Mục Lục

[Chương 1: Tổng quan về đề tài 2](#_Toc281898419)

[**1.1.** **Giới Thiệu** 2](#_Toc281898420)

[**1.2.** **Mục tiêu và phạm vi đề tài** 3](#_Toc281898421)

[**1.3.** **Các ứng dụng và nghiên cứu hiện nay** 8](#_Toc281898422)

[**1.3.1.** **Các ứng dụng liên quan** 8](#_Toc281898423)

[**1.3.2.** **Các nghiên cứu liên quan** 10](#_Toc281898424)

[Chương 2: Xây dưng hệ thống hỏi đáp để tìm kiếm bài báo khoa học 17](#_Toc281898425)

[**2.1** **Cơ sở lý thuyết và các thuật ngữ liên quan** 17](#_Toc281898426)

[**2.1.1** **Thực thể đặt tên** 17](#_Toc281898427)

[**2.1.2** **Pos Tagger** 17](#_Toc281898428)

[**2.1.3** **Bộ ba quan hệ** 17](#_Toc281898429)

[**2.1.4** **Wordnet và Similarity Wordnet** 17](#_Toc281898430)

[**2.2** **Tổng quan về hệ thống** 17](#_Toc281898431)

[**2.3** **Thiết kế cài đặt hệ thống** 17](#_Toc281898432)

[Tham Khảo 18](#_Toc281898433)

# Chương 1: Tổng quan về đề tài

* 1. **Giới Thiệu**

Hiện nay, cùng với sự phát triển mạnh mẽ của công nghệ thông tin nói chung và Internet nói riêng đã và đang phục vụ rất nhiều lợi ích cho con người. Song song với sự phát triển đó là lượng thông tin ngày càng lớn dần và kéo theo nó là nhu cầu tìm kiếm thông tin trên Internet ngày càng gia tăng. Để đáp ứng các nhu cầu ấy, hàng loạt các công cụ tìm kiếm đã được ra đời như MSN, Yahoo, Google ,… Tuy nhiên, các công cụ ấy hầu hết dựa trên từ khóa hay cụm từ khóa và không đưa ra câu cả lời cụ thể. Đây là cách tìm kiếm quen thuộc đối với người dùng hiện nay.

Với việc tìm kiếm từ khóa như vậy, các công cụ chỉ đưa về các trang liên kết hoặc các văn bản liên quan có chứa từ khóa cần tìm kiếm. Người dùng đôi khi cần phải tối ưu hóa câu truy vấn để tìm được thông tin chính xác nhất và phải chi trả thời gian cho việc tham khảo từng trang hoặc tài liệu để tìm được câu trả lời phù hợp. Điều đó vẫn chưa làm người dùng thực sự hài lòng.

Ngược lại, một hệ thống hỏi đáp sẽ phục vụ tốt hơn trong vấn đề tìm kiếm thông tin. Nó có thể tìm kiếm chính xác câu trả lời bằng việc đưa vào hệ thống một câu hỏi dưới dạng ngôn ngữ tự nhiên. Kết quả trả về là những thông tin thỏa mãn yêu cầu người dùng. Mục đích cuối cùng của người dùng là mong muốn tìm kiếm được thông tin một cách chính xác. Hơn thế nữa, điều này có thể làm máy tính “thông minh” hơn. Người dùng sẽ có giảm giác như nói chuyện với máy tính.

Với nhận định như thế, nhóm muốn xây dựng một hệ thống tìm kiếm theo hướng hỏi đáp bằng ngôn ngữ tự nhiên. Dữ liệu mà nhóm hướng tới là các bài báo khoa học thuộc lĩnh vự công nghệ thông tin. Hệ thống sẽ phục vụ riêng vấn đề tìm kiếm bài báo khoa học.

Trên thực tế, việc tìm kiếm các bài báo bằng từ khóa cũng chiếm không ít thời gian của những người nghiên cứu và tìm kiếm. Để lấy được bài báo ta mong muốn, họ vẫn phải lục lọi trong các kết quả trả về. Nếu các công cụ tìm kiếm có hỗ trợ chức năng tìm kiếm nâng cao, thì người dùng vẫn tốn công nhập vào từng dòng dữ liệu.

Một hệ thống hỏi đáp có thể sẽ phục vụ tốt hơn trong vấn đề tìm kiếm các bài báo khoa học. Nó sẽ tiết kiệm thời gian trong vấn đề nghiên cứu. Thay vì nhập vào từ khóa, câu hỏi tự nhiên sẽ được người dùng nhập vào theo các nội dung thuộc về bài báo. Kết quả trả về sẽ phù hợp theo những gì mà câu hỏi người dùng đưa vào. Việc nhập vào nhiều vùng dữ liệu để có một kết quả chính xác đã không cần thiết. Thay vào đó là một câu hỏi duy nhất bằng ngôn ngữ tự nhiên.

Từ những ý kiến trên, mục tiêu đề tài hướng tới là xây dựng một hệ thống hỏi đáp dựa trên ngôn ngữ tự nhiên phục vụ cho việc tìm kiếm bài báo khoa học. Chức năng tìm kiếm sẽ được hỗ trợ thêm chức năng hỏi đáp. Chính vì mục đích trên, nhóm quyết định chọn đề tài:

**“Xây dựng hệ thống tìm kiếm bài báo khoa học dựa trên hỏi đáp bằng ngôn ngữ tự nhiên”**

Với hệ thống này, người dùng có một lựa chọn khác để tìm kiếm bài báo. Đó là nhập vào hệ thống một câu hỏi dưới dạng ngôn ngữ tự nhiên thay vì từ khóa. Kết quả trả về của hệ thống là các thông tin và bài báo tương ứng với nội dung của câu hỏi. Việc này khiến người dùng giảm bớt thao tác lọc thông tin trong các kết quả trả về theo như cách tìm kiếm cũ là từ khóa.

Tóm lại, nhóm đề xuất cách thức hỏi đáp trong việc tìm kiếm tài liệu nhằm thỏa mãn được nhu cầu tìm kiếm của người dùng. Kết quả đánh giá sẽ dựa trên việc bài báo mà hệ thống trả về có đúng với câu hỏi tìm kiếm hay không và số lượng, loại câu hỏi mà hệ thống hỗ trợ.

* 1. **Mục tiêu và phạm vi đề tài**

Hệ thống được xây dựng mục đích để phục vụ tìm kiếm các bài báo khoa học chỉ thuộc lĩnh vực công nghệ thông tin. Vậy đối tượng người dùng là các nhà nghiên cứu, các kỹ sư, sinh viên công nghệ thông tin, ... Cách thức tìm kiếm sẽ bao gồm cả việc tìm kiếm bằng từ khóa và câu hỏi dưới dạng ngôn ngữ tự nhiên. Câu hỏi người dùng nhập vào sẽ là ngôn ngữ tiếng Anh.

Nguồn dữ liệu chứa thông tin về các bài báo sẽ được lấy từ nguồn DBLP (Digital Bibliography & Library Project). Đây là một kho thư viện điện tử lưu trữ thông tin của hơn 1,4 triệu bài báo [12]. Đây là một kho dữ liệu lớn và thích hợp để nhóm thực hiện đề tài. Ngòai ra, kho dữ liệu này còn được cung cấp miễn phí và tải về tại trang chủ của DBLP[[1]](#footnote-2). Do đó, phạm vi của thông tin chỉ nằm trong kho dữ liệu của DBLP.

Trong hệ thống, người dùng nhập vào một câu hỏi bằng ngôn ngữ tự nhiên và hệ thống trả về các bài báo liên quan đến câu hỏi. Các câu hỏi của người dùng nhóm giới hạn chỉ bao gồm các vấn đề liên quan đến thông tin của các bài báo khoa học : tiêu đề, tên tác giả, năm công bố, nhà xuất bản.

Tuy nhiên, số lượng câu hỏi mà hệ thống có thế giải quyết giới hạn bởi hai điều sau. Thứ nhất, nó phải có nội dung nằm trong các thông tin của bài báo. Thứ hai, các câu hỏi chỉ thuộc về một số loại mà nhóm đã phân ra sẵn.

Trong tiếng Anh, câu hỏi cơ bản được chia làm ba loại chính: Yes/No Question, Question Word Question và Choice Question**.** Đây là các lọai câu hỏi thông dụng mà ta thường thấy. Ngòai ra, ta còn có thể thấy một số lọai câu hỏi khác như tag question, open question, closed question, direct/ indirect question [13-16]**.** Tuy nhiên, trong phạm vi đề tài, ta chỉ xét tới 3 dạng chính nêu trên, vì tùy theo mục đích sử dụng mà còn các loại câu hỏi khác nhau thêm nữa. Nhưng nhìn chung các loại khác cũng quy về 3 dạng trên.

Theo TREC[[2]](#footnote-3), nơi chứa một tập các câu hỏi liên quan đến vấn đề hỏi đáp, các câu hỏi lại được phân làm 3 lọai : FACTOID, LIST và OTHER. Câu hỏi thuộc lọai FACTOID là một loại câu hỏi dành cho những vấn đề chưa được xác minh về thông tin và không chắc là có thông tin để trả lời. Câu hỏi LIST là câu hỏi thuộc dạng liệt kê một danh sách theo yêu cầu của câu hỏi và còn lại là các câu hỏi OTHER.

Đặt biệt trong các hệ thống hỏi đáp, đa số các nghiên cứu thì không phân loại theo như trên. Theo một số nghiên cứu [5,6]về việc phân lọai câu hỏi mà nhóm biết được, thì câu hỏi được phân lọai theo 6 mục thô và 50 mục mịn. Các mục này được dưa ra bởi Li và Roth (2002)[4]. Đây là hình thức phân lọai mà được sử dụng riêng cho vấn đề hỏi đáp.

|  |  |
| --- | --- |
| Mục thô | Mục mịn |
| ABBR | abbreviation,expansion |
| DESC | definition, description, manner, reason |
| ENTY | animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, subtance, symbol, technique, term, vehicle, word |
| HUM | description, group, individual, title |
| LOC | city, country, mountain, other, state |
| NUM | Code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight |

Bảng 1 – Phân loại 6 mục thô và 50 mục mịn theo Li và Roth

Vậy, tùy theo mục đích sư dụng mà câu hỏi còn đựơc phân lọai theo nhiều mục. Đối với 3 lọai câu hỏi cơ bản trong tiếng Anh thì hệ thống của nhóm hiện tại không giải quyết các câu hỏi choice question. Đó là câu hỏi thuộc dạng cho người dùng lựa chọn đáp án đã có sẵn. Chúng không phù hợp cho vấn đề tìm kiếm. Hệ thống chỉ giải quyết được một số các câu hỏi thuộc loại Question Word Question và Yes/No question. Đối với câu hỏi Yes/No question hệ thống xử lý tương tự câu hỏi Question Word Question. Vì câu hỏi loại này chỉ nhằm xác minh một vấn đề, cho nên, nếu hệ thống kiếm được kết quả thì sẽ trả về và ngược lại.

Theo nhận định, đối với việc phân loại của TREC thì các phân loại như FACTOID, LIST và OTHER vẫn chưa được rõ ràng. Nếu phân loại theo 3 mục này thì sẽ khó xác định được mục đích của câu hỏi muốn hỏi tới vấn đề gì.

Vì thế, các câu hỏi đặt ra trong đề tài này sẽ được phân loại theo 6 mục thô (ABBR, DESC, ENTY, HUM, LOC, NUM). Đây là cách phân loại hợp lý nhất đây là cách phân loại mà các hệ thống hỏi đáp thường dùng.

Như đã trình bày, hệ thống chủ yếu xử lý các câu hỏi dạng Question Word Question, tức là có các từ để hỏi như: what, who, when, where… Câu hỏi trong hệ thống sẽ được phân theo 6 mục thô của Li và Roth (2002). Việc phân lọai theo 50 mục mịn là không cần thiết, vì số lượng câu hỏi giải quyết được trong đề tài chỉ thuộc về một phạm vi xung quanh các bài báo công nghệ thông tin. Hơn nữa, số lượng câu hỏi không nhiều để phân vào 50 mục mịn. Việc phân lọai sẽ được phân lọai bằng tay theo ý kiến của các thành viên trong nhóm.

Nguồn câu hỏi được tham khảo từ :

* Dựa trên nguồn dữ liệu của DBLP ( các thuộc tính trong các bảng)
* Nhu cầu người dùng : các câu hỏi liên quan đến các vấn đề mà người dùng cần biết.
* Một số các mẫu câu trong tập TREC 2003 và 2007

Bảng phân loại các câu hỏi mà hệ thống xử lý:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Loại** | **Xử lý được** | **Không xử lý** | **Ví dụ** | **Chú thích** |
| Yes/No Question | X |  | *Are they good jobs?*  *Did he know just how deeply his father was opposed to things?* | Xử lý tương tự question-word question |
| Question -word question | X |  | *Why are you doing this?’*  *Who has the papers?* |  |
| Choice question |  | X | *Do you want a strong cup or a weak one?* | Không có giá trị trong tìm kiếm |
| Tag question | X |  | The weather's bad, isn't it? | Xử lý tương tự question-word question |
| Open/closed question |  | X | *Do you know what i mean?*  *How does it work?* | Không xử lý |
| Direct/Indirect question |  | X | Could you tell me where he is? (where is he? | Không xử lý |
|  |  |  |  |  |

Bảng 1 – Phân loại các câu xử lý/ không xử lý theo cú pháp Tiếng Anh.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Loại** | **Xử lý được** | **Không xử lý** | **Ví dụ** | **Chú thích** |
| ABBR |  | X | What does the abbreviation AIDS stand for ? | các câu hỏi về các từ viết tắt, từ mở rộng |
| DESC |  | X | What is ethology ?  What is titanium ? | các câu hỏi về miêu tả và định nghĩa |
| ENTY | X |  | What kind of animals were in the Paleozoic era ?  What is a female rabbit called ? | các câu hỏi liên quan đến các thực thể, con người, vật |
| HUM | X |  | What actor first portrayed James Bond ? | các câu hỏi liên quan tới người và tổ chức |
| LOC |  | X | Which city in Canada is the least-populated ? | các câu hỏi liên quan tới địa điểm |
| NUM | X |  | How many member states are in the UN ? | các câu hỏi liên quan tới số lượng, ngày tháng, mã, tiền... |

Bảng 2 – Phân loại các câu xử lý/ không xử lý theo các hệ thống hỏi đáp thông dụng (theo ngữ nghĩa).

Vậy, hệ thống chỉ giải quyết được các câu hỏi mang ngữ nghĩa liên quan đến ENTY, HUM và NUM. Tức là, các câu hỏi liên quan đến cuốn sách, bài báo (ENTY), các câu hỏi về tác giả, nhà xuất bản (HUM) và các câu hỏi về số lượng (NUM). Đối với cú pháp thì chỉ xử lý cho những dạng câu hỏi Question-Word question (tức câu hỏi có các tư hỏi như What, When, Where, Why...), Yes/No question và tag question. Các loại câu hỏi khác hiện nhóm chưa thể xử lý được do nhiều lý do về giải pháp, ngữ nghĩa và mục đích sử dụng. (Các loại câu hỏi chưa được đề cập trong báo này mặc định là chưa xử lý được ).

* 1. **Các ứng dụng và nghiên cứu hiện nay**
     1. **Các ứng dụng liên quan**

Hiện nay đã có một số trang web phục vụ riêng cho vấn đề tìm kiếm các bài báo khoa học như CiteXeer, IEEE, ACM, ScienceDirect, SSRN, PaperCube … Các trang web này đều có hệ thống tìm kiếm bài báo dựa trên từ khóa. Ngoài ra, người dùng còn có thể chọn lựa các tùy chọn để có kết quả tốt hơn (tìm kiếm theo tên tác giả, tìm kiếm theo tiêu đề, tìm kiếm theo nội dung abstract …). Một số trang như PaperCube có khả năng hiển thị bài báo theo dạng mô hình đồ thị. Điều này giúp người dùng định hướng được trong việc tìm kiếm.

Sau khi khảo sát web trên, nhóm tóm tắt lại ở bảng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Tên trang | Hình thức  tìm kiếm | Cách thức tìm kiếm | Hiển thị |
| CiteXeer | Từ khóa | Theo tiêu đề, theo tên tác giả, tìm các bảng (table) có trong nội dung bài báo. | Danh sách kết quả và nội dung chi tiết |
| IEEE | Từ khóa | Theo tiêu đề | Danh sách kết quả và nội dung chi tiết |
| ACM | Từ khóa | Theo tiêu đề | Danh sách kết quả và nội dung chi tiết |
| Sciencedirect | Từ khóa | Tất cả các trường, tên tác giả, tiêu đề | Danh sách kết quả và nội dung chi tiết và outline của bài báo |
| SSRN | Từ khóa | Tiêu đề, tên tác giả, nội dung tóm tắt (abstract) , từ khóa , ngày. | Danh sách kết quả và nội dung chi tiết |
| PaperCube | Từ khóa | Tiêu đề, nội dung tóm tắt, năm, chủ đề, tên tác giả | Danh sách kết quả, nội dung chi tiết bài báo, hiển thị bài báo dưới dạng đồ thị, cây … |

Như bảng đánh giá trên thì hầu hết các trang đều hỗ trợ việc tìm kiếm dựa trên tiêu đề và tên tác giả. Các kết quả từ việc tìm kiếm đều trả một danh sách các bài báo có liên quan đến từ khóa. Vì thế một số trang phải hỗ trợ thêm chức năng tìm kiếm nâng cao (advance search) để nâng cao kết quả tìm kiếm dựa trên việc nhập thêm nhiều thông tin tìm kiếm và ràng buộc. Việc này, như đã trình bày, việc tìm kiếm bài báo vẫn còn khá tốn thời gian do lượng kết quả trả về khá nhiều. Đối với các chức năng tìm nâng cao thì buộc người dùng phải nhập thêm một số thông tin để có được đáp án chắc chắn.

Đó là những chủ kiến về các hệ thống tìm kiếm hiện nay. Thay vào đó, nhóm có tìm hiểu về một số hệ thống hỏi đáp. Hiện theo như khảo sát của nhóm, đã có một số hệ thống hỏi đáp như Answer.com, NSIR, START, Answerbus.com … Các hệ thống này đã và đang được sử dụng trên internet. Tuy nhiên, các hệ thống hỏi đáp này phục vụ cho việc tìm kiếm thông tin như địa danh, nơi chốn, các vấn đề khoa học, lịch sử, địa lý …

* Question Answering System – NSIR [21]**.** Được phát triển bởi nhóm nghiên cứu CLAIR tại đại học Michigan. Hệ thống trả lời câu hỏi trong hầu hết các lĩnh vực, từ đơn giản đến phức tạp. Bên cạnh đó, hệ thống cũng lưu trữ một ngân hàng câu hỏi cho người sử dụng chọn lựa.
* [The START Natural Language Question Answering System](http://start.csail.mit.edu/) [22]**.** Là hệ thống hỏi và trả lời dựa trên nền Web đầu tiên trên thế giới. Không giống như các hệ thống rút trích thông tin khác, mục đích của START là hỗ trợ cho người sử dụng tìm kiếm được thông tin chính xác thay cho việc chỉ đơn thuần cung cấp một danh sách các kết quả tìm kiếm được. Hiện tại, hệ thống có thể trả lời hàng triệu câu hỏi bằng tiếng Anh về địa điểm, phim ảnh, con người, khái niệm…
* Ask.com [23] Là hệ thống hỗ trợ tìm kiếm các trang Web, hình ảnh, thông tin, bản đồ, lời chỉ dẫn, tìm kiếm cục bộ và mua sắm. Ask.com đang có những cải tiến đáng kể như thu hẹp phạm vi tìm kiếm để nâng cao chất lượng kết quả tìm được chứ không tìm kiếm trên tất cả các lĩnh vực. Nó cung cấp khả năng tìm kiếm, trả lời các câu hỏi tập trung vào các vấn đề cơ bản như các công thức món ăn, sở thích, bài tập cho trẻ em, giải trí và sức khỏe.

Đối với một hệ thống hỏi đáp để tìm kiếm các bài báo khoa học, nhóm chưa tìm thấy một hệ thống nào có hỗ trợ chức năng này. Có thể tạm kết luận là do người dùng vẫn còn quen thuộc với việc tìm kiếm bằng từ khóa hoặc các hệ thống còn đang trong thời gian nghiên cứu.

* + 1. **Các nghiên cứu liên quan**

Với những hệ thống hỏi đáp nêu trên cho thấy rằng có những nghiên cứu về vấn đề xử lý câu hỏi tự nhiên đề tìm kiếm thông tin. Đã có nhiều bài báo trình bày về hệ thống hỏi đáp bằng ngôn ngữ tự nhiên [1,2,7,8,9].

Vấn đề chung của các nghiên cứu là hiểu được câu hỏi mà người dùng đưa vào. Câu hỏi bằng nhiều cách sẽ được chuyển thành dạng ngôn ngữ mà máy có thể hiểu được. Ví dụ như câu truy vấn SQL.

* Một nghiên cứu gần đây của Wael Salloum, 2009 [7]. Tác giả đưa ra hướng tiếp cận bằng cách mô hình các văn bản và các câu hỏi thành một dạng đồ thị gọi là Conceptual Graph Formalism (CGF).

Nguồn dữ liệu của đề tài này là từ các văn bản(document). Một văn bản được tách thành nhiều câu. Và mỗi câu dựa trên việc phân tích cú pháp, mỗi nút trong cây cú pháp sẽ được rút khái niệm từ wordnet kết hợp với công cụ verbnet để xác định mối quan hệ giữa động từ với chủ từ, đối tượng. Câu sẽ được chuyển thành các Đồ thị khái niệm (gọi là Concept Graph (CG)). Cuối cùng các đồ thị này sẽ được lưu xuống cơ sở dữ liệu. Tương tự cho câu hỏi cũng chuyển thành một CG.

Cuối cùng, tác giả có đề cập tới một phép chiếu (projection operator) dùng để so sánh giữa CG của một câu hỏi với CG của một câu trong văn bản. Đó là cách mà câu trả lời được rút ra, xếp hạng và trả về cho người dùng.

Ví dụ : với câu “Mark Twain wrote Tom Sawyer”, thì đồ thị khái niệm sẽ được tạo như sau:



Hình 1 – Ví dụ về một đồ thị khái niệm

( được lấy từ Figure 3. tài liệu tham khảo [2])

Trong đồ thị ở hình 1, theo cách hiểu của nhóm, *Author* là một lớp đối tượng và một thể hiện của nó là *Mark Twain*. Cũng như thế, *Novel* có thể hiện là Tom Sawyer. Giữa *Author* và *Novel,* ta có mối quan hệ *Write*. *Agent* và *Patient* được coi là các mối quan hệ trung gian: *Author* là tác nhân (agent) của hành động *write* và *Novel* là một thể bị động (patient) so với hành động *write*.

Tương tự câu hỏi cũng được chuyển thành đồ thị khái niệm. Vấn đề ở các câu hỏi là cần xác định đối tượng cần hỏi.

Ví dụ: *“ Who invented the light bulb? “* sẽ đuợc chuyển thành như sau:

*[Invent]-*

*(Agnt) -> [Person: \*] ?*

*(Ptnt) -> [Light-Bulb]*

Ở đây, *Person* là đối tượng cần biết trong câu hỏi , giữa *person* và *Light-Bulb* có mối quan hệ *Invent*. *Agnt* và *Ptnt* là hai từ viết tắt của *Agent, Patient*. Đây là một cách thể hiện khác của đồ thị so với hình 1. Sau đó, công việc tiếp theo là so sánh ( projection operator) giữa đồ thị CG của câu hỏi với các đồ thị khái niệm CG mỗi câu để tìm câu trả lời.

Kết quả bài báo trên đạt được là biểu diễn văn bản và câu hỏi dưới dạng các CG. Bằng cách so sánh giữa các CG, tìm ra câu trả lời. Hướng đi tiếp theo của tác giả là tìm cách tổng hợp lại các câu trả lời với mục đích là đưa ra một câu trả lời duy nhất.

* Với nghiên cứu trong nước, nhóm tìm hiểu được một đề tài “Xây dựng công cụ tìm kiếm tài liệu học tập bằng các truy vấn ngôn ngữ tự nhiên trên kho học liệu mở tiếng Việt” [2]. Với tên đề tài thì hệ thống hỏi đáp mà cô hướng tới là một hệ thống hỏi và trả lời trên ngôn ngữ tiếng việt.

Với cách tiếp cận của đề tài trên, câu hỏi tiếng việt được đưa vào hệ thống phân tích. Một tập 40 luật cú pháp được tác giả tạo ra sẵn. Nếu phân tích được, thành phần sẽ tạo ra một cây cú pháp tương ứng. Từ đó, cây cú pháp được chuyển thành cây sinh mã truy vấn bằng cách ánh xạ lên một số nút của cây cú pháp vào ontology. Cuối cùng, câu truy vấn SPARQL được tạo ra và lấy dữ liệu trả lời.

Ontology có thể xem là một kho dữ liệu được lưu trữ dưới dạng ngữ nghĩa. SPARQL là ngôn ngữ để truy vấn vào kho dữ liệu này [17,18].

Dữ liệu trong đề tài này được lấy từ kho học liệu mở vocw[[3]](#footnote-4). Các metadata được rút từ các trang của vocw và cập nhật dữ liệu vào ontology.

Kết quả của đề tài trên được đánh giá như sau: với 40 câu hỏi chuẩn thì hệ thống chính xác 100%, với 91 câu hỏi ngẫu nhiên thì độ chính xác là 84,62%. Hướng đi tiếp theo của đề tài là cải thiện bộ phân tích cú pháp với mục đích tạo ra bộ phân tích mạnh hơn và nhanh hơn. Đồng thời, việc mở rộng không gian tìm kiếm cũng được tác giả lưu ý tới.

* Một đề tài khác là “ Dịch câu truy vấn có cấu trúc sang đồ thị ý niệm: cách tiếp cận ít phụ thuộc vào cú pháp”, 2008 [3]. Theo ý kiến của tác giả, việc phân tích cú pháp để hình thành đồ thị ý niệm gặp khó khăn khi các câu truy vấn không đúng cú pháp. Vì thế tác giả đưa ra một phương pháp dịch câu truy vấn sang đồ thị ý niệm mà không dựa trên việc phân tích cú pháp câu.

Ở đề tài này, phương pháp tiếp cận dựa trên việc nhận diện các thực thể đặt tên, không đặt tên và từ quan hệ. Sau khi nhận diện được, thao tác tiếp theo là nhận diện lớp của thực thể mà không xác định được nó thuộc lớp nào ( như What, Who ...). Bước kế tiếp là gom các thực thể được cho là giống nhau, xác định những mối quan hệ ẩn mà không được nêu trong câu hỏi. Cuối cùng, xác định quan hệ giữa các thực thể, xóa bỏ những mối quan hệ không thích hợp và xây dựng đồ thị ý niệm.

Công cụ đựơc sử dụng trong đề tài này là GATE [20]. Do GATE có khả năng làm việc với Ontology, công cụ OCAT (Ontology-based Corpus Annotation tool) được sử dụng để chú thích các thực thể đặt tên. Ngòai ra, công cụ ANNIE được dùng để nhận diện các từ quan hệ và thực thể không đặt tên bằng các lập ra các danh sách trong Gazetteer. Ontology được sử dụng ở đây có tên là PROTON.

Tập dữ liệu câu hỏi kiểm nghiệm được lấy từ 440 câu hỏi của TREC 2002 và 445 câu hỏi của TREC 2007.

Việc đánh giá của đề tài dựa trên độ chính xác của việc tạo đồ thị ý niệm. Kết quả của đề tài trên nhóm xin trình bày theo bảng sau:

(Quá trình làm giàu Ontology trong nghiên cứu này chủ yếu thủ công bằng cách bổ sung các lớp và mối quan hệ)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tổng Số câu | Trước khi làm giàu  Ontology | Sau khi làm giàu  Ontology |
| TREC 2002 | 440 | 15.23% | 78.64% |
| TREC 2007 | 445 | 8.54% | 60.45% |

* Với hệ thống eLSSNL (eLibrary Searching System by Natural Language) của tài liệu [1], câu truy vấn dưới dạng ngôn ngữ tự nhiên không được xây dựng thành đồ thị ý niệm như tài liệu [3] và [7].

Đối với hệ thống eLSSNL, tác giả dùng một phương pháp khác so với việc xây dựng đồ thị ý niệm. Câu truy vấn tự nhiên của người dùng được nhận dạng các giá trị thuộc tính tương ứng với các thuộc tính của bảng trong cơ sở dữ liệu. Do hệ thống này sử dụng dữ liệu miễn phí từ dự án Gutenberg[[4]](#footnote-5) chứa thông tin của khoảng 33,000 ebooks miễn phí.

Cách để hiểu câu truy vấn dưới dạng ngôn ngữ tự nhiên của người dùng được tài liệu [1] giải quyết bằng việc xây dựng một tập từ khóa dựa trên các chỉ định từ mà tác giả đã liệt kê. Sau đó, các giá trị hay từ khóa đó được nhận diện thuộc về thuộc tính nào trong bảng cơ sở dữ liệu.

Ví dụ:

*Nhận dạng giá trị cho thuộc tính Title/FriendlyTitle:*

+  DOM(Title) hoặc  DOM(FriendlyTitle) (DOM(fileld): là miền giá trị của một thuộc tính (field) trong table BOOKS )

+ Đi kèm theo các chỉ định từ: title/heading

*Nhận dạng giá trị cho thuộc tính NumDownload:*

+ Là số *(isNumber)*, ưu tiên tìm theo ID trước

+ Đi kèm theo các chỉ định từ: [more] than/>=/>; less/<=/<; top; best; most; numdownload.

Đối với từng thuộc tính của bảng sẽ có những điều kiện để nhận dạng như ví dụ trên. Cuối cùng bằng các phép kết hợp, câu truy vấn được sinh ra và tìm các dữ liệu thỏa mãn.

Ngoài ra, hệ thống eLSSNL còn hỗ trợ lưu lại các các câu truy vấn mà người dùng nhập vào và lưu trong một bảng dưới cơ sở dữ liệu. Người dùng còn được hỗ trợ gợi ý cho người dùng các câu truy vấn tương tự. Việc đó dựa vào việc so sánh các tập từ khóa được lưu lại trong bảng vừa đề cập và so sánh với tập từ khóa trong câu hỏi người dùng vừa nhập vào.

Theo đánh giá về hiệu quả tìm kiếm, hạn chế lớn nhất của eLSSNL là về mặt tốc độ. Nếu dùng người nhập vào một câu truy vấn tự nhiên mới hoàn toàn – tức là tập từ khóa không trùng với các tập từ khóa được lưu trữ- thì tốc độ tìm kiếm sẽ khá lâu nếu người dùng tìm trên nhiều giá trị thuộc tính.

**Nhận xét:**

* Đối với các phương pháp tạo đồ thị ý niệm từ câu hỏi [3,7], thì việc tạo đồ thị dựa trên nhận diện các thực thể đặt tên, không đặt tên. Các công cụ để nhận diện có thể dùng là Wordnet hay một Ontology (PROTON). Với cách phân tích cú pháp để tạo đồ thị ý niệm của [7] thì có gặp khó khăn nếu người dùng nhập một câu có nhập nhằng về cú pháp (hoặc sai cú pháp). Vì thế nhóm thấy hướng tiếp cận không theo việc phân tích cú pháp của [3] sẽ có thể khắc phục được điều này.
* Tương tự với nghiên cứu của tài liệu [2], theo hướng hỏi đáp trên ngôn ngữ tiếng việt, việc phân tích cú pháp của dựa theo 40 luật của tác giả. Nếu câu không có cú pháp đúng theo 40 luật này thì quá trình phân tích cú pháp thất bại.
* Hệ thống eLSSNL có hướng đi gần giống với nhóm, do hệ thống phục vụ tìm kiếm các sách báo trong kho dữ liệu Gutenberg bằng câu truy vấn tự nhiên. Tuy nhiên, việc xác định các thuộc tính bằng các chỉ định từ chỉ cho ra các kết quả là các cuốn sách. Nếu người dùng muốn biết các thông tin thêm về tác giả, các cuốn sách mà tác giả đó đã viết thì hệ thống eLSSNL không hỗ trợ.
* Ngòai ra, để nhận diện được thực thể đặt tên, là các từ (cụm từ) chỉ về tên người, địa danh , tổ chức...[19] thì ta cần có một từ điển như Wordnet để nhận diện. Hoặc với một cách khác, ta có thể dùng một ontology như theo tài liệu [3]. Đối với đề tài của nhóm, thì các thực thể đặt tên trong dữ liệu DBLP là các title, publisher, source, ... có số lượng rất lớn. Do dữ liệu của DBLP chứa hơn 1,4 triệu bài báo. Vì thế, việc nhận diện thực thể đặt tên theo các cách trên khó thực hiện được.

# Chương 2: Xây dưng hệ thống hỏi đáp để tìm kiếm bài báo khoa học

1. **Cơ sở lý thuyết và các thuật ngữ liên quan**
   1. **Thực thể đặt tên**

Thực thể đặt tên là những thành phần của văn bảng thuộc vào những mục sau: tên người, tên tổ chức, địa điểm, thời gian, số lượng, phần trăm … Trong câu hỏi, các thực thể đặt tên cũng thường hay xuất hiện. Ví dụ trong câu *“What books were written by Philip K. Chan”* thì *Philip K. Chan* được xem là một thực thể đặt tên (tên tác giả).

* 1. **Pos Tagger**

Pos Tagger hay còn gọi là gán nhãn từ loại. Việc sử dụng gán nhãn từ loại sẽ giúp ta biết được vai trò của mỗi từ trong câu (danh từ, động từ hay tính từ ... ). Từ đó, sẽ hỗ trợ ta cho xử lý về sau như việc phân tích cú pháp. Một số công cụ gán nhãn từ loại thường được biết đến: Wordnet, Gate pos tagger, OpenNLP ...

* 1. **Bộ ba quan hệ**

Trong các web ngữ nghĩa hiện nay thường sử dụng mô hình lưu trữ RDF ([Resource Description Framework](http://www.w3.org/RDF/)) [24]. RDF là một mô hình lưu trữ các đối tượng (resource) và các mối quan hệ giữa chúng. Đây là một mô hình ngữ nghĩa đơn giản và được biểu diễn trong cú pháp XML.

Theo [24] và [2] thì RDF là một mô hình lưu trữ theo dạng các bộ ba subject-predicate- object. Trong đó Subject là một chủ từ bắt đầu cho một bộ ba. Predicate là các vị từ ở giữa biểu diễn cho mối quan hệ giữa Subject và Object. Object là đối tượng còn là mà subject có quan hệ tới.

Việc hình thành các bộ ba như trên sẽ có lợi cho việc lưu trữ ngữ nghĩa. Các bộ ba có thể kết hợp lại thành một đồ thị ngữ nghĩa (Hình 2)



Hình 2 – Thông tin được biểu diễn dưới dạng đồ thị liên kết bởi các bộ ba

(Hình lấy từ tài liệu tham khảo [24])

Với hình trên, các bộ ba được liên kết thành một đồ thị. Một độ thị như thế sẽ có thể biểu diễn cho một đoạn văn bản, một thông tin trên web, một câu nói ... Ta có thể dễ dàng thấy rằng “*the\_thirteenth\_floor*” là một “*movie*” với mối quan hệ “*is\_a*”. Đồng thời nó cũng có mối quan hệ để thể hiện ngày phát hành là “*released\_in*” và có giá trị là năm “*1999*”. Vậy tóm lại, với cách hiểu như thế thì câu nói có thể là: *The Thirteenth Floor is a movie which are realesed in 1999*.

Dựa vào các bộ ba quan hệ này mà đã có một số nghiên cứu đã áp dụng để phục vụ cho vấn đề hỏi đáp như [2], [7], [8], [9] và [10]. Chủ yếu các nghiên cứu trình bày cách xây dựng các bộ ba từ câu hỏi và sử dụng chúng để truy vấn xuống một cơ sở dữ liệu đã lưu trữ sẵn các bộ ba như cơ sở dữ liệu hay Ontology. Ontology một dạng mô hình lưu trữ dữ liệu theo dạng RDF.

* 1. **Wordnet và Similarity Wordnet**

**WordNet**: là một dạng từ điển ngữ nghĩa tiếng Anh, được tạo ra từ năm 1985 và đang được duy trì ở các phòng thí nghiệm khoa học nhận thức của Đại học Princeton dưới sự hướng dẫn của giáo sư tâm lý học A. George Miller. Các cơ sở dữ liệu và phần mềm công cụ đã được phát hành theo một giấy phép kiểu BSD và có thể được tải về và sử dụng một cách tự do. Các cơ sở dữ liệu cũng có thể được duyệt trực tuyến. WordNet đã được hỗ trợ từ các Quỹ khoa học quốc gia, DARPA, các DTO (trước đây là Advanced Research and Development Activity), và REFLEX.

Trong WordNet, các từ được xếp vào các nhóm đồng nghĩa gọi là synsets (tập các nghĩa có thể thay thế nhau trong một ngữ cảnh nào đó), cung cấp các định nghĩa ngắn, tổng quát và ghi lại những quan hệ ngữ nghĩa khác nhau giữa các nhóm từ đồng nghĩa. Các nhóm đồng nghĩa liên kết với nhau thông qua các quan hệ ngữ nghĩa được xây dựng thông qua các nghiên cứu về cách sử dụng, cách lưu trữ các tri thức về ngôn ngữ trong bộ não con người. Nhờ cách tổ chức như vậy, WordNet đã cung cấp nhiều tri thức hữu dụng cho việc xử lý ngôn ngữ tự nhiên. Nó được sử dụng với 2 mục đích: để tạo ra sự kết hợp giữa từ điển và từ điển đồng nghĩa có thể sử dụng hợp lý hơn, và để hỗ trợ phân tích văn bản tự động và ứng dụng trí tuệ nhân tạo.

**Similarity Wordnet:** là một công cụ phát triển dựa trên bộ từ điển Wordnet. Nó tính toán mức độ tương tự ngữ nghĩa giữa hai từ và cho ra kết quả là một giá trị trong khoảng [0,1]. Việc sử dụng công cụ này giúp ta vẫn hiểu được ngữ nghĩa của từ nếu người dùng nhập vào một từ đồng nghĩa.

Ví dụ: *author* và *writer* là hai từ có cùng ngữ nghĩa. Nếu sử dụng công cụ này, độ tương tự giữa hai từ sẽ là *1.0*.

1. **Tổng quan về hệ thống**
   1. **Mô hình hệ thống:**

Hệ thống được xây dựng gồm hai luồng chính: người dùng nhập vào từ khóa để tìm bài báo và người dùng nhập câu hỏi đễ tìm kiếm.

1. 
2. Hình 3 – Mô hình hệ thống tìm kiếm bài báo

**Tìm kiếm theo từ khóa:** Người dùng nhập vào các từ khóa để tìm kiếm. Từ khóa có thể là tên bài báo, tên nhà xuất bản và nguồn của bài báo. Kết quả trả về là các bài báo có thuộc tính chứa từ khóa mà người dùng đã nhập.

**Dùng câu hỏi để tìm kiếm:** Thay vì nhập từ khóa. Hệ thống cho phép người dùng nhập vào một câu hỏi tự nhiên để tìm kiếm. Kết quả trả về sẽ là các bài báo phù hợp với các thông tin trong câu hỏi.

* 1. **Các chức năng trong hệ thống**

|  |  |  |
| --- | --- | --- |
| Chức năng | Đầu vào | Đầu ra |
| Tìm kiếm chung theo  từ khóa | Từ khóa là tên tiêu đề bài báo  (hoặc tên nguồn hoặc tên nhà xuất bản) | Danh sách các bài báo liên quan có tiêu đề (tên nguồn, tên nhà xuất bản) chứa từ khóa. |
| Tìm kiếm theo tên  tác giả | Tên tác giả | Danh sách các tác giả có tên tương tự, click chọn tác giả sẽ hiển thị các bài báo mà tác giả đó đã viết |
| Tìm kiếm nâng cao | Các thuộc tính mà người dùng có thể nhập: Năm (lớn hơn, nhỏ hơn, bằng), nhà xuất bản, nguồn, loại. | Danh sách các bài báo liên quan |
| Đa ngôn ngữ |  | Hiển thị đa ngôn ngữ |
| Hỏi đáp | Câu hỏi dạng ngôn ngữ tự nhiên | Các bài báo liên quan đến nội dung câu hỏi |

* Chức năng tìm kiếm: cung cấp chức năng tìm kiếm bài báo dựa vào từ khóa. Chức năng cho phép người dùng nhập vào từ khóa có liên quan đến tiêu đề, nguồn, nhà xuất bản.
* Chức năng tìm kiếm nâng cao: Cho phép người dùng nhập nhiều dữ liệu hơn để cho kết quả chính xác như mong muốn. Các thuộc tính mà người dùng có thể nhập: Năm (lớn hơn, nhỏ hơn hoặc bằng), nhà xuất bản, nguồn, loại.
* Chức năng tìm kiếm theo tên tác giả: Nhập vào tên tác giả, chươn trình trả về các tác giả có tên tương tự. Khi chọn vào một tác giả sẽ hiển thị các bài báo mà tác giả đó đã viết (nhóm theo từng năm).
* Đa ngôn ngữ: Hỗ trợ giao diện hiển thị đa ngôn ngữ
* Chức năng hỏi đáp: Người dùng thay vì tìm kiếm bằng từ khóa, nội dung nhập vào sẽ là một câu hỏi tự nhiên. Dựa trên phân tích câu hỏi chuyển thành câu truy vấn xuống cơ sở dữ liệu DBLP, chương trình trả về thông tin hoặc bài báo liên quan câu đến câu hỏi.

Ví dụ:

*Câu hỏi:* How many publications were written by *Philip K. Chan*?

Kết quả: <Danh sách các bài báo mà tác giả Philip K. Chan đã viết>

*Câu hỏi:* Who wrote paper “[*Incrementally Learning Rules for Anomaly Detection*](http://localhost:8080/QAWeb/showPubDetail.do?id=235040)*.”*?

Kết quả:

*\_* [*Denis Petrussenko*](http://localhost:8080/QAWeb/showPubsByAuthor.do?authorName=Denis%20Petrussenko)*,*

*\_* [*Philip*](http://localhost:8080/QAWeb/showPubsByAuthor.do?authorName=Denis%20Petrussenko)K. Chan,

1. **Thiết kế cài đặt hệ thống**
   1. **Môi trường xây dựng hệ thống**
   2. **Câu trúc các lớp**
   3. **Phương pháp đề xuất xử lý câu hỏi chuyển thành câu truy vấn SQL**
   4. **Cơ sở dữ liệu DBLP**

Dữ liệu của đề tài lấy từ nguồn dữ liệu DBLP. Cấu trúc các bảng được trình bày như bên hình dưới:

****

* **Table Publication (dblp\_pub\_new)**

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Loại | Giải thích |
| id | int(8) | Khóa toàn cục trong CSDL |
| dblp\_key | varchar(150) | Khóa trong file xml (một dạng khóa khác tổ chức trong dữ liệu xml) |
| title | longtext | Tiêu để bài báo |
| source | varchar(150) | Nguồn bài báo |
| source\_id | varchar(50) | Tham chiếu đến nguồn bài báo (đường dẫn trong dblp\_key) |
| series | varchar(100) | Series của bài báo ( chỉ dành cho kiểu book và proceedings |
| year | int(4) | Năm công bố |
| type | varchar(20) | Kiểu bài báo (vd: article, proceedings, book …) |
| volume | varchar(50) | Khối lượng của nguồn, nơi tác phẩm được xuất bản |
| number | varchar(20) | Số lượng nguồn nơi mà tác phẩm được công bố |
| month | varchar(30) | Tháng xuất bản |
| pages | varchar(100) | Số trang |
| ee | varchar(200) | Đường dẫn tới phiên bản điện tử |
| ee\_PDF | varchar(200) | Đường dẫn tới phiên bản điện tử dạng PDF |
| url | varchar(150) | Địa chỉ tới trang dblp ( bắt đầu với db/…) |
| publisher | varchar(250) | Tên nhà xuất bản |
| isbn | varchar(25) | Mã ISBN |
| crossref | varchar(50) | Khóa tham chiếu chéo tới một bài báo khác |
| titleSignature | varchar(255) | Tiêu đề bài báo (không có khoảng trắng) |
| doi | varchar(255) | Digital Object Identifier |
| mdate | date | Ngày chỉnh sửa |

* **Table Author (dblp\_author\_ref\_new)**

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Loại | Giải thích |
| pub\_id | int(8) | Khóa tương ứng với id trong bảng dblp\_pub\_new |
| author | varchar(70) | Tên tác giả |
| editor | int(1) | Biến xác nhận tác giả có phải là người chỉnh sửa |
| author\_num | int(3) | Số thứ tự của tác giả |

* **Table Publication Reference (dblp\_ref\_new)**

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Loại | Giải thích |
| id | int(8) | Khóa tương ứng với id trong bảng dblp\_pub\_new |
| dblp\_key | varchar(150) | Khóa dblp\_key của những bài báo được trích dẫn |

* 1. **Giao diện chương trình**
* Giao diện chính khi khởi động chương trình

****

* Giao diện tìm kiếm theo từ khóa



# Tham Khảo

**Tiếng Việt**

1. Đỗ Thị Thanh Tuyền (2008), “XÂY DỰNG HỆ THỐNG TRA CỨU THƯ VIỆN ĐIỆN TỬ BẰNG NGÔN NGỮ TỰ NHIÊN”.
2. Lương Quý Tịnh Hà (2009), “Xây dựng công cụ tìm kiếm tài liệu học tập bằng các truy vấn ngôn ngữ tự nhiên trên kho học liệu mở tiếng Việt”.
3. Cao Duy Trường (2008), “ Dịch câu truy vấn có cấu trúc sang đồ thị ý niệm: cách tiếp cận ít phụ thuộc vào cú pháp”.

**Tiếng Anh**

1. Li and Roth (2002), Learning Question Classifier.
2. LiXin, Huang Xuan Jing, Wu Lid-de (2006), Question Classification by Ensemble Learning.
3. Dell Zhang and Wee Sun Lee (2003), Question classification using support vector machines.
4. Wael Salloum (2009), “A Question Answering System based on Conceptual Graph Formalism”.
5. Shiyan Ou, Constantin Orasan, Dalila Mekhaldi and Laura Hasler (2008), “Automatic Question Pattern Generation for Ontology-based Question”,.
6. Meltem Serhatli and Ferda N. Alpaslan (2009), “An Ontology based Question Answering System on Software Test Document Domain”.
7. Lorand Dali, Delia Rusu, Blaz Fortuna, Dunja Mladenic and Marko Grobelnik (2009) “ Question Answering Based on Semantic Grahps”.
8. Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, Dunja Mladenic (2007) “Triplet Extraction from Sentences”.

**Online**

1. Digital Bibliography & Library Project

<http://www.informatik.uni-trier.de/~ley/db/>

1. Wikipadis Question Types

[http://en.wikipedia.org/wiki/Question#Types](http://en.wikipedia.org/wiki/Question)

1. English Club

<http://www.englishclub.com/grammar/verbs-questions_types.htm>

1. Oxford

[http://oxforddictionaries.com/view/entry/m\_en\_gb0680290#m\_en\_gb0680290](http://oxforddictionaries.com/view/entry/m_en_gb0680290)

1. Learn English

<http://www.learnenglish.de/grammar/questiontext.htm>

1. Ontology

<http://www.w3schools.com/RDF/rdf_owl.asp>, <http://www.phpvn.org/index.php?topic=133.0;wap2>

1. SPARQL

<http://en.wikipedia.org/wiki/SPARQL>

1. Named Entity

<http://en.wikipedia.org/wiki/Named_entity_recognition>

1. GATE tool

<http://gate.ac.uk/>

1. NSIR

<http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi>

1. **The START Natuaral Language Question Answering System**

[http://start.csail.mit.edu](http://start.csail.mit.edu/)

1. Ask.com

<http://www.ask.com>

1. RDF ([Resource Description Framework](http://www.w3.org/RDF/))

<http://www.xml.com/pub/a/2001/01/24/rdf.html?page=2>

1. http://dblp.uni-trier.de/xml/ [↑](#footnote-ref-2)
2. http://trec.nist.gov/ [↑](#footnote-ref-3)
3. <http://www.vocw.edu.vn/> [↑](#footnote-ref-4)
4. <http://www.gutenberg.org/wiki/Main_Page> [↑](#footnote-ref-5)