**2 .Mục tiêu và phạm vi đề tài**

Mục tiêu của đề tài là xây dựng một hệ thống hỏi đáp dựa trên ontology [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”?
* How many papers were written by “Philip K. Chan”?
* Which papers were written by “Philip K. Chan”?
* 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”?
* How many concepts in “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* How many papers have keyword “Information Extraction”?
* Who is the author of “A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”?
* Which papers relate to topic “Software Engineering” ?
* What papers include this concept “Software Engineering”?

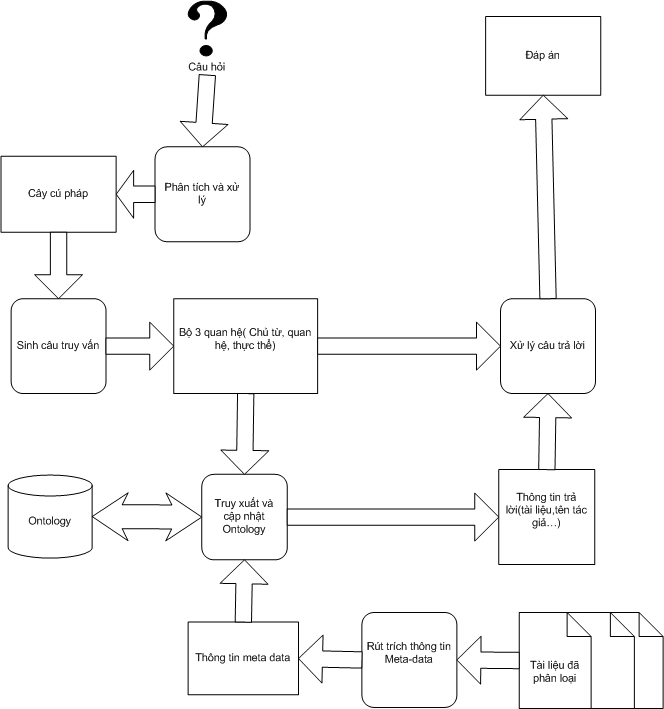
**3.Các ứng dụng và nghiên cứu liên quan**

Hiện nay, đã có một số trang web hỗ trợ hỏi đáp bằng ngôn ngữ tự nhiên như : <http://start.csail.mit.edu/>, <http://www.answerbus.com/about/index.shtml>, <http://www.answers.com>, …

**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 [1] và Berkerley Parser [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 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 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 |  | Đọc các tài liệu dự kiến tham khảo + tìm thêm một số tài liệu liên quan (nếu cần) |
| 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. 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