|  |
| --- |
| ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  đẠI HỌC CÔNG NGHỆ THÔNG TIN |
| Đề Cương  Đồ Án Tốt Nghiệp |
| Đề tài : Xây dựng hệ thống hỏi đáp  dựa trên Ontology |
|  |
| |  |  | | --- | --- | | **GV :** | **Ths. Huỳnh Ngọc Tín** | | **Thời gian dự kiến :** | **3 tháng** | | **Thành viên nhóm :** | **Nguyễn Thanh Hoàng - 06520182** | |  | **Nguyễn Thuận Hưng - 06520194** | |
|  |
|  |

Mục lục

[**1. Đặt vấn đề 3**](#_Toc268297442)

[**2. Mục tiêu và phạm vi đề tài 4**](#_Toc268297443)

[**3.Các ứng dụng và nghiên cứu liên quan 5**](#_Toc268297444)

[**4. Nội dung và kế hoạch triển khai 9**](#_Toc268297445)

[**5.1 nội dung 9**](#_Toc268297446)

[**5.2 kế hoạch triển khai 11**](#_Toc268297447)

[**6. Kết quả dự kiến 12**](#_Toc268297448)

[**7. Các tài liệu tham khảo dự kiến 12**](#_Toc268297449)

[**8. Tham khảo 13**](#_Toc268297450)

# 1. Đặt vấn đề

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

Với việc tìm kiếm từ khóa như vậy, nên các công cụ chỉ đưa về các trang hoặc các văn bản liên quan có chứa từ khóa cần tìm kiếm. Chính vì thế đòi hỏi người dùng phải tối ưu hóa câu truy vấn tìm kiếm để tìm được thông tin chính xác nhất và phải chi trả thời gian cho việc tham khảo từng trang hoặc tài liệu để tìm được câu trả lời phù hợp.

Mục đích cuối cùng của người dùng là mong muốn tìm kiếm được thông tin một cách chính xác. Chính vì nhu cầu đó, nhóm quyết định chọn đề tài:

**“Xây dựng hệ thống hỏi đáp dựa trên Ontology”**

Với hệ thống hỏi đáp này, người dùng nhập vào câu hỏi dưới dạng ngôn ngử tự nhiên thay vì từ khóa. Kết quả trả về của hệ thống là câu trả lời tương ứng thay cho những trang web chứa từ khóa. Tóm lại, hệ thống sẽ thõa mãn được nhu cầu tìm kiếm thông tin của người dùng.

# 2. Mục tiêu và phạm vi đề tài

Mục tiêu của đề tài là xây dựng một hệ thống hỏi đáp dựa trên ontology . Một ontology là một mô hình dữ liệu biểu diễn một lĩnh vực và được sử dụng để suy luận về các đối tượng trong lĩnh vực đó và mối quan hệ giữa chúng [5]. Ontology này sẽ lưu trữ thông tin thuộc lĩnh vực các bài báo khoa học .Người dung nhập vào một câu hỏi bằng ngôn ngữ tự nhiên và hệ thống trả về câu trả lời. Các câu hỏi của người dùng chỉ bao gồm các vấn đề liên quan đến thông tin của bài báo khoa học : tiêu đề, tên tác giả, ngày công bố , từ khóa ,chủ đề , miêu tả tổng quan ,khái niệm có trong bài báo .

Từ đó , nhóm dự kiến sẽ xử lý một số câu hỏi như sau :

* Who is “Philip K. Chan”?
* Who is the author of “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* Who writes “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data” ?
* How many papers were written by “Philip K. Chan”?
* How many concepts in “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* How many papers have keyword “Information Extraction”?
* What are the keywords of the “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* What is the topic of “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* What concepts are in “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* What papers include this concept “Software Engineering”?
* Which papers relate to topic “Software Engineering” ?
* Which papers were written by “Philip K. Chan”?

# 3.Các ứng dụng và nghiên cứu liên quan

Hiện nay đã có một số công cụ tìm kiếm tìm kiếm ra đời thay thế bằng cách trả lời câu hỏi thay cho từ khóa như <http://start.csail.mit.edu/>, <http://www.answerbus.com> , <http://www.answers.com>, … Qua việc sử dụng, nhóm nhận thấy các công cụ này thõa mãn được thông tin tìm kiếm dựa trên việc nhập vào một câu hỏi thay cho từ khóa. Hơn thế nữa, ngoài việc đưa ra câu trả lời, một số trang còn trả về hình ảnh minh họa hoặc một đoạn video. Điển hình là trang web START, khi ta gõ vào một câu hỏi “Where is California” thì câu trả lời là “California is a state in the USA” đồng thời kèm theo một bản đồ vị trí của bang này .

Sau khi sử dụng một số công cụ, nhóm đề ra bảng so sánh chủ quan đối với một số công cụ như sau :

|  |  |  |  |
| --- | --- | --- | --- |
| Công cụ | Data base | Hình thức  tìm kiếm | Khả năng tìm kiếm |
| <answer.com> | Toàn bộ văn bản của các trang web, các câu trả lời tương ứng với câu hỏi người dùng đưa vào. | Hỏi đáp | -Đưa ra danh sách nội dung các page gần đúng nhất có chứa câu trả lời đối thay vì trả về kết quả ngắn gọn nhất (Với câu hỏi đơn giản).  -Đối với câu trả lời phức tạp, hệ thống cho người dùng nhập câu trả lời và lấy kết quả đó truy vấn lần sau.  -Khả năng hỗ trỡ video.  -Khả năng trả về trang liên quan.  \*Đánh giá: Độ chính xác cao. |
| <start.csail.mit.edu> | Các trang web | Hỏi đáp | -Trả về một số ít kết quả đúng nhất và nguồn chứa tương ứng.  -Kết quả trả vê hỗ trợ hình ảnh video  \*Đánh giá: Độ chính xác cao. |
| <answerbus.com> | Các trang web | Hỏi đáp | -Đưa về một tập các câu trả lời là các đoạn văn bản trích từ các nguồn khác nhau.  \*Đánh giá: Độ chính xác không cao. |
| [tangra.si.umich.edu/clair/NSIR/html/nsir.cgi](http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi) | Lấy những mẫu thông tin từ  Google-snippet | Hỏi đáp | -Đưa một câu trả lời ngắn gọn (1-2 từ).  \*Đánh giá: Chỉ đáp ứng cho một số câu hỏi có câu trả lời ngắn gọn . |

**Các nghiên cứu liên quan**

* Một nghiên cứu gần đây của Wael Salloum “A Question Answering System based on Conceptual Graph Formalism” năm 2009 [6]. Tác giả đưa ra hướng tiếp cận bằng cách mô hình các văn bản và các câu hỏi thành một dạng đồ thị gọi là Conceptual Graph Formalism (CGF). Một văn bản được tách thành nhiều câu. Và các câu đó sẽ được chuyển thành các Đồ thị khái niệm (gọi là Concept Graph (CG)). Tương tự cho câu hỏi cũng chuyển thành một CG.

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

Theo tìm hiểu, đồ thị khái niệm gần như tương đồng với với bộ ba quan hệ (Chủ từ, quan hệ, thực thể). Tuy nhiên, nó có một vài điểm khác trong cách biểu diễn.

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



Hình 1 – Ví dụ về một đồ thị khái niệm

( được lấy từ Figure 3. tài liệu tham khảo [6])

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

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

*[Invent]-*

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

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

Ở đây , *Person* là đối tượng cần biết trong câu hỏi , giữa *person* và *Light-Bulb* có mối quan hệ *Invent*. Sau đó , công việc tiếp theo là so sánh ( projection operator) giữa đồ thị CG của câu hỏi với các đồ thị khái niệm CG mỗi câu để tìm câu trả lời.

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

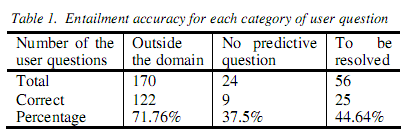
* Một nghiên cứu khác năm 2008 “Automatic Question Pattern Generation for Ontology-based Question Answering” của các tác giả Shiyan Ou, Constantin Orasan, Dalila Mekhaldi và Laura Hasler [7]. Với nghiên cứu này, các tác giả đề ra một phương pháp gọi là tự động tạo các khuôn mẫu cho các câu hỏi (Automatic Question Pattern generation) kèm theo việc sử dụng một phương pháp textual entailment (tạm dịch là kế thừa nguyên bản).

Phương pháp textual entailment theo tài liệu trình bày dùng để suy đoán câu hỏi mới đưa vào có phải là một dạng của câu hỏi đã có trước .

Ý tưởng chủ yếu là đưa ra một tập các câu hỏi dự kiến mà người dùng có thể hỏi cho hệ thống. Với các câu hỏi dự kiến này sẽ có các câu truy vấn tương ứng và rút ra câu trả lời . Các dữ liệu trong hệ thống này nằm trong một ontology mà tác giả giới hạn trong lĩnh vực “Movies and Cinema”.

Kết quả đánh giá dựa trên chất lượng các câu hỏi dự đoán được tạo ra và độ chính xác của phương pháp textual entailment .Một tập 250 câu hỏi tác giả rút ngẫu nhiên từ 4501 câu hỏi chuẫn liên quan đến lĩnh vực du lịch. Một bộ engine được áp dụng gọi là *textual entailment engine,* tạm gọi là bộ suy luận kế thừa. Ở đây , phương pháp bag-of-words được các tác giả sử dụng.

Theo nhóm thấy, kết quả cuối cùng cũng không được cao , và theo nhận xét của tác giả cũng vậy. Bảng kết quả được lấy từ trong tài liệu bao gồm các câu hỏi không thuộc lĩnh vực (170), các câu hỏi không nằm trong phần dư đoán trước (24) và số còn lại (56).



Hình 2 – Bảng đánh giá độ chính xác của textual entailemt engine

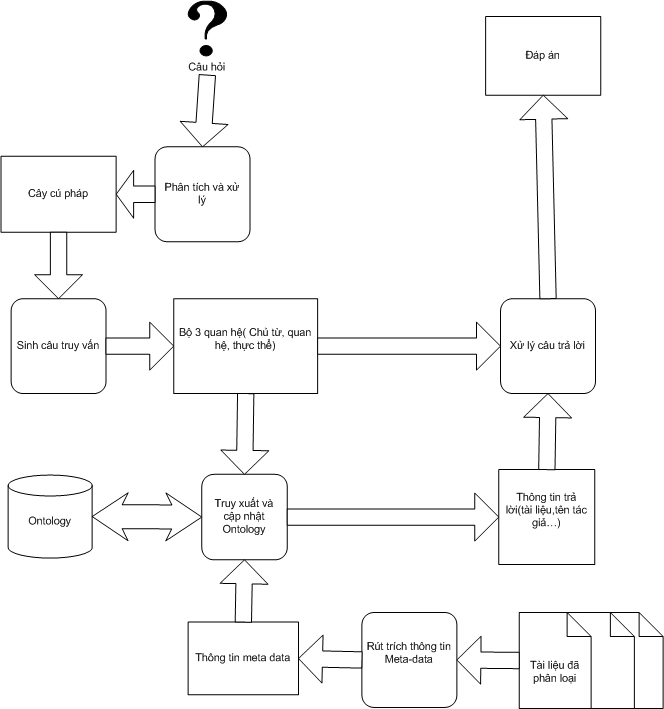
Lý do được đưa ra là với phương pháp bag-of-words , bộ *textual entailment engine* chưa đủ để suy đoán một số câu hỏi như : *“who directed 300”* và *“Who is the director of 300”.* Ngoài ra , còn do các câu hỏi dự đoán khá ngắn gọn , trong khi câu hỏi tự nhiên của người dùng đôi khi dài dòng và thừa từ .

Hướng đi tiếp theo của bài báo là tạo ra các câu hỏi dự đoán mang tính phức tạp hơn và cải tiến lại bộ *textual entailment engine* để đáp ứng được tính đa dạng của ngôn ngữ.

# 4. Nội dung và kế hoạch triển khai

## 5.1 nội dung

Từ mục tiêu trên nhóm đưa ra mô hình dự kiến hệ thống như sau :



Các thành phần chính

* Thành phần phân tích và xử lý câu hỏi
* Thành phần tạo câu truy vấn
* Thành phần truy xuất và cập nhật Ontology
* Thành phần xử lý câu trả lời
* Thành phần rút trích thông tin

Theo mô hình trên , chúng ta có hai luồng xử lý:

* Một luồng lấy thông tin vào ontology dựa trên một tập các tài liệu lấy về. Các tài liệu này là các bài báo khoa học có thể lấy trên các trang như : IEE <http://www.computer.org/portal/web/guest/home> , CiteXeer <http://citeseerx.ist.psu.edu/> và SSRN <http://papers.ssrn.com/sol3/displayabstractsearch.cfm> . Với mỗi tài liệu, ta đưa vào thành phần rút trích thông tin để lấy về các thông tin cần thiết như : tên tác giả , tiêu đề , ngày công bố , … Sau đó, các thông tin này được chuyển cho thành phần cập nhật dữ liệu để đưa vào Ontology .
* Luồng thứ hai là luồng mà nhóm hướng tới và tập trung vào đó là luồng xử lý câu truy vấn.
  + Câu hỏi người dùng nhập vào qua thành phần phân tích xử lý câu hỏi sẽ trả về một cây cú pháp . Cây cú pháp có thể thể hiện được mối quan hệ giữa các từ trong câu hỏi. Một số công cụ có thể áp dụng cho việc tạo cây cú pháp mà nhóm biết đến như Stanford Parser, Berkerley Parser và OPENNLP [1][2].
  + Từ cây cú pháp, thành phần tạo câu truy vấn sẽ rút ra bộ ba quan hệ RDF (resource – property –value) phục vụ cho việc truy vấn dữ liệu ontology. Vì ontology thường được lưu trữ dưới dạng tập tin OWL (Ontology Web Language) xây dựng dựa trên bộ ba RDF [4][5]. Nên ta cần rút ra được bộ ba này để truy vấn ontology thông qua ngôn ngữ SPARQL [3] .Đây là một ngôn ngữ phục vụ cho truy vấn dữ liệu dựa theo các bộ ba RDF.
  + Kết quả trên sẽ đưa vào để truy vấn ontology lấy dữ liệu về. Các dữ liệu lấy được sau khi truy vấn là các thông tin mà câu hỏi cần biết. ( có thể là tên tác giả, tiêu đề , …)
  + Cuối cùng, nhóm đề ra một bước xử lý câu trả lời , dự kiến thành phần xử lý câu trả lời này sẽ trả về một câu trả lời thân thiện theo ngôn ngữ tự nhiên phù hợp với câu hỏi.

## 5.2 kế hoạch triển khai

Bản dưới là kế hoạch dự kiến của nhóm cho việc triển khai hệ thống nói trên .Thời gian có thể sớmhoặc trễ hơn , tùy vào tiến độ hoạt dộng của nhóm. Nhưng nhóm sẽ nỗ lực để theo đúng tiến độ dự kiến.

|  |  |  |
| --- | --- | --- |
| Ngày | Tuần | Công việc thực hiện |
| 02/08 – 06/09/2010 | 4 tuần | * Đọc các tài liệu dự kiến tham khảo * tìm thêm một số tài liệu liên quan (nếu cần) * Bổ sung thêm số lượng câu hỏi cần xử lý |
| 06/09 – 20/09/2010 | 2 tuần | * Hiện thực module rút trích bộ ba quan hệ cho câu hỏi. |
| 20/09 – 11/10/2010 | 3 tuần | * Hoàn tất module rút trích bộ ba, thực hiện truy vấn lấy dữ liệu trà về . |
| 11/10 – 25/10/2010 | 2 tuần | * Xử lý module xử lý câu trả lời |
| 25/10 – 01/11/2010 | 1 tuần | * Kiểm tra lại nhửng thiếu sót và kiểm thử |

# 6. Kết quả dự kiến

Kết quả nhóm mong muốn đạt được là một hệ thống hỏi đáp thân thiện với người dùng . Các câu trả lời được xử lý và hiển thị theo ngôn ngữ tự nhiên . Ngoài ra, hệ thống đáp ứng được khoảng 80-90% các mẫu câu hỏi dự kiến bên trên .

# 7. Các tài liệu tham khảo dự kiến

* “A Question Answering System based on Conceptual Graph Formalism”, Wael Salloum , 2009.
* “Ontology-based Question Answering for digital Library”, Stephan Bloehdorn, Philipp Cimiano, Alistair Duke2, Peter Haase, J¨ org Heizmann, Ian Thurlow and Johanna Volker, 2007.
* “Automatic Question Pattern Generation for Ontology-based Question”, Shiyan Ou, Constantin Orasan, Dalila Mekhaldi and Laura Hasler, 2008.
* “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”, Lương Quý Tịnh Hà, 2009.
* “An Ontology based Question Answering System on Software Test Document Domain”, Meltem Serhatli and Ferda N. Alpaslan, 2009.

# 8. Tham khảo

Tài liệu tham khảo từ web :

[1] <http://nlp.stanford.edu/software/lex-parser.shtml>

[2] <http://nlp.cs.berkeley.edu/Main.html>

[3] <http://en.wikipedia.org/wiki/SPARQL>

[4] <http://www.w3schools.com/RDF/rdf_owl.asp> , phần Introduction to OWL

[5] ,<http://www.phpvn.org/index.php?topic=133.0;wap2> ,phần Ngôn ngữ Web ontology và khái niệm về Ontology

Các bài báo :

[6] “A Question Answering System based on Conceptual Graph Formalism”, Wael Salloum , 2009.

[7] “Automatic Question Pattern Generation for Ontology-based Question”, Shiyan Ou, Constantin Orasan, Dalila Mekhaldi and Laura Hasler, 2008.