**Chương 1: Tổng quan về đề tài**

1. **Giới Thiệu**

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

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

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

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

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

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

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

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

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

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

1. **Mục tiêu và phạm vi đề tài**

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

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

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

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

Trong tiếng Anh, câu hỏi cơ bản được chia làm ba loại : Yes/No Question, Question Word Question và Choice Question **(REF).** Đây là các lọai câu hỏi thông dụng mà ta thường thấy. Ngòai ra, ta còn có thể thấy một số lọai câu hỏi khác như tag question, open question, closed question, direct/ indirect question **(REF).**

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

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

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

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

Vậy, tùy theo mục đích sư dụng mà câu hỏi còn đựơc phân lọai theo nhiều mục. Đối với 3 lọai câu hỏi cơ bản trong tiếng Anh thì hệ thống của nhóm hiện tại không giải quyết các câu hỏi yes/no question và choice question. Đó là các câu hỏi thuộc dạng xác minh vấn đề (có hoặc không) và câu hỏi lựa chọn. Chúng không phù hợp cho vấn đề tìm kiếm. Hệ thống chỉ giải quyết được một số các câu hỏi thuộc loại Question Word Question.

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

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

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

Nguồn câu hỏi được tham khảo từ :

- Dựa trên nguồn dữ liệu của DBLP ( các thuộc tính trong các bảng)

- Nhu cầu người dùng : các câu hỏi liên quan đến các vấn đề mà người dùng cần biết.

- Một số các mẫu câu trong tập TREC 2003 và 2007

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Phân loại** | **Câu hỏi** | **Kết quả** |
| 1 | ENTY | Which books were written by Rafiul Ahad and Amelia Carlson ? | *Answer:*  *1.Object SQL - A Language for the Design and Implementation of Object Databases.(1995)* |
| 2 | ENTY | Which books were written by Rafiul Ahad and Amelia Carlson in 1995 ? | *Answer:*  *1.Object SQL - A Language for the Design and Implementation of Object Databases.(1995)* |
| 3 | ENTY | Which books were written by Rafiul Ahad from 1999 to 2010 ? | *Answer:*  *1. A Scalable and Highly Available Networked Database Architecture.(1999)* |
| 4 | ENTY | Which books were published by O'Reilly in 1999 ? | *Answer:*  *1. Java Power Reference (1999)*  *2. Java Foundation Classes in a Nutshell (1999)*  *3. Java 2D Graphics (1999)*  *...* |
| 5 | ENTY | What are titles of books written by Marcus Thint ? | *Answer:*  *1.* *Ranking Answers by Hierarchical Topic Models..(2009)*  *2. Investigation of Question Classifier in Question Answering.(2009)*  *…* |
| 6 | ENTY | What books did Jennifer Widom write ?  or What books has Jennifer Widom written ? | *Answer:*  *1.Synthesizing view definitions from data.(2010)*  *2.* *LIVE: A Lineage-Supported Versioned DBMS.(2010)*  *…* |
| 7 | ENTY | What book did Philip K. Chan write in 1999? | *Answer:*  *1.AdaCost: Misclassification Cost-Sensitive Boosting.(1999)*  *2.* *Constructing Web User Profiles: A non-invasive Learning Approach.(1999)*  *3.Guest Editors' Introduction.(1999)* |
| 8 | ENTY | What book did Philip K. Chan write from 1999 to 2000? | *Answer:*  *1.AdaCost: Misclassification Cost-Sensitive Boosting.(1999)*  *2.* *Constructing Web User Profiles: A non-invasive Learning Approach.(1999)*  *3....* |
| 9 | ENTY | What are the titles of the books published by O’reilly in 1999 ? | *Answer:*  *1. Java Power Reference (1999)*  *2. Java Foundation Classes in a Nutshell (1999)*  *3. ...* |
| 10 | ENTY | What books has isbn 1-56592-484-3 | *Answer:*  *1.Java 2D Graphics (1999)* |
| 11 | ENTY | What books has doi 10.1145/360271.360274 | *Answer:*  *1.* *Java 2 Distributed Object Middleware Performance Analysis and Optimization.(2000)* |
| 12 | ENTY | What books were written by “Philip K. Chan” from ACM? | *Answer:* |
| 13 | ENTY | What publications have resulted from TREC (in 1999)? | *Answer:* |
| 14 | ENTY | What books were published by ACM and Springer in 2010? | *Answer:*  *1.* *Advances in Music Information Retrieval*  *2.* *Advances in Multi-Objective Nature Inspired Computing*  *3. …* |
| 15 | NUM | How many papers were written by Rafiul Ahad ? | *Answer:*  *11 papers were written by Rafiul Ahad.*  *1. A Scalable and Highly Available Networked Database Architecture.(1999)*  *2. Object SQL - A Language for the Design and …*  *3. …* |
| 16 | NUM | How many papers were written by Rafiul Ahad in 1999 ? | *Answer:*  *1 papers were written by Rafiul Ahad*  *1.* *A Scalable and Highly Available Networked Database Architecture.(1999)* |
| 17 | NUM | How many publisher did “Thomas Hermann” write for? | *Answer:*  1 publisher did “Thomas Hermann” work with?  1. Springer |
| 18 | NUM | How many papers were written by Raul Rato and Rui Gomes in 2010? | *Answer:*  1 papers were written by Raul Rato and Rui Gomes in 2010?  1.Railscan: A Tool for the Detection and Quantification of Rail Corrugation. |
| 19 | HUM | Who write books in 1999 ? | *Answer:*  *-James Gosling*  *-Ken Arnold*  *-Berthier A. Ribeiro-Neto*  *-Ricardo A. Baeza-Yates*  *…* |
| 20 | HUM | Who write books from 1999 to 2010 ? | *Answer:*  *-Leslie Lamport*  *-Arthur J. Bernstein*  *-Michael Kifer*  *…* |
| 21 | HUM | Who published books from 1999 to 2000 ? | *Answer:*  *-CSREA Press*  *-Metropolis*  *-World Scientific*  *-CSLI Publications*  *-Vieweg*  *-ACM*  *…* |
| 22 | HUM | Who published books in 1999 ? | *Answer:*  *-AAAI*  *-AAAI Press*  *-ACM*  *-ACM Press*  *-ACM Press / Addison-Wesley*  *-Addison-Wesley*  *-APress*  *….* |
| 23 | HUM | Who is the author of “Working Models for Uncertain Data” | *Answer:*  *-Anish Das Sarma*  *-Omar Benjelloun*  *-Alon Y. Halevy*  *-Jennifer Widom* |
| 24 | HUM | What composer wrote “ *Java 2D Graphics*” | *-* *Jonathan Knudsen* |
| 25 | HUM | What composer wrote books from 1999 in ACM? | *Answer:* |
| 26 | HUM | Who is the author of the paper “ *Question Classification using Head Words and their Hypernyms.”?* | *Answer:* |
| 27 | HUM | Who wrote “ *Question Classification using Head Words and their Hypernyms.”?* | *Answer:* |

Vậy, hệ thống chỉ giải quyết được các câu hỏi thuộc các thể lọai ENTY, HUM và NUM. Tức là, các câu hỏi liên quan đến cuốn sách, bài báo (ENTY), các câu hỏi về tác giả, nhà xuất bản (HUM) và các câu hỏi về số lượng (NUM).

1. http://dblp.uni-trier.de/xml/ [↑](#footnote-ref-2)
2. http://trec.nist.gov/ [↑](#footnote-ref-3)