MỤC LỤC

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 3](#_Toc288434521)

[**1.1** **Giới thiệu** 3](#_Toc288434522)

[**1.2** **Mục tiêu và phạm vi đề tài** 4](#_Toc288434523)

[**1.3** **Kết quả dự kiến** 5](#_Toc288434524)

[**1.4** **Cấu trúc nội dung trình bày** 6](#_Toc288434525)

[**CHƯƠNG 2 : KHẢO SÁT CÁC NGHIÊN CỨU LIÊN QUAN** 7](#_Toc288434526)

[**2.1** **Cơ sở lý thuyết và các thuật ngữ liên quan** 7](#_Toc288434527)

[**2.1.1** **Giới thiệu về hỏi đáp** 7](#_Toc288434528)

[**2.1.2** **Nhận diện thực thể có tên** 8](#_Toc288434529)

[**2.1.3** **Gán nhãn từ loại** 9](#_Toc288434530)

[**2.1.4** **Bộ ba quan hệ và đồ thị khái niệm** 10](#_Toc288434531)

[**2.1.4.1.** **Bộ ba quan hệ** 10](#_Toc288434538)

[**2.1.4.2.** **Đồ thị khái niệm** 11](#_Toc288434539)

[**2.1.5** **Từ điển Wordnet** 13](#_Toc288434540)

[**2.2** **Các nghiên cứu liên quan đến hỏi đáp và tìm kiếm** 14](#_Toc288434541)

[**2.2.1** **Hệ thống hỏi đáp dựa trên đồ thị ý niệm** 14](#_Toc288434542)

[**2.2.2** **Tra cứu thư viện điện tử bằng ngôn ngữ tự nhiên** 19](#_Toc288434543)

[**2.2.3** **Nghiên cứu rút trích bộ ba trong câu** 20](#_Toc288434544)

[**2.2.4** **Nhận xét** 22](#_Toc288434545)

[CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT XỬ LÝ CÂU HỎI 24](#_Toc288434546)

[**3.1** **Mở đầu** 24](#_Toc288434547)

[**3.2** **Tạo tập tin cấu hình ngữ nghĩa cho cơ sở dữ liệu** 26](#_Toc288434548)

[**3.3** **Các bước xử lý câu hỏi người dùng** 30](#_Toc288434549)

[**3.3.1** **Gán nhãn từ loại** 30](#_Toc288434550)

[**3.3.2** **Đơn giản hóa danh từ, động từ, cụm động từ** 33](#_Toc288434551)

[**3.3.3** **Rút các bộ ba** 35](#_Toc288434552)

[**3.3.4** **Nhận diện thực thể** 39](#_Toc288434553)

[**3.3.5** **Phân loại câu hỏi** 47](#_Toc288434554)

[**3.3.6** **Sinh câu truy vấn SQL** 48](#_Toc288434555)

[**3.3.7** **Một số ví dụ** 56](#_Toc288434556)

[CHƯƠNG 4: HIỆN THỰC HỆ THỐNG 64](#_Toc288434557)

[**4.1** **Tổng quan về hệ thống** 64](#_Toc288434558)

[**4.1.1** **Mô hình hệ thống** 64](#_Toc288434559)

[**4.1.2** **Các chức năng trong hệ thống** 65](#_Toc288434560)

[**4.2** **Thiết kế cài đặt hệ thống** 67](#_Toc288434561)

[**4.2.1** **Cấu trúc các lớp** 67](#_Toc288434562)

[**4.2.2** **Cơ sở dữ liệu DBLP** 68](#_Toc288434563)

[**4.2.3** **Giao diện chương trình** 71](#_Toc288434564)

[**4.2.4** **Đánh giá và nhận xét** 74](#_Toc288434565)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 75](#_Toc288434566)

[5.1 Kết luận 75](#_Toc288434567)

[5.2 Hướng phát triển 75](#_Toc288434568)

[TÀI LIỆU THAM KHẢO 76](#_Toc288434569)

[Tiếng Việt 76](#_Toc288434570)

[Tiếng Anh 76](#_Toc288434571)

[Trang web 77](#_Toc288434572)

[Phụ lục A : Tập 102 câu hỏi kiểm thử 79](#_Toc288434573)

[Phụ lục B : Danh sách các nhãn từ loại Penn Treebank 82](#_Toc288434574)

[Phụ lục C : Tập 19 luật từ loại 83](#_Toc288434575)

[Phụ lục D : Tập tin cấu hình ngữ nghĩa cho cơ sở dữ liệu BDBLP 84](#_Toc288434576)

[Phụ lục E: Những bộ ba rút được từ tập tin cấu hình ngữ nghĩa. 88](#_Toc288434577)

# 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. Để đáp ứng nhu cầu tìm kiếm của người dùng trên lượng thông tin ngày càng lớn, hàng loạt các công cụ tìm kiếm đã được ra đời như Bing[[1]](#footnote-2), Yahoo![[2]](#footnote-3), Google[[3]](#footnote-4), … Nhưng vấn đề đặt ra ở đây là: Các công cụ tìm kiếm trên, hầu hết đều thực hiện việc tìm kiếm dựa trên các từ hay cụm từ khóa. Kết quả trả về của các hệ thống trên thường là các trang liên kết (hoặc văn bản) có nội dung chứa từ khóa tìm kiếm. Vì thế để có được thông tin chính xác nhất, người dùng cần tốn thời gian để duyệt rất nhiều các kết quả trả về từ công cụ tìm kiếm, đôi khi kết quả trả về không phù hợp với mục đích của người dùng.

Để khắc phục vấn đề tìm kiếm dựa trên từ khóa, một số nghiên cứu đã đề xuất ra các hệ thống hỏi đáp [1, 6, 7, 8, 13], mục đích của những nghiên cứu là làm cho hệ thống có thể hiểu đúng được ngữ nghĩa trong câu hỏi mà người dùng đưa vào, như vậy hệ thống tìm kiếm có thể đưa ra câu trả lời chính xác nhất, thỏa mãn được nhu cầu người dùng. Hơn thế nữa, cách thức hỏi đáp sẽ làm máy tính trở nên “thông minh” hơn, tạo ra một hướng giao tiếp thân thiện giữa người và máy.

Trong lĩnh vực khoa học, nhu cầu tìm kiếm thông tin bài báo phục vụ cho các nghiên cứu và tham khảo là rất cần thiết. Hiện nay để tìm kiếm một bài báo, người dùng có thể tìm kiếm trên các thư viện số phổ biến như: CiteSeerX[[4]](#footnote-5), IEEE[[5]](#footnote-6), ACM[[6]](#footnote-7), ScienceDirect[[7]](#footnote-8), SSRN[[8]](#footnote-9), PaperCube[[9]](#footnote-10), Google Scholar[[10]](#footnote-11) ... Tuy nhiên, các thư viện này vẫn còn tìm kiếm chủ yếu dựa trên từ khóa do người dùng nhập vào. Việc duyệt tìm để có được bài báo mong muốn trong rất nhiều kết quả trả về sẽ phải hao tốn một khoảng thời gian, tương tự như trường hợp các công cụ tìm kiếm thông tin ở trên.

Với nhận định như thế, đề tài hướng tới xây dựng một hệ thống hỏi đáp phục vụ cho nhu cầu tìm kiếm các bài báo khoa học. Nó sẽ cung cấp cho người dùng một giao diện hỏi đáp để tìm kiếm bài báo. Người dùng sẽ nhập vào một câu hỏi dưới dạng ngôn ngữ tự nhiên thay cho từ khóa. Kết quả trả về là các thông tin hoặc bài báo tương ứng với nội dung của câu hỏi, thõa mãn đúng theo câu hỏi người dùng đặt ra.

* 1. **Mục tiêu và phạm vi đề tài**

Đề tài hướng tới xây dựng một trang web với mục đích phục vụ tìm kiếm các bài báo khoa học 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 bao gồm :

* Tìm kiếm bằng từ khóa thông thường.
* Tìm kiếm bằng hình thức đặt ra một câu hỏi cho hệ thống thông qua giao diện hỏi đáp.

Câu hỏi và từ khóa người dùng nhập vào sẽ ở dạng ngôn ngữ là tiếng Anh.

Nguồn dữ liệu chứa thông tin về các bài báo sẽ được lấy từ nguồn DBLP[[11]](#footnote-12) (Digital Bibliography & Library Project). Đây là một kho thư viện điện tử lưu trữ thông tin của hơn 1,5 triệu bài báo. Dữ liệu DBLP luôn được cập nhật theo định kỳ (mới nhất tính đến 1/2011). Đây là một kho dữ liệu lớn và thích hợp để chúng em thực hiện đề tài. Ngoài ra, kho dữ liệu này còn được cung cấp miễn phí và tải về tại trang chủ của DBLP.

Dựa vào nguồn dữ liệu đã có trong DBLP, các câu hỏi của người dùng đặt ra chỉ giới hạn trong phạm vi các thông tin : tiêu đề, tên tác giả, năm công bố, nguồn gốc và nhà xuất bản. Các dạng câu hỏi dự kiến có thể xử lý được bao gồm:

* Yes/No Question: Câu hỏi nghi vấn.
  + Ví dụ: Are there any books written by Mike Holcombe in 2000 ?
* Wh-Question: Các câu hỏi bắt đầu với who, what, which và how many.
  + Ví dụ: What book did Philip K. Chan write in 1999 ?
* List question : Câu hỏi dạng liệt kê.
  + Ví dụ: Name all publications were publish in 2010.

Các dạng trên dựa theo phân loại chuẩn trong tiếng Anh [15, 16, 17] (Yes/no question và Wh-question) và phân loại List Question của tập câu hỏi chuẩn TREC[[12]](#footnote-13).

Cuối cùng, để đánh giá kết quả đề tài, một tập 102 câu hỏi thuộc 3 dạng câu hỏi trên được tạo ra để kiểm thử. Các câu hỏi trong tập này được xây dựng bằng tay dựa vào một số mẫu câu hỏi chuẩn trong TREC, và một số câu hỏi khác được cho là cần thiết trong tìm kiếm bài báo. (giới hạn bởi thông tin dữ liệu của DBLP)

* 1. **Kết quả dự kiến**

Xây dựng một hệ thống phục vụ tìm kiếm các thông tin bài báo với giao diện thân thiện và dễ dàng. Người dùng nhanh chóng nắm bắt được cách sử dụng và định hướng được thông tin tìm kiếm dựa trên cách tổ chức dữ liệu trên giao diện web.

Cung cấp một giao diện hỏi đáp cho người dùng để tìm các bài báo. Dựa trên thuật toán đề xuất, hệ thống dự kiến xử lý được 90% các câu hỏi trong tập kiểm thử.

* 1. **Cấu trúc nội dung trình bày**

Ở chương đầu, chúng em trình bày tổng quan về đề tài, nêu rõ mục đích và phạm vi thực hiện cũng như các kết quả dự kiến.

Chương tiếp theo - Chương 2 - sẽ trình bày nội dung khảo sát một số nghiên cứu liên quan đến đề tài. Trước đó, chương sẽ giải thích một số khái niệm để làm cơ sở để hiểu rõ các vấn đề trong phần khảo sát. Tiếp theo, sẽ giới thiệu những nghiên cứu của một số các tác giả liên quan đến vấn đề hỏi đáp và kèm theo các kết quả mà họ đạt được.

Chương 3 trình bày về phương pháp đề xuất để giải quyết vấn đề đặt ra trong đề tài là sử dụng câu hỏi để tìm kiếm các bài báo trong DBLP. Chương sẽ nêu ra từng bước mà câu hỏi được chuyển thành câu truy vấn SQL xuống cơ sở dữ liệu.

Chương 4 và 5 sẽ trình bày nội dung về hiện thực hệ thống, các đánh giá thực nghiệm. Cuối cùng sẽ là kết luận của đề tài và đưa ra hướng phát triển.

**CHƯƠNG 2 : KHẢO SÁT CÁC NGHIÊN CỨU LIÊN QUAN**

Nội dung chính của chương này trình bày kết quả khảo sát một số nghiên cứu liên quan đến đề tài. Ở mục **2.1**, chương sẽ giải thích một số khái niệm để làm cơ sở để hiểu rõ các vấn đề trong phần khảo sát. Tiếp theo, mục **2.2** sẽ giới thiệu những nghiên cứu của một số các tác giả liên quan đến vấn đề hỏi đáp, tìm kiếm dựa trên ngôn ngữ tự nhiên, các phương pháp mà họ sử dụng và kèm theo đó là các kết quả mà nghiên cứu đạt được.

* 1. **Cơ sở lý thuyết và các thuật ngữ liên quan**
     1. **Giới thiệu về hỏi đáp**

Theo [1, 2, 30, 31], thì hỏi đáp (QA) là một loại rút trích thông tin, cung cấp một tập các tài liệu (ví dụ như các trang Web hay tài liệu cục bộ). Hệ thống có thể đưa ra câu trả lời cho những câu hỏi đặt ra bằng ngôn ngữ tự nhiên. Bởi vì phải xử lý ngôn ngữ tự nhiên (NLP) nên QA phức tạp hơn các loại rút trích thông tin khác (như rút trích văn bản chẳng hạn) và đôi khi nó được xem như là bước kế tiếp của bộ máy tìm kiếm.

QA nghiên cứu giải quyết một phạm vi rộng lớn các loại câu hỏi khác nhau, bao gồm: sự việc, danh sách, sự xác định, như thế nào, tại sao, giả thuyết, ràng buộc ngữ nghĩa và các câu hỏi đa ngôn ngữ. QA tìm kiếm trên phạm vi rộng lớn các tập hợp khác nhau, từ tập hợp văn bản cục bộ đến nội dung trên các trang Web. Bao gồm hai phân hệ chính:

* **Closed-domain QA:** giải quyết những câu hỏi với những lĩnh vực xác định. Ví dụ như: y tế, giáo dục, khoa học, lịch sử ...
* **Open-domain QA:** giải quyết những câu hỏi về mọi lĩnh vực và có thể chỉ cung cấp kiến thức tổng quát chung chung. Những hệ thống như thế này thường có nhiều dữ liệu sẵn có để rút trích câu trả lời.
  + 1. **Nhận diện thực thể có tên**

Theo [5] và [21], nhận diện thực thể có tên (NER-Named Entity Recognition) là một công việc thuộc lĩnh vực trích xuất thông tin nhằm tìm kiếm, xác định và phân lớp các thành tố trong văn bản không cấu trúc thuộc vào các nhóm thực thể được xác định trước như tên người, tổ chức, vị trí, biểu thức thời gian, con số, giá trị tiền tệ, tỉ lệ phần trăm, v.v. Thực thể có tên (Named Entity) có rất nhiều ứng dụng, đặc biệt trong các lĩnh vực như hiểu văn bản, dịch máy, truy vấn thông tin, và hỏi đáp tự động.

Hiện nay, hầu hết các hệ thống nhận diện thực thể có tên áp dụng các kỹ thuật khai thác dữ liệu văn bản, xử lý ngôn ngữ tự nhiên và tiếp cận theo các hướng chính sau:

* Kỹ thuật dựa trên văn phạm ngôn ngữ: qui tắc, luật văn phạm được xây dựng bằng tay nhờ ý kiến chuyên gia ngôn ngữ, và tốn nhiều thời gian cho việc xây dựng qui tắc văn phạm. Qui tắc văn phạm sẽ phải thay đổi khi có sự thay đổi vễ lĩnh vực ứng dụng hay ngôn ngữ.
* Các mô hình học thống kê: ít phụ thuộc ngôn ngữ, và cũng không phụ thuộc vào chuyên gia lĩnh vực nhưng cần chuẩn bị tập dữ liệu huấn luyện thật tốt vả đủ lớn để có thể xây dựng được một bộ phân lớp tối ưu.
* Kết hợp máy học và các kỹ thuật xử lý ngôn ngữ tự nhiên.

Hệ thống nhận diện thực thể có tên phổ biến: có thể kể đến các hệ thống phổ biến hiện nay như:

* Hệ thống Standford NER[[13]](#footnote-14): xây dựng bộ phân lớp CRFClassifier dựa trên mô hình thuộc tính ngẫu nhiên có điều kiện (CRF-Condictional Random Field)
* Hệ thống GATE-ANNIE [[14]](#footnote-15): là một hệ thống con của GATE Framework (General Architecture of Text Engineering) một trong các dự án lớn nhất thuộc khoa Khoa học Máy tính, Đại học Sheffield của Anh. Đây là hệ thống dựa trên các từ điển, Ontology và việc xây dựng luật để đánh dấu (annotation) các thành tố trong văn bản. Việc xác định các thực thể có tên trong văn bản thực hiện trong quá trình đánh dấu văn bản.

Trong đề tài, các thực thể đặt tên là các tên tác giả, nhà xuất bản hay tiêu đề … xuất hiện trong câu hỏi của người dùng.

Ví dụ: trong câu *“What books were written by Philip K. Chan”* thì *Philip K. Chan* (tên một tác giả) sẽ được xem là một thực thể đặt tên.

Trong phần xử lý câu hỏi người dùng, bước xử lý gán nhãn từ loại, các thực thể đặt tên sẽ được gán nhãn bằng tay là danh từ riêng số ít (NNP) (xem mục **3.3.1**).

* + 1. **Gán nhãn từ loại**

Trong phạm trù ngôn ngữ học, gán nhãn từ loại (part-of-speech tagging) còn được gọi là gán nhãn ngữ pháp hay định hướng từ loại. Đó là quá trình đánh dấu lên những từ trong một văn bản thuộc về một phần của bài phát biểu (tài liệu) [29]. Quá trình gán nhãn dựa trên định nghĩa của từ, cũng như ngữ cảnh (Ví dụ: mối quan hệ của nó với các từ lân cận hoặc liên quan trong cùng một câu, cụm từ, hoặc một đoạn. Một hình thức đơn giản của việc này là việc xác định các từ như danh từ, động từ, tính từ, trạng từ ...

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, hoặc loại bỏ hay nhận biết các từ loại cần thiết trong câu, so sánh ngữ nghĩa giữa hai từ có cùng từ loại ...

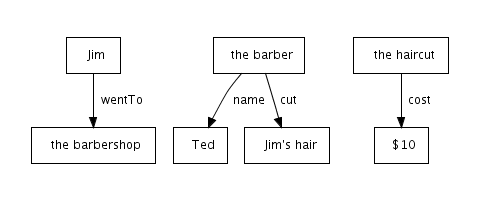
Một số công cụ gán nhãn từ loại thường được biết đến: Wordnet[[15]](#footnote-16), Gate pos tagger[[16]](#footnote-17), OpenNLP[[17]](#footnote-18) ... Trong đề tài, mỗi từ trong câu hỏi người dùng sẽ được gán nhãn từ loại, phục vụ cho quá trình rút các bộ ba dựa trên những luật từ loại mà đề tài đề ra (xem mục **3.3.3**).

* + 1. **Bộ ba quan hệ và đồ thị khái niệm**

1. 1. 4. 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/)) [26]. Theo [1] và [26], 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. RDF lưu trữ thông tin dữ liệu theo dạng các bộ ba subject-predicate- object. Chúng ta có thể hiểu như sau : *subject* là một chủ từ bắt đầu cho một bộ ba, tiếp đó *predicate* là các vị ngữ (ví dụ: động từ ) ở giữa biểu diễn cho mối quan hệ giữa *subject* và *object*, cuối cùng *object* là đối tượng mà chủ từ (*subject*) có quan hệ tới. Từ đó, ta có được cái nhìn đầu tiên về bộ ba quan hệ.

Ví dụ: Một số ví dụ về bộ ba quan hệ



Hình 2.1 – hình minh họa cho các bộ ba quan hệ

**(**Hình lấy từ [32]**)**

Theo hình trên thì ta có các bộ ba sau :

1. <Jim, went\_to, the\_barbershop>

2. <the\_barber, name, Ted>

3. <the\_barber, cut, Jim’s hair>,

4. <the\_haircut, cost, $10>

Mỗi bộ ba sẽ biểu diễn cho một mối quan hệ ngữ nghĩa. Ví dụ giữa Jim (tên người) và *the\_barbershop* (đối tượng) có mối quan hệ là *went\_to* mang ý nghĩa là : *Jim went to the barbershop.* Đặc biệt ở bộ ba thứ 2 và thứ 3 có cùng chủ ngữ (subject) là *the\_barber*, nhưng khác nhau ở mối quan hệ. Do đó ta có thể kết hợp hai bộ ba lại như trên hình vẽ. Các dấu mũi tên chỉ chiều của mối quan hệ.

* + - 1. **Đồ thị khái niệm**

Đồ thị khái niệm là một cách để biểu diễn tri thức được John F. Sowa đưa ra năm 1976 [27]. Theo tài liệu [2] thì đồ thị khái niệm có khả năng diễn đạt ngữ nghĩa một cách chính xác, dễ hiểu đối với con người và khả năng xử lý đối với máy tính.Chính vì vậy, nó thường được xử dụng như một ngôn ngữ trung gian để chuyển đổi giữa ngôn ngữ tự nhiên và nhưng mô hình xử lý hướng máy tính. Một ví dụ của đồ thị khái niệm được minh hoạt ở hình 2.



Hình 2.2 – Thông tin được biểu diễn dưới dạng đồ thị liên kết bởi các bộ ba

(Hình lấy từ tài liệu tham khảo [26])

Với hình trên, chúng ta có thể hiểu, đồ thị được hình thành từ các bộ ba có cùng chủ ngữ (subject) hay đối tượng (object). Một đồ thị như thế sẽ có thể biểu diễn cho một đoạn văn bản, một thông tin trên web, một câu nói ...

Hình 2.2 là một đồ thị thể hiện cho một thông tin về phim ảnh. Trong đó gồm các đối tượng như: tên diễn viên (*vincent\_donofrio, chris\_noth*), tên các phim (*the\_thirteen\_floor, the\_matrix, sex\_and\_the\_city, law\_&\_order\_criminal\_intent*) và một đối tượng khác (*movie, tv\_show*). Các đối tượng liên kết với nhau qua các mũi tên có hướng chỉ chiều quan hệ và đi kèm theo các mũi tên là các cụm từ chỉ mối quan hệ như: *started\_in, is\_a, has\_name, release\_in và similar\_plot\_as*.

Trong hình, ta có thể dễ dàng thấy rằng “*the\_thirteenth\_floor*” là tên một cuốn phim “*movie*”. Do đó nó có 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*. Tương tự, “chris\_noth” là tên một diễn viên tham gia vai chính trong hai bộ phim “*sex\_and\_the\_city” và “law\_&\_order\_criminal\_intent”,* cho nên giữa diễn viên này và hai bộ phim có mối quan *hệ là “starred\_in”.* Vậy, câu thông tin dưới dạng văn bản có thể là: Chris Noth starred in “Sex and City” and “*Law & Order Criminal Intent”.* Đối với các mối quan hệ giữa các đối tượng còn lại trong hình, chúng ta cũng thể thể hiểu theo cách diễn đạt như trên.

Dựa vào các bộ ba quan hệ và đồ thị khái niệm mà đã có một số nghiên cứu đã áp dụng phục vụ cho vấn đề hỏi đáp như [1], [2], [6], [8]và [13]. 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 MSQL/MYSQL hay Ontology[[18]](#footnote-19). Ontology là một dạng mô hình có thể lưu trữ dữ liệu theo dạng RDF.

* + 1. **Từ điển 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 của BSD ([Berkeley Software Distribution](http://en.wikipedia.org/wiki/Berkeley_Software_Distribution)) và có thể được tải về và sử dụng một cách tự do. Các cơ sở dữ liệu cũng có thể được truy cập trực tuyến.

Trong WordNet, các từ được xếp vào các nhóm đồng nghĩa gọi là synsets (tập các từ 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.

Đề tài sử dụng bộ từ điển Wordnet để tìm kiếm các từ đồng nghĩa, đưa các động từ về dạng nguyên mẫu trong tiếng Anh ( ví dụ : nguyên mẫu của từ *wrote* là *write*), và một số công cụ phát triển dựa trên Wordnet như Similarity Wordnet[[19]](#footnote-20). Similarity Wordnet có thể tính toán mức độ tương tự ngữ nghĩa giữa hai từ dựa trên đường đi ngữ nghĩa của chúng trong Wordnet 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. Trong đề tài, công cụ này được sử dụng để tính toán độ tương đồng ngữ nghĩa giữa hai danh từ khi người dùng nhập các từ có ngữ nghĩa tương tự nhau.

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*. Người dùng có thể nhập một trong hai từ mà hệ thống vẫn có thể hiểu được.

* + 1. **Các nghiên cứu liên quan đến hỏi đáp và tìm kiếm**

Trong mục này, chúng em sẽ trình bày một số nghiên cứu liên quan đến hỏi đáp và sử dụng ngôn ngữ tự nhiên trong vấn đề tìm kiếm thông tin (hoặc sách báo). Qua khảo sát, nhiều công trình nghiên cứu đã trình bày những kết quả của họ về cách thức hỏi đáp bằng ngôn ngữ tự nhiên. Một số nghiên cứu điển hình như: [1], [2], [3], [6], [8] và [13]. 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.

Các khảo sát của chúng em tập trung vào một số nghiên cứu sau:

* Hỏi đáp để tìm kiếm thông tin dựa trên đồ thị ý niệm [2, 6, 8, 13].
* Tra cứu thư viện điện tử bằng ngôn ngữ tự nhiên [3].
* Phương pháp rút trích bộ ba trong câu , phục vụ cho vấn đề hỏi đáp [9, 11].

1. **Hệ thống hỏi đáp dựa trên đồ thị ý niệm**

* Một nghiên cứu gần đây của Wael Salloum, 2009 [8]. Mục tiêu nghiên cứu là xây dựng hệ thống hỏi đáp dựa trên phương pháp đồ thị ý niệm, phục vụ cho vấn đề tìm kiếm thông tin. Tác giả đưa ra hướng tiếp cận bằng cách mô hình các văn bản và câu hỏi người dùng thành một dạng đồ thị gọi là Conceptual Graph Formalism (CGF).

Nguồn dữ liệu của nghiên cứu này là từ các văn bản (document) nói chung, không thuộc riêng về một lĩnh vực nào, và có độ tin cậy cao. 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. Theo cách hiểu, thao tác này chỉ đơn giản là so sánh mức độ liên quan giữa đồ thị câu hỏi và đồ thị các câu trả lời. 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 2.3 – Ví dụ về một đồ thị khái niệm

( được lấy từ Figure 3. tài liệu tham khảo [8])

Trong đồ thị ở hình 2.3, chúng ta có thể hiểu như sau, *Author* là một lớp đối tượng liên quan đến tác giả và một thể hiện của nó là *Mark Twain*. Cũng như thế, *Novel* là một lớp đối tượng liên quan đến tiểu thuyết và có thể hiện là Tom Sawyer (tên một cuốn sách tiểu thuyết). 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) gây ra 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 xác định đối tượng cần hỏi đến.

Ví dụ: *“ Who write Tom Sawyer? “* sẽ đuợc chuyển thành như sau:

*[write]-*

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

*(Ptnt) -> [Novel:Tom Sawyer]*

*Hay trình bày theo cách hiểu:*

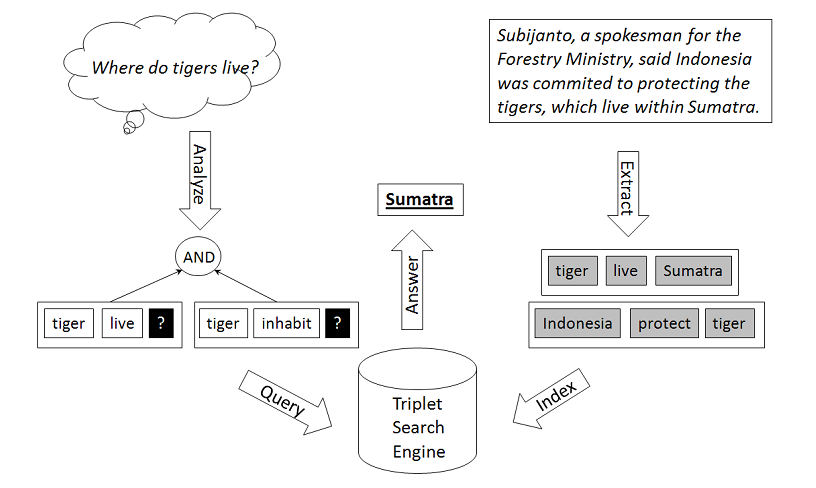
*[Person: ?] -(Agnt)-[write]-(Ptnt)- [Novel:Tom Sawyer]*

Theo cách ghi ở đây, là một cách khác để thể hiện, *Person* là đối tượng cần biết trong câu hỏi , giữa *person* và *Tom Sawyer* có mối quan hệ *write*. *Agnt* và *Ptnt* là hai từ viết tắt của *Agent, Patient*. Đây là một cách thể hiện khác của đồ thị so với hình 2.3. 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 nghiên cứu 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. Tuy nhiên, các câu trả lời được rút ra dựa trên so sánh và xếp hạng. Do đó, 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.

* Tương tự như [8], Lorand Dali cũng nghiên cứu một hệ thống hỏi đáp để tìm kiếm thông tin dựa trên đồ thị ngữ nghĩa (hay ý niệm)[6]. Trong hệ thống, câu hỏi người dùng sẽ được phân tích cú pháp thành cây cú pháp bằng công cụ OpenNLP. Sau đó, cây cú pháp sẽ được phân tích tạo thành các bộ ba. Việc phân tích chủ yếu để duyệt qua các nút và rút ra các đối tượng cần thiết (ví dụ như rút danh từ chính trong các ngữ danh từ (NP), rút động từ chính trong ngữ động từ (VP)). Chi tiết thuật toán phân tích cây cú pháp thành bộ ba được trình bày trong [6].

Dữ liệu mà [6] sử dụng là các tài liệu (document). Hệ thống hỏi đáp này không thuộc về một lĩnh vực nào cả (open-domain). Các tài liệu được rút trích các bộ ba được đánh chỉ mục vào cơ sở dữ liệu (hình 2.4). Hệ thống sử dụng một bộ công cụ tìm kiếm bộ ba tên là Text Garden library[[20]](#footnote-21) có khả năng truy xuất theo các bộ ba phục vụ cho việc hỏi đáp.

Hình 2.4 – Mô hình hệ thống hỏi đáp của Lorand Dali

Quá trình hỏi đáp của hệ thống được thể hiện ở hình 2.4, câu hỏi người dùng: *“Where do tigers live”* qua quá trình phân tích sẽ được chuyền thành bộ ba *<tiger, live, ?>.*Ở đây, nghiên cứu sử dụng công cụ Wordnet để mở rộng thêm số lượng bộ ba, dựa trên mối quan hệ đồng nghĩa (Synonymy) giữa động từ *“live”* và *“inhabit”*. Vì thế, ta được hai bộ ba sau khi phân tích là: *<tiger, live, ?>* và *<tiger, inhabit, ?>* .Sau đó các bộ ba này sẽ được đưa cho bộ công cụ tìm kiếm (Triplet Search Engine) để rút ra kết quả từ dữ liệu lưu trữ. Cuối cùng, hệ thống trả về câu trả lời là “Sumatra”.

Hướng đi tiếp theo của [6] là bổ sung thêm một số mô-đun về nhận diện thực thể đặt tên, cũng như một số mô-đun khác trong việc rút trích bộ ba. Ngoài ra, nghiên cứu còn muốn mở rộng một mô-đun hỏi đáp trên nguồn dữ liệu Ontology.

* Một nghiên cứu khác hướng đến vấn đề dịch câu truy vấn có cấu trúc sang đồ thị ý niệm: cách tiếp cận ít phụ thuộc vào cú pháp, 2008 [2]. 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. Hệ thống chỉ xây dựng đồ thị cho các câu truy vấn đúng qui định sau: các từ quan hệ phải đứng giữa hai thực thể, các từ quan hệ không đứng liền kề nhau. 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 (các mối quan hệ có trong câu hỏi mà không xác định được từ quan hệ. Ví dụ như câu *“What country is Modesto, California in?”*, giữa thực thể *“Modesto”(*CITY) và *“California”*(PROVINCE) tồn tại mối quan hệ nhưng không xuất hiện từ quan hệ nào). 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. 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. Ngoài 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 được 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% |

Bảng 1- Kết quả nghiên cứu tạo đồ thị ý niệm từ tài liệu [2]

Tuy nhiên, nghiên cứu chỉ dừng lại ở mức tạo ra các đồ thị ý niệm. Nếu có sẵn tập tập dữ liệu để truy vấn như tài liệu [8]. Thì công việc hỏi đáp cũng dễ dàng được thực hiện dựa trên so sánh đồ thị của câu hỏi với các đồ thị lưu trữ. Ngoài ra, nghiên cứu cho thấy: ngoài sử dụng phương pháp phân tích cú pháp để tạo đồ thị ý niệm, ta có thể theo phương pháp mà [2] trình bày.

1. **Tra cứu thư viện điện tử bằng ngôn ngữ tự nhiên**

Trong nghiên cứu [3], tác giả xây dựng hệ thống eLSSNL (eLibrary Searching System by Natural Language) phục vụ tìm kiếm các cuốn sách trong thư viện Gutenberg[[21]](#footnote-22). Câu truy vấn dưới dạng ngôn ngữ tự nhiên sẽ không được xây dựng thành đồ thị ý niệm như các nghiên cứu bên trên [2, 6, 8].

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 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 [3] 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.

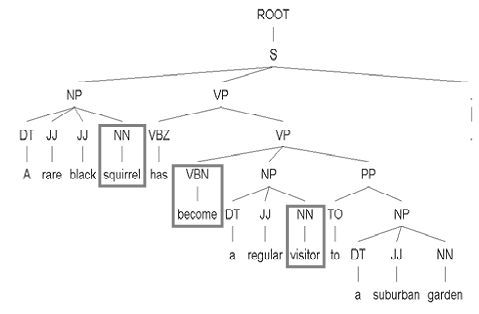
Ngoài ra, hệ thống eLSSNL còn hỗ trợ lưu lại các các câu truy vấn mà người dùng nhập vào và lưu trong một bảng dưới cơ sở dữ liệu. Người dùng còn được hỗ trợ gợi ý cho người dùng các câu truy vấn tương tự. Việc đó dựa vào việc so sánh các tập từ khóa được lưu lại trong bảng vừa đề cập và so sánh với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

1. **Nghiên cứu rút trích bộ ba trong câu**

Mục này, đề tài khảo sát một số nghiên cứu riêng cho việc rút trích các bộ ba trong các câu nằm trong văn bản. Theo [9, 11] thì việc rút ra các bộ ba sẽ phục vụ cho một số vấn đề như tóm tắt văn bản, hỏi đáp, cập nhật dữ liệu vào ontology... .

Khảo sát cho thấy các nghiên cứu thường áp dụng việc phân tích cú pháp để rút ra các bộ ba trong câu hỏi [6, 8, 9,11]. Với nghiên cứu trong tài liệu [11] trình bày về cách rút bộ ba trong câu, các tác giả đã sử dụng nhiều bộ công cụ khác nhau để phân tích cú pháp trong câu như: Stanford Parser[[22]](#footnote-23), OpenNLP, Link Parser[[23]](#footnote-24), MiniPar[[24]](#footnote-25). Phương pháp rút trích bộ ba được miêu tả ở hình sau:



Hình 2.5– Minh họa việc rút trích bộ ba bằng công cụ phân tích cú pháp

Stanford Parser. (Hình lấy từ tài liệu [11])

Cây cú pháp trên là kết quá mà công cụ Stanford parser trả về sau khi phân tích câu: *“A rare black squirrel has become a regular visitor to a suburban garden”*. Trong đó:

* S: là đại diện cho câu (sentence)
* NP : ngữ danh từ
* VP : ngữ động từ
* PP : ngữ giới từ
* DT, JJ, NN, VBZ, TO: các nhãn từ loại (tham khảo phụ lục B).

Một câu sau khi được phân tích cú pháp được biểu diễn bởi ba nhánh con: ngữ danh từ (NP), ngữ động từ (VP) và dấu chấm câu (.). Để rút ra được bộ ba <subject, predicate, object>, thuật toán của [11] thực hiện các bước sau:

* Đầu tiên chủ ngử (subject) được tìm bằng phương pháp vét cạn trên nhánh ngữ danh từ (NP) tìm ra nút con đầu tiên là danh từ (NN, NNP, NNPS hay NNS).
* Tiếp theo là thành phần vị ngữ (predicate) sẽ được tìm trong nhánh ngữ động từ (VP), lấy động từ (VB, VBD, VBG, VBN, VBP, VBZ) nằm trong nút con sâu nhất của nhánh là thành phần thứ hai của bộ ba.
* Cuối cùng, đối tượng (object) được tìm trong các nhánh con của ngữ động từ (VP). Các nhánh con có thể là ngữ danh từ (NP), ngữ giới từ (PP) hay ngữ tính từ (ADJP). Nếu trong các nhánh ngữ danh từ (NP) và ngữ giới từ, ta tìm kiếm danh từ đầu tiên thấy được trong nhánh, trong khi với nhánh ngữ tính từ (ADJP) thì tìm tính từ đầu tiên thấy được (JJ, JJR, JJS).

Kết quả của [11] là đưa ra các thuật toán để rút trích các bộ ba trong câu hỏi dựa trên nhiều công cụ phân tích cú pháp khác nhau. Bên trên chỉ là thuật toán rút trích bộ ba áp dụng cho công cụ Stanford Parser. Với một số thuật toán trên các công cụ khác như MiniPar và Link Parser, ta có thể tham khảo chi tiết trong tài liệu [11].

1. **Nhận xét**

+ Đối với các phương pháp tạo đồ thị ý niệm từ câu hỏi [2, 6, 8], thì việc tạo đồ thị cần dựa trên một số thao tác nhận diện các thực thể đặt tên, không đặt tên. Để nhận diện được thực thể đặt tên, là các từ (cụm từ) chỉ về tên người, địa danh , tổ chức...thì ta cần có một công cụ như GATE để nhận diện. Hoặc với một cách khác, ta có thể dùng một ontology như theo tài liệu [2]. Đối với đề tài này, 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 1,5 triệu bài báo. Vì thế, việc nhận diện thực thể đặt tên (như tên tiêu đề bài báo) theo các cách trên có thể sẽ ảnh hưởng nhiều tới thời gian xử lý và nhận diện. Và hơn thế nữa là đề tài chưa khảo sát thấy một công cụ nào để nhận diện các thực thể đặt tên như tiêu đề bài báo hay tên các hội nghị, hoặc một ontology nào lưu trữ thông tin của các bài báo.

+ Trong những nghiên cứu hỏi đáp đã khảo sát, việc tạo đồ thị ý niệm từ câu trong văn bản và việc tạo đồ thị ý niệm cho câu hỏi là cùng một phương pháp [6, 8]. Đồ thị giữa câu và câu hỏi chỉ khác nhau ở phần thực thể bị khuyết trong câu hỏi, đó chính là đối tượng mà câu hỏi muốn biết đến.

+ Việc rút ra các bộ ba hay tạo thành đồ thị ý niệm được cho là như nhau, vì đồ thị ý niệm được hình thành từ sự liên kết giữa các bộ ba có cùng chủ ngữ (subject) hay đối tượng (object). Nếu có thể rút đúng được các bộ ba ngữ nghĩa trong câu hỏi, chuyển chúng thành một câu truy vấn duy nhất xuống cơ sờ dữ liệu thì ta vẫn có thể lấy được chính xác được thông tin người dùng đang cần hỏi tới.

+ Với cách phân tích cú pháp để tạo đồ thị ý niệm của [6, 8] có thể sẽ gặp khó khăn nếu người dùng nhập một câu có nhập nhằng về cú pháp (hoặc sai cú pháp). Theo đề xuất của [2] cho ta thấy có một hướng đi khác để xây dựng đồ thị ý niệm mà ít phụ thuộc vào phân tích cú pháp. Vậy, trong vấn đề này, ta có thể tạm chia thành hai hướng để xây dựng đồ thị ý niệm là : phân tích cú pháp câu và ít phụ thuộc vào phân tích cú pháp câu.

+ Hệ thống eLSSNL phục vụ tìm kiếm các ebook trong kho dữ liệu Gutenberg bằng câu truy vấn tự nhiên, có hướng đi gần giống với đề tài. Tuy nhiên, việc xác định từ khóa và các thuộc tính bằng các chỉ định từ chỉ cho ra các kết quả là các cuốn sách hay ebook. 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.

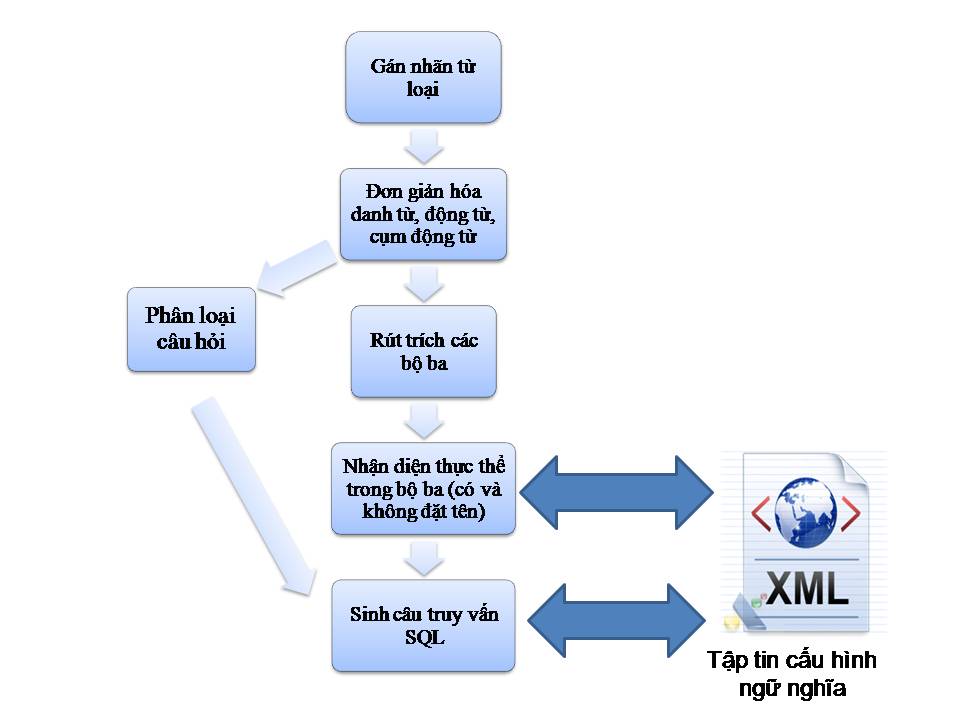
# CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT XỬ LÝ CÂU HỎI

1. **Mở đầu**

Trong chương này, đề tài trình bày các bước xử lý mà đề tài áp dụng để chuyển câu hỏi tự nhiên thành câu truy vấn SQL. Câu truy vấn này sẽ được thực thi để tìm kiếm các bài báo trong thư viện DBLP, trả về cho người dùng.

Đề tài đề xuất một phương pháp rút trích các bộ ba mà không theo phân tích cú pháp câu. Vì trong phân tích cú pháp có thể xảy ra các trường hợp nhập nhằng cú pháp và sai cú pháp. Phương pháp rút trích bộ ba trong đề tài sẽ ít phụ thuộc vào cú pháp tương tự như tài liệu [2]. Quá trình rút trích các bộ ba sẽ đựa trên một số các luật về nhãn từ loại để rút ra các đối tượng cho một bộ ba.

Các bước xử lý được minh họa ở hình sau:



Hình 3.1 – Mô hình minh họa các bước xử lý câu hỏi người dùng

Đầu tiên, câu hỏi sẽ qua hai mô-đun xử lý để thực hiện thao tác gán nhãn từ loại và đơn giản hóa các danh từ, động từ, cụm động từ (đưa động từ trong tiếng Anh về dạng nguyên mẫu) trong câu. Kết quả của hai mô-đun sẽ là một danh sách các từ, cụm từ được gán nhãn. Sau đó, danh sách sẽ được đưa vào mô-đun rút trích để lấy ra các bộ ba dựa vào các luật về từ loại (ví dụ: danh từ (NN)- động từ(VB) - danh từ (NN)). Từ các bộ ba rút được, mô-đun nhận diện các thực thể được thực thi và nhận biết các thực thể trong bộ ba. Ở bước này, đề tài đã xây dựng trước một tập tin XML cấu hình ngữ nghĩa cho cơ sở dữ liệu. Tập tin này miêu tả các mối quan hệ giữa các bảng và thuộc tính với nhau. Trong tập tin còn chứa các từ thể hiện cho các mối quan hệ (Ví dụ: từ *“write”*thể hiện cho mối quan hệ giữa bảng Publication và bảng Author). Việc nhận diện sẽ tiến hành bằng cách so khớp giữa từ quan hệ có trong bộ ba với các từ chỉ mối quan hệ lưu trữ sẵn trong tập tin XML để nhận biết các mối quan hệ giữa bảng và thuộc tính (mục **3.2**). Cuối cùng, Các bộ ba sau khi được nhận diện các thực thể (có thể đóng vai trò là thuộc tính hay là đại diện của một bảng trong cơ sở dữ liệu) sẽ được mô-đun sinh câu truy vấn chuyển thành câu lệnh SQL. Việc hình thành câu truy vấn cần có sự hỗ trợ của tập tin cấu hình ngữ nghĩa XML, vì trong tập tin này có lưu thông tin kết nối các khóa giữa các bảng trong cơ sở dữ liệu.

Bước phân loại câu hỏi giúp xác định loại câu hỏi mà người dùng nhập vào (các loại câu hỏi nằm trong phạm vi đề tài), và xác định đối tượng mà câu hỏi đề cập tới (như hỏi về tác giả, tiêu đề hay nhà xuất bản). Từ đó sẽ là thông tin đầu vào cho bước cuối cùng tạo câu truy vấn.

Ở mục sau sẽ trình bày về cách để tạo ra tập tin cấu hình ngữ nghĩa phục vụ cho bước nhận diện thực thể. Và tiếp theo đó, chi tiết các bước để chuyển một câu hỏi thành câu truy vấn sẽ được nêu rõ.

1. **Tạo tập tin cấu hình ngữ nghĩa cho cơ sở dữ liệu**

Bởi vì thông tin trên cơ sở dữ liệu quan hệ thông thường chưa thật sự miêu tả rõ các mối quan hệ ngữ nghĩa giữa các bảng với nhau hoặc giữa bảng với các thuộc tính của bảng đó. Chúng chỉ chỉ liên kết với nhau qua các từ khóa và quan hệ thuộc tính của bảng. Trên thực tế khảo sát, các hệ thống hỏi đáp dựa trên đồ thị ý niệm và bộ ba quan hệ đều thực hiện trên một cơ sở dữ liệu lưu trữ sẵn các bộ ba [6, 8, 13] hay một ontology ngữ nghĩa [1, 7]. Từ lý do trên, nhóm tạo ra một tập tin cấu hình ngữ nghĩa cho cơ sở dự liệu nhằm lưu các mối quan hệ giữa bảng và thuộc tính, giữa bảng với bảng. Đồng thời, nó còn lưu trữ một số các từ (cụm từ) quan hệ được nhập bằng tay. Chẳng hạn, với một bảng có tên là Publication và nó có thuộc tính là publisher, ta có thể định nghĩa quan hệ giữa bảng Publication và thuộc tính publisher là “is published”. Ta có các từ, cụm từ quan hệ là : be publish by, be release by, from, in ... Như thế, ta có thể hiểu là ta có các bộ ba sau:

(Publication, be publish by, publisher)

(Publication, be realese by, publisher)

(Publication, from, publisher)

(Publication, in, publisher) (Tham khảo thêm tại phụ lục E)

Việc lưu trữ bằng tập tin cấu hình sẽ cho ta một tập các bộ ba giống như trên, các bộ ba này thể hiện các mối quan hệ trong cơ sở dữ liệu, từ đó phục vụ cho các thao tác so sánh và đối chiếu với các bộ ba rút được trong câu hỏi của người dùng ở các bước xử lý sau.

Hơn nữa, tên của các bảng hay thuộc tính trong cơ sở dữ liệu đôi lúc được đặt bởi người dùng, điều đó gây ra sự khó khăn cho việc nhận dạng các thực thể của câu hỏi.

Ví dụ: với câu hỏi *Who write “*Active Database Systems*”*. Giả sử trong cơ sở dữ liệu, bảng chứa thông tin để trả lời câu hỏi vừa rồi có tên là dblp\_pub\_new. Đây là một tên được đặt bởi người tạo ra cơ sở dữ liệu, cho nên nó không mang một ngữ nghĩa nào hết. Vì thế, việc tận dụng các từ điển đồng nghĩa để xác định sự xuất hiện của tên bảng trong câu trở nên khó khăn. Nhưng nếu ta đặt tên bảng *dblp\_pub\_new* bằng một tên tượng trưng khác có nghĩa như *publication* thì việc nhận dạng sẽ dễ dàng hơn nhiều.

Đồng thời, cần có một thông tin để miêu tả các mối quan hệ giữa bảng và các thuộc tính trong bảng, ánh xạ giữa các bảng nếu các bảng đó có quan hệ với nhau. Điều này sẽ giúp ích nhiều trong việc phát sinh câu truy vấn SQL tự động. Từ một câu hỏi được đưa ra, dựa vào các đối tượng trong câu hỏi đó và những mối quan hệ ánh xạ của các bảng với nhau trong database, hệ thống sẽ sinh ra câu truy vấn.

Chính vì thế, đề tài đề xuất một lược đồ XML để mô tả thông tin của database và các mối quan hệ (hình 3).



Hình 3.2 – Lược đồ XML biểu diễn mối quan hệ ngữ nghĩa cho cơ sở dữ liệu

Trong đó: (giá trị đầu là tên của nút cha, giá trị trong dấu “[]” là tên của nút con)

* Database[database-name]: tên của database.
* database [mapping-table]: là thông tin của table ở giữa dùng để ánh xạ mối quan hệ n-n giữa 2 table với nhau.
* mapping-table [table-key]: là khóa chính của table đầu tiên.
* mapping-table [related-table-key]: là khóa chính của table quan hệ.
* mapping-table [mapping-table-name]: là tên của table ánh xạ đã có sẵn trong database.
* mapping-table [name] tên của mapping-table do người dùng đặt.
* database[table]: thông tin của table trong database
* table [primary-key]: Khóa chính của table.
* table [table-name]: Tên của table.
* table[visible]: là thuộc tính dùng để thể hiện table đó có được hiện ra cho người dùng thấy hay không.
* table[table-alias]: tên gọi khác của table đó. Ví dụ: Một table có table-name là dblp\_pub\_new. Ta có thể đặt table-alias là Publication để người dùng dễ hiểu.
* table[field] thông tin thuộc tính của table.
* field[field-name] tên của thuộc tính trong database.
* field[field-alias] tên gọi khác của thuộc tính.
* field[type]: Kiểu dữ liệu của thuộc tính đó.
* field[visible] thể hiện thuộc tính đó có hiện cho người dùng thấy hay không.
* field[relation]: thể hiện mối quan hệ giữa table và thuộc tính.
* relation[relation-name] tên của mối quan hệ.
* relation[related-table] tên của table quan hệ. Thông tin này sẽ hiểu rõ hơn trong ví dụ.
* relation[reversed-relation] tên quan hệ đảo.
* relation[mapping-table] tên của table-mapping ở giữa 2 table

1. **Các bước xử lý câu hỏi người dùng**
   1. **Gán nhãn từ loại**

Bước xử lí này chủ yếu sẽ thực hiện công việc gán nhãn từ loại cho các từ trong câu hỏi mà ngươì dùng nhập vào.

* Đầu vào: câu hỏi người dùng nhập vào.
* Đầu ra: Danh sách các từ đã được gán nhãn từ loại.

Việc gán nhãn từ loại sẽ giúp ta dễ dàng xác định vai trò các từ trong câu. Từ đó, ta dễ dàng ánh xạ với các bộ từ điển ngữ nghĩa như Wordnet để so sánh đối chiếu, hay tìm đồng nghĩa, tìm khái niệm, tổng quá hóa...

Hiện nay, có nhiều công cụ gán nhãn từ loại được đưa ra như OpenNLP[[25]](#footnote-26), Annie Pos Tagger[[26]](#footnote-27) … Mặc dù các công cụ đó đã giúp người dùng có thể gán nhãn hầu như chính xác các từ loại trong câu tiếng anh, tuy nhiên với những câu có nội dung chứa những cụm từ là tên tác giả: Philip K. Chan, Mark Twain…hay tên bài báo: “Active database system”,… thì việc gán nhãn từ loại từng từ trong câu lại không chính xác về mặt ý nghĩa của cụm từ đó**.** Điển hình: Các công cụ đó chưa có thể hiểu được những từ viết hoa liên tiếp như Philip K. Chan, Mark Twain… hay cụm từ trong dấu ngoặc kép như “Active database system”,… mà thay vào đó là những từ riêng lẽ trong cụm từ đó.

Chẳng hạn với câu hỏi:

*Who is the author of* ***“A Distributed Object Manager for the Smalltalk-80 System”****?*

*Thì kết quả xử lý Pos Tagger với công cụ OpenNLP sẽ cho ra kết quả như sau:*



Kết quả sẽ là:

*Who(*WP) *is(*VBZ) *the(DT) author(NN) of(IN) “(NN)* ***A(DT) Distributed(NNP) Object(NNP) Manager(NNP) for(IN) the(DT) Smalltalk-80(JJ) System(NNP) ”(NNP) ?***

Trong đó:

WP: Đại từ câu hỏi.

VBZ: Động từ số ít.

JJ: Tính từ

NN: Danh từ số ít

NNP: Danh từ viết hoa số ít.

DT: Từ hạn định

IN: Giới từ.

(Tham khảo thêm các nhãn từ loại thông dụng ở phụ luc B)

Công cụ ANNIE POS TAGGER sẽ cho kết quả tương tự với khác biệt là dấu ngoặc kép thứ 2 được gán nhãn là NN (danh từ số ít) trong khi với công cụ OpenNLP là NNP(danh từ viết hoa số ít) và với từ ***Smalltalk-80*** thìxuất hiện nhãn NNP (danh từ viết hoa số ít) CD (bảng số) thay vì JJ( tính từ) như ở OpenNLP.



Kết quả:

*Who(*WP) *is(*VBZ) *the(DT) author(NN) of(IN) “(NN)* ***A(DT) Distributed(VBN) Object(NNP) Manager(NNP) for(IN) the(DT) Smalltalk-(NNP) 80(CD) System(NNP) ”(NN) ?***

Với kết quả chạy thử 2 công cụ trên, đề tài nhận thấy kết quả của 2 công cụ hầu như giống nhau. Cho nên đề tài chọn công cụ OpenNLP làm công cụ chính trong quá trình cài đặt. Đồng thời, với 2 kết quả đó cho thấy các cụm từ như ***“A Distributed Object Manager for the Smalltalk-80 System”*** đã bị mất đi ý nghĩa của nó.

Để không làm mất đi ý nghĩa của những cụm từ như tên tác giả, tên bài báo,…đề tài sẽ không đưa những cụm từ đó tham gia vào quá trình gán nhãn từ loại từ động mà gán nhãn bằng tay là danh từ viết hoa số ít(NNP). Những cụm từ được xác định theo điều kiện sau:

* Những từ viết hoa hoặc những cụm từ viết hoa liên tiếp.

Ví dụ: Mark Twain, Philip K. Chan...

* Những từ hoặc cụm từ trong dấu ngoặc kép.

Ví dụ: “Active Database Systems”.

* Những từ viết hoa tòan bộ

Ví dụ: ACM, IEEE …

Ví dụ: Với câu hỏi *Who is the author of* ***“A Distributed Object Manager for the Smalltalk-80 System.”****?*, ta gom những từ trong chuỗi “***A Distributed Object Manager for the Smalltalk-80 System.***” thành một cụm từ. Gán nhãn nó lại thành NNP(danh từ viết hoa số ít).

*Kết quả gán nhãn từ lọai(bằng OpenNLP): Who(WP) is(VBZ) the(DT) author(NN) of(IN)* ***“A Distributed Object Manager for the Smalltalk-80 System.”*** *(NNP)*

Các từ sau khi được gán nhãn sẽ được đưa vào danh sách. Ngoài ra, các mạo từ như: *the, a, an* và một số ký hiệu đặc biệt như *(?,:,;...)* được lọai bỏ khỏi danh sách do chúng không có ý nghĩa gì trong thao tác truy vấn.

* 1. **Đơn giản hóa danh từ, động từ, cụm động từ**
* Đầu vào: danh sách các từ đã được gán nhãn từ loại.
* Đầu ra: danh sách các từ đã được gán nhãn từ loại. Trong đó, các từ, cụm động từ đã được đơn giản hóa.

Mặc dù đã gán nhãn từ loại ở bước trên, nhưng các động từ trong câu vẫn cần được đơn giản hóa. Do đôi khi trong câu xuất hiện các cụm động từ hay động từ như has written, wrote,…có vai trò như nhau về mặt quan hệ. Như trong ví dụ sau:

* Philip K. Chan has written “Active database System”.
* Philip K. Chan wrote “Active database System”.

Ta nhận thấy hai động từ (cụm động từ) “wrote” và “has written” có mối quan hệ như nhau giữa Philip K. Chan và “Active database System”. Và 2 động từ(cụm động từ) này có thể được đơn giản hóa thành từ “write” để chỉ mối quan hệ trong 2 câu trên.

Đối với các danh từ, trong trường hợp danh từ ở trạng thái là danh từ số nhiều trong tiếng Anh (ví dụ: books, authors …), ta cần đưa danh từ trở về trạng thái là danh từ số ít (ví dụ: book, author …). Thao tác này chỉ để đưa các danh từ về một dạng thống nhất tiện cho việc xử lý và so sánh sự tương đồng ngữ nghĩa giữa hai danh từ.

Do vậy, đề tài đưa ra một số quy tắc để đơn giản hóa động từ (cụm động từ) như sau:

* VBP/VBZ(1) + VBN(2) 🡪 VB (2).

Ví dụ: has written 🡪 write.

* VBP/VBZ (1) + VBN (2) + VBN (3) 🡪 VB (2) + VB(3)

Ví dụ : has been written 🡪 be write

* VBP/VBZ (1) + VBN (2) + VBG (3) 🡪 VB(3)

Ví dụ: has been writing 🡪 write.

* VBP/VBZ(1) + VBG(2) 🡪 VB (2).

Ví dụ: is writing 🡪 write.

* VBP/VBZ(1)-be + VBN(2) 🡪 VB(1) + VB(2) (VBP/VBZ là động từ TOBE)

Ví dụ: are written 🡪 be write

* VBP/VBZ(1) + VBG(2) + VBN(3) 🡪 VB(1) + VB(3)

Ví dụ: is being written 🡪 be write

* VBD(1) 🡪 VB(1).

Ví dụ: wrote🡪 write.

* VBD(1) + VBG(2) 🡪 VB(2).

Ví dụ: was writing 🡪 write.

* VBD(1) + VBN(2) 🡪 VB(1) + VB(2)

Ví dụ: was written 🡪 be write

* VBD(1) + VBG(2) + VBN(3) 🡪 VB(1) + VB(3)

Ví dụ: was being written 🡪 be write.

* MD(1) + VBN(2) 🡪 VBP(3)

Ví dụ: can write 🡪 write.

* MD(1) +VB(2) + VBN(2) 🡪VB(2) + VB(3)

Ví dụ: can be written 🡪 be write

Trong đó:

* VBZ, VBD, VBN, MD… là những tự loại trong phụ lục B.
* 1,2,3 …là số thứ tự của động từ

Để phục vụ cho bước đơn giản hóa danh từ, động từ (cụm động từ) ở trên, đề tài sử dụng công cụ Wordnet để chuyển các chúng về dạng nguyên mẫu.

* 1. **Rút các bộ ba**
* Đầu vào: sẽ là danh sách từ loại tạo ra từ bước **3.3.2**.
* Đầu ra: sẽ là tập các bộ ba quan hệ chứa trong câu hỏi. Tuy nhiên các thực thể trong bộ ba này chưa được nhận diện.

Bước này chủ yếu sẽ rút ra các bộ ba chứa trong câu hỏi, việc rút ra chính xác các bộ ba sẽ giúp ta có được các đối tượng và mối quan hệ giữa chúng trong câu hỏi. Với cách thông thường trong các bài khảo sát, bộ ba được rút dựa trên phân tích thành phần cú pháp của câu. Tuy nhiên, với phương pháp này, việc rút bộ ba có thể sẽ bị ảnh hưởng nếu người dùng nhập sai cú pháp (hay nhập nhằng về cú pháp). Vậy, đề tài đề xuất một hướng tiếp cận khác dựa trên các nhãn từ loại để rút ra các bộ ba trong câu. Cách tiếp cận này sẽ phục thuộc ít vào phân tích cú pháp, tương tự như tài liệu [2].

Ý tưởng dựa trên các nhãn từ loại trong câu, ta có thể thấy các danh từ (NN) thường đại diện cho các đối tượng, các động từ (hay cụm động từ) có thể biểu diễn cho quan hệ giữa các danh từ với nhau, và trong [2] thì các giới từ (IN) cũng thể hiện mối quan hệ như động từ. Ví dụ: *author(****NN****) write(****VB****) book(****NN****)* hay *book(****NN****) in(****IN****) 1999(****CD****).*

Đề tài sử dụng những luật từ loại như mẫu sau để rút bộ ba (tổng cộng gồm 18 luật):

1. (NN|NNP)-1 VB-2 (NN)-3 NNP-4-NA VB-5 --> NNP-4-NA,VB-4,(NN|NNP)-1 -->DEL
2. (NN)-1 WDT-2 (NN|NNP)-3-NA VB-4 -->(NN|NNP)-3-NA,VB-4,(NN)-1 -->DEL
3. (NN)-1 WDT-2 VB-3 VB-4 IN-5 (NN|NNP|CD)-6-NA --> NN-1,VB-3 VB-4 IN-5,(NN|NNP|CD)-6 -->DEL
4. ...

(Tham khảo 18 luật tại phụ lục C)

Trong luật, vế trái và vế phải phân cách bởi -->. Trong đó, vế trái là đầu vào, vế phải là đầu ra.Vế trái là một dãy các từ loại mà danh sách các từ loại trong vế trái sẽ so khớp với các từ loại trong câu ở từng vị trí đã được đánh số.

Vế phải sẽ có 3 vế nhỏ phân cách với nhau bởi dấu phẩy, 3 vế nhỏ này lần lượt là chủ thể, từ quan hệ, đối tượng.

Trong mỗi vế sẽ có những từ loại, thông tin trong từ lọai sẽ cách nhau bởi dấu “–”. Trong đó:

* Vế thứ 1 (NPP, NN, VBN...) là từ loại của từ.
* Vế thứ 2 (1, 2…) chỉ vị trí của từ loại trong luật.
* Vế thứ 3 (be, have, NA…) (có thể có hoặc không) chỉ giá trị của từ loại đó. Nếu là NA thì có nghĩa là từ loại tại vị trí tương ứng sẽ bao gồm một hay nhiều từ (cụm từ) có từ loại giống với từ loại ở cùng vị trí trong luật , và các từ loại được nối với nhau bởi các từ quan hệ như And, Or. (NA chỉ là ký hiệu chứ không mang một ý nghĩa nào hết).

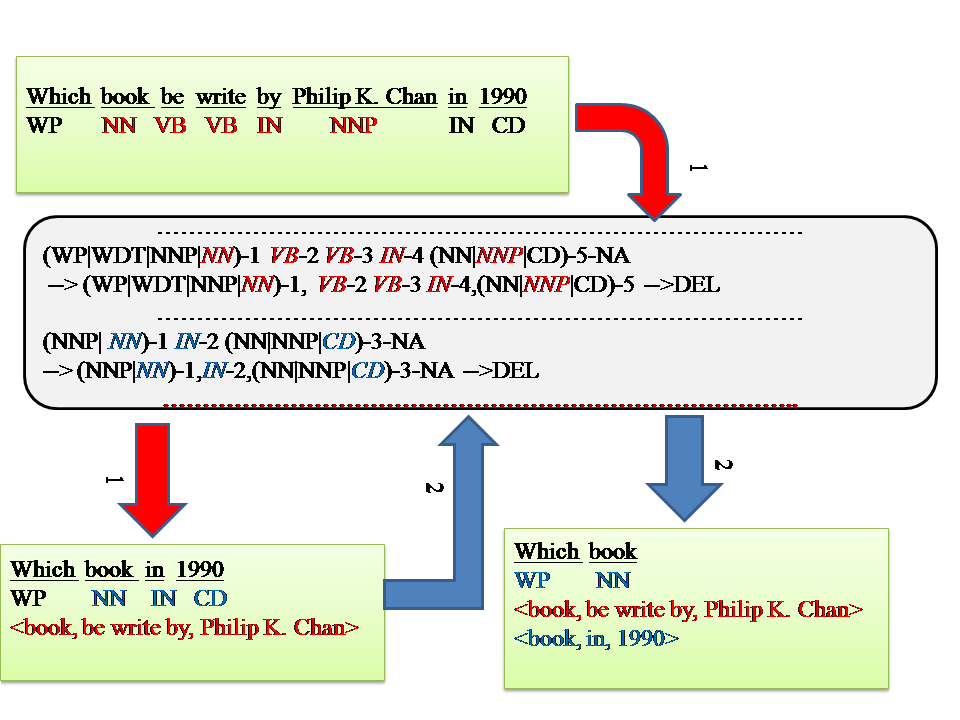
Một khi danh sách từ loại thỏa mãn được một trong các luật mà đề tài sử dụng thì bộ ba sẽ được rút ra. Quá trình này được lặp lại cho tới khi không còn có thể rút được nữa, vì sau khi rút một bộ ba thì các thực thể rút được đồng thời cũng được đưa ra khỏi danh sách.

Ví dụ: Who(WP) write(VB) DataBase(NNP) and(CC) Network Sytem(NNP).

Trong câu trên thì NNP sẽ có giá trị là *NA* vì nó bao gồm 2 danh từ liên tiếp nhau (là Database và Network System) và giữa chúng không có bất kỳ từ quan hệ là động từ hay giới từ nào.

* Từ DEL nói lên rằng các từ loại trong câu có được loại bỏ đi hay không sau khi rút được bộ ba. Việc xóa sẽ bắt đầu từ từ ví trí sau danh từ đầu tiên trong luật.

**Ví dụ:** Which book was written by Philip K. Chan in 1990?



Hình 3.3 - Ví dụ rút bộ ba.

Câu trên qua bước **3.3.1** và **3.3.2** sẽ được chuyển thành:

Which(WP) book(**NN**) be(**VB**) write(**VB**) by(**IN**) Philip K. Chan(**NNP**) in(IN) 1990(CD)

Đầu tiên danh sách sẽ thõa mãn luật thứ 6 là:

(WP|WDT|NNP|***NN***)-1 ***VB***-2 ***VB***-3 ***IN***-4 (NN|***NNP***|CD)-5-NA --> (WP|WDT|NNP|***NN***)-1, ***VB***-2 ***VB***-3 ***IN***-4,(NN|***NNP***|CD)-5 -->DEL

và nó sẽ rút ra được bộ ba

<Book, be write by, Philip K.Chan>.

Tiếp theo luật này sẽ tiếp tục xóa các từ loại sau từ *book* (theo luật nó sẽ bắt đầu xóa các từ loại: be(VB), write(VB),by(IN), Philip K. Chan(NNP) ). Danh sách còn lại:

Which(WP) book(**NN**) in(**IN**) 1990(**CD**)

Tiếp theo nó sẽ chạy các luật kế tiếp nó và so khớp với luật thứ 14 là :

(NNP**|NN**)-1 **IN**-2 (NN|NNP|**CD**)-3-NA --> (WP|WDT|NNP|NN)-1,IN-2,(NN|NNP|CD)-3-NA -->DEL

và rút ra được bộ ba

<book, in, 1990>.

Tiếp tục, luật này sẽ lại tiếp tục xóa các từ loại sau từ *book* (theo luật nó sẽ bắt đầu xóa các từ loại: in(IN), 1990(CD) ). Danh sách còn lại:

Which(WP) book(NN)

Lúc này nó lại tiếp tục chạy tiếp các luật còn lại cho đến hết và không còn luật nào so khớp. Cuối cùng nó rút ra được 2 bộ ba:

<Book, be write by, Philip K.Chan>.

<book, in, 1990>.

* 1. **Nhận diện thực thể**
* Đầu vào: chính là những bộ ba rút được từ bước trên, tập tin cấu hình ngữ nghĩa.
* Đầu ra: những bộ ba có các thực thể đã được nhận diện.

Ở bước này sẽ nhận diện loại thực thể của mỗi thực thể trong từng bộ ba rút được từ bước **3.3.3**. Với việc nhận diện thực thể, ta sẽ nhận biết mỗi thực thể là một trường hợp của bảng (table) hay thuộc tính (properties) trong cơ sở dữ liệu, từ đó sẽ phục vụ việc phát sinh câu truy vấn SQL cho bước sau.

Để nhận diện loại thực thể của từng thực thể trong bộ ba, ta cần sử dụng kết hợp các phương pháp: nhận diện thực thực thể đặt tên (xem **PP1**), nhận diện thực thể không đặt tên (xem **PP2**) mà đề tài sử dụng ở bên dưới theo các bước sau:

Gọi tập bộ ba có được từ tập tin cấu hình ngữ nghĩa là tập Q (tập các bộ ba ở phụ lục E). Ta cần tìm tất cả bộ ba trong Q mà từ (cụm từ) quan hệ giống hay đồng nghĩa (hoặc gần nghĩa) với từ (cụm từ ) quan hệ nằm trong bộ ba cần nhận diện. Ta tạm gọi tập bộ ba tìm được này là tập R.

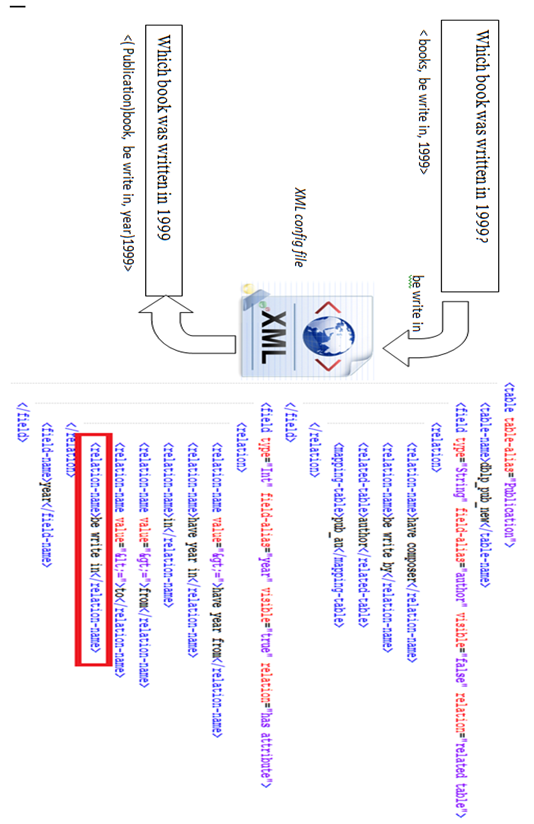
Dựa vào từ điển Wordnet, ta dễ dàng xác định 2 động từ có đồng nghĩa hay gần nghĩa với nhau. Chẳng hạn ta có từ quan hệ là *compose*, thì từ *write* sẽ được xác định là đồng nghĩa khi tra vào từ điển Wordnet. Điều này có nghĩa khi tra mối quan hệ *compose* vào tập Q, đồng thời ta cũng có thể tra theo một mối quan hệ khác tương đương là *write*.

Giá trị tập R lúc này sẽ có 3 trường hợp:

* **Số lượng bộ ba tìm thấy cho tập R bằng 1 (size(R)=1)**. Ta có 2 trường hợp sau:

+ Bộ ba duy nhất không phải quan hệ ngược : Loại thực thể cho mỗi thực thể trong bộ ba cần nhận diện tương ứng với mỗi loại thực thể trong bộ ba duy nhất tìm thấy.

Ví dụ: Với câu hỏi: Which book is written in 1999? (hình 3.4)



**(publication, be write in, year)**

Hình 3.4 - Nhận diện bộ ba

bằng từ quan hệ.

Bằng bước rút bộ ba, ta rút được bộ ba cần nhận diện:

<book, be write in, 1999>

Từ quan hệ “be write in” sẽ được so sánh với tất cả các từ quan hệ của các bộ ba tập Q. Ta sẽ tìm được 1 bộ ba duy nhất là:

(Publication, be write in, year)

Lúc này các thực thể trong bộ ba cần nhận diện <book, be write in, 1999> sẽ được nhận diện tương ứng với các loại thực thể trong bộ ba duy nhất (Publication, be write in, year). Ta sẽ có bộ ba được nhận diện sau:

<book: Publication, be write in, 1999: year>

+ Bộ ba duy nhất là quan hệ ngược: Ta hoán vị các thực thể trong bộ ba cần nhận diên. Sau đó, ta gán loại thực thể cho mỗi thực thể trong bộ ba cần nhận diện (đã được hoán vị) tương ứng với mỗi loại thực thể trong bộ ba duy nhất tìm được.

Ví dụ: Với câu hỏi: Who publish “Active Database System” ?

Bằng bước rút bộ ba, ta rút được bộ ba sau:

<who, publish, “Active Database System”>

Từ quan hệ “publish” sẽ được so sánh với tất cả các từ quan hệ của các bộ ba ở phụ lục E. Ta sẽ tìm được 1 bộ ba duy nhất là:

(Publication, publish, publisher) (Quan hệ ngược)

Bộ ba tìm thấy ở dạng quan hệ ngược, nên ta hoán vị bộ ba cần nhận diện thành <who, publish, “Active Database System”> thành:

<“Active Database System”, publish, who>

Lúc này loại thực thể của từng thực thể trong bộ ba cần nhận diện <“Active Database System”, publish, who> (đã được hoán vị) sẽ tương ứng với các loại thực thể trong bộ ba:(Book, publish, publisher). Ta sẽ có bộ ba được nhận diện sau:

<“Active Database System”: Publication, publish, who: publisher>

* **Số lượng bộ ba tìm thấy cho tập R nhiều hơn 1( size(R)>1):**

Đầu tiên ta cần nhận diện từng thực thể trong bộ ba cần nhận diện. Với cách thức nhận diện thực thể đặt tên (có các nhãn từ loại là NNP,CD) ở mục PP1 và không đặt tên (có nhãn từ loại là NN) ở mục PP2 bên dưới. Ta nhận thấy việc nhận diện thực thể chưa đặt tên chỉ là so sánh ngữ nghĩa giữa 2 danh từ, trong khi nhận diện thực thể đặt tên là đem giá trị của thực thể truy xuất vào database để tìm ra loại thực thể. Nếu database cần truy xuất có dữ liệu lớn thì việc nhận diện thực thể đặt tên sẽ tốn chi phí thời gian rất nhiều. Vì vậy, việc nhận diện sẽ ưu tiên cho thực thể chưa đặt tên trước hơn là thực thể đặt tên. Do đó, giả sử với một bộ ba cần nhận diện được thể hiện là <obj1, relation, obj2> thì thứ tự ưu tiên sẽ là:

* Nhận diện obj1 (nếu obj1 là thực thể chưa đặt tên).
* Nhận diện obj2 (nếu obj2 là thực thể chưa đặt tên).
* Nhận diện obj1 (nếu obj1 là thực thể đặt tên).
* Nhận diện obj2 (nếu obj2 là thực thể đặt tên).

Sau mỗi lần nhận diện thực thể trong bộ ba (đặt tên hoặc chưa đặt tên), ta lại lấy loại thực thể vừa nhận diện được đem so sánh giống nhau với loại thực thể của các bộ ba trong tập R ở vị trí tương ứng. Sau đó loại bỏ trong tập R những bộ ba không thõa mãn điều kiện so sánh. (Chi tiết xem ở ví dụ bên dưới)

Ví dụ với câu hỏi:

Which books are written by Philip K. Chan in 1999?

Câu trên sẽ được chuyển sang danh sách từ loại:

Which(WP) book(NN) be(VB) write(VB) by(IN) Philip K. Chan(NNP) in(IN) 1999(CD) ?

Ta rút được 2 bộ ba cần nhận diện:

<book, be write by, Philip K. Chan>

<book, in 1999>

Với bộ ba cần nhận diện <book, in, 1999> thì khi so sánh từ quan hệ vào tập Q, ta tìm được tập R có ba bộ ba sau:

(Publication, in, year)

(Publication, in, publisher)

(Publication, in, source)

Theo thứ tự ưu tiên vừa nêu, trong bộ ba <book, in, 1999>, ta nhận diện danh từ “book” (có nhãn từ loại là NN) trước tiên. Bằng phương pháp nhận diện thực thể chưa đặt tên được nêu ở mục **PP2**, ta nhận diện được “book” có loại thực thể là ***Publication***. Ta đem ***Publication*** đem so sánh với loại thực thể trong các bộ ba của tập R ở vị trí đầu tiên( vì book giữ vị trí đầu tiên trong bộ ba <book, in, 1999>).

Ta rút ra được tập R có 3 bộ ba:

(Publication, in, year)

(Publication, in, publisher)

(Publication, in, source)

Tiếp tục, ta nhận diện thực thể thứ 2 là 1999 (có nhãn từ loại là CD). Bằng phương pháp nhận diện thực thể đặt tên được nêu ở mục **PP1**, ta nhận diện được “1999” có loại thực thể là ***year***. Ta đem ***year*** đem so sánh với loại thực thể trong các bộ ba của tập R ở vị trí thực thể thứ 2 (vì 1999 giữ vị trí thực thể thứ 2 trong bộ ba <book, in, 1999>). Ta rút gọn được tập R còn 1 bộ ba:

(Publication, in, year)

Nếu trong tập R có một bộ ba mà loại thực thể đầu tiên và loại thực thể thứ 2 của bộ ba đó lần lượt giống với loại thực thể đầu tiên và thứ 2 của các bộ ba còn lại. Ta xóa bỏ các bộ ba còn lại và giữ lại 1 bộ ba duy nhất. Sau đó xử lý giống trường hợp 1.

Ví dụ:

Nếu tập R là 3 bộ ba:

(Publication, have publisher, publisher)

(Publication, be publish by, publisher)

(Publication, be release by, publisher)

Thì ta chỉ giữ lại 1 bộ ba duy nhất là:

(Publication, have publisher, publisher)

* **Nếu số lượng bộ ba tìm thấy cho tập R bằng 0 ( size(R)=0):**

Gán R= Q. Rồi bắt đầu xử lý giống trường hợp **size(R)>1**.

Ngoài ra, việc nhận diện thực thể sẽ được thực hiện tốt hơn bằng cách nhận diện các thực thể có cùng loại (cùng là một loại thực thể như nhau). Như vậy sẽ giảm được chi phí cho quá trình nhận diện do chỉ cần nhận diện một trong hai thực thể. Việc nhận diện này theo một số quy tắc sau:

+ Trước thực thể đặt tên là một thực thể chưa đặt tên thì thực thể đặt tên được nhận diện là loại của thực thể chưa đặc tên.

Ví dụ: *Who write book “Active Database System”?* Với câu này thì book và “Active Database System” có cùng loại thực thể.

+ Các thực thể đặt tên được nối bởi các từ quan hệ and, or.

Ví dụ: *Who* *write “Active Database System” and “My SQL Database”?* Với câu hỏi này thì *“Active Database System”* và *“My SQL Database”* được nhận dạng cùng một loại thực thể.

**Phương pháp 1 (PP1) :Nhận diện thực thể đặt tên**

Trong phạm vi đề tài, thực thể đặt tên là những từ (hay cụm từ) được xác định với nhãn từ loại là danh từ viết hoa số ít (NNP) hoặc số(CD). Để nhận diện được thực thể đặt tên, thông thường các nghiên cứu thường sử dụng các công cụ như GATE hay một ontology. Tuy nhiên, thực thể đặt tên trong đề tài như tiêu đề có số lượng khá lớn (hơn 1,5 triệu), và đề tài chưa tìm thấy một công cụ nào để xác định thực thể đặt tên là tiêu đề bài báo. Vì thế,đề tài ánh xạ giá trị của thực thể đặt tên đó trực tiếp vào cơ sở dữ liệu và theo một số điều kiện.

Việc ánh xạ này sẽ được ưu tiên cho các thuộc tính đại diện của cho tên bảng (ví dụ: thuộc tính đại diện của Author sẽ là Name, của Publication là title) rồi mới xét đến các thuộc tính khác.

Ví dụ: Với câu hỏi What book did James Brown write in 1999 from ACM?

Ta sẽ có được danh sách từ loại như sau:

Which(WP) book() do(VB) James Brown(NNP) write(VB) in(IN) 1999(CD) from(IN) ACM(NNP)

Qua bước rút bộ ba, ta sẽ rút được 3 bộ ba:

< James Brown, write , book>

<book, in, 1999>

<book, from, ACM>

Xét 2 bộ ba <book, in, 1999> và <book, from, ACM>.

Để nhận diện loại thực thể của ACM là bảng(table) hay thuộc tính(property) nào trong cơ sở dữ liệu, ta đem giá trị “ACM” đem tra vào tất cả các thuộc tính của các bảng, ta tìm thấy ACM sẽ ở thuộc tính ***publisher*** của bảng ***Publication***. Do đó, ACM sẽ được nhận diện là ***publisher***. Tương tự thực thể với giá trị là 1999 sẽ được nhận diện là ***year***.

Nếu giá trị của thực thể có kiểu dữ liệu là chuỗi(String) thì chỉ tìm ở các thuộc tính có kiểu dữ liệu là chuỗi(String), nếu kiểu dữ liệu là số(Integer) thì sẽ tìm ở những thuộc tính có kiểu dữ liệu là số(Integer)...

Ví dụ với 2 bộ ba <book, in, 1999>, <book, from, ACM>. Giá trị 1999 chỉ ánh xạ vào những thuộc tính có kiểu dữ liệu là Integer như year, còn giá trị ACM sẽ được ánh xạ vào các thuộc tính có kiểu dữ liệu là String như title,publisher, source…

**Phương pháp 2 (PP2): Nhận diện thực thể chưa đặt tên**

Ngược lại với thực thể đặt tên là thực thể không đặt tên, nó thường là các danh từ xuất hiện trong câu như writer, author, publisher, year…Trong phạm vi đề tài, các loại thực thể này được xác định với nhãn từ loại là danh từ số ít(NN). Với loại thực thể này, chúng đóng vai trò để nhận diện ra các bảng và thuộc tính phục vụ cho tạo câu lệnh SQL. Vì thế, thực thể không đặt tên được nhận biết để đánh dấu sự xuất hiện của những bảng hoặc thuộc tính trong cơ sở dữ liệu.

Tuy nhiên, trong câu hỏi tự nhiên, người dùng có thể tùy ý nhập vào một từ nào đó có cùng nghĩa với từ mình đã xác định (như tên của bảng, thuộc tính). Ví dụ:

Từ “writer” sẽ đồng nghĩa với từ “author” . Nếu như không nhận biết chúng là hai từ đồng nghĩa trả lời được những câu hỏi liên quan đến writer.

Đề tài sử dụng công cụ Similarity Wordnet để hỗ trợ cho quá trình nhận dạng thực thể chưa đặt tên. Similarity Wordnet là một công cụ có thể dùng để đánh giá độ tương đồng về nghĩa giữa hai danh từ. Nếu giữa 2 danh từ hoàn tòan cùng nghĩa, thì mức đánh giá giữa hai danh từ đó là 1. Mức thấp nhất là 0, tức là hai danh từ đó không có liên quan gì về nghĩa với nhau.

Ví dụ: Với câu hỏi *Who is the writer of “*Active Database Systems*” ?*

Thực thể chưa đặt tên ở đây (ngoại trừ từ who) là writer. Giả sử trong cơ sở dữ liệu đã có chứa một bảng có tên là Author.

Bằng công cụ Similarity Wordnet, ta có thể nhận biết được từ *writer* là *author* (có độ tương đồng là 1) thay vì tên bảng nào khác trong cơ sở dữ liệu.

Như vậy, với cách này đầu tiên ta chỉ cần tìm các bảng mà tên của nó có độ tương đồng về nghĩa với thực thể không đặt tên trong một mức giới hạn. Theo quá trình thực nghiệm, thì nhóm chỉ xét độ tương đồng trên *6.5*. Theo một thứ tự ưu tiên, đầu tiên sẽ dùng công cụ so sánh với các tên bảng (tên allias) trong cơ sở dữ liệu. Nếu kết quả đều không thỏa, ta so sánh tiếp với các thuộc tính trong các bảng. Nếu kết quả vẫn tiếp tục không thỏa, ta chọn thuộc tính hay tên bảng có độ tương đồng cao nhất.

* 1. **Phân loại câu hỏi**
* Đầu vào: Danh sách từ loại tạo ra từ bước **3.3.2**.
* Đầu ra: loại câu hỏi và đối tượng câu hỏi.

Bước này nhằm phân loại câu hỏi người dùng thuộc loại này trong các loại câu hỏi mà phạm vi đề tài đặt ra: yes/no question, wh-word question và list question. Việc xác định loại câu hỏi giúp hệ thống trả lời đúng với mục đích hay đối tượng mà người dùng nhắm tới. Đối tượng được xác định dựa trên loại câu hỏi của người dùng. Nhóm dùng một số quy tắc đơn giản tương tự như tài liệu [2] trình bày.

+ Với câu hỏi Who : đối tượng có thể là Person (Author) hoặc Organization (publisher). Tùy theo thực thể nhận diện được trong bộ ba mà quyết định đối tượng cho câu SELECT ở bước sinh câu truy vấn.

+ Với các hỏi What và Which: xác định dựa trên danh từ đứng sau nó (vd: Which books ... thì lấy từ books là đối tượng SELECT).

+ Với câu hỏi How many: là danh từ đứng sau “how many”. (vd: How many books ... thì lấy từ books là đối tượng SELECT).

Các trường hợp khác (ví dụ câu hỏi List question) nếu không xác định được thì mặc định là đối tượng *Publication* trong cơ sở dữ liệu. Việc nhận diện đối tượng câu hỏi cũng theo phương pháp 2 (PP2) trong mục **3.3.4**.

* 1. **Sinh câu truy vấn SQL**
* Đầu vào: các bộ ba được mà các thực thể được nhận diện, đối tượng câu hỏi và lớp câu hỏi từ bước phân loại câu hỏi, tập tin cấu hình ngữ nghĩa.
* Đầu ra: câu lệnh SQL duy nhất.

Ở bước cuối cùng, từ các bộ ba, ta phát sinh ra câu lệnh SQL để có thể tiến hành truy vấn lấy thông tin đúng theo các mối quan hệ trong bộ ba. Bước này thao tác chỉ đơn giản duyệt qua các bộ ba, rút các phần tử và đưa vào câu truy vấn SQL. Câu truy vấn được chia làm 3 phần SELECT, FROM và WHERE.

**SELECT**: Chọn đối tượng để lấy dữ liệu. Đối tượng này được lấy từ bước phân loại câu hỏi.

**FROM**: Xác định các bảng xuất hiện trong các bộ ba và bỏ vào câu truy vấn FROM.

**WHERE**: Bỏ các thuộc tính nhận diện được trong bộ ba vào câu truy vấn. Ngoài ra, còn xác định các kết nối giữa các bảng dựa trên mối quan hệ được đinh nghĩa sẵn trong tập tin cấu hình.

Từ các bộ ba, ta có thể phát sinh ra câu truy vấn như sau:

* **Đối tượng truy vấn (SELECT):**
* Nếu thực thể câu hỏi là thuộc tính:
* Thuộc tính đó có *related-table* (Xem tập tin cấu hình ngữ nghĩa ở phụ lục D):

Ta có câu truy vấn: SELECT `related-table[table-alias]`.\*

Ví dụ: Với câu hỏi: *Who write “Database”?* Ta sẽ rút và nhận diện được bộ ba <Who: *author*,write, “Database”: *publication* >. Dựa vào thông tin cấu hình ở file XML (Xem thêm ở phụ lục D), author là thuộc tính và thuộc tính đó đó có chứa related-table là Author, Author có table-alias là “Author”. Nên ta có câu truy vấn:

SELECT Author.\*

* Thuộc tính không có *related-table*:

Ta có câu truy vấn: SELECT `table[table-alias]`.field[field-alias]

Ví dụ: Với câu hỏi: *Who publish “Active Database System”?* Ta sẽ rút và nhận diện được bộ ba <“*Active Database System*”: *publication*, publish, Who: *publisher* >. Dựa vào thông tin cấu hình ở file XML, *publisher* là thuộc tính và thuộc tính đó đó không có chứa related-table nào. Nên ta có câu truy vấn:

SELECT `Publication`.publisher

* Nếu thực thể câu hỏi là table:

Ta có câu truy vấn: SELECT ` table[table-alias]`.\*

Ví dụ: Với câu hỏi: *Which book is published by O’really?* Ta sẽ rút và nhận diện được bộ ba <book: *publication*, is publish by, “ O’really’ ”: *publisher* >. Dựa vào thông tin cấu hình ở file XML, *publication* là table. Nên ta có câu truy vấn:

SELECT `publication`.\*

* **Nguồn truy vấn (FROM):**

Ta cần lập danh sách tất cả các tên table từ danh sách các bộ ba, nếu có thuộc tính có mapping-table hoặc related-table thì thêm mapping-table hoặc related-table đó vào nguồn truy xuất (mapping-table là thông tin của bảng ánh xạ trong tập tin cấu hình ngữ nghĩa). Một bảng(table) sẽ được đưa vào nguồn truy vấn FROM theo cấu trúc:

`table[table-name]` `table[table-alias]`

Ví dụ: Với câu hỏi: Who write “Active Database Systems” ?

Ta có bộ 3 quan hệ đã được nhận diện < Who:*Author* , write, “Active Database Systems” : *Author*.*publication* >

Ở đây, “Active Database Systems” là thuộc tính *publication* của *Author*, và nó có related-table là *Publication*, mapping-table là *au\_pub*. Bởi vì mapping-table *au\_pub* có mapping-table [mapping-table-name] là `dblp\_author\_ref\_new` trùng với table[table-name] của Author nên ta sẽ không thêm vào.

Nên ta có được mẫu truy vấn sau:

SELECT DISTINCT `Author`.\*

FROM

`dblp\_author\_ref\_new` `Author`,`dblp\_pub\_new` `Publication`

Trường hợp nhiều bộ ba được nhận diện giống nhau, có cùng từ quan hệ, và chỉ khác nhau ở giá trị thực thể (thực thể ở dạng giống nhau).

Ví dụ: Who is the author of "Working Models for Uncertain Data" and "Active Database Systems"?

Ta có các bộ ba đã được nhận diện là:

<author: *Author*, of, "Working Models for Uncertain Data": *Author .publication*>, <author:*Author*, of, "Active Database Systems": *Author .publication*>

Ở đây 2 bộ ba được nhận dạng giống nhau, chỉ khác nhau ở giá trị là "Working Models for Uncertain Data" và "Active Database Systems". Ta gọi Author là thực thể *không* *ở dạng giống nhau* và thực thể còn lại thì ngược lại.

Với trường hợp này thì nguồn *From* lại bao gồm những câu truy vấn con chứa *Select*, *From*, *Where*. Ta tiếp tục đặt tên cho những câu truy vấn con đó theo quy tắc:

table[table-alias] của\_thực\_thể\_không\_ở\_dạng\_giống\_nhau + Số thứ tự n€[1,+∞] (Ngoại trừ trường hợp đầu tiên không đánh số để tối ưu hóa câu truy vấn chính)

Trong đó :

+ Select: là table hay thuộc tính của thực thể đặt tên ở dạng giống nhau. Bước này giống như quá trình tạo câu select ở trên.

+ From: giống như bước tạo From ở trên: xác định tất cả các table trong bộ ba.

+Where: giống như bước tạo Where ở mục dưới.

Sau đó INNER JOIN tất cả các câu truy vấn con đó với nhau với điều kiện là

table[table-alias] + Số thứ tự n. table[primary-key] = table[table-alias] + Số thứ tự n+1. table[primary-key]

Ví dụ: Với ví dụ trên, ta có câu truy vấn cho from như sau:

FROM

(SELECT `Author`.\*

FROM `dblp\_author\_ref\_new` `Author`,`dblp\_pub\_new` `Publication`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Publication`.`title` = "Working Models for Uncertain Data" )) Author

INNER JOIN

(SELECT `Author`.\*

FROM `dblp\_author\_ref\_new` `Author`,`dblp\_pub\_new` `Publication`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Publication`.`title` = "Active Database Systems" )) Author1 ON Author.author = Author1.author

* **Điều kiện truy vấn (WHERE):**

Nếu thực thể là một table và có giá trị, ta dựa vào thuộc tính đại diện của bảng(table) đó trong tập tin cấu hình ngữ nghĩa. Ta tạm gọi thuộc tính đại diện đó là presentation\_field. Ta có cấu trúc sau:

`table[table-alias]`.`presentation\_field[field\_name]` = giá\_trị

Ví dụ: Với thực thể “Active Database System” được nhận diện là ***Publication***, ***Publication*** có:

table[table-alias]:Publication

presentation\_field: title

presentation\_field[field\_name]: title

Giá trị là “Active Database System”. Ta có đoạn query sau:

Publication.title = “Active Database System”

Nếu thực thể là một thuộc tính:

Nếu một thuộc tính có related\_table và có mapping table(quan hệ n-n), ta có thêm điều kiện sau(dựa vào thông tin tại XML):

table[table-alias].table[primary\_key] = mapping\_table[mapping-table-name]. mapping\_table[table-key]

AND mapping\_table[mapping-table-name].mapping\_table[related-table-key] = related-table[table-alias].related-table[primary-key]

Nếu mapping\_table[mapping-table-name] trùng với bất kì table[table-name] thì mapping\_table[mapping-table-name] trong câu truy vấn sẽ được thay bằng table[table-alias] của table đó.

Ví dụ: Với bộ ba <book:Publication,be write by, Philip K. Chan:author> thì:

Bảng Publication có (Xem thêm ở tập tin cấu hình ngữ nghĩa):

table[table-alias]: Publication

table[primary\_key]: id

mapping\_table: pub\_au

mapping\_table[table-key]: pub\_id

mapping\_table[related-table-key]: author

mapping\_table[mapping-table-name]:dblp\_ author\_ref\_new

Bảng Author có:

table[table-alias]: Author

table[primary\_key]: author

table[name]: dblp\_author\_ref\_new

Thuộc tính author có:

related-table: Author

related-table[table-alias]: Author

related-table[primary-key]: Author

Do đó, một đoạn câu truy vấn được sinh ra như sau:

`Publication`.`id` = `dblp\_author\_ref\_new`.`pub\_id` and `dblp\_author\_ref\_new`.`author` = `Author`.`author`

Vì bảng Author và mapping\_table có cùng chung một bảng trong cơ sở dữ liệu là ***dblp\_author\_ref\_new***, nên câu truy vấn trên được rút gọn thành:

`Publication`.`id` = `Author`.`pub\_id`

+ Nếu thuộc tính có related-table đồng thời thuộc tính đó có giá trị, ta làm tương tự với thực thể là table, ta tạm gọi thuộc tính đại diện cho related-table đó là related\_presentation\_field, ta có thêm điều kiện sau:

`related-table[table-alias]`.`related\_presentation\_field[field-name]` = giá\_trị\_thuộc\_tính

Ví dụ: Với bộ ba <book:Publication,be write by, Philip K. Chan:author>. Ta có thêm một số thông tin sau:

Với bảng Author:

presentation\_field:Author

presentation\_field[field-name]:author

Như vậy ta có đoạn câu truy vấn sau:

`Author`.`author` = ‘Philip K. Chan’

Nếu một thuộc tính chỉ có mỗi related\_table (quan hệ n-1) , ta có điều kiện:

table[table-alias].field[field\_name] = related\_table[table\_alias].related-table[primary-key]

Trong đề tài không có ví dụ cụ thể cho trường hợp này, ở đây field[field\_name] được xem là khóa ngoại cho một bảng khác.

Nếu thuộc tính đó đã có giá trị, ta chỉ cần thêm điều kiện:

`table[table-alias]`.`field[field\_name]` toán\_tử giá\_trị\_thuộc\_tính

Toán\_tử: là những toán tử được định nghĩa sẵn cho mỗi từ quan hệ.

from(>=),in( =),to( <=), ….

Ví dụ: Với bộ ba đã được nhận diện: <book:Publication,from,1999:year>

Với thực thể 1999 có:

table[table-alias]: Publication

toán\_tử: >=

Ta có đoạn truy vấn sau:

`Publication`.`year` >= 1999

Ví dụ :Với câu hỏi : Who write “Active Database Systems”?

Với các bộ 3 đã được nhận diện < Who:*Author* , write, “Active Database Systems” :*Publication* >

, “Active Database Systems” là thuộc tính publication của author và có mapping-table-name là ref\_author\_book, có related-table là publication, đồng thời thuộc tính đại diện của publication là title có giá trị là “Active Database Systems”, ta sinh được câu truy vấn sau:

SELECT DISTINCT `Author`.\*

FROM

`dblp\_author\_ref\_new` `Author`,`dblp\_pub\_new` `Publication`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND `Publication`.`title` = `Active Database Systems `

Ví dụ: Với câu hỏi “Who published “Active Database Systems”

Các bộ ba được tìm thấy <“Active Database Systems” : *Publication*,publish, Who : *publisher*. Ta có câu truy vấn:

SELECT Publication.publisher

FROM

`dblp\_pub\_new` `Publication`

WHERE

(`Publication`.`title` = “Active Database Systems” )

Sau khi kết hợp 3 phần, ta có được câu truy vấn hoàn chỉnh và lấy dữ liệu từ DBLP. Như mô hình hệ thống ở mục sau, thành phần xử lý câu trả lời được nêu ra chủ yếu để trình bày nội dung kết quả trên giao diện web. Ví dụ: thông tin trả về là các bài báo thì người dùng có thể nhấp chuột vào các bài báo để xem thông tin chi tiết (đường link).

* 1. **Một số ví dụ**

Một vài ví dụ mà hệ thống giải quyết được với thông tin trong cơ sở dữ liệu DBLP.

***Ví dụ 1:*** Câu hỏi “Who published “Active Database Systems”?

* Bước 1: Gán nhãn từ loại

Đầu vào : Câu hỏi.

Đầu ra: Danh sách các từ đã gán nhãn từ loại.

Kết quả: WP/Who VBD/published NNS/“Active Database Systems”.

* Bước 2: Đơn giản hóa danh từ, động từ, cụm động từ

Đầu vào: Danh sách nhãn từ loại ở bước 1

WP/Who VBD/published NNS/“Active Database Systems”.

Đầu ra: Danh sách nhãn từ loại đã được đơn giản hóa.

Kết quả: WP/Who VB/publish NNS/“Active Database Systems”.

* Bước 3: Rút bộ ba

Đầu vào: danh sách từ được gán nhãn ở bước 2.

**WP**/Who **VB/**publish **NNP**/“Active Database Systems”.

Đầu ra: Tập các bộ ba quan hệ.

Xử lý: Danh sách các từ được gán nhãn sẽ khớp với luật 13 trong tập luật (Xem thêm ở phụ lục C):

(**WP**|WDT|NNP|NN)-1 **VB**-2 (NN|**NNP**)-3-NA --> (WP|WDT|NNP|NN)-1,VB-2,(NN)-3

Kết quả: <Who, publish, “Active Database Systems”>

* Bước 4: Nhận diện thực thể

Đầu vào: Các bộ 3 quan hệ chưa được nhận diện ở bước 3.

<Who, publish, “Active Database Systems”>

Đầu ra: Bộ ba quan hệ đã được nhận diện.

Xử lý: Từ quan hệ “publish” sẽ được ánh xạ vào danh sách những bộ ba rút được từ file cấu hình ngữ nghĩa (những bộ ba này xem ở phụ lục E). Lúc này sẽ tìm thấy 1 bộ ba duy nhất:

(Publication, publish, publisher). (Quan hệ ngược trong tập tin XML cấu hình ngữ nghĩa <reverse-relation>)

Vì bộ ba duy nhất tìm thấy ở quan hệ ngược (reverse-relation) nên ta sẽ hoán vị các thực thể trong bộ ba <Who, publish, “Active Database Systems”> rồi gán loại thực thể tương ứng với bộ ba duy nhất tìm được:

<“Active Database Systems”:Publication, publish, Who: publisher>

Kết quả:

<“Active Database Systems”:Publication, publish, publisher>

* Bước 5: Sinh câu truy vấn  
  Đầu vào: Danh sách các bộ 3 quan hệ đã nhận diện ở bước 4.

<“Active Database Systems”:Publication, publish, publisher>

Đầu ra: Câu truy vấn SQL

*Đối tượng truy vấn :*Select `Publication`.`publisher`

*Nguồn truy vấn:*From `dblp\_pub\_new` `Publication`

*Điều kiện:* Where `Publication`.title = “Active Database Systems”

*Câu truy vấn cuối cùng:*

Select `Publication`.`publisher`

From `dblp\_pub\_new` `Publication`

Where `Publication`.title = “Active Database Systems”

***Ví dụ 2:*** Câu hỏi *Which books were written by Rafiul Ahad and Amelia Carlson in 2010 ?*

* Bước 1: Gán nhãn từ loại

Đầu vào : Câu hỏi.

Đầu ra: Danh sách các từ đã gán nhãn từ loại.

Kết quả:

WDT/Which books/NNS were/VBD written/VBN by/IN Rafiul Ahad/NNP and/CC Amelia Carlson/NNP in/IN 2011/CD

* Bước 2: Đơn giản hóa danh từ, động từ, cụm động từ

Đầu vào: Danh sách nhãn từ loại ở bước 1

WDT/Which books/NNS were/VBD written/VBN by/IN Rafiul Ahad/NNP and/CC Amelia Carlson/NNP in/IN 2011/CD

Đầu ra: Danh sách nhãn từ loại đã được đơn giản hóa.

Kết quả:

WP/Which book/NN be/VB write/VB by/IN Rafiul Ahad/NNP and/CC Amelia Carlson/NNP in/IN 2011/CD

* Bước 3: Rút bộ ba

Đầu vào: danh sách từ đã gán nhãn từ loại ở bước 2

WP/Which book/NN be/VB write/VB by/IN Rafiul Ahad/NNP and/CC Amelia Carlson/NNP in/IN 2011/CD

Đầu ra: Tập các bộ ba quan hệ.

Xử lý:

Danh sách các từ được gán nhãn đầu tiên sẽ khớp với luật 8 trong tập luật (Xem thêm ở phụ lục C):

(WP|WDT|NNP|NN)-1 VB-2 VB-3 IN-4 (NN|NNP|CD)-5-NA --> (WP|WDT|NNP|NN)-1,VB-2 VB-3 IN-4,(NN|NNP|CD)-5 -->DEL

Nên ta rút được 2 bộ ba:

<book, be write by, Rafiul Ahad>

<book, be write by, Amelia Carlson >

Đồng thời, trong luật có từ “DEL”, do đó sẽ xóa bỏ tạm thời các từ từ sau danh từ đầu tiên là book. Do đó, danh sách tạm các từ được gán nhãn từ loại là:

WP/Which book/NN in/IN 2011/CD

Tiếp tục nó lại so khớp với luật thứ 17 trong tập lục ở phụ lục C:

(NNP|NN)-1 (IN|TO)-2 (NN|NNP|CD)-3-NA --> (WP|WDT|NNP|NN)-1,(IN|TO)-2,(NN|NNP|CD)-3-NA -->DEL

Lúc này nó sẽ rút ra được 1 bộ ba nữa là:

<book,in, 2011>

Tiếp tục nó sẽ xóa các từ được gán nhãn từ loại từ sau danh từ đầu tiên và trở thành danh sách tạm thời sau:

WP/Which book/NN

Danh sách này không còn phụ hợp với luật nào khác, do đó kết quả trả về sẽ là ba bộ ba vừa rút được.

Kết quả:

<book, be write by, Rafiul Ahad>

<book, be write by, Amelia Carlson >

<book,in, 2011>

* Bước 4: Nhận diện thực thể

Đầu vào: Các bộ 3 quan hệ chưa được nhận diện ở bước 3.

<book, be write by, Rafiul Ahad>

<book, be write by, Amelia Carlson >

<book,in, 2011>

Đầu ra:Các bộ ba quan hệ đã được nhận diện.

Xử lý:

Với bộ ba: <book, be write by, Rafiul Ahad>

Từ quan hệ “be write by” sẽ được ánh xạ vào danh sách những bộ ba rút được từ file cấu hình ngữ nghĩa (những bộ ba này xem ở phụ lục E). Lúc này sẽ tìm thấy 1 bộ ba duy nhất:

(Publication, be write by, author).

Vì bộ ba tìm được là duy nhất nên ta gán loại thực thể cho bộ ba chưa nhận diện tương ứng với bộ ba duy nhất tìm được:

<book: Publication, be write by, Rafiul Ahad: author>

Tương tự, với bộ ba <book, be write by, Amelia Carlson > ta sẽ có bộ ba được nhận diện như sau:

<book: Publication, be write by, Amelia Carlson: author>

Với bộ ba chưa nhận diện <book, in, year>, từ quan hệ “in” sẽ được ánh xạ vào danh sách những bộ ba rút được từ file cấu hình ngữ nghĩa (những bộ ba này xem ở phụ lục E). Lúc này sẽ tìm thấy 3 bộ ba:

R={

<Publication,in, year>,

<Publication,in, source>,

<Publication,in, publisher>,

}

Lúc này, ta nhận diện thực thể “book”, vì thực thể “book” đã được nhận diện là “Publication” ở 2 bộ ba trên, nên ta không cần phải nhận diện thực thể “book” lần nữa. Ta đem so sánh loại thực thể “Publication” so sánh với các loại thực thể đầu tiên ở tập R. Lúc này, ta sẽ rút gọn tập R thành

R={

<Publication,in, year>,

<Publication,in, source>,

<Publication,in, publisher>,

}.

Tiếp theo, ta nhận diện thực thể 2001. Ta đem giá trị 2011 đem tra vào các thuộc tính year, source, publisher. Vì 2001 là kiểu số nên ta tra vào các thuộc tính có kiểu số (Integer) trong cơ sở dữ liệu (chỉ có year thỏa mãn), ta suy ra được 2001 có loại thực thể là year. Tiếp tục, ta đem loại thực thể year đem so sánh với các loại thực thể thứ 2 trong tập R. Ta rút được tập R chỉ có 1 bộ ba duy nhất:

R={

<Publication, in, year>

}.

Vì chỉ có 1 bộ ba duy nhất trong tập R, nên ta gán loại thực thể ở bộ ba duy nhất này cho bộ ba cần nhận diện <book, in, 2001>. Ta có bộ ba đã được nhận diện như sau:

<book:Publication, in, 2001:year>

Kết quả:

<book: Publication, be write by, Rafiul Ahad: author>

<book: Publication, be write by, Amelia Carlson: author>

<book:Publication, in, 2001:year>

* Bước 5: Sinh câu truy vấn  
  Đầu vào: Danh sách các bộ 3 quan hệ đã nhận diện ở bước 4.

<book: Publication, be write by, Rafiul Ahad: author>

<book: Publication, be write by, Amelia Carlson: author>

<book:Publication, in, 2001:year>

Đầu ra: Câu truy vấn MySQL

*Đối tượng truy vấn :*Select `Publication`.\*

*Nguồn truy vấn:*

Vì *Rafiul Ahad* và *Amelia Carlson* là 2 bộ ba có từ quan hệ và loại thực thể giống nhau, chỉ khác nhau ở giá trị. Nên nguồn truy vấn của nó là một câu truy vấn con:

FROM

(SELECT `Publication`.\*

FROM `dblp\_pub\_new` `Publication`,`dblp\_author\_ref\_new` `Author`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Author`.`author` = ‘*Rafiul Ahad’* )) Publication

INNER JOIN

(SELECT `Publication`.\*

FROM `dblp\_pub\_new` `Publication`,`dblp\_author\_ref\_new` `Author`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Author`.`author` = ‘*Amelia Carlson’* )) Publication1 ON Publication.id = Publication1.id

*Điều kiện:* (`Publication`.`year` = 1999 )

*Câu truy vấn:*

SELECT DISTINCT `Publication`.\*

FROM

(SELECT `Publication`.\*

FROM `dblp\_pub\_new` `Publication`,`dblp\_author\_ref\_new` `Author`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Author`.`author` = ‘*Rafiul Ahad’* )) Publication

INNER JOIN

(SELECT `Publication`.\*

FROM `dblp\_pub\_new` `Publication`,`dblp\_author\_ref\_new` `Author`

WHERE

`Publication`.`id` = `Author`.`pub\_id` AND (`Author`.`author` = ‘*Amelia Carlson’* )) Publication1 ON Publication.id = Publication1.id

WHERE

(`Publication`.`year` = 1999 )

# CHƯƠNG 4: HIỆN THỰC HỆ THỐNG

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.



Hình 4.1 – Mô hình hệ thống tìm kiếm bài báo

Chương trình hệ thống gồm hai luồng chính:

-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 |
| Tìm kiếm  hỏi đáp | Câu hỏi dạng ngôn ngữ tự nhiên | Các bài báo liên quan đến nội dung câu hỏi |

Bảng 2 – bảng các chức năng trong hệ thống

* 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, chương 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. Các kết quả nhóm theo từng năm, giao diện hiển thị các bài báo mà tác giả đã viết trong mỗi năm một cách trực quan ( tương tự DBLP).
* Đa ngôn ngữ: Hỗ trợ giao diện hiển thị đa ngôn ngữ tiếng Anh-Tiếng Việt
* Chức năng hỏi đáp: Đây là một chức năng mới của chương trình. Người dùng nhập vào một câu hỏi tự nhiên duy nhất để tìm kiếm. 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. Ngoài ra, người dùng được cung cấp một số câu hỏi mẫu có sẵn trên giao diện chức năng.

Ví dụ:

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

-Kết quả:

|  |
| --- |
| 1. [Experiments on Multi-Strategy Learning by Meta-Learning](http://localhost:8080/QAWeb/showPubDetail.do?id=107298)  2. [Learning Useful System Call Attributes for Anomaly Detection](http://localhost:8080/QAWeb/showPubDetail.do?id=234312)  3. [Incrementally Learning Rules for Anomaly Detection](http://localhost:8080/QAWeb/showPubDetail.do?id=235040)  …  36. [Systems for Knowledge Discovery in Databases](http://localhost:8080/QAWeb/showPubDetail.do?id=2138381) |

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

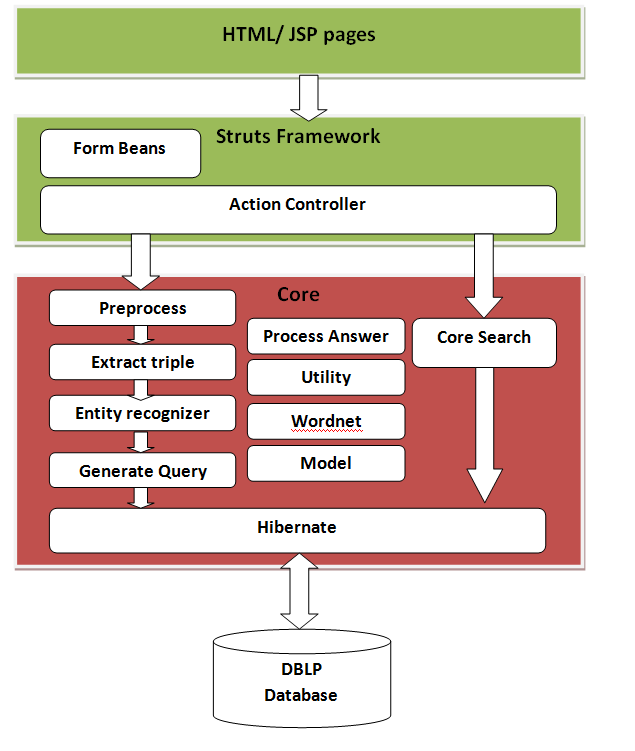
-Kết quả:

[*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**
2. **Cấu trúc các lớp**

* ***com.myapp.struts:*** gói chứa các lớp ngôn ngữ (gồm tiếng Anh và tiếng Việt).
* ***uit.qadbpss.core.search:*** nhân chứa các lớp xử lý tìm kiếm theo từ khóa, lớp thao tác với Wordnet.
* ***uit.qadbpss.core.wordnet:*** nhân chứa các lớp thao tác với bộ từ điển Wordnet và lớp thao tác các so sánh độ tương đồng ngữ nghĩa.
* ***uit.qadbpss.dbconfig:*** gói chứa các lớp cấu hình, lớp đại diện cho các thuộc tính trong bảng của cơ sở dữ liệu, lớp thao tác với file cấu hình XML.
* ***uit.qadbpss.model:*** lớp đối tượng đại diện cho một đối tượng dưới cơ sở dữ liệu ( dùng để truy xuất Hibernate).
* ***uit.qadbpss.preprocess:*** các lớp tiền xử lý câu hỏi của người dùng.
* ***uit.qadbpss.extracttriple:*** các lớp hỗ trợ rút các bộ ba trong câu hỏi.
* ***uit.qadbpss.entityrecog:*** Lớp nhận diện các đối tượng trong bộ ba.
* ***uit.qadbpss.generatequery:*** Lớp tạo câu truy vấn SQL từ các bộ ba rút ra được từ các lớp trên.
* ***uit.qadbpss.util:*** Lớp toàn cục cho cả hệ thống*.*Các lớp về kho ký tự CharPool, String Pool.



Hình 4.2– Cấu trúc các lớp trong hệ thống

1. **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): bảng chứa thông tin các ấn bản (Publication) trong DBLP, chứa các thông tin như tiêu đề, năm, nhà sản xuất, ...**

|  |  |  |
| --- | --- | --- |
| **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): Bảng lưu thông tin các tác giả và pub\_id (Publication ID) của ấn bản mà tác giả đó đã viết.**

|  |  |  |
| --- | --- | --- |
| **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): Bảng lưu thông tin các tài liệu tham khảo của một ấn bản (Publication).**

|  |  |  |
| --- | --- | --- |
| **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**

****

* **Giao diện chạy thực nghiệm**

**Chạy thực nghiệm với câu:** *Who is the author of "Working Models for Uncertain Data"*

**Kết quả:**

****

**Chạy thực nghiệm với câu:** *Which books were written by Rafiul Ahad and Amelia Carlson ?*

**Kết quả:**

****

**Chạy thực nghiệm với câu:** *Which papers cited to "Foundations of Databases"?*

**Kết quả:**

**

1. **Đánh giá và nhận xét**

Với phương pháp xử lý câu hỏi đề xuất trên,việc xử lý khá dễ dàng do thao tác trên danh sách từ loại. Việc bổ sung các mẫu luật từ loại khá đơn giản và số lượng luật không quá lớn (18 luật).

Về cách nhận diện thực thể trong bộ ba, phương pháp so sánh bộ ba rút được với các bộ ba trong tập tin cấu hình sẽ ít bị ràng buộc theo các phương pháp khác [1, 3, 5, 6], vì thông thường ta cần phải có một công cụ như GATE để nhận diện các thực thể đặt tên. Đối với một kho dữ liệu xác định (closed-domain) về các bài báo khoa học về công nghệ thông tin, thì GATE không thể nhận diện hết được các thực thể đặt tên, do chúng chỉ nhận diện được các đối tượng thông dụng như (tên người, địa điểm, ngày tháng ...). Tuy nhiên, trong trường hợp xấu nhất, hệ thống ánh xạ trực tiếp giá trị của thực thể vào trong cơ sở dữ liệu để nhận diện loại thực thể. Thao tác này chiếm thời gian nhiều trong quá trình xử lý câu hỏi.

Tập kiếm thử gồm 102 câu dựa tên thông tin hiện có của cơ sở dữ liệu, nhu cầu tìm kiếm và một số dạng câu hỏi trong TREC. Độ chính xác khi rút bộ ba được trình bày ở bảng sau:

|  |  |
| --- | --- |
|  | Khả năng rút trích bộ ba |
| Độ chính xác | 96% |
| Số câu hỏi rút được bộ ba | 99 |
| Số câu hỏi rút đúng | 98 |
| Số câu hỏi rút sai | 3 |
| Số câu hỏi không rút được bộ ba | 1 |
| Tổng số câu | 102 |

Bảng 3 – Đánh giá khả năng rút trích bộ ba

Với kết quả này:

+ Các trường hợp không rút được bộ ba và rút thiếu chứng tỏ tập 19 mẫu luật từ loại vẫn chưa hoàn chỉnh và bao quát hết được tập mẫu. Điều này có thể giải quyết bằng cách bổ sung và chỉnh sửa lại luật.

+ Với trường hợp bộ ba bị rút sai, nguyên nhân do một số nhập nhằng về ngữ nghĩa (từ quan hệ không nhất thiết phải nằm giữa các đối tượng) khiến bộ ba rút ra không đúng với ngữ nghĩa.Vấn đề này hiện vẫn chưa giải quyết được và đang tìm hướng khắc phục.

+ Một số trường hợp rút sai do bộ gán nhãn từ loại không phân biệt được giữa động từ và danh từ tùy theo ngữ cảnh (ví dụ từ *“work”* vừa là động từ, vừa là danh từ. Cho nên trong câu *“Which publisher does Marcus Thint work with”* thì từ *“work”* sẽ bị gán nhãn sai là danh từ thay vì động từ).

# 

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

# 5.1 Kết luận

Hệ thống được chúng em xây dựng nên nhằm thỏa mãn một phần nào nhu cầu của người dùng trong vấn đề tìm kiếm các bài báo khoa học. Cách thức hỏi đáp được đưa vào để khai thác nâng cao hiệu quả trong vấn đề tìm kiếm. Điều này còn khá xa lạ với người dùng hiện nay.

Với sự phát triển của công nghệ thông tin hiện nay, việc sử dụng một câu hỏi dưới dạng ngôn ngữ tự nhiên để tìm kiếm một thông tin hay tài liệu nào đó có thê sẽ là một phương pháp tìm kiếm trong tương tai, giúp cho khoảng cách giữa con người và máy tính ngày càng gần nhau hơn.

# 5.2 Hướng phát triển

Hướng đi tiếp theo của hệ thống sẽ phát triển thêm các mô-đun:

+ Gợi ý các câu hỏi tương tự cho người dùng, mô-đun đưa ra một số hướng xử lý khi người dùng nhập không chính xác tên tác giả, tổ chức hay tiêu đề (với trường hợp như thế có thể chỉnh sửa lại hoặc gợi ý giá trị đúng cho người dùng), và

+ Mô-đun có thể chỉnh lỗi chính tả khi người dùng nhập sai chính tả, vì đôi khi người dùng có thể nhập sai một từ nào đó trong câu hỏi.

Đề tài có dự định áp dụng phương pháp hỏi đáp đề xuất trên cho một cơ sở dữ liệu khác là Ontology – một cơ sở dữ liệu đã tổ chức theo ngữ nghĩa.

Ngoài ra, đề tài còn hướng tới bổ sung một số câu hỏi kiểm thử dựa trên tham khảo từ một số nguồn câu hỏi khác và từ người sử dụng, để có được một nguồn kiểm thử phong phú, đa dạng.

# TÀI LIỆU THAM KHẢO

## Tiếng Việt

1. *Lương Quý Tịnh Hà****, 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****,**luận văn Thạc Sĩ, trường Đại học Công Nghệ Thông Tin, Tp. HCM, 2009.*
2. *Cao Duy Trường,* ***Dịch câu truy vấn có cấu trúc sang đồ thị ý niệm: cách tiếp cận ít phụ thuộc vào cú pháp****, luận văn Thạc Sĩ,**trường Đại học Bách Khoa,**Tp.HCM****,*** *2008****.***
3. *Đỗ Thị Thanh Tuyền,* ***Xây dựng hệ thống tra cứu thư viên điện tử bằng ngôn ngữ tự nhiên****, luận văn Thạc Sĩ Công Nghệ Thông Tin, Tp. HCM, 2008.*
4. *TS.Đinh Điền,* ***Giáo trình Xử lý ngôn ngữ tự nhiên****, NXB Đại học quốc gia TP HCM, 2006.*
5. *Huỳnh Ngọc Tín,* ***Báo cáo chuyên đề rút trích thông tin****, Đại Học Công Nghệ Thông Tin, 2010.*

## Tiếng Anh

1. *Lorand Dali, Delia Rusu, Blaz Fortuna, Dunja Mladenic and Marko Grobelnik,* ***Question Answering Based on Semantic Grahps****, Department of Knowledge Technologies, Jožef Stefan Institute, 2009.*
2. *Meltem Serhatli and Ferda N. Alpaslan,* ***An Ontology based Question Answering System on Software Test Document Domain****, World Academy of Science, Engineering and Technology 54 2009.*
3. *Wael Salloum,* ***A Question Answering System based on Conceptual Graph Formalism****, Conference: The 2nd International Symposium on Knowledge Acquisition and Modeling (KAM 2009), IEEE Computer Society Press, 2009.*
4. *DongHyun Choi and Key-Sun Choi,* ***Automatic Relation Triple extraction by dependency parse tree ,*** *Computer Science Department**Semantic Web Research Center, KAIST Daejeon, Korea, 2008.*
5. *Shiyan Ou, Constantin Orasan, Dalila Mekhaldi and Laura Hasler,* ***Automatic Question Pattern Generation for Ontology-based Question****. In Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS2008). Menlo Park, CA: AAAI Press., 2008, pp. 183 – 188.*
6. *Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, Dunja Mladenic ,* ***Triplet Extraction from Sentences****, Department of Knowledge Technologies, Jožef Stefan Institute, 2007.*
7. [*Stephan Bloehdorn*](http://www.springerlink.com/content/?Author=Stephan+Bloehdorn)*,* [*Philipp Cimiano*](http://www.springerlink.com/content/?Author=Philipp+Cimiano)*,* [*Alistair Duke*](http://www.springerlink.com/content/?Author=Alistair+Duke)*,* [*Peter Haase*](http://www.springerlink.com/content/?Author=Peter+Haase)*,* [*Jörg Heizmann*](http://www.springerlink.com/content/?Author=J%c3%b6rg+Heizmann)*,* [*Ian Thurlow*](http://www.springerlink.com/content/?Author=Ian+Thurlow) *and* [*Johanna Völker*](http://www.springerlink.com/content/?Author=Johanna+V%c3%b6lker)*,* ***Ontology-Based Question Answering for Digital Libraries,*** [Lecture Notes in Computer Science](http://www.springerlink.com/content/0302-9743/) , Volume 4675/2007.
8. *Kenneth C. Litkowski,* ***Question-Answering using Semantic Relation Triples*** *, In Proceedings of the 8th Text Retrieval Conference (TREC-8), 1999.*

## Trang web

1. Digital Bibliography & Library Project

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

1. Wiki 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 Recognition (NER)

<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=1>

1. Conceptual graph

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

1. Wordnet

<http://wordnet.princeton.edu/>

1. Pos tagging

<http://en.wikipedia.org/wiki/Part-of-speech_tagging>

1. Định nghĩa Question answering

<http://www.wordiq.com/definition/Question_answering>

1. Định nghĩa Question answering

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

1. Triples

[**http://syntopika.wordpress.com/2010/08/02/n-triples-example/**](http://syntopika.wordpress.com/2010/08/02/n-triples-example/)

# Phụ lục A : Tập 102 câu hỏi kiểm thử

|  |
| --- |
| 1. Who is the author of the book "Question Classification using Head Words and their Hypernyms"? 2. Who is the author of "Working Models for Uncertain Data" 3. Who is the author of the paper "Question Classification using Head Words and their Hypernyms"? 4. Who is the composer of the book "Question Classification using Head Words and their Hypernyms"? 5. Who is the creator of the book "Question Classification using Head Words and their Hypernyms"? 6. Who write books in 1999 ? 7. Who composed books in 2010? 8. Who write books from 1999 to 2010 ? 9. Who published books from 1999 to 2000 ? 10. Who published books in 1999 ? 11. Who released books in 2006 ? 12. Who wrote books for Springer in 2010 ? 13. Who wrote papers for ACM from 2009 to 2010 ? 14. Who wrote "Question Classification using Head Words and their Hypernyms"? 15. Who published "Question Classification using Head Words and their Hypernyms"? 16. Who is the publisher of book "Working Models for Uncertain Data"? 17. Who is the distributor of publication "Working Models for Uncertain Data"? 18. Which books were written by Rafiul Ahad and Amelia Carlson in 2010 ? 19. Which books were written by Rafiul Ahad from 1999 to 2010 ? 20. Which papers were written by Amelia Carlson from 1999? 21. Which books were published by O'Reilly or ACM in 1999 ? 22. Which publications were published by O'Reilly or ACM in 1999 ? 23. Which books were published by O'Reilly in 1999 ? 24. Which books were published by Springer from 1999 to 2000? 25. Which publications were published by O'Reilly or ACM from 2009 to 2010? 26. Which books did Richard L. Muller or James Brown write in ACM ? 27. Which books did Richard L. Muller write in ACM ? 28. Which authors write books for ACM? 29. Which authors wrote books for Springer from 1999? 30. Which authors composed books for O'Reilly ? 31. Which authors composed books in 1999 ? 32. Which authors write books in 2010 ? 33. Which books does Rafiul Ahad wrote with Amelia Carlson? 34. Which publication have source from TREC in 2010? 35. Which papers refer to "Foundations of Databases"? 36. Which publications refer to "Foundations of Databases"? 37. Which documents cited to "Foundations of Databases"? 38. Which books are published in 1999 and written by "Ken Arnold"? 39. Which books are written by "Ken Arnold" and published by IEEE ? 40. Which books are written by "Ken Arnold" and have source from DBPL ? 41. Which papers are composed by "Ken Arnold" and have source from DBPL ? 42. Which publishers released books in 2010 ? 43. Which publishers released book "Data Structures and Algorithms"? 44. Which publishers released book from 2009 to 2010? 45. How many papers were written by Rafiul Ahad ? 46. How many papers were written by Rafiul Ahad in 2010 ? 47. How many publisher did "Philip K. Chan" cooperate with? 48. How many publisher did "Philip K. Chan" work with? 49. How many papers were written by Raul Rato and Rui Gomes in 2010? 50. How many books have source from TREC ? 51. How many papers derived source from TREC ? 52. How many authors wrote book in 2009 ? 53. How many publisher released book in 2010? 54. How many publications refer to "Foundations of Databases"? 55. How many books cited to "Foundations of Databases"? 56. What are titles of books written by Marcus Thint ? 57. What are headings of books written by Marcus Thint ? 58. What are names of books written by William Kent ? 59. What papers did Jennifer Widom write ? 60. What books did Jennifer Widom write ? 61. What book did Philip K. Chan write in 1999 ? 62. What book did Philip K. Chan write from 1999 to 2000? 63. What book did James Brown write in 1999 from ACM? 64. What are the titles of the books published by O’reilly in 1999 ? 65. What composer wrote "Java 2D Graphics" 66. What books has isbn 1-56592-484-3 67. What books has doi 10.1145/360271.360274 68. What composer wrote books from 1999 in ACM? 69. What books were written by "Philip K. Chan" from ACM? 70. What books were published by ACM or Springer in 2010? 71. What publications have resulted from TREC in 1999? 72. What books refer to "Foundations of Databases" 73. What publications have resulted from TREC? 74. What year is "Foundations of Databases" written in? 75. What year is "Java 2D Graphics" written in? 76. What references of the book "Foundations of Databases" ? 77. What citations of the book "Foundations of Databases" ? 78. Does Rafiul Ahad write books in 2010 ? 79. Does Rafiul Ahad write books from 2005 to 2010 ? 80. Did you find me some books publish by O'Reilly ? 81. Do you know any books written in 2010 and publish by ACM ? 82. Did IEEE publish books in 2010 ? 83. Did you have any books were publish by ACM ? 84. Are there any books published by IEEE in 2000 ? 85. Are there any books written by Mike Holcombe in 2000 ? 86. Name some books Richard L. Muller write for Springer 87. Name all publications were publish in 2010 88. Name papers which were published from 1999 to 2001 in ACM ? 89. Name publications which result from DBPL in 2010 ? 90. List some books which are written by Jennifer Widom from 1999 91. List all books were published by Springer in 2010 92. List some books were published by ACM 93. List authors Who wrote books for Springer in 2010 94. Find books which are composed by Rafiul Ahad and Amelia Carlson 95. Find books are composed by Rafiul Ahad and Amelia Carlson 96. Find authors who wrote books for ACM from 2008 to 2009 97. Find publications have resulted from ACM Annual Conference 98. Show me some books were written by Lorand Dali 99. Show me any books which published in ACM and written by Lorand Dali. 100. Alfred V. Aho is author of "Data Structures and Algorithms" , isn't it? |

# 

# Phụ lục B : Danh sách các nhãn từ loại Penn Treebank

(<http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html> )

|  |  |  |  |
| --- | --- | --- | --- |
| Số TT | Nhãn | Mô tả | Ví dụ |
| 1.  2.  3.  4.  5.  6.  7.  8.  9.  10.  11.  12.  13.  14.  15.  16.  17.  18.  19.  20.  21.  22.  23.  24.  25.  26.  27.  28.  29.  30.  31.  32.  33.  34.  35.  36. | CC  CD  DT  EX  FW  IN  JJ  JJR  JJS  LS  MD  NN  NNS  NNP  NNPS  PDT  POS  PRP  PRP$  RB  RBR  RBS  RP  SYM  TO  UH  VB  VBD  VBG  VBN  VBP  VBZ  WDT  WP  WP$  WRB | Liên từ (Coordinating conjunction )  Số từ (Cardinal number)  Định từ (Determiner)  “có” (Existential *there)*  Từ nước ngoài (Foreign word)  Giới từ (Preposition or subordinating conjunction)  Tính từ (Adjective)  Tính từ so sánh hơn (Adjective, comparative)  Tính từ so sánh hơn cực cấp (Adjective, superlative)  Dấu liệt kê (List item marker)  Từ hình thái (Modal)  Danh từ số ít (Noun, singular or mass)  Danh từ số nhiều (Noun, plural)  Danh từ riêng số ít (Proper noun, singular)  Danh từ riêng số nhiều (Proper noun, plural)  Tiền chỉ định từ (Predeterminer)  Sở hữu cách (Possessive ending)  Đại từ nhân xưng (Personal pronoun)  Đại từ sở hữu (Possessive pronoun)  Trạng từ (Adverb)  Trạng từ so sánh hơn (Adverb, comparative)  Trạng từ so sánh cực cấp (Adverb, superlative)  Tiểu từ (Particle)  Ký hiệu (Symbol)  *Từ to*  Thán từ (Interjection)  Động từ nguyên thể (Verb, base form)  Động từ quá khứ (Verb, past tense)  Phân từ (Verb, gerund or present participle)  Quá khứ phân từ (Verb, past participle)  Động từ không phải ngôi ba số ít hiện tại  (Verb, non-3rd person singular present)  Động từ ngôi ba số ít hiện tại  (Verb, 3rd person singular present)  Định từ bắt đầu bằng Wh- (Wh-determiner)  Đại từ bắt đầu bằng Wh- (Wh-pronoun)  Đại từ sở hữu bắt đầu bằng Wh-  (Possessive wh-pronoun)  Trạng từ bắt đầu bằng Wh- (Wh-adverb) | and, or, but  1, 2, ..  the, a, an...  there  in, on , at...  big, good, hard  bigger, better ..  biggest, best...  :  can, may, might ..  book, sugar,...  books, children  John, Hanoi..  IBMs, Fords, ...  This, each...  ‘s  I, you, he ...  mine, yours ...  slow, hardly ...  slower, faster ...  slowest, fastest..  on,off  < [ | > /  to  oh !  work, write...  worked, wrote ..  working, writing  written, stolen ...  work,write ...  works, writes ...  which, what, ..  who, where ...  whose, ...  when , where ... |

# Phụ lục C : Tập 19 luật từ loại

|  |
| --- |
| 1. (NN|NNP)-1 VB-2 (NN)-3 NNP-4-NA VB-5 --> NNP-4-NA,VB-4,(NN|NNP)-1 -->DEL 2. (NN)-1 WDT-2 (NN|NNP)-3-NA VB-4 -->(NN|NNP)-3-NA,VB-4,(NN)-1 -->DEL 3. (NN)-1 WDT-2 VB-3 VB-4 IN-5 (NN|NNP|CD)-6-NA --> NN-1,VB-3 VB-4 IN-5,(NN|NNP|CD)-6 -->DEL 4. (NN|NNP)-1 VB-2 (NN|NNP)-3-NA VB-4 --> (NN|NNP)-3-NA,VB-4,(NN|NNP)-1 -->DEL 5. (WP|WDT|NNP|NN)-1 VB-2 VB-3 IN-4 (NN)-5 (NNP|CD)-6-NA --> (WP|WDT|NNP|NN)-1,VB-2 VB-3 IN-4,(NNP|CD)-6 -->DEL 6. (WP|WDT|NNP|NN)-1 VB-2 VB-3 IN-4 (NN|NNP|CD)-5-NA --> (WP|WDT|NNP|NN)-1,VB-2 VB-3 IN-4,(NN|NNP|CD)-5 -->DEL 7. (WP|WDT|NNP|NN)-1 (VBN)-2 IN-3 (NN|NNP|CD)-4-NA --> (WP|WDT|NNP|NN)-1,VB-2 IN-3,(NN|NNP|CD)-4 -->DEL 8. (NN)-1 NNP-2-NA VB-3 --> NNP-2-NA,VB-3,(NN)-1 -->DEL 9. (WP|WDT|NNP|NN)-1 (VB)-2 (NN)-3 NNP-4-NA --> (WP|WDT|NNP|NN)-1,VB-2,NNP-4 10. (WP|WDT|NNP|NN)-1 (VB)-2 (NN|NNP)-3-NA --> (WP|WDT|NNP|NN)-1,VB-2,(NN)-3 11. (WP|WDT|NNP|NN)-1 (VB)-2 IN-3 (NN)-4 NNP-5-NA --> (WP|WDT|NNP|NN)-1,VB-2,NNP-5 12. (WP|WDT|NNP|NN)-1 (VB)-2 (IN|TO)-3 (NN|NNP)-4-NA --> (WP|WDT|NNP|NN)-1,VB-2,(NN|NNP)-4 13. (NNP|NN)-1 IN-2 (NN)-3 (NNP|CD)-4-NA --> (WP|WDT|NNP|NN)-1,IN-2,(NNP|CD)-4-NA 14. (NNP|NN)-1 IN-2 (NN|NNP|CD)-3-NA --> (WP|WDT|NNP|NN)-1,IN-2,(NN|NNP|CD)-3-NA -->DEL 15. (NNP|NN)-1 TO-2 (NN|NNP|CD)-3-NA --> (WP|WDT|NNP|NN)-1,TO-2,(NN|NNP|CD)-3-NA -->DEL 16. (NNP|NN)-1 IN-2 (NN|NNP)-3 IN-4 (NN|NNP|CD)-5-NA --> (NNP|NN)-3,IN-4,(NN|NNP|CD)-5-NA 17. (NNP|NN)-1 IN-2 CD-3 IN-4 (NN|NNP)-5-NA --> (NNP|NN)-1,IN-4,(NN|NNP|CD)-5-NA 18. VB-1 (NN|NNP)-2 CD-3 --> (NN|NNP)-2,has,CD-3 19. (WP)-1 (NN)-2 VB-3 NNP-4 VBN-5 IN-6 --> NNP-4,VB-5 IN-6,(NN)-2-->DEL |

# Phụ lục D : Tập tin cấu hình ngữ nghĩa cho cơ sở dữ liệu BDBLP

|  |
| --- |
| <?xml version="1.0" encoding="utf-8"?>  <database xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="config.xsd">  <table table-alias="Publication">  <table-name>dblp\_pub\_new</table-name>  <field type="String" field-alias="author" visible="false" relation="related table">  <relation>  <relation-name>have author</relation-name>  <relation-name>have writer</relation-name>  <relation-name>have composer</relation-name>  <relation-name>be write by</relation-name>  <relation-name>be compose by</relation-name>  <related-table>author</related-table>  <mapping-table>pub\_au</mapping-table>  </relation>  </field>    <field type="String" field-alias="title" visible="true" relation="has attribute" presentation="true">  <relation>  <relation-name>have title</relation-name>  <relation-name>have heading</relation-name>  </relation>  <field-name>title</field-name>  </field>  <field type="String" field-alias="publisher" visible="true" relation="has attribute">  <relation>  <relation-name>have publisher</relation-name>  <relation-name>be publish by</relation-name>  <relation-name>be release by</relation-name>  <relation-name>for</relation-name>  <relation-name>from</relation-name>  <relation-name>in</relation-name>  <reversed-relation>publish</reversed-relation>  <reversed-relation>have publish</reversed-relation>  <reversed-relation>release</reversed-relation>  <reversed-relation>have release</reversed-relation>  </relation>  <field-name>publisher</field-name>  </field>    <field type="String" field-alias="source" visible="true" relation="has attribute">  <relation>  <relation-name>have source</relation-name>  <relation-name>have source from</relation-name>  <relation-name>have source in</relation-name>  <relation-name>from</relation-name>  <relation-name>in</relation-name>  <relation-name>result from</relation-name>  </relation>  <field-name>source</field-name>  </field>    <field type="Int" field-alias="year" visible="true" relation="has attribute">  <relation>  <relation-name value="&gt;=">have year from</relation-name>  <relation-name>have year in</relation-name>  <relation-name>in</relation-name>  <relation-name value="&gt;=">from</relation-name>  <relation-name value="&lt;=">to</relation-name>  <relation-name>be write in</relation-name>  </relation>  <field-name>year</field-name>  </field>  <field type="String" field-alias="isbn" visible="true" relation="has attribute">  <relation>  <relation-name>have isbn</relation-name>  </relation>  <field-name>isbn</field-name>    </field>  <field type="String" field-alias="doi" visible="true" relation="has attribute">  <relation>  <relation-name>have doi</relation-name>  </relation>  <field-name>doi</field-name>    </field>    <field type="String" field-alias="reference" visible="false" relation="mapping table">  <relation>    <relation-name>cite</relation-name>  <relation-name>cite to</relation-name>  <relation-name>refer to</relation-name>  <relation-name>refer</relation-name>  <mapping-table>pub\_ref</mapping-table>  <related-table>reference</related-table>  </relation>  </field>  <primary-key>id</primary-key>    </table>  <table table-alias="Author">  <table-name>dblp\_author\_ref\_new</table-name>  <field type="String" field-alias="publication" visible="false" relation="related table">  <relation>  <relation-name>write</relation-name>  <related-table>publication</related-table>  <mapping-table>au\_pub</mapping-table>  </relation>  </field>  <field type="String" field-alias="publisher" visible="false" relation="related table">  <relation>  <relation-name>write for</relation-name>  <relation-name>cooperate</relation-name>  <related-table>publication</related-table>  <mapping-table>au\_pub</mapping-table>  </relation>  <field-name>publisher</field-name>  </field>  <field type="String" field-alias="name" visible="true" relation="has attribute" presentation="true">  <relation>  <relation-name>have name</relation-name>  </relation>  <field-name>author</field-name>  </field>  <primary-key>author</primary-key>  </table>  <table table-alias="Reference">  <table-name>dblp\_pub\_new</table-name>  <field type="String" field-alias="title" visible="true" relation="has attribute" presentation="true">  <relation>  <relation-name>have title</relation-name>  <relation-name>have heading</relation-name>  </relation>  <field-name>title</field-name>  </field>  <primary-key>dblp\_key</primary-key>  </table>  <database-name>dblp</database-name>  <mapping-table name="ref\_pub">  <table-key>dblp\_key</table-key>  <related-table-key>id</related-table-key>  <mapping-table-name>dblp\_ref\_new</mapping-table-name>  </mapping-table>  <mapping-table name="pub\_ref">  <table-key>id</table-key>  <related-table-key>dblp\_key</related-table-key>  <mapping-table-name>dblp\_ref\_new</mapping-table-name>  </mapping-table>  <mapping-table name="pub\_au">  <table-key>pub\_id</table-key>  <related-table-key>author</related-table-key>  <mapping-table-name>dblp\_author\_ref\_new</mapping-table-name>  </mapping-table>  <mapping-table name="au\_pub">  <table-key>author</table-key>  <related-table-key>pub\_id</related-table-key>  <mapping-table-name>dblp\_author\_ref\_new</mapping-table-name>  </mapping-table>  </database> |

# Phụ lục E: Những bộ ba rút được từ tập tin cấu hình ngữ nghĩa.

Với cách thức cấu hình trong tập tin XML ở phụ lục D. Mặc định khi lấy lên thông tin ta sẽ có được các bộ ba miêu tả các mối quan hệ giữa bảng và thuộc tính như sau:

(Publication, have author, Author)

(Publication, have writer, Author)

(Publication, have composer, Author)

(Publication, be write by, Author)

(Publication, be compose by, Author)

(Publication, have title, title)

(Publication, have heading, title)

(Publication, have publisher, publisher)

(Publication, be publish by, publisher)

(Publication, be release by, publisher)

(Publication, for, publisher)

(Publication, from, publisher)

(Publication, in, publisher)

(Publication, publish, publisher) (Quan hệ ngược)

(Publication, have publish, publisher) (Quan hệ ngược)

(Publication, release, publisher) (Quan hệ ngược)

(Publication, have release, publisher) (Quan hệ ngược)

(Publication, have source, source)

(Publication, have source from, source)

(Publication, have source in, source)

(Publication, from, source)

(Publication, in, source)

(Publication, result from, source)

(Publication, have year from, year)

(Publication, have year in, year)

(Publication, in, year)

(Publication, from, year)

(Publication, to, year)

(Publication, be write in, year)

(Publication, have isbn, isbn)

(Publication, have doi, doi)

(Publication, cite, Reference)

(Publication, cite to, Reference)

(Publication, refer to, Reference)

(Publication, refer, Reference)

(Author, write, Author.Publication)

(Author, write for, Author.publisher)

(Author, cooperate, Author.publisher)

(Author, have name, Author.name)

(Reference, have title, Reference.title)

(Reference, have heading, Reference.title)

1. <http://www.bing.com/> [↑](#footnote-ref-2)
2. <http://www.yahoo.com/> [↑](#footnote-ref-3)
3. <http://www.google.com/> [↑](#footnote-ref-4)
4. <http://citeseer.ist.psu.edu/> [↑](#footnote-ref-5)
5. <http://www.ieee.org/> [↑](#footnote-ref-6)
6. <http://portal.acm.org/> [↑](#footnote-ref-7)
7. <http://www.sciencedirect.com/> [↑](#footnote-ref-8)
8. <http://www.ssrn.com/> [↑](#footnote-ref-9)
9. <http://papercube.peterbergstrom.com/> [↑](#footnote-ref-10)
10. <http://scholar.google.com/> [↑](#footnote-ref-11)
11. <http://dblp.uni-trier.de/xml/> [↑](#footnote-ref-12)
12. <http://trec.nist.gov/> [↑](#footnote-ref-13)
13. <http://nlp.stanford.edu/ner/index.shtml> [↑](#footnote-ref-14)
14. <http://gate.ac.uk/ie/annie.html> [↑](#footnote-ref-15)
15. <http://wordnet.princeton.edu/wordnet/> [↑](#footnote-ref-16)
16. <http://gate.ac.uk/sale/tao/splitch6.html#sec:annie:tagger> [↑](#footnote-ref-17)
17. <http://incubator.apache.org/opennlp/download.html> [↑](#footnote-ref-18)
18. <http://www.w3schools.com/RDF/rdf_owl.asp>, [↑](#footnote-ref-19)
19. <http://sourceforge.net/projects/wn-similarity/> [↑](#footnote-ref-20)
20. <http://ailab.ijs.si/dunja/textgarden/> [↑](#footnote-ref-21)
21. <http://www.gutenberg.org/wiki/Main_Page> [↑](#footnote-ref-22)
22. <http://nlp.stanford.edu/software/lex-parser.shtml> [↑](#footnote-ref-23)
23. <http://www.link.cs.cmu.edu/link/> [↑](#footnote-ref-24)
24. <http://webdocs.cs.ualberta.ca/~lindek/minipar.htm> [↑](#footnote-ref-25)
25. <http://opennlp.sourceforge.net/> [↑](#footnote-ref-26)
26. <http://gate.ac.uk/gate/doc/plugins.html> [↑](#footnote-ref-27)