho phù hợp với ngôn ngữ của thông tin cần tìm (Query Translation) ; xử lý theo hướng chuyển ngữ cho thông tin cần tìm phù hợp với ngôn ngữ của truy vấn (Document Translation) và dạng kết hợp hai loại trên (interlingual techniques cho). Phương pháp của chúng tôi theo hướng thứ nhất có hỗ trợ tiếng Việt và ngôn ngữ thông tin cần tìm là tiếng Anh. Phần này gồm các bước sau.

a) Truy vấn tiếng Việt nhập vào được xử lý nhằm rút trích các cụm danh từ, mang nghĩa của truy vấn. Vì theo thống kê hơn 85% các truy vấn là ở dạng cụm danh từ. Để thực hiện rút trích các cụm danh từ cần tiến hành tiền xử lý: phân đoạn từ, gán nhãn từ loại, xác định các cụm danh từ. Từ các cụm danh từ hệ thống xác định các cụm từ đặc trưng ngữ nghĩa của câu truy vấn.

b) Một số đặc trưng của tiếng Việt sẽ gây khó khăn khi xây dựng hệ thống IR. Vì vậy từ một số kết quả thử nghiệm, chúng tôi đã chọn loại chỉ mục phù hợp cho tiếng Việt là các cụm từ với 2 đến 3 từ kết hợp với danh mục từ sẽ cho độ chính xác cao nhất.

c) Xây dựng từ điển song ngữ phù hợp cho phương pháp Query Translation để thực hiện việc chuyển ngữ cho cụm từ tiếng Việt sang tiếng Anh trước khi thực hiện việc truy xuất thông tin.

d) Chuyển ngữ cụm danh từ tiếng Việt sang tiếng Anh.

e) Sau khi thực hiện ***bước d)*** có thể có nhiều hơn một cụm danh từ truy vấn tiếng Anh được tạo ra. Vì vậy cần xử lý nhập nhằng cho cụm danh từ tiếng Anh.
f) Mở rộng truy vấn tiếng Anh.
g) Tích hợp các cụm danh từ tiếng Anh ở ***bước d)-f)*** theo dạng diễn dịch ngữ nghĩa cho truy vấn tiếng Anh.
h) Thực hiện tìm kiếm thông tin tương ứng cho các câu truy vấn tiếng Anh ở ***bước g)*** thông qua động cơ tìm kiếm thông tin như Google, Yahoo …

• ***Phương tiện thực hiện:***

- Xây dựng từ điển song ngữ Việt-Anh để thực hiện ***bước c):***
Để thiết kế cấu trúc và nguồn dữ liệu từ điển, với tiêu chí là một từ điển không thuộc dạng từ điển giải nghĩa, có thể hỗ trợ việc truy xuất nội dung một cách tự động thông qua chương trình, được gọi là từ điển máy khả đọc (Machine Readable Dictionary-MRD, chúng tôi dựa vào từ điển Hồ Ngọc Đức *(http://www.informatik.uni-leipzig.de/~duc/Dict/index.html)* , vì đây là mã nguồn mở nên có thể phát triển cho phù hợp hệ thống CIR có hỗ trợ tiếng Việt. Từ diển song ngữ Việt-Anh do chúng tôi xây dựng bao gồm: danh mục các từ, cụm từ và thành ngữ chuyên dụng (collocation) nhằm tăng khả năng đáp ứng của chúng trong quá trình chuyển ngữ truy vấn tiếng Việt.
Theo cấu trúc này, kho ngữ liệu mà chúng tôi đề xuất tích hợp hai phần từ điển Anh và Việt, các phần này liên kết với nhau thông qua một bảng ánh xạ chung (*dựa theo nguyên tắc thiết kế cơ sở dữ liệu nhằm giải quyết trường hợp một từ tiếng Việt liên kết với nhiều từ tiếng Anh và ngược lại*).
- Để thực hiện chuyển ngữ câu truy vấn cho CLIR chúng tôi sử dụng từ điển MRD và áp dụng và mở rộng phương pháp xử lý của C.Monz và Bonnie J.Dorr để chuyển ngữ từ tiếng Việt sang Anh - thực hiện ***bước d) và e)***
Giải thuật xử lý chuyển ngữ của chúng tôi có tên ***KeyPhrases Disambiguation and Translation (KPDT)***. Giải thuật này là sự mở rộng cho thuật toán Iterative Translation Disambiguation của C.Monz và Bonnie J.Dorr:

+ Trong trường hợp kết hợp với từ điển đặc trưng (song ngữ)

+ Truy vấn có các cụm danh từ đặc trưng

+ Sử dụng kết quả tìm kiếm từ *search engines* trong việc tính toán trọng số liên kết

+ Mở rộng mạng đồng hiện thành đồ thị liên kết "sao"

+ Giải pháp chọn kết quả cực đại cục bộ và mức ngưỡng

+ Xử lý tổ hợp các từ thành cụm sau khi chuyển ngữ đơn giản

- Danh sách các truy vấn tiếng Anh sau khi được chuyển ngữ sẽ dùng cho ứng dụng CIR, được biểu diễn ở dạng tổ hợp các cụm danh từ tiếng Anh với các phép toán luận lý AND, OR- thực hiện ***bước g)***

• ***Điều kiện thực hiện:***

- Nghiên cứu cấu trúc cụm từ: danh từ, động từ để xác định các cụm từ chức năng ngữ nghĩa tiếng Việt. Nhóm tác giả nghiên cứu lý thuyết ngôn ngữ tiếng Việt dựa vào tài liệu của "*Diệp Quang Ban, 2004. Ngữ pháp tiếng Việt , tập 1,2. NXB Giáo dục"* và " Hồ Lê, 1993. *Cú pháp tiếng Việt – Cú pháp tình huống*. NXB Khoa học Xã hội" tạo điều kiện thực hiện ***bước a)***

- Nghiên cứu cơ sở lý thuyết về cụm từ đặc trưng (phát triển kết quả nghiên cứu của đề tài Vietnamese Information Retrieval của nhóm tác giả)- ***tạo điều kiện thực hiện bước a)***

- Nghiên cứu cơ sở lý thuyết về truy vấn thông tin hướng đến ngữ nghĩa và khả năng hỗ trợ tiếng Việt- ***tạo điều kiện thực hiện bước g) và h).***

- Nghiên cứu một số ứng dụng mã nguồn mở GATE, Lucene và xác định giải pháp cho mở rộng cho tiếng Việt- ***tạo điều kiện thực hiện bước f)***

- Phát triển phương pháp tạo chỉ mục hướng đến ngữ nghĩa và tìm kiếm thông tin trong chỉ mục- ***tạo điều kiện thực hiện bước h:***

Tạo hai khối chức năng chính là bộ tạo chỉ mục hướng đến ngữ nghĩa (Semantic Indexer, SI) và bộ tìm kiếm truy vấn (Query Searcher, QS).

+ Bộ tạo chỉ mục hướng đến ngữ nghĩa (Semantic Indexer, SI) xử lý hai bước:

- Tạo chỉ mục tài liệu (Document Indexing) bằng các công cụ tạo chỉ mục sẵn có như Lucene, Lemure, …. để tạo hệ thống chỉ mục thông thường.

- Tạo chỉ mục hướng đến ngữ nghĩa (Semantic Indexing) để xây dựng hệ thống chỉ mục hướng đến ngữ nghĩa bằng việc bổ sung các thông tin từ các mạng ngữ nghĩa vào mỗi mục từ trong chỉ mục vừa được tạo. Những mạng ngữ nghĩa này được xây dựng tự động từ việc chọn lọc các mục từ trong chỉ mục ban đầu cùng với các quan hệ liên kết.

+ Bộ tìm kiếm truy vấn (Query Searcher, QS):

Trong khối này, người dùng có thể nhận được dữ liệu thông qua việc xem xét hay tìm kiếm không chỉ cho truy vấn dạng đơn giản mà còn cho truy vấn dạng phức tạp chứa đựng các từ khóa và thuộc tính ngữ nghĩa.

Trong hệ thống truy xuất thông tin hướng đến ngữ nghĩa (SIRS), bộ chỉ mục hóa ngữ nghĩa đảm trách tác vụ chỉ mục hóa để tạo chỉ mục hướng đến ngữ nghĩa. Tác vụ này được thực hiện dựa theo thuật toán **SemanticIndexCreating** do chúng tôi xây dựng.

- Để hiện thực việc mở rộng truy vấn, có khả năng liên kết với động cơ tìm kiếm thông tin chúng tôi đã xây dựng giải thuật "***Mở rộng cụm danh từ tương đương"- SNPE*** cho phép đầu vào là dạng cụm danh từ tiếng Anh. Giải thuật này bổ sung một số tính chất đặc trưng cho từng đối tượng trong truy vấn. Việc bổ sung này sẽ thêm các phần tử mới vào cụm danh từ của truy vấn, tạo ra sự mở rộng truy vấn-***tạo điều kiện thực hiện bước f)***

3. Suy diễn tìm kiếm nội dung trả lời cho câu hỏi (có hỗ trợ tiếng Việt) dạng văn bản trong một miền/lĩnh vực thông tin xác định trước.

● ***Cách thức thực hiện***

a) Thiết kế cơ sở tri thức tiếng Việt (**VKB**):

* *Phương pháp tổ chức cơ sở tri thức*

Chúng tôi chọn Protont làm cơ sở để phát triển cơ sở tri thức (ontology) cho hỏi đáp và truy vấn thông tin dạng văn bản có hỗ trợ tiếng Việt.

Cấu trúc của ontology thể hiện như sau:
Đối tượng có 2 thành phần: chủ thể, sản phẩm
Sản phẩm có 2 thành phần: tài liệu truyền thông, tài liệu được xuất bản.
Tài liệu được xuất bản gồm có: bài báo, sách, tài liệu xuất bản định kỳ.
Sách gồm có: xuất bản độc lập, kỷ yếu hội nghị
Tài liệu xuất bản định kỳ
Các lớp trong ontology được kế thừa từ Protont và được chúng tôi thay đổi, mở rộng để phù hợp với yêu cầu của hệ thống. Ví dụ: lớp Tác_giả và các quan hệ của nó được thêm vào ontology để lưu trữ thông tin về tác giả của các tài liệu.

* *Thiết kế cơ sở tri thức*

Cơ sở tri thức của hệ thống gồm các thông tin về tác giả, sách, bài báo, chủ đề, … cũng như các tài liệu dạng điện tử. Các thông tin này được lưu trữ trong cơ sở dữ liệu bên dưới tương ứng với các lớp trong ontology. Ví dụ: ứng với lớp Tác_giả trong ontology là bảng t_Tac_gia lưu trữ các thông tin về tác giả đó như tên, tuổi, …
Để truy xuất cơ sở tri thức, hệ thống sử dụng ngôn ngữ truy vấn cơ sở tri thức: ngôn ngữ truy vấn cơ sở tri thức phụ thuộc vào cấu trúc cơ sở tri thức và ngôn ngữ biểu diễn tri thức. Trong VKB, ngôn ngữ truy vấn sử dụng là SQL.

Mô hình hướng đối tượng: cung cấp các lớp hướng đối tượng tương ứng với các lớp được mô tả trong ontology. Các lớp này hiện thực các phương thức truy xuất kết quả trả về từ truy vấn, các phương thức thể hiện hành vi, quan hệ của các đối tượng trong cơ sở tri thức.

Ví dụ: ứng với lớp Sách trong ontology, chúng ta có các lớp hướng đối tượng DBook và OBook. DBook cung cấp các phương thức thao tác với dữ liệu thông qua ngôn ngữ SQL như: getBooks, getAuthors, … OBook cung cấp các phương thức thể hiện hành vi của đối tượng Sách như: getAuthors, setAuthors, setPublisher, … Các thao tác với dữ liệu của OBook được thực hiện thông qua DB

** Huấn luyện cơ sở tri thức*

Một cơ sở tri thức phong phú với lượng dữ liệu tương đối lớn là cần thiết để nâng cao sức mạnh của hệ thống. Tuy nhiên, việc đưa một lượng dữ liệu lớn vào cơ sở tri thức bằng tay là không mấy khả thi. Vì vậy, đề tài hướng đến việc (bán) tự động xây dựng cơ sở tri thức từ các tài liệu điện tử, như các bài báo, sách điện tử, …

Quá trình huấn luyện được thực hiện như sau:

- Rút trích thông tin
- Phân loại chủ đề

Các tài liệu điện tử sẽ được *rút trích các thông* về tác giả, nhà xuất bản, cụm từ khoá. Mô-đun *học phân loại chủ đề* sẽ nhận các thông tin cần thiết và phân loại tài liệu. Sau khi được chỉnh sửa bằng tay, các tài liệu sẽ được lưu trữ trong cơ sở tri thức.

+ Rút trích thông tin:

Rút trích thông tin là rút trích các cụm từ khoá (dựa trên giải thuật Naïve Bayes)

Một vấn đề quan trọng trong việc rút trích thông tin từ các tài liệu điện tử là rút trích các cụm từ khoá, thuật ngữ mô tả nội dung, chủ đề của tài liệu. Rút trích thông tin được thực hiện 2 bước: rút trích cụm từ dự tuyển và rút trích cụm từ khóa:

Rút trích các cụm từ khoá dự tuyển : Cụm từ khoá dự tuyển là những từ mang đặc điểm: Không vượt quá ba từ; không phải là tên riêng; không bắt đầu hoặc kết thúc với những từ dừng (stop-word).

Đối với tiếng Anh, các cụm từ dự tuyển được xem xét sẽ là những cụm danh từ (được rút trích bằng công cụ GATE). Còn đối với tiếng Việt, ngoài các đặc điểm như trên, cụm từ dự tuyển phải bắt đầu bằng một danh từ, (cụ thể là các danh từ đơn thể, danh từ trừu tượng, danh từ tổng thể).

Rút trích các cụm từ khoá: Hai đặc trưng chúng tôi sử dụng trong quá trình huấn luyện là:

- Sự xuất hiện của cụm từ mô tả số lần cụm từ đó xuất hiện trong tập huấn luyện.

- **Mẫu từ loại của cụm từ** là mẫu từ loại của các từ trong cụm từ.

Ở bước huấn luyện, các tài liệu đã được xác định cụm từ khoá bằng tay sẽ được sử dụng để tạo ra mô hình huấn luyện rút trích cụm từ khoá. Các tài liệu huấn luyện này sẽ được rút trích các cụm từ khoá dự tuyển. Tiếp đó, các giá trị đặc trưng sẽ được tính toán (giải thuật bước đầu sử dụng là Naïve Bayes). Các cụm từ khoá dự tuyển sẽ được đánh dấu là cụm từ khoá hay không dựa trên kết quả được xác định bằng tay và được đưa vào dữ liệu huấn luyện.

Khi cần rút trích cụm từ khoá cho các tài liệu mới, các bước tương tự quá trình huấn luyện được thực hiện. Các cụm từ khoá dự tuyển sẽ được rút trích. Tiếp đó, các giá trị đặc trưng sẽ được tính toán.

+ *Phân loại theo chủ đề:*

Sau khi rút trích các từ khoá của tài liệu, sẽ là bước phân loại theo chủ đề. Chúng tôi sử giải thuật phân loại theo chủ đề được trình bày như sau:

Đầu tiên, các cụm từ khoá đặc trưng cho ngữ nghĩa của các chủ đề sẽ được lan truyền ngược từ nút lá đến nút gốc của cây chủ đề. Tiếp đó cây chủ đề sẽ được duyệt từ gốc để tìm nút thích hợp (chủ đề thích hợp của tài liệu).

b) Suy diễn tìm nội dung trả lời:

Quan hệ ngữ nghĩa giữa những đối tượng trong **VKB** đóng vai trò quan trọng xác định thông tin cần truy xuất trong hệ thống **VIRS** và **VQAS**. Để biểu diễn quan hệ này, chúng tôi sử dụng đồ thị. Cơ sở lý thuyết đồ thị sẽ hỗ trợ trong việc suy diễn để tìm thông tin trả lời .

Từ đồ thị biểu diễn quan hệ ngữ nghĩa nêu trên, dựa trên giải thuật A*, chúng tôi xây dựng cơ chế suy diễn tìm ra nội dung trả lời. Hơn nữa, đồ thị quan hệ này còn hỗ trợ xác định các mẫu câu hỏi liên quan đến **VKB**.

•Phương tiện thực hiện

- Cơ sở tri thức được xây dựng trong đề tài hướng đến việc lưu trữ dữ liệu cho hệ thống hỏi đáp thư viện điện tử. Do đó, cơ sở tri thức (ontology) phải nắm bắt ngữ nghĩa của các khái niệm liên quan đến thư viện điện tử như: Sách, Bài báo, Tạp chí, Tác giả, Nhà xuất bản … Sau khi nghiên cứu các dạng ontology nổi tiếng như Protont, Wordnet, OpenCyc, cơ sở tri thức của các hệ thống British Telecommunication và Aqualog, chúng tôi kế thừa các lớp trong ontology Protont và thay đổi, mở rộng để phù hợp với yêu cầu của hệ thống.

- Cài đặt giải thuật phân tích suy diễn tìm nội dung trả lời. Để đạt được hiệu quả tốt trong việc xây dựng giải thuật suy diễn, chúng tôi sử dụng các kỹ thuật tiên tiến, công cụ hỗ trợ trong lĩnh vực công nghệ phần mềm: công cụ MS SQL Server và tập lệnh T-SQL (ngôn ngữ lệnh truy xuất cơ sở dữ liệu).

• ***Điều kiện thực hiện:***

Điều kiện để thực hiện Thiết kế cơ sở tri thức tiếng Việt (**VKB**)-***bước a):***

- Nghiên cứu và áp dụng giải thuật Naïve Bayes để xác định cụm danh từ dự tuyển có là cụm từ khoá hay không:

$$P_{[yes]} = \frac{Y}{Y+N}.P_{[w|yes]}.P_{[p|yes]}$$

Trong đó:

(i) $P_{[yes]}$: xác suất cụm từ dự tuyển là cụm từ khoá
(ii) Y: số cụm từ khoá trong tập huấn luyện
(iii) N: số cụm từ dự tuyển không là khoá
(iv) $P_{[w|yes]}$: xác suất *w* là cụm từ khoá
(v) $P_{[p|yes]}$: xác suất cụm từ có mẫu *p* là cụm từ khoá

Tương tự, ta có xác suất cụm từ dự tuyển để có thể xác định không là cụm từ khoá.

- Trong **VKB**, việc phân loại văn bản theo chủ đề được thực hiện theo giải thuật ***" Phân loại văn bản trong VKB".***
- Để huấn luyện cơ sở tri thức chúng tôi đã đề xuất giải thuật "***Huấn luyện cơ sở tri thức VKB"*.**
- Để thực hiện suy diễn tìm nội dung trả lời, dựa trên ý tưởng giải thuật A*, chúng tôi xây dựng phương pháp suy diễn tìm nội dung trả lời phù hợp cho hệ thống có hỗ trợ tiếng Việt.

Mô tả vắn tắt các hình vẽ

Hình 1. Sơ đồ hệ thống VietSIRS

Hình 2. Cấu trúc chức năng hệ thống VietSIRS

Hình 3. Phương pháp phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt

Hình 4. Phân tích nội dung câu hỏi tiếng Việt trong một miền xác định

Hình 5. Minh họa cây phân tích của văn phạm phụ thuộc

Hình 6. Cây phân tích sau khi xử lý ngữ nghĩa

Hình 7. Các khối chức năng thực hiện công việc ở bước 2

Hình 8. Sơ đồ giải thuật chuyển ngữ các cụm từ khóa tiếng Việt sang tiếng Anh

Hình 9. Sơ đồ giải thuật tạo chỉ mục hướng đến ngữ nghĩa cho truy xuất thông tin (bước 2)

Hình 10. Chỉ mục hướng đến ngữ nghĩa trong truy xuất thông tin xuyên ngôn ngữ (bước 2)

Hình 11. Xác định dạng thức truy vấn quy ước

Hình 12. Cấu trúc liên kết các thông tin trong cơ sở tri thức

Hình13. Đường liên kết trên đồ thị G phục vụ cho suy diễn tìm nội dung thông tin

Hình 14. Hệ thống hỗ trợ cho huấn luyện Ontology: các khối chức năng

Hình 15. Sơ đồ giải thuật tìm nội dung câu trả lời

Hình 16. Sơ đồ giải thuật suy diễn tìm nội dung trả lời

Hình 17. Phương thức suy diễn tìm nội dung trả lời cho miền thông tin là các tài liệu khoa học

Mô tả chi tiết các phương án thực hiện sáng chế

• ***Phương án thực hiện***

Phương pháp phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt được thực hiện trên hệ thống ở hình 1, với ba khối chức năng là ba hệ thống con:

- Hệ thống truy xuất thông tin (gồm truy xuất thông tin xuyên ngôn ngữ Việt-Anh) **VIRS**
- Hệ thống hỏi đáp có hỗ trợ tiếng Việt **VQAS**
- Để phục vụ cho hai hệ thống trên hoạt động là Cơ sở tri thức có hỗ trợ tiếng Việt **VKB.**

Ngoài ba khối cấu trúc chức năng như đặc tả trong hình1 và hình 2 là VIRS, VQAS và VKB, hệ thống VietSIRS còn có một hệ thống hỗ trợ. Hệ thống này làm nhiệm vụ rút trích thông tin và phân tích những thông tin cần thiết để huấn luyện, làm giàu cơ sở tri thức.

Để đạt được các mục tiêu xây dựng hệ thống VietSIRS, các vấn đề cần thực hiện là:

1. Nghiên cứu cơ sở lý thuyết

1.1. Nghiên cứu lý thuyết ngôn ngữ

Nghiên cứu văn phạm phụ thuộc có ràng buộc ngữ nghĩa; nghiên cứu văn phạm tiếng Việt để xác định các dạng câu hỏi, bao gồm câu hỏi tổng quát, cấu hỏi thành phần, câu hỏi lựa chọn ...; nghiên cứu cấu trúc cụm từ: danh từ, động từ để xác định các cụm từ chức năng ngữ nghĩa tiếng Việt; nghiên cứu cơ sở lý thuyết về cụm từ đặc trưng và khả năng ứng dụng trong tiếng Việt (phát triển kết quả nghiên cứu của đề tài Vietnamese Information Retrieval của nhóm tác giả).

1.2 Nghiên cứu cơ sở lý thuyết về phân tích cú pháp, ngữ nghĩa

Nghiên cứu phương pháp xây dựng tập luật cú pháp và phương pháp xử lý ngữ nghĩa trên cơ sở văn phạm phụ thuộc và ràng buộc ngữ nghĩa giữa các đối tượng trong câu truy vấn để phân tích các dạng câu hỏi và xây dựng các câu trả lời có hỗ trợ tiếng Việt.

1.3. Nghiên cứu cơ sở lý thuyết về truy vấn thông tin hướng đến ngữ nghĩa và khả năng hỗ trợ tiếng Việt

Nghiên cứu lý thuyết để xây dựng phương pháp tạo chỉ mục hướng đến ngữ nghĩa và tìm kiếm thông tin trong chỉ mục.

1.4 Nghiên cứu phương pháp xây dựng cơ sở tri thức

Xây dựng cơ sở tri thức bao gồm: xác định phương pháp tổ chức cơ sở tri thức, thiết kế cơ sở tri thức, huấn luyện cơ sở tri thức.

Chúng tôi chọn Protont làm cơ sở để phát triển cơ sở tri thức (ontology) cho hỏi đáp và truy vấn thông tin dạng văn bản có hỗ trợ tiếng Việt.

Thiết kế cơ sở tri thức

Cơ sở tri thức của hệ thống gồm các thông tin về tác giả, sách, bài báo, chủ đề, ... cũng như các tài liệu dạng điện tử. Các thông tin này được lưu trữ trong cơ sở dữ liệu bên dưới tương ứng với các lớp trong ontology.

Huấn luyện cơ sở tri thức

Quá trình huấn luyện được thực hiện:

- Rút trích thông tin
- Phân loại chủ để

2. Nghiên cứu một số công nghệ hỗ trợ

2.1 Tham khảo các ontology của WordNet, OpenCyc, Prontont để hỗ trợ xây dựng cơ sở tri thức

2.2 Nghiên cứu một số ứng dụng mã nguồn mở

Sử dụng GATE để xác định các cụm từ, JAPE (thuộc GATE) để tạo bộ phân tích cú pháp có ràng buộc ngữ nghĩa; sử dụng GATE, Lucene hỗ trợ cho giải pháp mở rộng cho tiếng Việt.

3. Thiết kế tổ chức cơ sở tri thức tiếng Việt (VKB)

3.1 Thiết kế cơ sở tri thức và huấn luyện tiếng Việt

Sau khi nghiên cứu cứu phương pháp xây dựng cơ sở tri thức, chúng tôi chọn Protont làm cơ sở để phát triển cơ sở tri thức (ontology) cho hỏi đáp và truy vấn thông tin dạng văn bản có hỗ trợ tiếng Việt. Thiết kế cơ sở tri thức gồm các bước:

Thiết kế cơ sở tri thức :

Cơ sở tri thức của hệ thống gồm các thông tin về tác giả, sách, bài báo, chủ đề, ... cũng như các tài liệu dạng điện tử. Các thông tin này được lưu trữ trong cơ sở dữ liệu bên dưới tương ứng với các lớp trong ontology.

Huấn luyện cơ sở tri thức

Quá trình huấn luyện được thực hiện:

- Rút trích thông tin
- Phân loại chủ đề

Cơ sở tri thức được thiết kế theo cấu trúc tổ chức được mô tả trong hình 12:

Trong cơ sở tri thức này, các dữ liệu được lưu trữ trong những nhóm dữ liệu thành phần, ví dụ như Authors, Papers ...

Từ cấu trúc này, quan hệ giữa các thành phần được biểu diễn trong hình 13

Việc huấn luyện cho cơ sở tri thức có hỗ trợ tiếng Việt được mô tả ở hình 14 . Phương thức này có thể hoạt động độc lập trong một điều kiện bất kỳ.

3.2 Xây dựng phương thức khai thác cơ sở tri thức tiếng Việt

Việc khai thác cơ sở tri thức được cài đặt thành chương trình máy tính để phục vụ rút trích các dữ liệu cần thiết từ cơ sở tri thức.

3.3 Hiện thực động cơ huấn luyện cơ sở tri thức

Từ phương pháp trên, chúng tôi xây dựng chương trình máy tính và cài đặt như mã nguồn nhằm thực hiện giải pháp đã nêu, từ đó tạo nên động cơ huấn luyện cơ sở tri thức.

4. Xây dựng hệ thống xử lý truy vấn tương tác

4.1 Tiền xử lý các truy vấn

Xây dựng phương thức thực tiền xử lý, như phân đoạn từ, gán nhãn từ loại, xác định các cụm từ, phân tích nội dung câu hỏi tiếng Việt. Công việc này được mô tả trong hình 4.

4.2 Xây dựng dạng truy vấn quy ước

Định dạng truy vấn quy ước được nhóm tác giả đề xuất để hỗ trợ người dùng thực hiện việc tìm kiếm thông tin cần thiết từ mức độ đơn giản đến phức tạp. Các dạng truy vấn quy ước sau đây diễn đạt truy vấn ở dạng nhóm các từ khóa cùng tính chất tương ứng.

** Dạng cơ bản F1*

q=<[-]keyword$_1$>[<*/+[-]keyword$_2$>] … [<*/+[-]keyword$_n$>]

Ở đây keyword$_1$, keyword$_2$, …, keyword$_n$ là những trị từ vựng (với đa số trường hợp là danh từ hay cụm danh từ).

Dạng này cho phép người dùng tìm kiếm theo từ khóa như thông thường, và có hỗ trợ các phép toán * (and), + (or) và – (not) cho từ khóa.

** Dạng cơ bản F2*

q=<[-]keyword$_1$:[-]attribute$_1$> [<*/+[-]keyword$_2$:[-]attribute$_2$>]

…[<*/+[-]keyword$_n$:[-]attribute$_n$>]

Tương tự như trên, keyword$_1$, keyword$_2$, …, keyword$_n$ là những trị từ vựng (với đa số trường hợp là danh từ hay cụm danh từ), và attribute$_1$, attribute$_2$, …, attribute$_n$ cũng là những trị từ vựng mô tả tính chất tương ứng với những từ khóa.

Dạng này cho phép người dùng tìm kiếm theo từ khóa và tính chất ngữ nghĩa, và có hỗ trợ các phép toán * (and), + (or) và – (not) cho từ khóa cùng tính chất ngữ nghĩa.

** Dạng cơ bản F3*

q=<[-]keyword$_1$*/+…*/+[-]keyword$_n$:[-]attribute>

Dạng này cho phép người dùng tìm kiếm theo từ khóa và tính chất ngữ nghĩa, và có hỗ trợ các phép toán * (and), + (or) và – (not) cho từ khóa, nhưng chỉ hỗ trợ phép toán – (not) cho tính chất ngữ nghĩa.

** Dạng cơ bản F4*

q=<[-]keyword:[-]attribute$_1$*/+…*/+[-]attribute$_n$>

Dạng này cho phép người dùng tìm kiếm theo từ khóa và tính chất ngữ nghĩa, và có hỗ trợ các phép toán * (and), + (or) và – (not) cho tính chất ngữ nghĩa, nhưng chỉ hỗ trợ phép toán – (not) cho từ khóa.

** Dạng phức hợp Fc*

q=<[-]keyword$_1$*/+…*/+[-]keyword$_n$:[-]attribute$_1$*/+…*/+[-]attribute$_m$>

Dạng này cho phép người dùng tìm kiếm thông tin theo nhóm từ khóa và nhóm tính chất ngữ nghĩa, có sự hỗ trợ của phép toán * (and), + (or), – (not) cho từ khóa và tính chất ngữ nghĩa.

Dạng này có thể được thể hiện theo một trong hai kiểu sau:

$q=q_1$ */+ q_2 */+…*/+ q_n

trong đó:

q_i=<[-]keyword$_1$*/+… */+[-]keyword$_n$:[-]attribute$_i$> (dạng F3)

hay là:

q_i=<[-]keyword$_i$:[-]attribute$_1$*/+…*/+[-]attribute$_n$> (dạng F4)

4.3 Phát triển hệ thống xử lý truy vấn dựa trên cơ sở tri thức tiếng Việt (VKB).

Phương pháp tiền xử lý truy vấn được xây dựng thành chương trình máy tính và được cài đặt vào hệ thống để hiện thực giải pháp đã nêu, từ đó tạo nên hệ thống xử lý truy vấn dựa trên cơ sở tri thức tiếng Việt.

5. Thiết kế mô hình hệ thống truy vấn thông tin hướng đến ngữ nghĩa có hỗ trợ tiếng Việt (VietSIRS)

a) Hệ thống Truy xuất thông tin Anh-Việt trực tuyến (VIRS)

5.1 Xây dựng và phát triển phương thức tạo chỉ mục hướng đến ngữ nghĩa và tìm kiếm thông tin trong chỉ mục.

Chỉ mục hướng đến ngữ nghĩa được nhóm tác giả xây dựng nhằm bổ sung các thông tin hữu ích vào hệ thống chỉ mục sẵn có của một hệ thống truy hồi thông tin. Trong chỉ mục hướng đến ngữ nghĩa, các thông tin bổ sung giúp hình thành các nhóm liên kết giữa các đầu mục con trong chỉ mục. Mỗi nhóm mang một ý nghĩa khác nhau và các phần tử trong nhóm có ý nghĩa và vai trò tương đương nhau. Việc xác định nhóm với các phần tử khác từ một phần tử đã có sẽ giúp việc tìm kiếm thông tin nhận được nhiều hơn kết quả tương tự nhau. Hình 9 là sự minh họa một chỉ mục hướng đến ngữ nghĩa.

Việc huấn luyện một chỉ mục, gọi là *chỉ mục hóa* hay *tạo chỉ mục* (indexing), được *bộ chỉ mục hóa* (indexer) thực hiện, là tác vụ quan trong nhất trong bất kỳ hệ thống truy vấn thông tin nào.

Trong hệ thống truy xuất thông tin hướng ngữ nghĩa, bộ chỉ mục hóa ngữ nghĩa thực hiện tác vụ chỉ mục hóa để tạo chỉ mục hướng đến ngữ nghĩa. Tác vụ này được thực hiện theo giải thuật trong hình10. Mỗi bước của giải thuật thực hiện một nhiệm vụ và trả về kết quả cho những bước kế tiếp.

Với cơ chế huấn luyện này, việc xác định $c(r_k)$ đã góp phần xây dựng các liên kết giữa những mục từ cùng thoả điều kiện, hay nói cách khác thì $c(r_k)$ là một đồ thị chứa các

mục từ và các quan hệ ngữ nghĩa liên quan giữa chúng.

5.2 Xây dựng và hiện thực phương thức truy vấn dựa trên chỉ mục hướng đến ngữ nghĩa

Phương pháp tìm kiếm thông tin theo chỉ mục hướng đến ngữ nghĩa, được trình bày trong hình 10.

5.3 Phát triển phương thức truy vấn thông tin xuyên ngôn ngữ Việt – Anh

Phương thức truy vấn thông tin xuyên ngôn ngữ Việt – Anh (phát triển từ kết quả nghiên cứu đề tài Vietnamese Information Retrieval của nhóm tác giả) nhằm phục vụ người dùng chuyển ngữ một nội dung truy vấn trong tiếng Việt sang tiếng Anh tương ứng. Trình tự xử lý chuyển ngữ được mô tả ở hình 8. Từ điển song ngữ Việt Anh giúp ra cứu nội dung chuyển ngữ mức từ vựng giữa một mục từ trong tiếng Việt và trả về các mục từ trong tiếng Anh tương ứng.

b) Xây dựng hệ thống hỏi đáp tiếng Việt (VQAS)

5.4 xây dựng phương thức nhận dạng câu hỏi trên cơ sở tập mẫu câu hướng đến ngữ nghĩa

Phương thức nhận dạng câu hỏi được trình bày trong sơ đồ ở hình 4. Mỗi bước trong sơ đồ thực hiện tuần tự và nhận kết quả từ các bước trước đó. Danh sách từ đồng nghĩa được rút trích từ từ điển tiếng Việt nhằm bổ sung từ đồng nghĩa cho một mục từ xác định trước, phục vụ việc xác định dạng câu và mở rộng các truy vấn. Danh sách phụ từ, danh từ, động từ của điển tiếng Việt hỗ trợ bước phân tích cú pháp và xử lý ngữ nghĩa. Kho dữ liệu chứa tập cây phân tích phụ thuộc kết quả sẽ cung cấp dữ liệu cho bước xử lý ngữ nghĩa và bước kết thúc.

5.5 Phát triển phương thức tạo nội dung trả lời dựa trên cơ sở tri thức tiếng Việt (VKB)

Tác vụ tạo nội dung trả lời được thể hiện trong sơ đồ giải thuật ở hình 15, trong đó kho thông tin cơ sở tri thức đóng vai trò là nguồn cung cấp dữ liệu cho các bước xử lý Chuẩn bị và Phân tích.

5.6 Phát triển phương thức hỏi đáp dựa trên tập mẫu câu và cơ sở tri thức tiếng Việt

Phương thức này gồm các bước sau:

- Nhận dạng câu hỏi dựa trên tập các mẫu hướng đến ngữ nghĩa
- Xác định nội dung trả lời đơn giản
- Phương pháp truy xuất thông tin các tài liệu khoa học
- Phương pháp mở rộng truy vấn thông tin

Các phương pháp con truy xuất thông tin bài báo khoa học và mở rộng truy vấn thông tin

được trình bày lần lượt sau đây trong các hình 14 và 15.

6. Phát triển một số phần mềm hỗ trợ cho hệ thống VietSIRS

Các phương pháp của các công việc từ 3 đến 6 được trình bày đã được xây dựng và cài đặt như mã nguồn của chương trình máy tính để thực hiện các giải pháp đã nêu, từ đó tạo nên những chương trình máy tính (phần mềm) phục vụ các công việcsau.

6.1 Khai thác dữ liệu VKB phục vụ cho công việc ở 5.3 và 5.6

6.2 Truy vấn thông tin phục vụ cho công việc ở 5.3

6.3 Hỏi đáp thông tin phục vụ cho công việc ở 5.6

- ***Giải pháp kỹ thuật, và thiết kế nghiên cứu***

1. Tổ chức hệ thống truy vấn thông tin hướng đến ngữ nghĩa hỗ trợ tiếng Việt (VietSIRS): bao gồm nhóm các hệ thống con VIRS, VQAS và VKB (thể hiện trong hình 1) được đề xuất trong mục tiêu của đề tài:

2. Nghiên cứu và đề xuất các phương pháp trong Truy vấn thông tin và Xử lý ngôn ngữ tự nhiên:

a. Phân tích cú pháp và xử lý ngữ nghĩa các câu truy vấn và câu hỏi

b. Xác định dạng truy vấn dựa trên tập mẫu câu

c. Hiểu ngữ nghĩa và chuyển các câu truy vấn sang dạng diễn dịch ngữ nghĩa .

d. Chọn lọc tập câu trả lời

e. Tạo chỉ mục hướng đến ngữ nghĩa và thực hiện truy vấn dữ liệu

3. Kiến trúc hệ thống đề xuất được trình bày ở hình 2 về mặt kỹ thuật.

4. Các hệ thống thành phần trong hệ thống VietSIRS có chức năng như động cơ xử lý, góp phần thực hiện toàn bộ hoạt động của hệ thống, bao gồm:

a. ***Hệ thống Hỗ trợ*** với các mô-đun của nó thực hiện công việc rút trích thông tin từ kho ngữ liệu và đồng thời từ WWW để phân tích các thông tin và cung cấp cho mô-đun huấn luyện của VKB. Việc phát triển hệ thống này, đặc biệt là vấn đề khai thác thông tin trên mạng Internet, dựa trên các công nghệ tiên tiến như Java hoặc Microsoft DotNet và các kỹ thuật lập trình mạng (network programming) liên quan. Việc phân tích các dữ liệu thu được (với dạng chủ yếu các văn bản, câu, từ trong ngôn ngữ tự nhiên tiếng Anh) bao gồm các bài toán tiền xử lý: phân đoạn từ (word segmentation), gán nhãn từ loại (Part Of Speech Tagging), xác định các cụm từ (phrase chunker) rất cần sự hỗ trợ của các công cụ xử lý ngôn ngữ tự nhiên sẵn có như GATE, Jape. Các bài toán liên quan đến phân tích tài liệu tiếng Việt cũng cần sử dụng các công cụ - là kết quả đã thực hiện được trong các đề tài NCKH trước đó của tập thể nhóm nghiên cứu xử lý ngôn ngữ tự nhiên của khoa Khoa học và Kỹ thuật máy tính thuộc trường Đại học Bách khoa- ĐHQG

TP.HCM như vnSegmentation, vnPosTagger...

b. ***Cơ sở tri thức VKB*** gồm cơ sở tri thức và hai mô-đun *Huấn luyện thông tin* và *Khai thác cơ sở tri thức* đóng vai trò như hệ thống dữ liệu trung tâm với các tác vụ: thực hiện việc bổ sung các dữ liệu (trong quá trình huấn luyện) và cung cấp khả năng xuất dữ liệu cho hệ thống khác hay mô-đun khác. Vấn đề huấn luyện cơ sở tri thức dựa trên các dữ liệu nhập ở bước (a) đòi hỏi:

   - Rút trích dữ liệu dựa trên các mẫu huấn luyện (training pattern) hay các tập luật sinh của văn phạm với sự hỗ trợ của các công cụ GATE, Jape...;
   - Kết hợp các phương pháp như thống kê, học máy... để chọn lọc các từ dự tuyển đạt yêu cầu cho việc lưu trữ.
   - Kết hợp với WordNet (hoặc một số ontology cho tiếng Anh đã có) để khai thác các nét ngữ nghĩa, quan hệ ngữ nghĩa của các từ và cụm từ phục vụ cho đề tài nghiên cứu có hỗ trợ tiếng Việt.
   - Đồng thời sử dụng công cụ/phần mềm quản trị cơ sơ sở dữ liệu như mySQL hay SQL Server, hoặc công cụ khác như Protegre để tổ chức lưu trữ và vận hành cơ sở trì thức.

c. Khai thác thông tin trong cơ sở tri thức để cung cấp thông tin cho hai hệ thống **VIRS** và **VQAS**. Việc hiện thực các động cơ khai thác này dựa trên các thư viện hỗ trợ của phần mềm (mySQL, SQL Server, ...) kết hợp với công nghệ lập trình tiên tiến.

d. ***Các mô-đun trong VIRS*** sẽ thực hiện truy vấn thông tin hướng đến ngữ nghĩa. Điểm cốt lõi trong hoạt động của **VIRS** chính là hệ thống chỉ mục hướng đến ngữ nghĩa. Dựa trên hệ thống chỉ mục được tạo bởi bất kỳ công cụ tạo chỉ mục thông thường nào (Lucene, Lemur, IRtools, ...), mô-đun *Huấn luyện & khai thác chỉ mục hướng đến ngữ nghĩa* sẽ phát triển một tầng (layer) xử lý ngữ nghĩa cho chỉ mục bằng cách tạo các phân lớp ngữ nghĩa (semantic layer) tương đương dựa trên các nét ngữ nghĩa cần thiết. Hệ thống chỉ mục hướng ngữ nghĩa này một mặt khai thác sức mạnh của các tổ hợp chỉ mục con thông thường, nhưng đồnmg thời sẽ bổ sung thêm các yếu tố và quan hệ ngữ nghĩa từ các phân lớp ngữ nghĩa, nhằm cung cấp các kết quả truy vấn rộng và chính xác hơn. Các dạng thức truy vấn mở rộng (thay cho các truy vấn đơn giản chỉ gồm từ khoá) sẽ giúp người sử dụng có thể khai thác các ưu điểm của hệ thống chỉ mục hướng ngữ nghĩa. Ngoài ra, việc truy vấn bằng tiếng Việt cũng là một nhiệm vụ trọng tâm của hệ thống **VIRS** với mục tiêu chuyển ngữ truy vấn từ tiếng Việt sang Anh để giúp người dùng có cơ hội khai thác thông tin cần thiết. Các kết quả liên quan đã đạt được (trong phạm vi đề tài NCKH trong giai đoạn trước đây) của nhóm nghiên cứu sẽ là cơ sở cho việc phát triển hệ thống **VIRS**.

e. ***Các mô-đun trong VQAS*** thực hiện việc tìm nội dung trả lời cho các câu hỏi của người dùng thông qua các quá trình xử lý: phân tích cú pháp và ngữ nghĩa câu hỏi

như minh họa ở hình 5 và 6; tạo diễn dịch ngữ nghĩa là dạng bộ ba luận lý trên cơ sở các tập luật mẫu; xác định các câu trả lời và tạo nội dung trả lời phù hợp theo ngữ cảnh.Việc khai thác các dữ liệu, các nét ngữ nghĩa và quan hệ ngữ nghĩa trong cơ sở tri thức được sử dụng để phân tích ngữ nghĩa các dạng câu hỏi WH như what, when, why, who, ... how, cũng như nâng cao độ chính xác của các thông tin cần truy xuất. Các kết quả nghiên cứu liên quan trên thế giới về Q&A (đặc biệt là Q&A trong tiếng Trung hoa, Hàn, Nhật là những tham khảo hữu ích cho nhóm tác giả để thực hiện đề tài với mục tiêu xây dựng và triển khai hệ thống Q&A cho tiếng Việt.

▫ ***Ví dụ minh họa giải pháp kỹ thuật***

Hình 3. Ví dụ minh họa cho giải pháp kỹ thuật

Những lợi ích (hiệu quả) có thể đạt được.

• ***Đối với lĩnh vực khoa học công nghệ***

Bồi dưỡng, đào tạo cán bộ KH&CN:

- Thông qua đề tài liên quan giải pháp kỹ thuật này, các cán bộ trực tiếp tham gia thực hiện sẽ có cơ hội tự bồi dưỡng, nâng cao trình độ chuyên môn của mình; tạo các nhóm nghiên cứu độc lập và mở rộng hợp tác với các chuyên gia nước ngoài trong lĩnh vực liên quan.

- Đề tài này có nhiều vấn đề nghiên cứu chuyên sâu, sẽ là những đề tài cho luận văn thạc sĩ và tiến sĩ. Khi thực hiện đề tài này, chúng tôi đã đào tạo 2 tiến sĩ, 6 thạc sĩ, và một số kỹ sư, cử nhân. Các học viên cao học, nghiên cứuvà sinh viên đại học đã tham gia giải quyết các vấn đề khác nhau thông qua luận văn của mình. Các kỹ sư, cử nhân, thạc sĩ, tiến sĩ được đào tạo thông qua đề tài liên quan đến giải pháp kỹ thuật sẽ làm việc trong lĩnh vực xử lý ngôn ngữ tự nhiên – truy vấn và hỏi đáp thông tin, đặc biệt xử lý văn bản tiếng Việt và nghiên cứu các hệ thống khai thác thông tin văn bản.

• ***Đối với lĩnh vực khoa học:***

- Các kết quả của đề tài này thuộc lĩnh vực xử lí ngôn ngữ tự nhiên, đặc biệt là xử lý tiếng Việt dạng văn bản. Đề tài đã đóng góp một số cải tiến về giải pháp công nghệ cho các hệ thống truy vấn thông tin và hỏi đáp hướng đến ngữ nghĩa nói chung.

- Các kết quả của đề tài cũng sẽ tạo tiền đề cho những nghiên cứu mới có ý nghĩa về lý thuyết và ứng dụng, ...

- ***Đối với tổ chức chủ trì và các cơ sở ứng dụng kết quả nghiên cứu***

- Đối với cơ quan chủ trì, lợi ích trực tiếp chính là việc thụ hưởng được thành quả về mặt khoa học và con người. Thông qua đề tài, một lớp cán bộ nghiên cứu được đào tạo cả về lý thuyết và phương pháp làm nghiên cứu khoa học cũng như kinh nghiệm thực tiễn, giúp họ nâng cao khả năng nghiên cứu, kinh nghiệm và kiến thức khoa học hiện đại. Các bài báo và công trình khoa học mà các thành viên tham gia sẽ góp phần nâng cao uy tín về khoa học cho tổ chức chủ trì..

- Đối với cơ sở sử dụng kết quả nghiên cứu, mặc nhiên họ được thừa hưởng những sản phẩm với công nghệ mới, phục vụ cho hoạt động khoa học, kinh doanh, sản xuất.

- ***Đối với kinh tế - xã hội và môi trường***

- Hệ thống mà đề tài sẽ xây dựng và phát triển, khi ứng dụng sẽ mang lại những lợi ích không chỉ về kinh tế mà còn về xã hội. Cụ thể, hệ thống sẽ hỗ trợ người dùng truy vấn thông tin một cách thông minh và uyển chuyển hơn, cho kết quả mang độ chính xác cao hơn. Đặc biệt hơn là hệ thống sẽ giúp các thư viện của các cơ quan trường học, viện nghiên cứu, các tòa soạn báo, triển khai phục vụ người dùng trong việc khai thác thông tin được hiệu quả hơn.

- Các kết quả của đề tài sẽ góp phần xây dựng và củng cố uy thế cạnh tranh cho các sản phẩm và công nghệ nội địa về Web có ngữ nghĩa, hệ thống hỏi đáp tự động, truy vấn thông tin đa phương tiện hướng đến ngữ nghĩa có hỗ trợ tiếng Việt trong tương lai với các sản phẩm cùng chủng loại của nước ngoài.

Yêu cầu bảo hộ

1. Phương pháp phục vụ hỏi đáp và truy xuất thông tin dạng văn bản có hỗ trợ tiếng Việt bao gồm các bước:

phân tích nội dung câu hỏi tiếng Việt trong một miền/lĩnh vực thông tin xác định trước thực hiện như sau: xác định miền/lĩnh vực thông tin để xác định các khái niệm, các đối tượng trừu tượng, vật lý và các mối quan hệ chức năng ngữ nghĩa giữa các đối tượng; xây dựng tập mẫu các dạng câu hỏi, các câu truy vấn tiếng Việt trên miền thông tin được xác định; xây dựng từ điển các từ trong phạm vi miền thông tin bao gồm cả các từ/các cụm từ đồng nghĩa; tiền xử lý câu truy vấn: phân đoạn từ, gán nhãn từ loại; xác định cấu trúc cụm danh từ, động từ, trạng từ; xác định các cụm từ chức năng ngữ nghĩa trong câu tiếng Việt; xây dựng tập luật phân tích cú pháp và ngữ nghĩa trên cơ sở văn phạm phụ thuộc và ràng buộc ngữ nghĩa giữa các đối tượng trong câu truy vấn; xác định dạng diễn dịch ngữ nghĩa cho các câu truy vấn; xây dựng giải thuật giải quyết diễn dịch ngữ nghĩa các câu truy vấn tiếng Việt;

truy vấn thông tin (có hỗ trợ tiếng Việt) dạng văn bản trong hệ thống truy xuất thông tin xuyên/đa ngôn ngữ thực hiện như sau: tiền xử lý: phân đoạn từ, gán nhãn từ loại; xác định các cụm danh từ; từ các cụm danh từ hệ thống xác định các cụm từ đặc trưng ngữ nghĩa của câu truy vấn; truy vấn tiếng Việt nhập vào được xử lý nhằm rút trích các cụm danh từ mang nghĩa của truy vấn; xây dựng hệ thống IR với chỉ mục phù hợp cho tiếng Việt là các cụm từ với 2 đến 3 từ kết hợp với danh mục từ; xây dựng từ điển song ngữ phù hợp cho phương pháp Query Translation để thực hiện việc chuyển ngữ cho cụm từ tiếng Việt sang tiếng Anh; chuyển ngữ cụm danh từ tiếng Việt sang tiếng Anh dùng giải thuật KeyPhrases Disambiguation and Translation; xử lý nhập nhằng cho các cụm danh từ tiếng Anh được chuyển ngữ từ cụm danh từ tiếng Việt tương ứng; mở rộng truy vấn tiếng Anh; tích hợp các cụm danh từ tiếng Anh theo dạng diễn dịch ngữ nghĩa cho truy vấn; thực hiện tìm kiếm thông tin tương ứng cho các câu truy vấn tiếng Anh thông qua động cơ tìm kiếm thông tin như Google, Yahoo …;

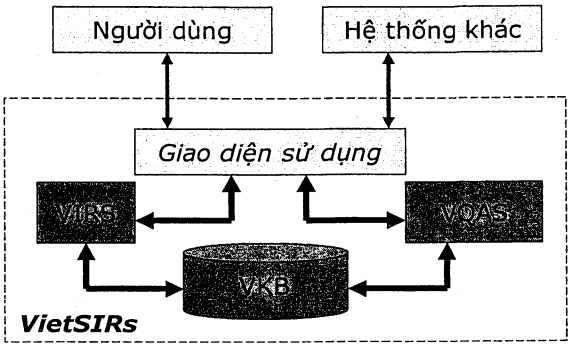
suy diễn tìm kiếm nội dung trả lời cho câu hỏi (có hỗ trợ tiếng Việt) dạng văn bản trong một miền/lĩnh vực thông tin xác định trước như sau: thiết kế cơ sở tri thức; truy vấn cơ sở tri thức; rút trích thông tin; phân loại theo chủ đề; các cụm từ khoá đặc trưng cho ngữ nghĩa của các chủ đề sẽ được lan truyền ngược từ nút lá đến nút gốc của cây chủ đề, cây chủ đề sẽ được duyệt từ gốc để tìm nút thích hợp (chủ đề thích hợp của tài liệu); suy diễn tìm nội dung trả lời.

2. Phương pháp theo điểm 1, trong đó các dạng câu truy vấn bao gồm câu hỏi tổng quát; câu hỏi xác định thành phần; câu hỏi lựa chọn.

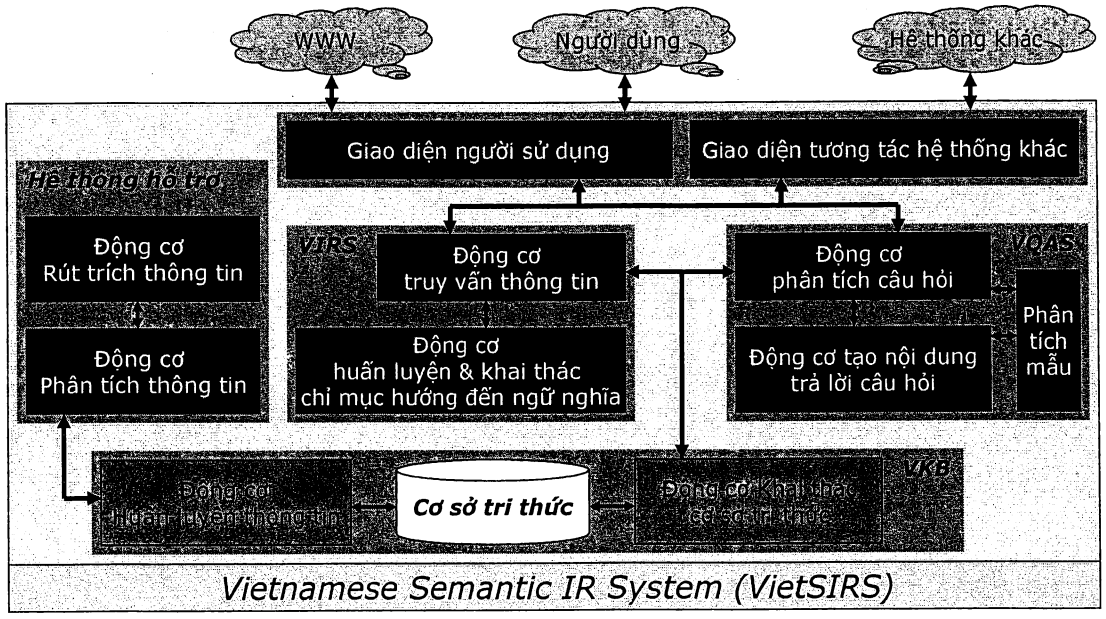
3. Phương pháp theo điểm 1, trong đó việc tạo chỉ mục hướng đến ngữ nghĩa và tìm kiếm thông tin trong chỉ mục sử dụng hai khối chức năng chính là bộ tạo chỉ mục hướng đến ngữ nghĩa (Semantic Indexer, SI), dựa theo thuật toán SemanticIndexCreating và bộ tìm kiếm truy vấn (Query Searcher, QS).

4. Phương pháp theo điểm 1, trong đó để truy xuất cơ sở tri thức, hệ thống sử dụng ngôn ngữ truy vấn cơ sở tri thức ngôn ngữ truy vấn sử dụng là SQL.

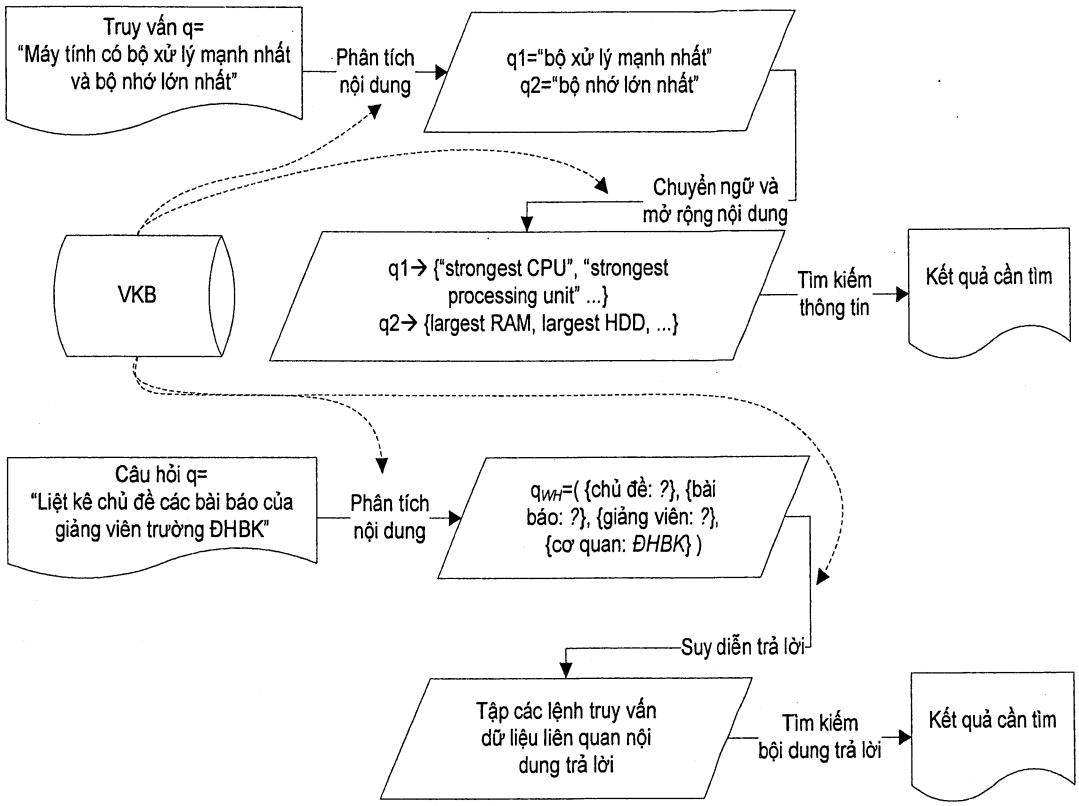
PHỤ LỤC



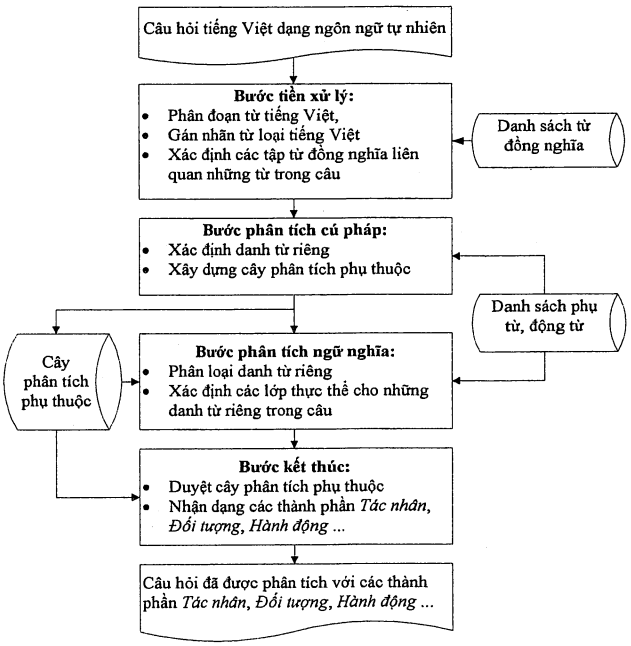
Hình 1.



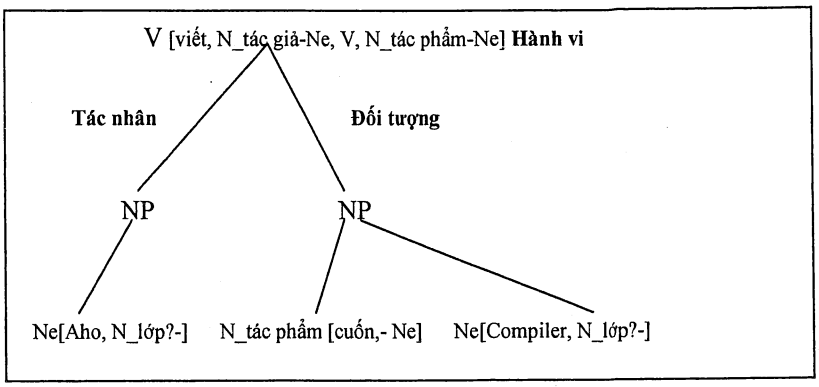
Hình 2.


Truy vấn q=
"Máy tính có bộ xử lý mạnh nhất và bộ nhớ lớn nhất"
Phân tích nội dung
q1="bộ xử lý mạnh nhất"
q2="bộ nhớ lớn nhất"
Chuyển ngữ và mở rộng nội dung
VKB
q1→ {"strongest CPU", "strongest processing unit" ...}
q2→ {largest RAM, largest HDD, ...}
Tìm kiếm thông tin
Kết quả cần tìm
Câu hỏi q=
"Liệt kê chủ đề các bài báo của giảng viên trường ĐHBK"
Phân tích nội dung
q_{WH}=({chủ đề: ?}, {bài báo: ?}, {giảng viên: ?}, {cơ quan: ĐHBK})
Suy diễn trả lời
Tập các lệnh truy vấn dữ liệu liên quan nội dung trả lời
Tìm kiếm bội dung trả lời
Kết quả cần tìm

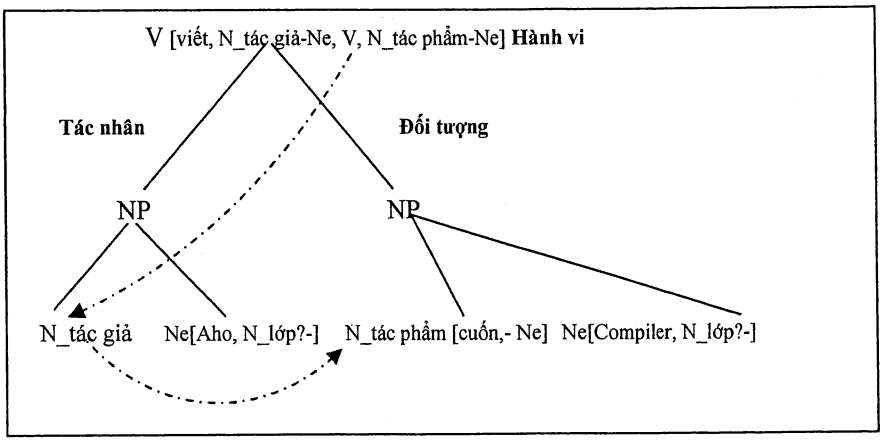
Hình 3.


Câu hỏi tiếng Việt dạng ngôn ngữ tự nhiên
Bước tiền xử lý:
Phân đoạn từ tiếng Việt,
Gán nhãn từ loại tiếng Việt
Xác định các tập từ đồng nghĩa liên quan những từ trong câu
Danh sách từ đồng nghĩa
Bước phân tích cú pháp:
Xác định danh từ riêng
Xây dựng cây phân tích phụ thuộc
Danh sách phụ từ, động từ
Cây phân tích phụ thuộc
Bước phân tích ngữ nghĩa:
Phân loại danh từ riêng
Xác định các lớp thực thể cho những danh từ riêng trong câu
Bước kết thúc:
Duyệt cây phân tích phụ thuộc
Nhận dạng các thành phần Tác nhân, Đối tượng, Hành động ...
Câu hỏi đã được phân tích với các thành phần Tác nhân, Đối tượng, Hành động ...

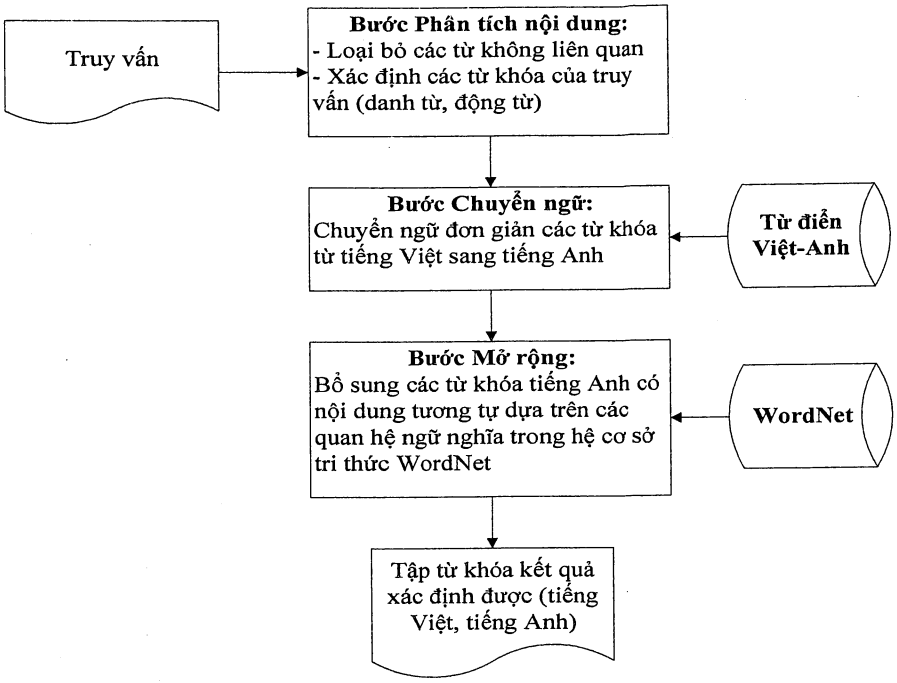
Hình 4


V [viết, N_tác giả-Ne, V, N_tác phẩm-Ne] Hành vi
Tác nhân
Đối tượng
NP
NP
Ne[Aho, N_lớp?-]
N_tác phẩm [cuốn,- Ne]
Ne[Compiler, N_lớp?-]

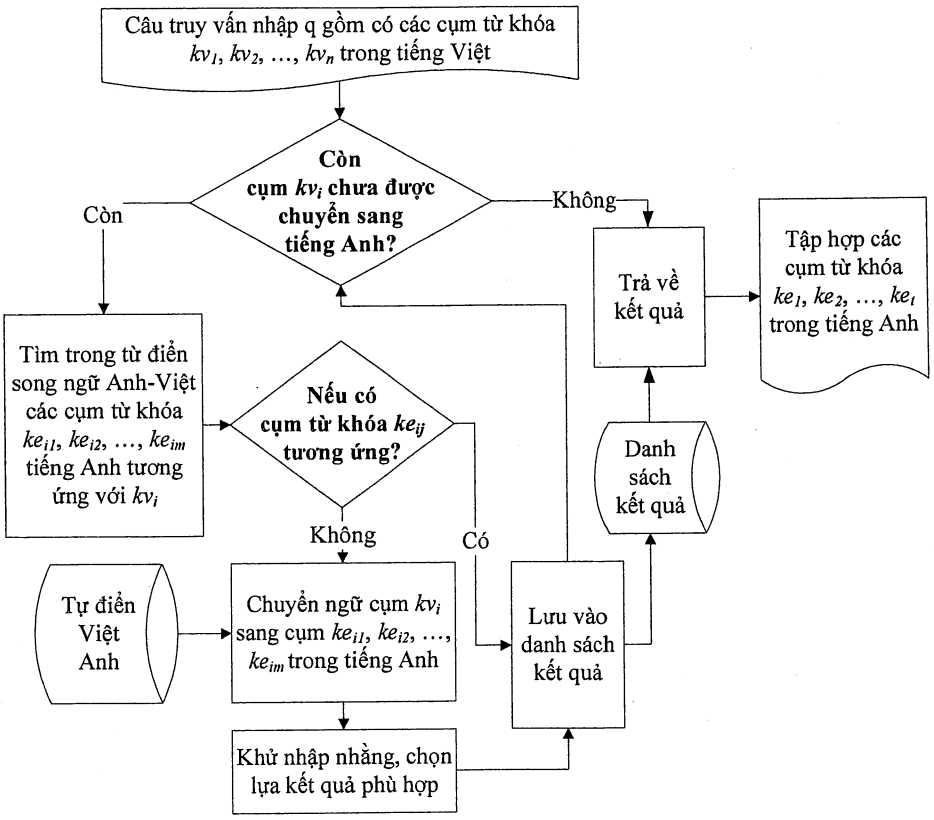
Hình 5.


V [viết, N_tác giả-Ne, V, N_tác phẩm-Ne] Hành vi
Tác nhân
Đối tượng
NP
NP
N_tác giả
Ne[Aho, N_lớp?-]
N_tác phẩm [cuốn,- Ne]
Ne[Compiler, N_lớp?-]

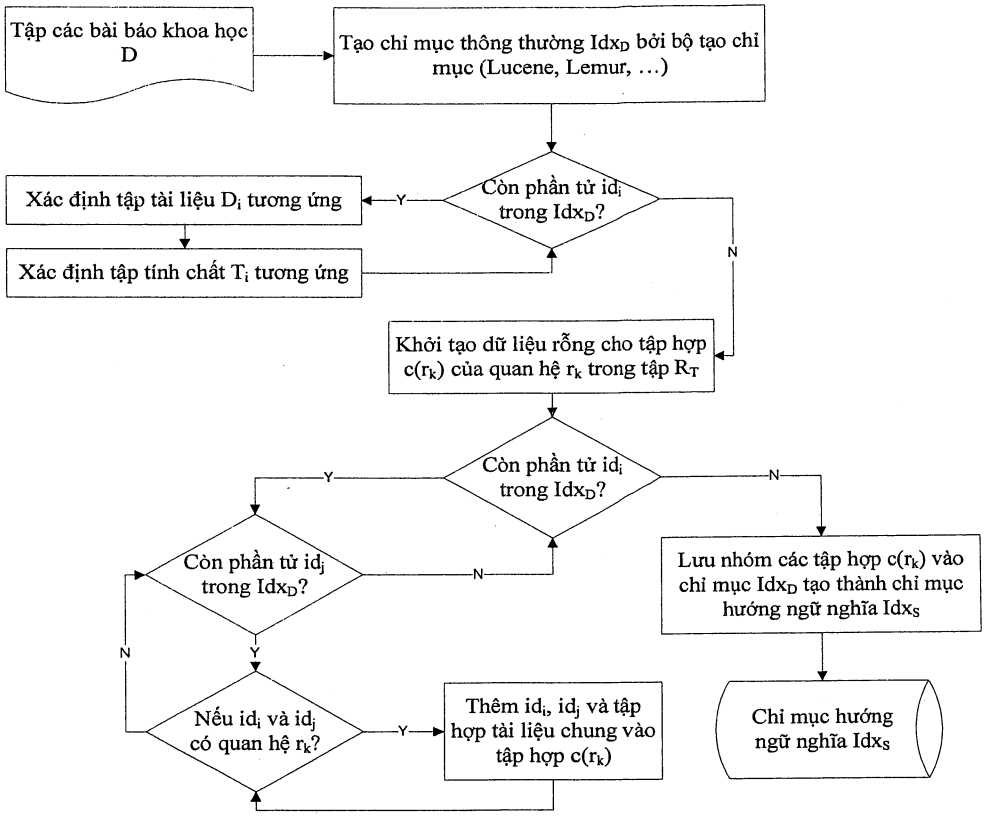
Hình 6.


Truy vấn
Bước Phân tích nội dung:
- Loại bỏ các từ không liên quan
- Xác định các từ khóa của truy vấn (danh từ, động từ)
Bước Chuyển ngữ:
Chuyển ngữ đơn giản các từ khóa từ tiếng Việt sang tiếng Anh
Từ điển Việt-Anh
Bước Mở rộng:
Bổ sung các từ khóa tiếng Anh có nội dung tương tự dựa trên các quan hệ ngữ nghĩa trong hệ cơ sở tri thức WordNet
WordNet
Tập từ khóa kết quả xác định được (tiếng Việt, tiếng Anh)

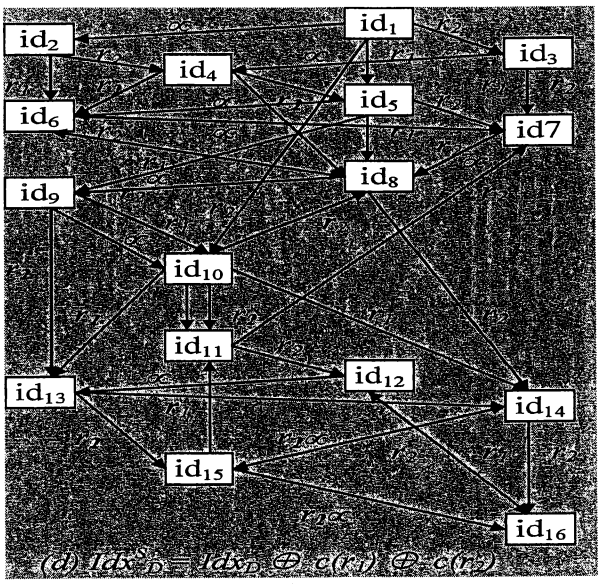
Hình 7.


Câu truy vấn nhập q gồm có các cụm từ khóa $kv_1, kv_2, \ldots, kv_n$ trong tiếng Việt
Còn cụm kv_i chưa được chuyển sang tiếng Anh?
Còn
Không
Tìm trong từ điển song ngữ Anh-Việt các cụm từ khóa $ke_{i1}, ke_{i2}, \ldots, ke_{im}$ tiếng Anh tương ứng với kv_i
Nếu có cụm từ khóa ke_{ij} tương ứng?
Không
Có
Tự điển Việt Anh
Chuyển ngữ cụm kv_i sang cụm $ke_{i1}, ke_{i2}, \ldots, ke_{im}$ trong tiếng Anh
Khử nhập nhằng, chọn lựa kết quả phù hợp
Lưu vào danh sách kết quả
Danh sách kết quả
Trả về kết quả
Tập hợp các cụm từ khóa $ke_1, ke_2, \ldots, ke_t$ trong tiếng Anh

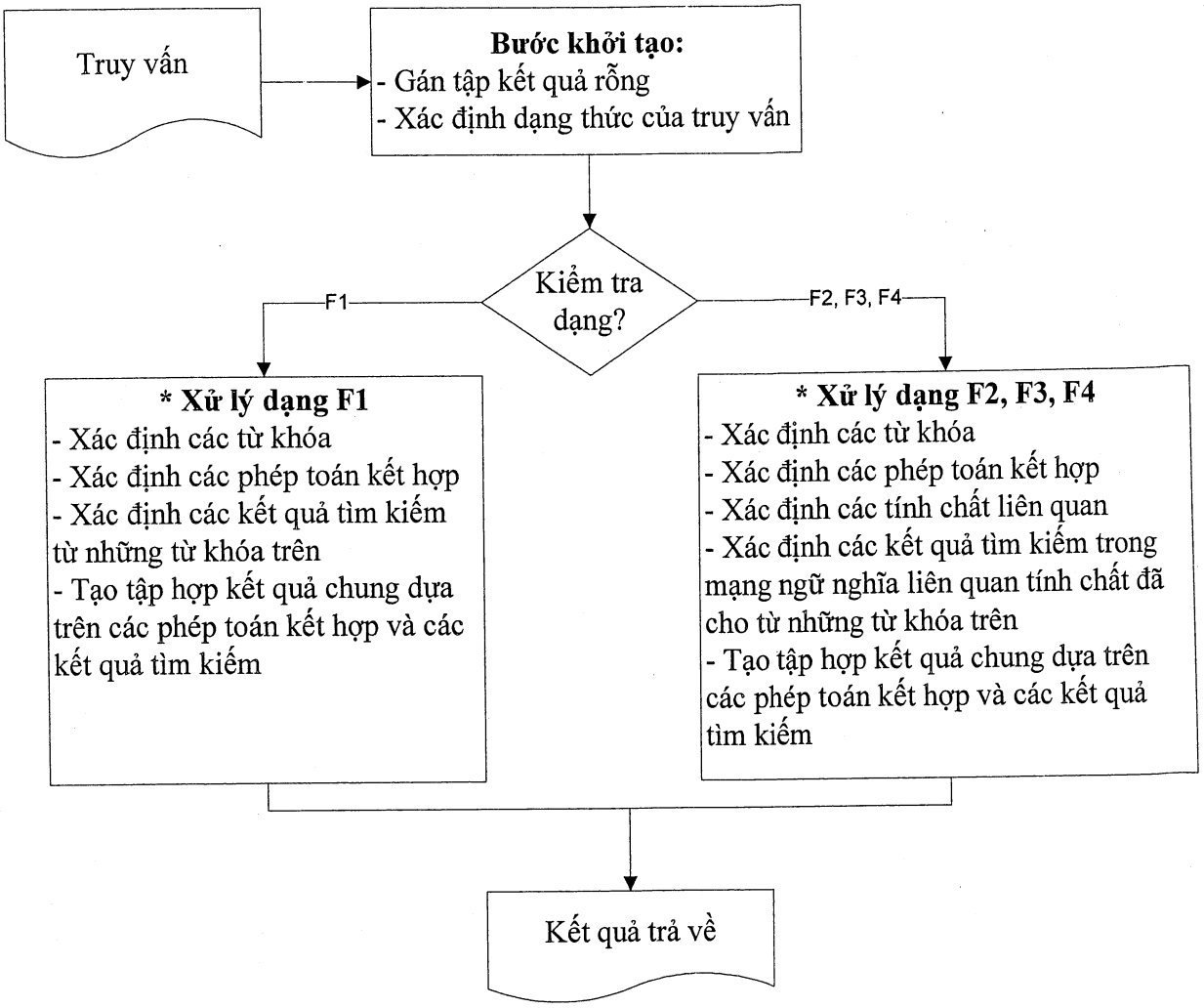
Hình 8.



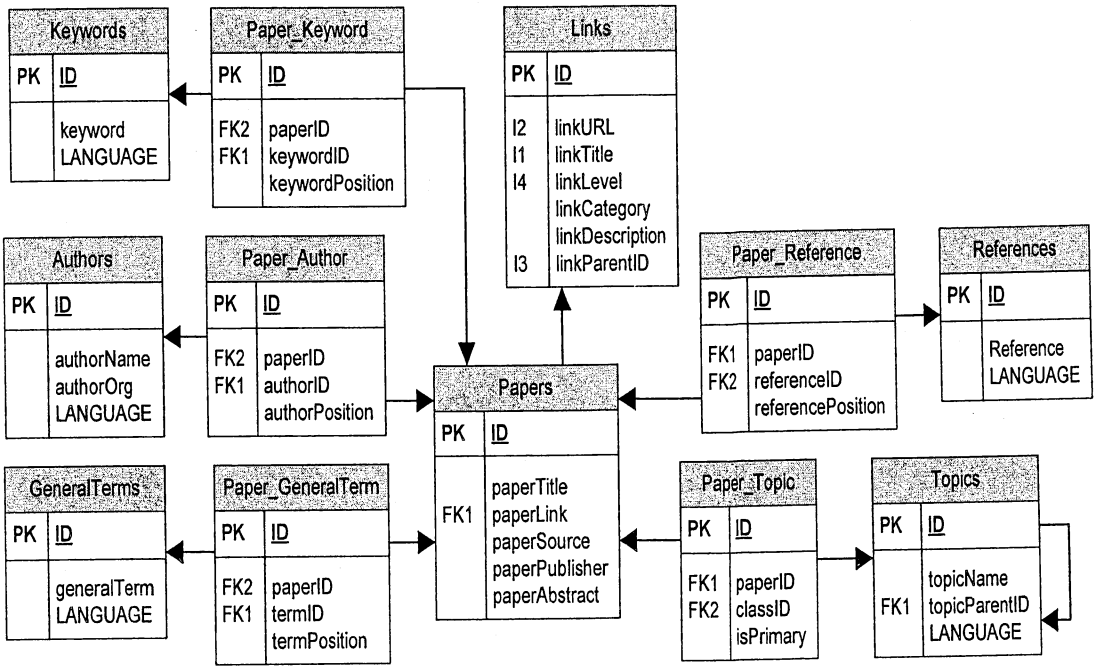
Hình 9.



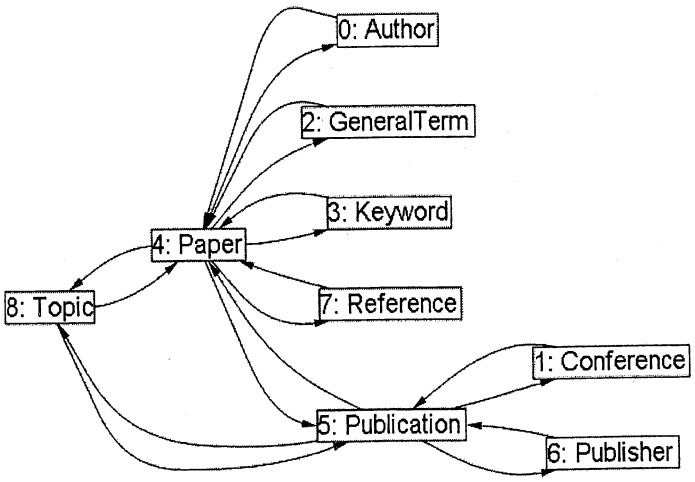
Hình 10


Truy vấn
Bước khởi tạo:
- Gán tập kết quả rỗng
- Xác định dạng thức của truy vấn
Kiểm tra dạng?
F1
F2, F3, F4
* Xử lý dạng F1
- Xác định các từ khóa
- Xác định các phép toán kết hợp
- Xác định các kết quả tìm kiếm từ những từ khóa trên
- Tạo tập hợp kết quả chung dựa trên các phép toán kết hợp và các kết quả tìm kiếm
* Xử lý dạng F2, F3, F4
- Xác định các từ khóa
- Xác định các phép toán kết hợp
- Xác định các tính chất liên quan
- Xác định các kết quả tìm kiếm trong mạng ngữ nghĩa liên quan tính chất đã cho từ những từ khóa trên
- Tạo tập hợp kết quả chung dựa trên các phép toán kết hợp và các kết quả tìm kiếm
Kết quả trả về

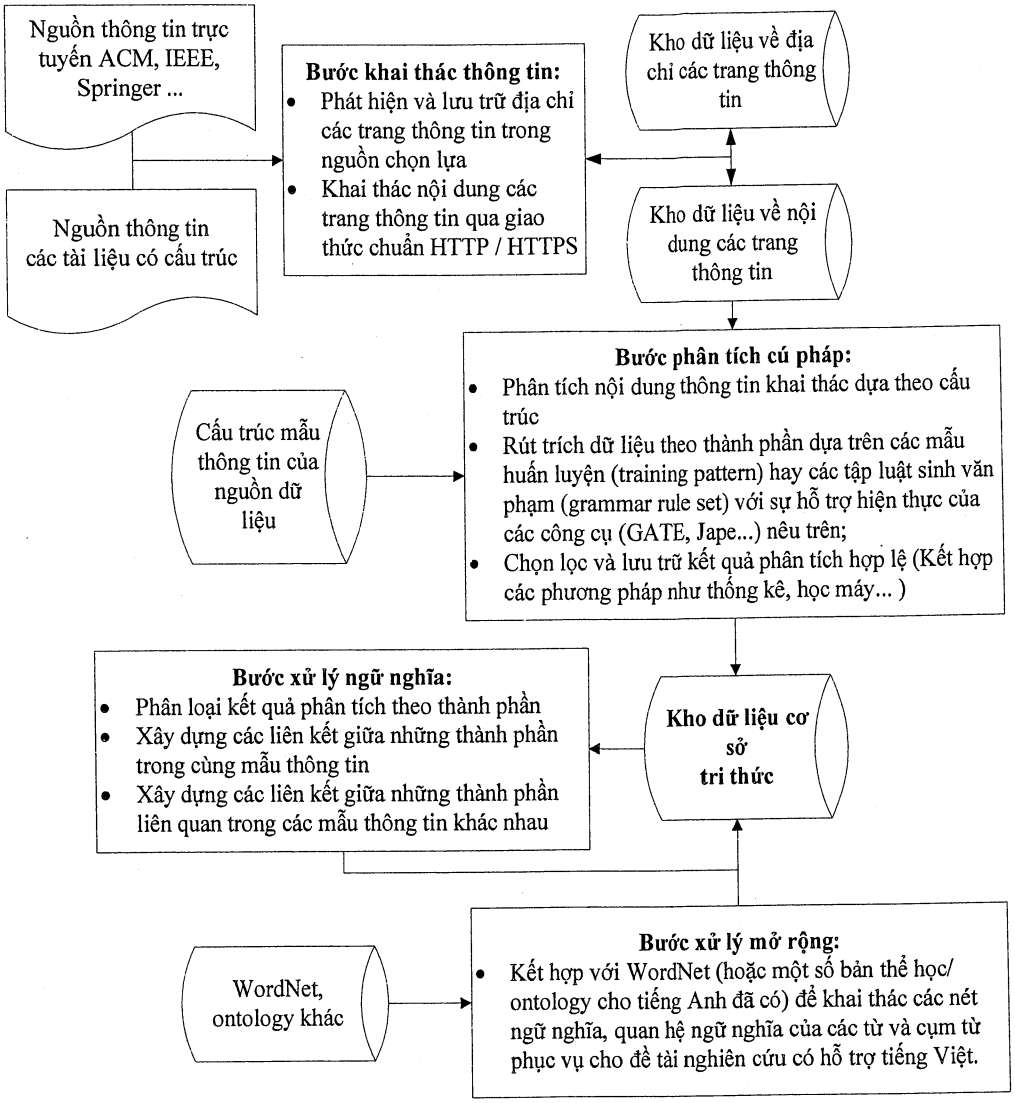
Hình 11.



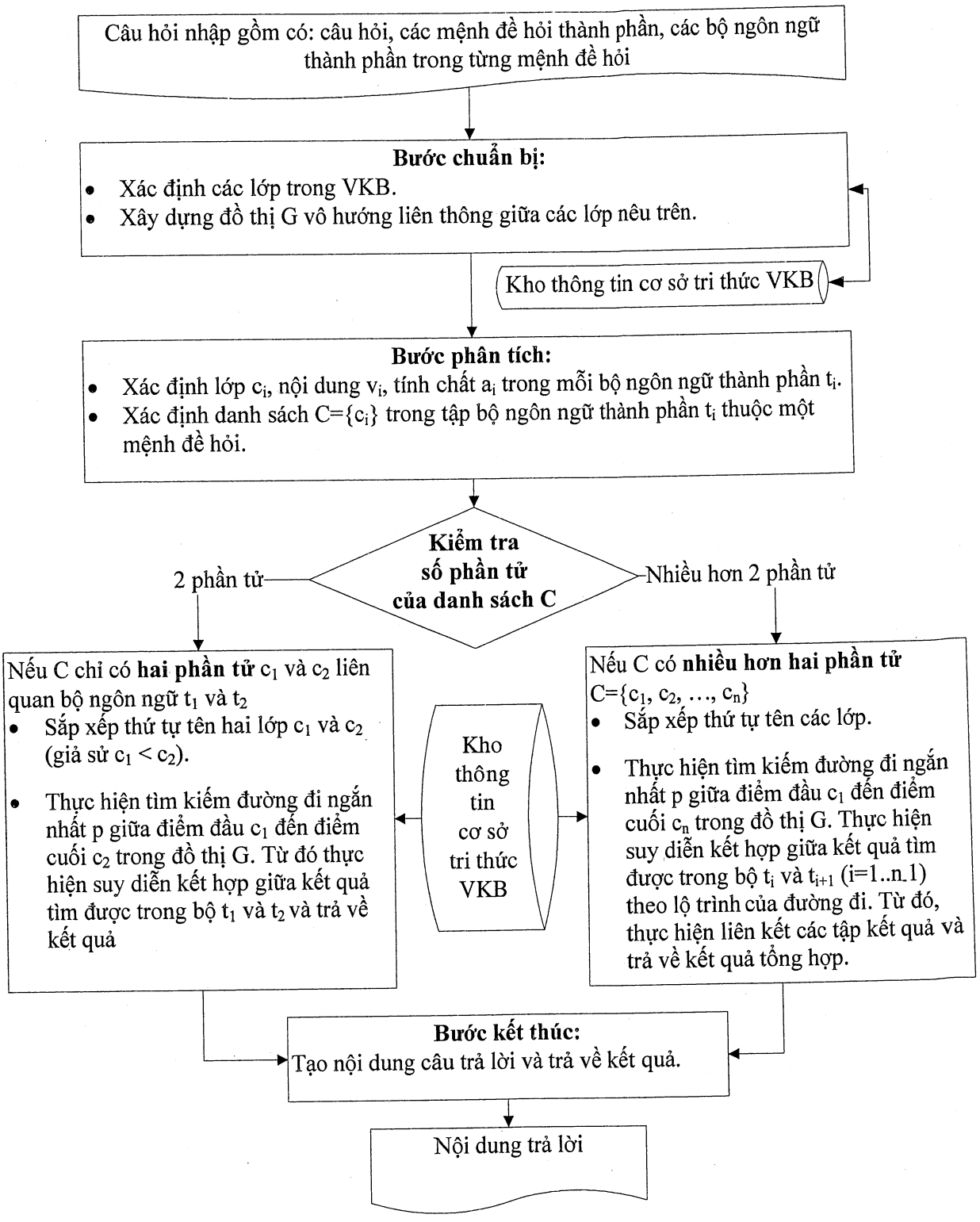
Hình 12.



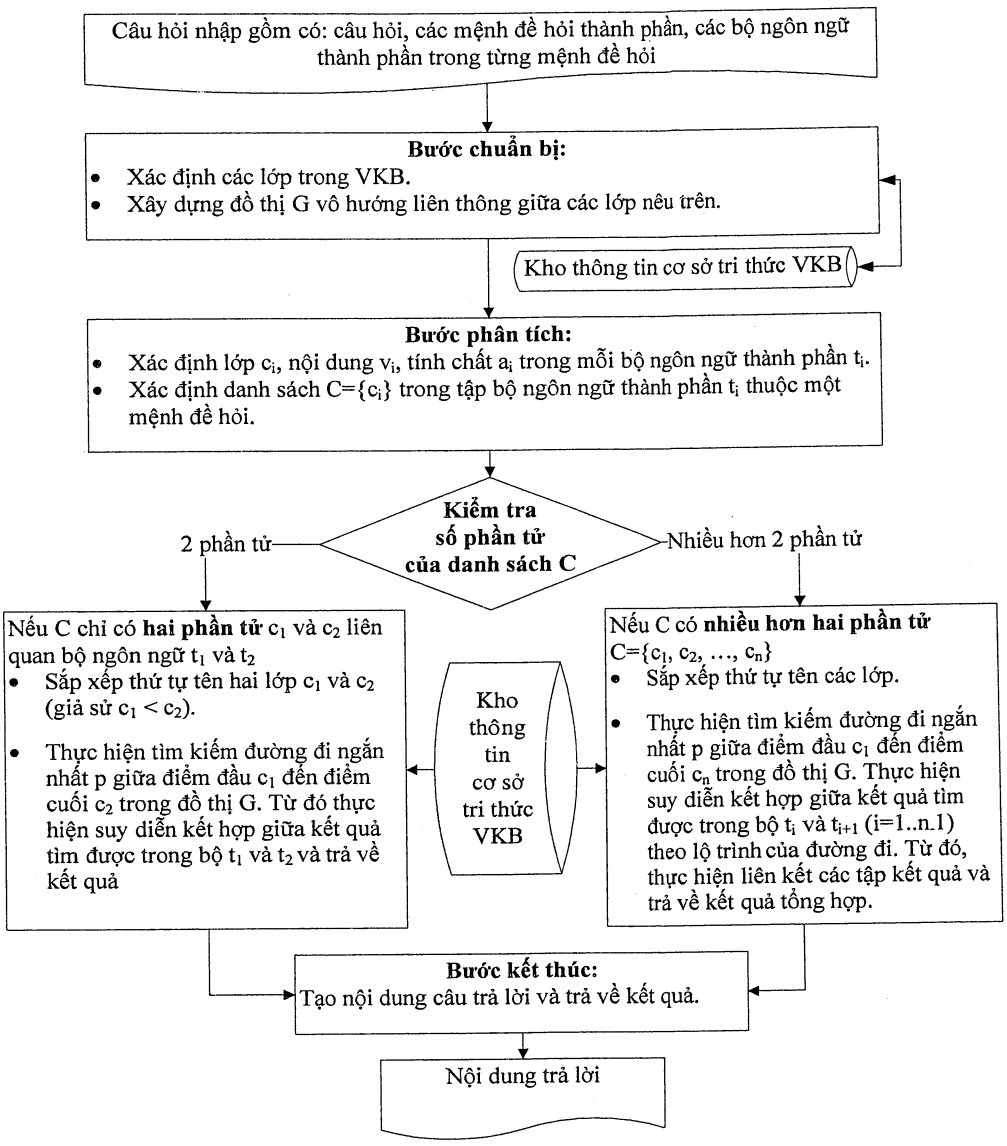
Hình 13.


Nguồn thông tin trực tuyến ACM, IEEE, Springer ...
Nguồn thông tin các tài liệu có cấu trúc
Bước khai thác thông tin:
• Phát hiện và lưu trữ địa chỉ các trang thông tin trong nguồn chọn lựa
• Khai thác nội dung các trang thông tin qua giao thức chuẩn HTTP / HTTPS
Kho dữ liệu về địa chỉ các trang thông tin
Kho dữ liệu về nội dung các trang thông tin
Cấu trúc mẫu thông tin của nguồn dữ liệu
Bước phân tích cú pháp:
• Phân tích nội dung thông tin khai thác dựa theo cấu trúc
• Rút trích dữ liệu theo thành phần dựa trên các mẫu huấn luyện (training pattern) hay các tập luật sinh văn phạm (grammar rule set) với sự hỗ trợ hiện thực của các công cụ (GATE, Jape...) nêu trên;
• Chọn lọc và lưu trữ kết quả phân tích hợp lệ (Kết hợp các phương pháp như thống kê, học máy...)
Bước xử lý ngữ nghĩa:
• Phân loại kết quả phân tích theo thành phần
• Xây dựng các liên kết giữa những thành phần trong cùng mẫu thông tin
• Xây dựng các liên kết giữa những thành phần liên quan trong các mẫu thông tin khác nhau
Kho dữ liệu cơ sở tri thức
WordNet, ontology khác
Bước xử lý mở rộng:
• Kết hợp với WordNet (hoặc một số bản thể học/ ontology cho tiếng Anh đã có) để khai thác các nét ngữ nghĩa, quan hệ ngữ nghĩa của các từ và cụm từ phục vụ cho đề tài nghiên cứu có hỗ trợ tiếng Việt.

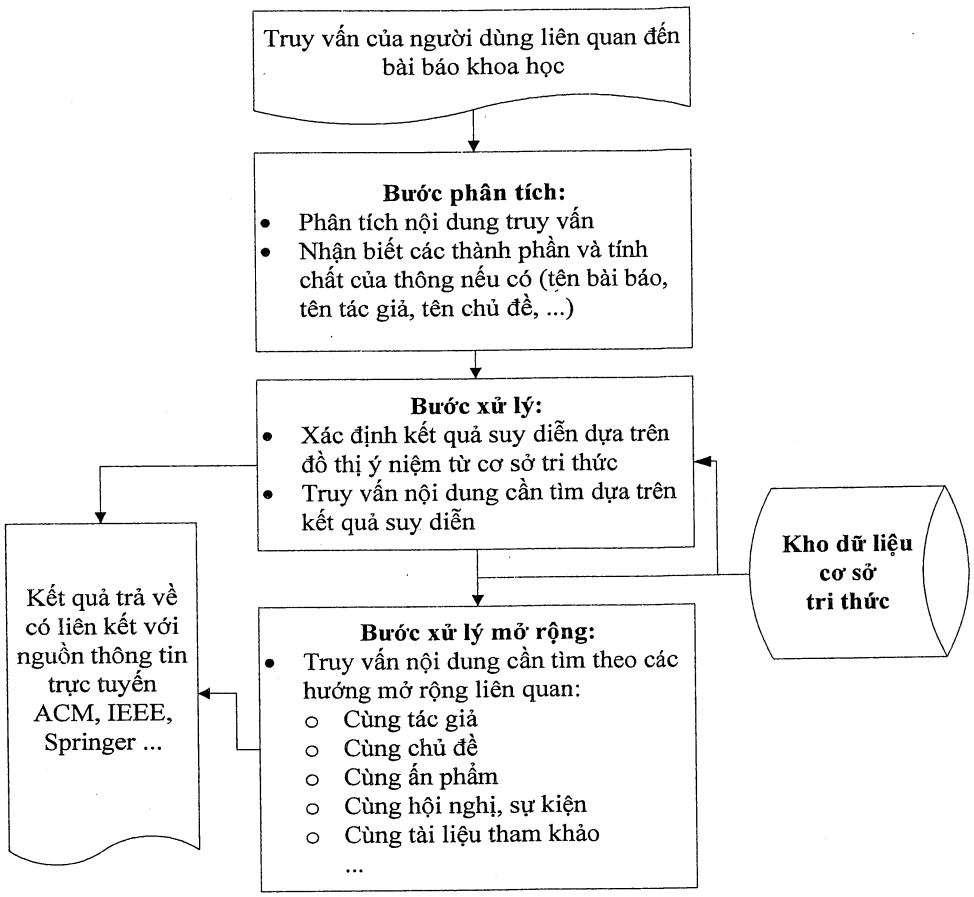
Hình 14.


Câu hỏi nhập gồm có: câu hỏi, các mệnh đề hỏi thành phần, các bộ ngôn ngữ thành phần trong từng mệnh đề hỏi
Bước chuẩn bị:
Xác định các lớp trong VKB.
Xây dựng đồ thị G vô hướng liên thông giữa các lớp nêu trên.
Kho thông tin cơ sở tri thức VKB
Bước phân tích:
Xác định lớp c_i, nội dung v_i, tính chất a_i trong mỗi bộ ngôn ngữ thành phần t_i.
Xác định danh sách $C=\{c_i\}$ trong tập bộ ngôn ngữ thành phần t_i thuộc một mệnh đề hỏi.
Kiểm tra số phần tử của danh sách C
2 phần tử
Nhiều hơn 2 phần tử
Nếu C chỉ có hai phần tử c_1 và c_2 liên quan bộ ngôn ngữ t_1 và t_2
Sắp xếp thứ tự tên hai lớp c_1 và c_2 (giả sử $c_1 < c_2$).
Thực hiện tìm kiếm đường đi ngắn nhất p giữa điểm đầu c_1 đến điểm cuối c_2 trong đồ thị G. Từ đó thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t_1 và t_2 và trả về kết quả
Kho thông tin cơ sở tri thức VKB
Nếu C có nhiều hơn hai phần tử $C=\{c_1, c_2, \ldots, c_n\}$
Sắp xếp thứ tự tên các lớp.
Thực hiện tìm kiếm đường đi ngắn nhất p giữa điểm đầu c_1 đến điểm cuối c_n trong đồ thị G. Thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t_i và t_{i+1} (i=1..n-1) theo lộ trình của đường đi. Từ đó, thực hiện liên kết các tập kết quả và trả về kết quả tổng hợp.
Bước kết thúc:
Tạo nội dung câu trả lời và trả về kết quả.
Nội dung trả lời

Hình 15.


Câu hỏi nhập gồm có: câu hỏi, các mệnh đề hỏi thành phần, các bộ ngôn ngữ thành phần trong từng mệnh đề hỏi
Bước chuẩn bị:
Xác định các lớp trong VKB.
Xây dựng đồ thị G vô hướng liên thông giữa các lớp nêu trên.
Kho thông tin cơ sở tri thức VKB
Bước phân tích:
Xác định lớp c_i, nội dung v_i, tính chất a_i trong mỗi bộ ngôn ngữ thành phần t_i.
Xác định danh sách $C=\{c_i\}$ trong tập bộ ngôn ngữ thành phần t_i thuộc một mệnh đề hỏi.
Kiểm tra số phần tử của danh sách C
2 phần tử
Nhiều hơn 2 phần tử
Nếu C chỉ có hai phần tử c_1 và c_2 liên quan bộ ngôn ngữ t_1 và t_2
Sắp xếp thứ tự tên hai lớp c_1 và c_2 (giả sử $c_1 < c_2$).
Thực hiện tìm kiếm đường đi ngắn nhất p giữa điểm đầu c_1 đến điểm cuối c_2 trong đồ thị G. Từ đó thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t_1 và t_2 và trả về kết quả
Kho thông tin cơ sở tri thức VKB
Nếu C có nhiều hơn hai phần tử $C=\{c_1, c_2, \ldots, c_n\}$
Sắp xếp thứ tự tên các lớp.
Thực hiện tìm kiếm đường đi ngắn nhất p giữa điểm đầu c_1 đến điểm cuối c_n trong đồ thị G. Thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t_i và t_{i+1} (i=1..n-1) theo lộ trình của đường đi. Từ đó, thực hiện liên kết các tập kết quả và trả về kết quả tổng hợp.
Bước kết thúc:
Tạo nội dung câu trả lời và trả về kết quả.
Nội dung trả lời

Hình 16.


Truy vấn của người dùng liên quan đến bài báo khoa học
Bước phân tích:
Phân tích nội dung truy vấn
Nhận biết các thành phần và tính chất của thông nếu có (tên bài báo, tên tác giả, tên chủ đề, ...)
Bước xử lý:
Xác định kết quả suy diễn dựa trên đồ thị ý niệm từ cơ sở tri thức
Truy vấn nội dung cần tìm dựa trên kết quả suy diễn
Kho dữ liệu cơ sở tri thức
Kết quả trả về có liên kết với nguồn thông tin trực tuyến ACM, IEEE, Springer ...
Bước xử lý mở rộng:
Truy vấn nội dung cần tìm theo các hướng mở rộng liên quan:
Cùng tác giả
Cùng chủ đề
Cùng ấn phẩm
Cùng hội nghị, sự kiện
Cùng tài liệu tham khảo
...

Hình 17.