HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HOÀNG ANH

XỬ LÝ NGỮ NGHĨA TRONG MÁY TÌM KIẾM

CHUYÊN NGÀNH : KHOA HỌC MÁY TÍNH

MÃ SỐ: 60.48.01.01 (Khoa học máy tính)

LUẬN VĂN THẠC SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN MẠNH HÙNG

HÀ NỘI – 2014

CHƯƠNG I: GIỚI THIỆU BÀI TOÁN MÁY TÌM KIẾM NGỮ NGHĨA CHO TIẾNG VIỆT

Chương này trình bày tổng quan về máy tìm kiếm, máy tìm kiếm ngữ nghĩa, các vấn đề của tìm kiếm ngữ nghĩa trong Tiếng Việt, các vấn đề liên quan và giải pháp đề xuất

1.1. Máy tìm kiếm và máy tìm kiếm ngữ nghĩa

1.1.1. Máy tìm kiếm

Máy tìm kiếm là một công cụ quan trọng dùng để tìm kiếm thông tin trên mạng Internet. Nếu không có máy tìm kiếm thì người dùng sẽ không biết làm thế nào để tìm kiếm thông tin cần thiết trên các trang web. Ngày nay với sự phát triển của Internet, có rất nhiều công cụ tìm kiếm đã ra đời nhằm giúp đỡ người dùng trong quá trình tìm kiếm thông tin cần thiết. Do nguồn thông tin trên internet ngày càng lớn nên rất khó để máy tìm kiếm trả về thông tin đúng mong muốn của người dùng. Do vậy việc tự động phân nhóm và tổ chức dữ liệu thành các miền để dễ dàng cho việc tìm kiếm đã trở nên phổ biến[3].

Máy tìm kiếm là công cụ phổ biến nhất để tìm ra những thông tin cần thiết trên mạng cho người dùng. Máy tìm kiếm có một đặc điểm chung là thu thập một tập hợp lớn các dữ liệu trên mạng internet để phục vụ cho người dùng tìm kiếm. Hầu hết tất cả các máy tìm kiếm đều chia làm 3 phần:

- ✓ Một cơ sở dữ liệu chứa các tài liệu trên Internet.
- ✓ Một máy tìm kiếm hoạt động trên cơ sở dữ liệu đó.
- ✓ Một tập các chương trình có nhiệm vụ xác định làm thế nào để kết quả tìm kiếm được hiển thị.

Một quá trình tìm kiếm của người dùng đơn giản là: nhập từ khóa và nhận kết quả trả ra đã được sắp xếp thứ tự từ máy tìm kiếm.

Kết quả tìm kiếm của một máy tìm kiếm có tốt hay không phụ thuộc vào 2 vấn đề:

- ✓ Chất lượng của hệ thống (là cách xây dựng hệ thống tìm kiếm như thế nào)
- ✓ Tập dữ liệu của máy tìm kiếm.

1.1.2. Máy tìm kiếm ngữ nghĩa

Khác với máy tìm kiếm truyền thống, một máy tìm kiếm ngữ nghĩa lưu trữ thông tin có ngữ nghĩa về dữ liệu và có thể trả lời những câu truy vấn phức tạp từ người dùng.

Quá trình xử lý của một máy tìm kiếm ngữ nghĩa khi người dùng nhập từ khóa:

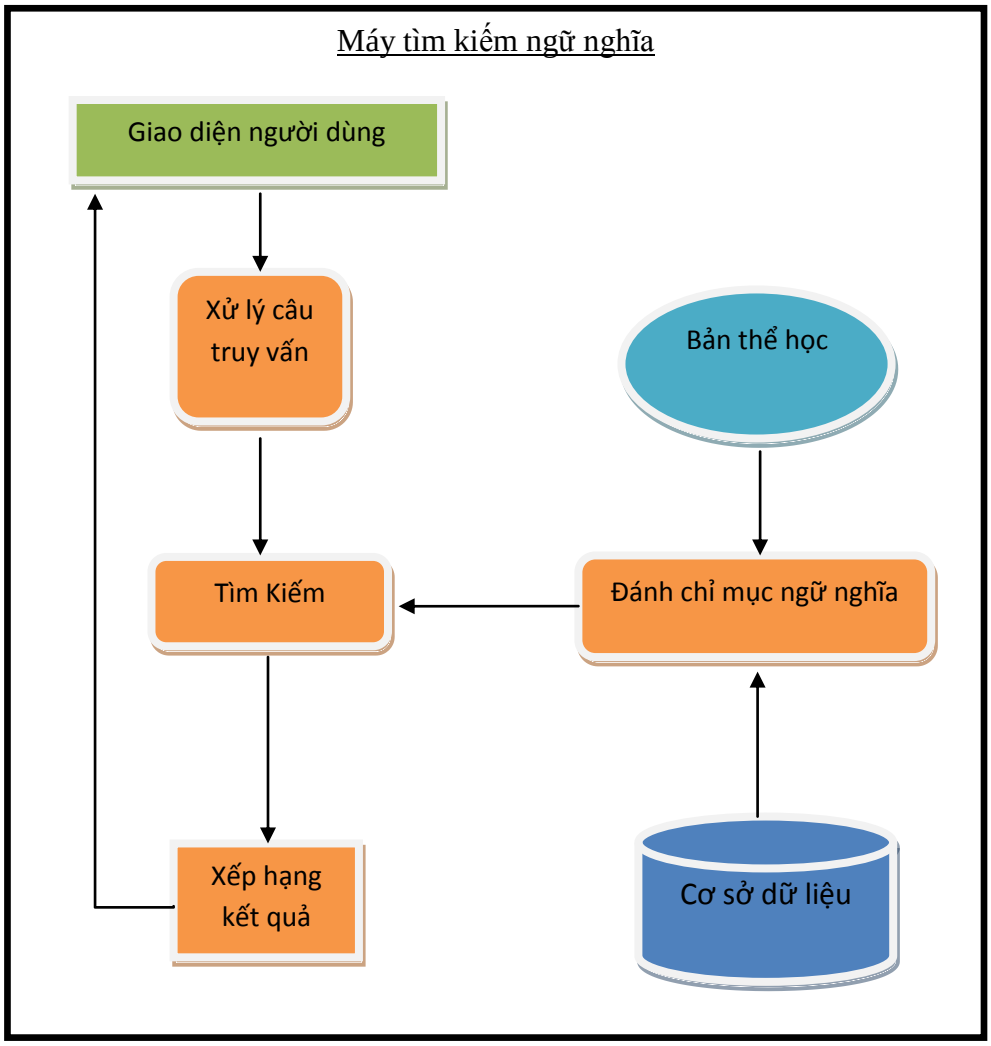
- ✓ Làm sáng tỏ câu hỏi của người dùng, trích chọn từ khóa thích hợp theo ngữ cảnh.
- ✓ Một tập các khái niệm được sử dụng để xây dựng câu truy vấn dựa vào bản thể học
- ✓ Trả về kết quả cho người dùng.

1.1.3. Kiến trúc tổng quan của máy tìm kiếm ngữ nghĩa

Một máy tím kiếm ngữ nghĩa về cơ bản cũng có cấu trúc giống như máy tìm kiếm thông thường bao gồm hai phần chính

- ❖ **Giao diện người dùng**: bao gồm hai phần chính
  - ✓ Giao diện nhập câu truy vấn: cho phép người dùng nhập câu truy vấn.
  - ✓ Hiển thị kết quả: Phần này hiển thị kết quả đã sắp xếp theo thứ tự ưu tiên cho người dùng
- ❖ **Kiến trúc bên trong:**

  Dưới đây là mô hình kiến trúc của một máy tìm kiếm ngữ nghĩa



Hình 1.1 Mô hình kiến trúc tổng quan của máy tìm kiếm ngữ nghĩa[9]

Trong mô hình kiến trúc của máy tìm kiếm ngữ nghĩa, chúng ta có thể nhìn thấy bốn bước chính:

- ✓ Đánh chỉ mục ngữ nghĩa (indexing)
- ✓ Xử lý câu truy vấn (query processing)
- ✓ Tìm kiếm (searching)
- ✓ Xếp hạng (ranking)

Khác so với máy tìm kiếm truyền thống, trong máy tìm kiếm ngữ nghĩa xử lý câu truy vấn và quá trình đánh chỉ mục đều dựa vào bản thể học. Về chi tiết chúng ta sẽ tìm hiểu trong chương II.

1.2. Tìm kiếm ngữ nghĩa trong tiếng Việt.

1.2.1. Đặc trưng của tiếng Việt

Theo như tác giả Vũ Xuân Lương [13] tiếng viết có một số đặc trưng về ngữ âm, từ vựng, ngữ pháp như sau

1.2.1.1. Đặc điểm ngữ âm

Trong tiếng Việt có một loại đơn vị đặc biệt gọi là "tiếng". Về mặt ngữ âm, mỗi tiếng là một âm tiết. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối, tạo ra tiềm năng của ngữ âm tiếng Việt trong việc thể hiện các đơn vị có nghĩa. Nhiều từ tượng hình, tượng thanh có giá trị gợi tả đặc sắc. Khi tạo câu, tạo lời, người Việt rất chú ý đến sự hài hoà về ngữ âm, đến nhạc điệu của câu văn.

1.2.1.2. Đặc điểm từ vựng

Mỗi tiếng, nói chung, là một yếu tố có nghĩa. Tiếng là đơn vị cơ sở của hệ thống các đơn vị có nghĩa của tiếng Việt. Từ tiếng, người ta tạo ra các đơn vị từ vựng khác để định danh sự vật, hiện tượng..., chủ yếu nhờ phương thức ghép và phương thức láy.

Việc tạo ra các đơn vị từ vựng ở phương thức ghép luôn chịu sự chi phối của quy luật kết hợp ngữ nghĩa, ví dụ: đất nước, máy bay, nhà lầu xe hơi, nhà tan cửa nát... Hiện nay, đây là phương thức chủ yếu để sản sinh ra các đơn vị từ vựng. Theo phương thức này, tiếng Việt triệt để sử dụng các yếu tố cấu tạo từ thuần Việt hay vay mượn từ các ngôn ngữ khác để tạo ra các từ, ngữ mới, ví dụ: tiếp thị, karaoke, thư điện tử (e-mail), thư thoại (voice mail), phiên bản (version), xa lộ thông tin, siêu liên kết văn bản, truy cập ngẫu nhiên, v.v.

Việc tạo ra các đơn vị từ vựng ở phương thức láy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng, chẳng hạn: chôm chỉa, chỏng chơ, đỏng đa đỏng đảnh, thơ thẩn, lúng lá lúng liếng, v.v.

Vốn từ vựng tối thiểu của tiếng Việt phần lớn là các từ đơn tiết (một âm tiết, một tiếng). Sự linh hoạt trong sử dụng, việc tạo ra các từ ngữ mới một cách dễ dàng đã tạo điều kiện thuận lợi cho sự phát triển vốn từ, vừa phong phú về số lượng, vừa đa dạng trong hoạt động. Cùng một sự vật, hiện tượng, một hoạt động hay một đặc trưng, có thể có nhiều từ ngữ khác nhau biểu thị. Tiềm năng của vốn từ ngữ tiếng Việt được phát huy cao độ trong các phong cách chức năng ngôn ngữ, đặc biệt là trong phong cách ngôn ngữ nghệ thuật. Hiện nay, do sự phát triển vượt

bậc của khoa học-kĩ thuật, đặc biệt là công nghệ thông tin, thì tiềm năng đó còn được phát huy mạnh mẽ hơn.

1.2.1.3. Đặc điểm ngữ pháp

Từ của tiếng Việt không biến đổi hình thái. Đặc điểm này sẽ chi phối các đặc điểm ngữ pháp khác. Khi từ kết hợp từ thành các kết cấu như ngữ, câu, tiếng Việt rất coi trọng phương thức trật tự từ và hư từ.

Việc sắp xếp các từ theo một trật tự nhất định là cách chủ yếu để biểu thị các quan hệ cú pháp. Trong tiếng Việt khi nói "Anh ta lại đến" là khác với "Lại đến anh ta" . Khi các từ cùng loại kết hợp với nhau theo quan hệ chính phụ thì từ đứng trước giữ vai trò chính, từ đứng sau giữ vai trò phụ. Nhờ trật tự kết hợp của từ mà "củ cải" khác với "cải củ" , "tình cảm" khác với "cảm tình" . Trật tự chủ ngữ đứng trước, vị ngữ đứng sau là trật tự phổ biến của kết cấu câu tiếng Việt.

Phương thức hư từ cũng là phương thức ngữ pháp chủ yếu của tiếng Việt. Nhờ hư từ mà tổ hợp "anh của em" khác với tổ hợp "anh và em" , "anh vì em" . Hư từ cùng với trật tự từ cho phép tiếng Việt tạo ra nhiều câu cùng có nội dung thông báo cơ bản như nhau nhưng khác nhau về sắc thái biểu cảm. Ví dụ, so sánh các câu sau đây:

- ✓ Ông ấy không hút thuốc.
- ✓ Thuốc, ông ấy không hút.
- ✓ Thuốc, ông ấy cũng không hút.

Ngoài trật tự từ và hư từ, tiếng Việt còn sử dụng phương thức ngữ điệu. Ngữ điệu giữ vai trò trong việc biểu hiện quan hệ cú pháp của các yếu tố trong câu, nhờ đó nhằm đưa ra nội dung muốn thông báo. Trên văn bản, ngữ điệu thường được biểu hiện bằng dấu câu. Chúng ta thử so sánh 2 câu sau để thấy sự khác nhau trong nội dung thông báo:

- ✓ Đêm hôm qua, cầu gãy.
- ✓ Đêm hôm, qua cầu gãy.

Qua một số đặc điểm nổi bật vừa nêu trên đây, chúng ta có thể hình dung được phần nào bản sắc và tiềm năng của tiếng Việt.

1.3. Một số phương pháp xử lý ngữ nghĩa trong máy tìm kiếm

Xử lý ngữ nghĩa trong máy tìm kiếm đã trở thành xu hướng hiện nay, có rất nhiều nghiên cứu liên quan đến xử lý ngữ nghĩa trong máy tìm kiếm

Trong nghiên cứu [1], tác giả đã trình bày việc tiếp cận xử lý ngữ nghĩa bằng cách xử lý câu truy vấn đầu vào dựa trên bản thể học, bằng cách xử dụng mối quan hệ ngữ nghĩa để tìm ra những từ ngữ tương quan với câu truy vấn.

Trong nghiên cứu [6], tác giả đã hướng đến xử lý ngữ nghĩa bằng cách làm sáng tỏ câu truy vấn bằng cách tạo ra tập các từ khóa có liên quan đến nhau.

Nhìn chung những bài báo trên đều đưa ra cách xử lý ngữ nghĩa bằng cách làm hiểu câu truy vấn của người dùng rồi từ đó truy vấn vào trong cơ sở dữ liệu để tìm tài liệu phù hợp nhất với câu truy vấn.

1.4. Giải pháp đề xuất của luận văn

Trong những hướng nghiên cứu trên, các tác giả đều tập chung vào xử lý câu truy vấn trước khi tìm kiếm. Luận văn sẽ trình bày một hướng khác dựa trên bản thể học để đánh chỉ mục ngữ nghĩa cho tài liệu, theo hướng này thì nội dung văn bản đã được xử lý trước khi tìm kiếm.

Kết Luận Chương

Như vậy trong **chương I** luận văn đã đi tìm hiểu về máy tìm kiếm, kiến trúc tổng quan của máy tìm kiếm ngữ nghĩa, đặc điểm của tiếng Viêt. Không giống như tiếng Anh và một số ngôn ngữ khác, từ trong tiếng Việt không đơn giản là được phân biệt bởi dấu cách, trong tiếng Việt một từ có thể có một âm tiết hoặc có thể có nhiều âm tiết. Điều này sẽ gây khó khăn trong quá trình đánh chỉ mục cho tài liệu, vấn đề này sẽ được giải quyết trong **chương II**. **Chương II** luận văn sẽ đi vào tìm hiểu cách đánh chỉ mục ngữ nghĩa cho tài liệu.

CHƯƠNG II: XỬ LÝ NGỮ NGHĨA TRONG MÁY TÌM KIẾM

Trong chương này luận văn đi vào nghiên cứu phương pháp tìm kiếm ngữ nghĩa dựa trên bản thể học. Đánh chỉ mục ngữ nghĩa và tìm kiếm dựa vào mô hình không gian vector và bản thể học.

2.1. Bản thể học

2.1.1. Định nghĩa bản thể học

Có nhiều định nghĩa khác nhau về Bản thể học. Theo Gruber: một bản thể học là một sự mô tả một cách hình thức và rõ ràng về các khái niệm.

Theo tài liệu tham khảo [11], các thuật ngữ giữ vai trò quan trọng trong bản thể hoc: "một bản thể học là một tập hợp có cấu trúc phân cấp các thuật ngữ dùng để mô tả một lĩnh vực nào đó và có thể dùng như một bộ khung cho một cơ sở tri thức".

2.1.2. Các thành phần của bản thể học

Các thành phần thường gặp của Bản thể học bao gồm:

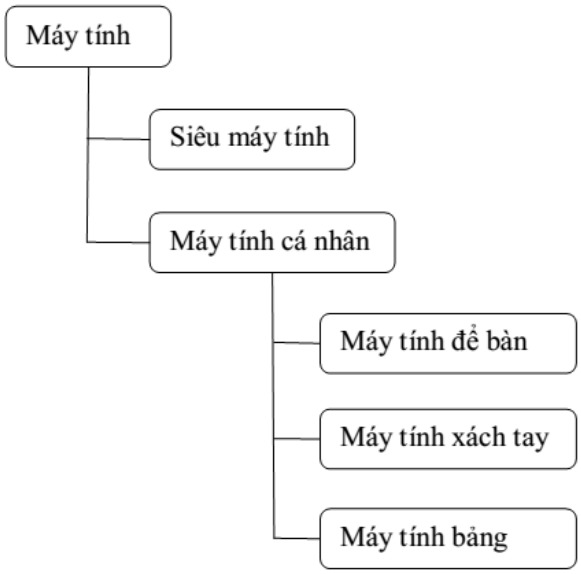
- *Thực thể (individual):* là thành phần cơ bản của một bản thể học. Các thực thể trong một bản thể học có thể bao gồm các đối tượng cụ thể như con người, động vật... Một bản thể học có thể không cần bất kỳ một thực thể nào.
- *Lớp (class):* là nhóm, tập hợp các đối tượng trừu tượng. Chúng có thể chứa các cá thể, các lớp khác.
- *Thuộc tính (attribute):* các đối tượng trong bản thể học có thể được mô tả thông qua việc khai báo các thuộc tính của chúng. Mỗi thuộc tính đề có tên và giá trị của thuộc tính đó. Các thuộc tính được sử dụng để lưu trữ các thông tin mà đối tượng có thể có. Ví dụ, một cá nhân có thể có các thuộc tính như họ tên, ngày sinh, quê quán, số CMND... Giá trị của một thuộc tính có thể là kiểu dữ liệu phức tạp.

❖ *Mối quan hệ (relationship):* quan hệ giữa các đối tượng trong một bản thể học cho biết các đối tượng liên hệ với đối tượng khác như thế nào. Sức mạnh của bản thể học nằm ở khả năng diễn đạt quan hệ. Tập hợp các quan hệ cùng nhau mô tả ngữ nghĩa của một miền. Tập các dạng quan hệ được sử dụng và cây phân cấp thứ bậc của chúng thể hiện sức mạnh diễn đạt của ngôn ngữ dùng để biểu diễn bản thể học. Sự xuất hiện của mối quan hệ is_a hay còn gọi là mối quan hệ cha con tạo ra một cấu trúc phân cấp thứ bậc, dạng cấu trúc cây này mô tả rõ ràng cách thức các đối tượng liên hệ với nhau.

Ví dụ **hình 2.1**, ta thấy lớp "máy tính cá nhân" là lớp cha của lớp "máy tính xách tay" nhưng lại là con của lớp "máy tính". Một dạng quan hệ phổ biến khác là quan hệ meronymy hay còn gọi là quan hệ "thành phần của", biểu diễn làm thế nào các đối tượng kết hợp với nhau để tạo nên một đối tượng tổng hợp. Ví dụ, nếu ta mở rộng bản thể học để chứa thêm khái niệm như "bộ nhớ", chúng ta có thể nói rằng lớp "bộ nhớ" là một thành phần của "máy tính". Khi đó cấu trúc cây đơn giản và nhẹ nhàng trước đó sẽ nhanh chóng trở nên phức tạp.

Bản thể học thường phân biệt các nhóm quan hệ như:

- ✓ Quan hệ giữa các lớp,
- ✓ Quan hệ giữa các thực thể,
- ✓ Quan hệ giữa thực thể và một lớp.



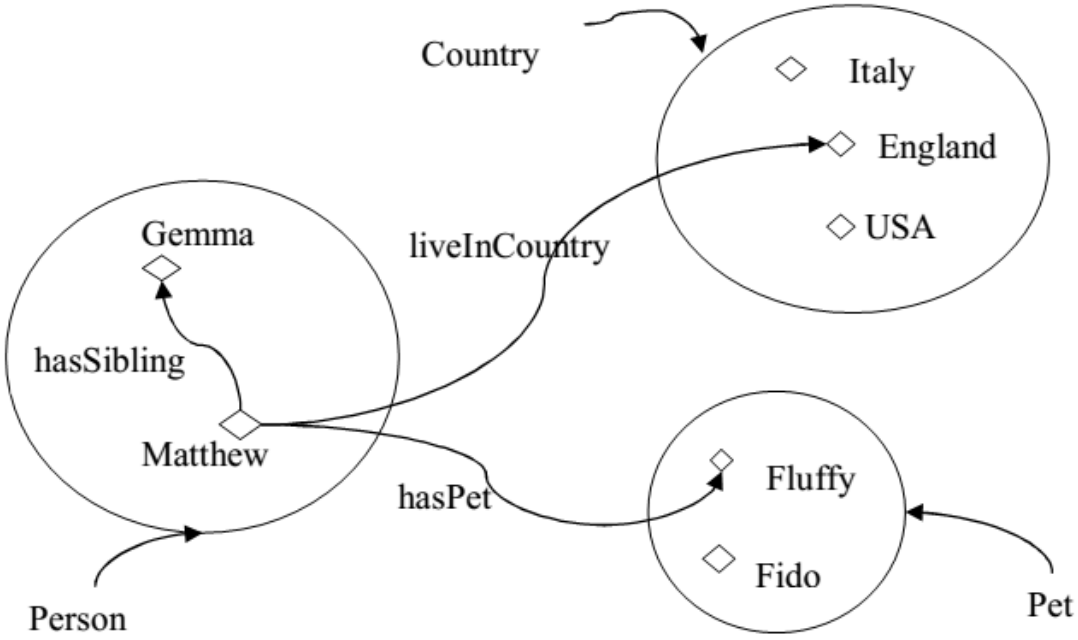
Hình 2.1 Thể hiện mối quan hệ cha con trong Ontology

2.1.3. Ngôn ngữ biểu diễn bản thể học

Cũng như các loại mô hình dữ liệu khác, bản thể học cũng cần một ngôn ngữ để biểu diễn. Ngữ nghĩa của bản thể học phụ thuộc rất nhiều vào khả năng biểu diễn của ngôn ngữ đó. Một số ngôn ngữ thường được sử dụng như: ngôn ngữ RDF, RDFS, OWL, CycL... Tuy nhiên ở đây chỉ giới thiệu những đặc điểm chính của ngôn ngữ OWL vì nó cung cấp tập từ vựng định nghĩa lớp và thuộc tính phong phú hơn nên có tính diễn đạt cao hơn và hỗ trợ khả năng suy diễn tốt hơn đáp ứng được yêu cầu khi xây dựng bản thể học của luận văn.

Theo tài liệu tham khảo [2] các thành phần chính của OWL Bản thể học gồm: lớp, thuộc tính và thực thể.

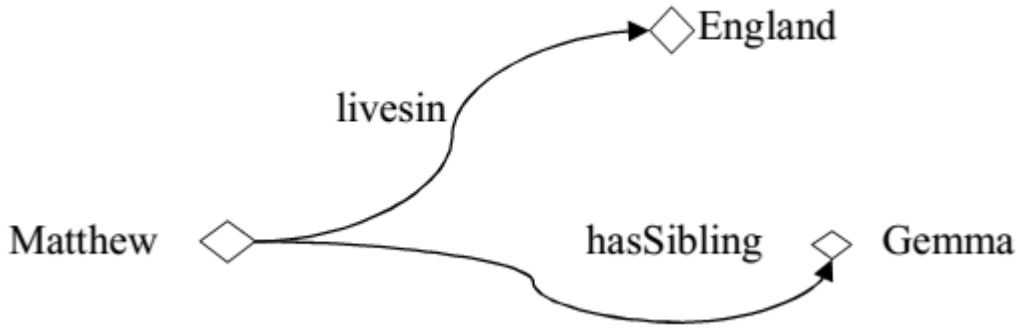
owl: Class - là một nhóm các thực thể có liên quan. Lớp có thể được xây dựng trong một hệ thống phân cấp bằng các sử dụng subClassOf. Thing là lớp của tất cả các thực thể và là lớp cha của tất cả các lớp OWL.



Hình 2.2. Mô tả lớp trong bản thể học (bao gồm cả thực thể)[2]

rdfs: subclassOf - là sự phân cấp lớp có thể được tạo ra bằng cách làm cho một hoặc nhiều khai báo rằng một lớp là một lớp con của lớp khác. Ví dụ, lớp người là một phân lớp của lớp động vật có vú.

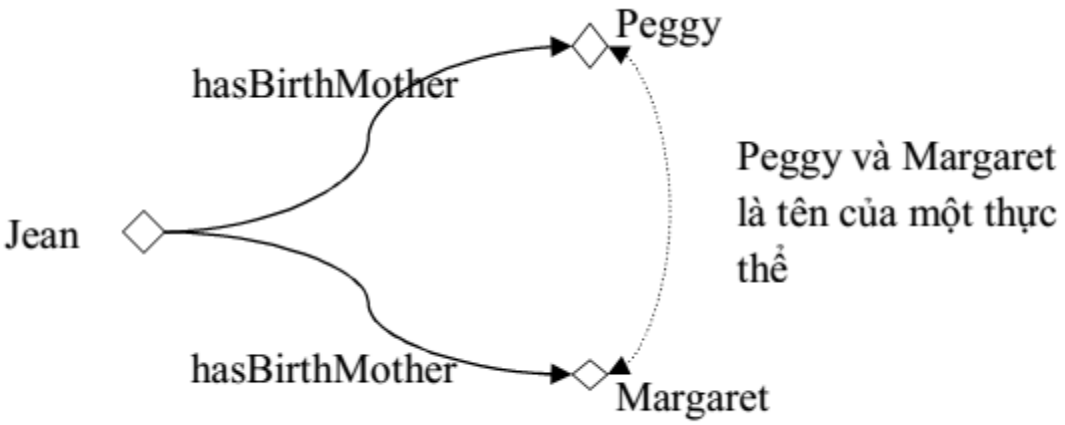
owl: ObjectProperty - là thuộc tính mô tả mối quan hệ giữa hai thực thể. Ví dụ, thuộc tính hasSibling liên kết thực thể Matthew với thực thể Gemma.



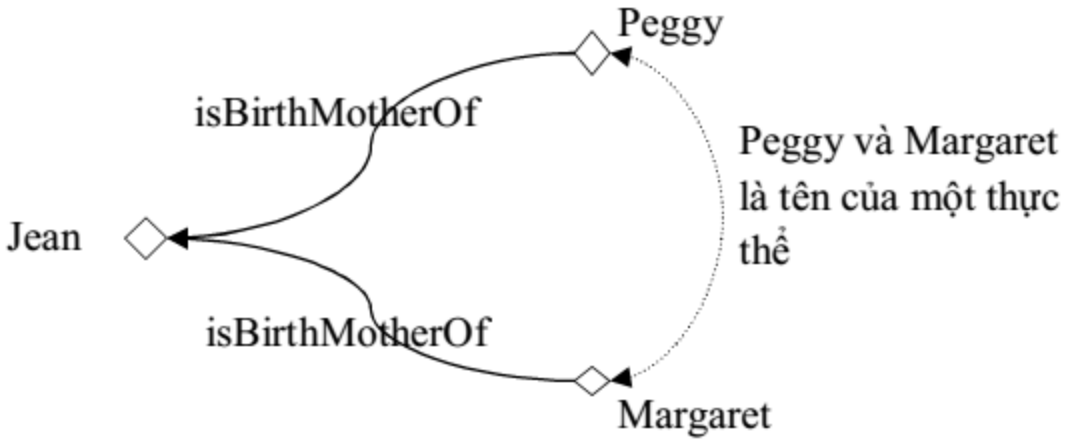
Hình 2.3. Mối quan hệ giữa các thực thể[2]

ObjectProperty có 4 tính chất sau:

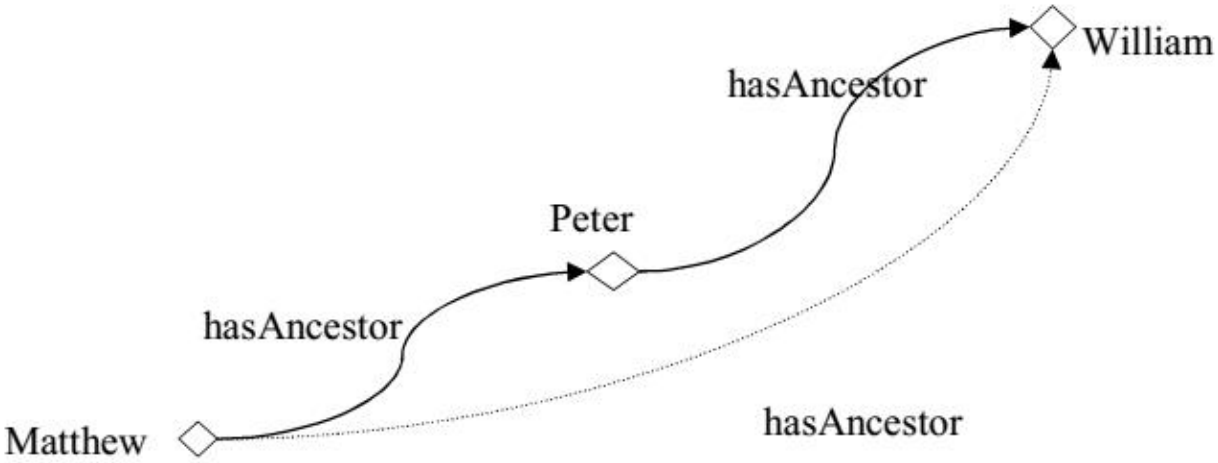
✓ Functional: một thực thể chỉ liên quan nhiều nhất đến một thực thể khác



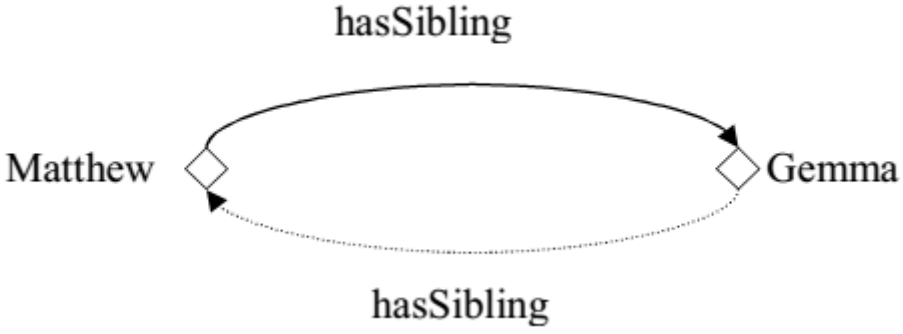
✓ Inverse Functional: là thuộc tính đảo ngược của Functional.



✓ Transitive: thực thể a quan hệ với thực thể b, thực thể b quan hệ với thực thể c → thực thể a quan hệ với thực thể c.



✓ Symmetric: thực thể a quan hệ với thực thể b → thực thể b quan hệ với thực thể a



owl: DatatypeProperty - mô tả mối quan hệ giữa thực thể và giá trị của nó. Ví dụ, thuộc tính hasAge có thể được sử dụng để chỉ quan hệ một thể hiện của lớp người với một thể hiện của kiểu dữ liệu số nguyên.

rdfs: Domain - của một thực thể giới hạn các thuộc tính mà thuộc tính đó có thể áp dụng. Nếu một thuộc tính dùng để kết nối một thực thể này với một thực thể khác, và thuộc tính thuộc một lớp trong miền của nó, thì các thực thể phải phụ thuộc vào lớp đó. Ví dụ, thuộc tính hasChild có thể khai báo có miền là động vật có vú. Từ bộ lập luận có thể suy luận rằng nếu Frank hasChild Anna, thì Frank là động vật có vú.

rdfs: range - phạm vi của một thuộc tính giới hạn thực thể là giá trị mà thuộc tính có thể có. Nếu thuộc tính dùng để tạo quan hệ từ thực thể này đến thực thể khác, và thuộc tính có lớp trong phạm vi của nó, thì các thực thể khác phải phụ thuộc vào phạm vi của lớp.

Ví dụ, thuộc tính hasChild có thể được khai báo trong phạm vi lớp động vật có vú. Từ một bộ lập luận có thể suy luận rằng nếu Luise được kết nối đến Deboral bằng thuộc tính hasChild, thì Deboral là một động vật có vú.

owl:NameIndividual - mô tả các đối tượng trong một lĩnh vực mà chúng ta quan tâm. Có thể có nhiều tên được sử dụng để nói về một thực thể.

Hình 2.4. Các thực thể trong bản thể học[2]

2.1.4. Cách xây dựng Bản thể học

Có nhiều phương thức khác nhau để xây dựng một *bản thể học*, nhưng nhìn chung các phương pháp đều thực hiện theo hai bước cơ bản là:

- ✓ Xây dựng cấu trúc lớp phân cấp
- ✓ Định nghĩa các thuộc tính cho lớp

Trong thực tế, việc phát triển *bản thể học* để mô tả lĩnh vực cần quan tâm là một việc không đơn giản, phụ thuộc vào rất nhiều công cụ sử dụng, tính chất, quy mô, sự thường xuyên biến đổi của miền cũng như các quan hệ phức tạp trong đó. Những khó khăn này đòi hỏi công việc xây dựng *bản thể học* phải là một quá trình lặp đi lặp lại, mỗi lần lặp cải thiện, tinh chế và phát triển dần. Công việc xây dựng *bản thể học* cũng cần phải tính đến khả năng mở rộng lĩnh vực quan tâm trong tương lai, khả năng kế thừa các hệ thống *Bản thể học* có sẵn, cũng như tính linh động để *bản thể học* có khả năng mô tả tốt nhất các quan hệ phức tạp trong thế giới thực.

Một số nguyên tắc xây dựng *bản thể học* thông qua các bước sau:

- ✓ Xác định miền quan tâm và phạm vi của *bản thể học*.
- ✓ Xem xét việc kế thừa các *bản thể học* có sẵn.
- ✓ Liệt kê các thuật ngữ quan trọng trong *bản thể học*.
- ✓ Xây dựng các lớp và cấu trúc lớp phân cấp.
- ✓ Định nghĩa các thuộc tính và quan hệ cho lớp.
- ✓ Định nghĩa các rằng buộc về thuộc tính và quan hệ của lớp.
- ✓ Tạo các thực thể cho lớp.

2.2. Tìm kiếm ngữ nghĩa dựa trên bản thể học

Trong phần này luận văn sẽ trình bày hai quá trình chính của máy tìm kiếm ngữ nghĩa đó là đánh chỉ mục ngữ nghĩa và xử lý câu truy vấn dựa vào bản thể học.

2.2.1. Đánh chỉ mục ngữ nghĩa dựa trên bản thể học

Khác với phương pháp đánh chỉ mục thông thường đó là trọng số của một từ được tính toán chỉ dựa vào sự xuất hiện chính xác của từ trong văn bản mà không quan tâm đến những từ có ngữ nghĩa tương tự. Phương pháp đánh chỉ mục ngữ nghĩa dựa trên bản thể học sẽ khắc phục được thiếu sót này.

Trong phần này sẽ trình bày cách đánh chỉ mục ngữ nghĩa cho mỗi tài liệu dựa vào bản thể học.

2.2.1.1. Trọng số của từ

Thông thường, trọng số của từ được tính dựa vào thuật toán TF-IDF. TF-IDF là viết tắt của từ "*term frequency-inverse document frequency"*. Ý tưởng của thuật toán này là một từ mang ý nghĩa càng lớn nếu nó có độ phân bố xuất hiện trong một văn bản lớn đồng thời xuất hiện ít trong các văn bản còn lại.[7]

Khi truy vấn trong các tài liệu, để tính toán tìm ra những tài liệu thích hợp với câu truy vấn, ta sẽ coi mỗi câu truy vấn là tập của các từ và tính toán giữa các từ của câu truy vấn với tài liệu.

Chúng ta sẽ gán cho mỗi từ trong một tài liệu một ***trọng số***, trọng số của từ sẽ phụ thuộc vào số lần xuất hiện của từ đó trong một tài liệu. Chúng ta sẽ tính toán độ thích hợp của từ t trong câu truy vấn với một tài liệu d dựa trên trọng số của t trong d. Để đơn giản ta sẽ gán trọng số của từ bằng chính số lần xuất hiện của từ đó trong tài liệu. Trọng số này chính là tần suất xuất hiện của từ trong một tài liệu ký hiệu là $TF_{t,d}$

Như vậy, mỗi tài liệu d sẽ có một tập các từ và mỗi từ sẽ có một trọng số và chúng ta có thể định lượng mỗi tài liệu dựa trên trọng số của các từ trong tài liệu đó.

Ví dụ chúng ta có 2 tài liệu

- ✓ d1: "lập trình viên Java"
- ✓ d2: "lập trình viên PHP"

Chúng ta tính toán trọng số cho mỗi từ trong d1, d2

	lập	trình	viên	Java	PHP
d1	1	1	1	1	0
d2	1	1	1	0	1

Tính trọng số dựa vào tần xuất xuất hiện của từ trong một tài liệu có nghĩa là khi truy vấn thì tất cả các từ trong các tài liệu đều được coi là quan trọng như nhau.

Nhưng trong thực tế, những từ quan trọng thường ít xuất hiện. Ví dụ như một tập các tài liệu liên quan đến lĩnh vực công nghệ thông tin thì từ "máy tính" thường xuất hiện trong tất cả các tài liệu. Như vậy khi truy vấn, nếu tính toán dựa vào trọng số được tính toán như trên sẽ đưa ra kết quả thiếu chính xác, để làm giảm mức độ ảnh hưởng của từ thường xuyên xuất hiện khi truy vấn chúng ta sẽ tính toán trọng số của từ dựa trên tần suất xuất hiện của từ trong một tài liệu $TF_{t,d}$ và số lần xuất hiện của từ trên tất cả các tài liệu DF_t.

Giả sử số tài liệu trong cơ sở dữ liệu là N, ta sẽ tính giá trị nghịch đảo của tần suất xuất hiện của từ trong các tài liệu bằng công thức[7]:

$$IDF_t = log \frac{N}{DFt}$$

Như vậy giá trị IDF của một từ hiếm xuất hiện trong tất cả các tài liệu và xuất hiện nhiều trong một số các tài liệu (từ quan trọng) sẽ cao và ngược lại. Ta sẽ tính lại trọng số của từ t trong tài liệu d bằng công thức sau[7]:

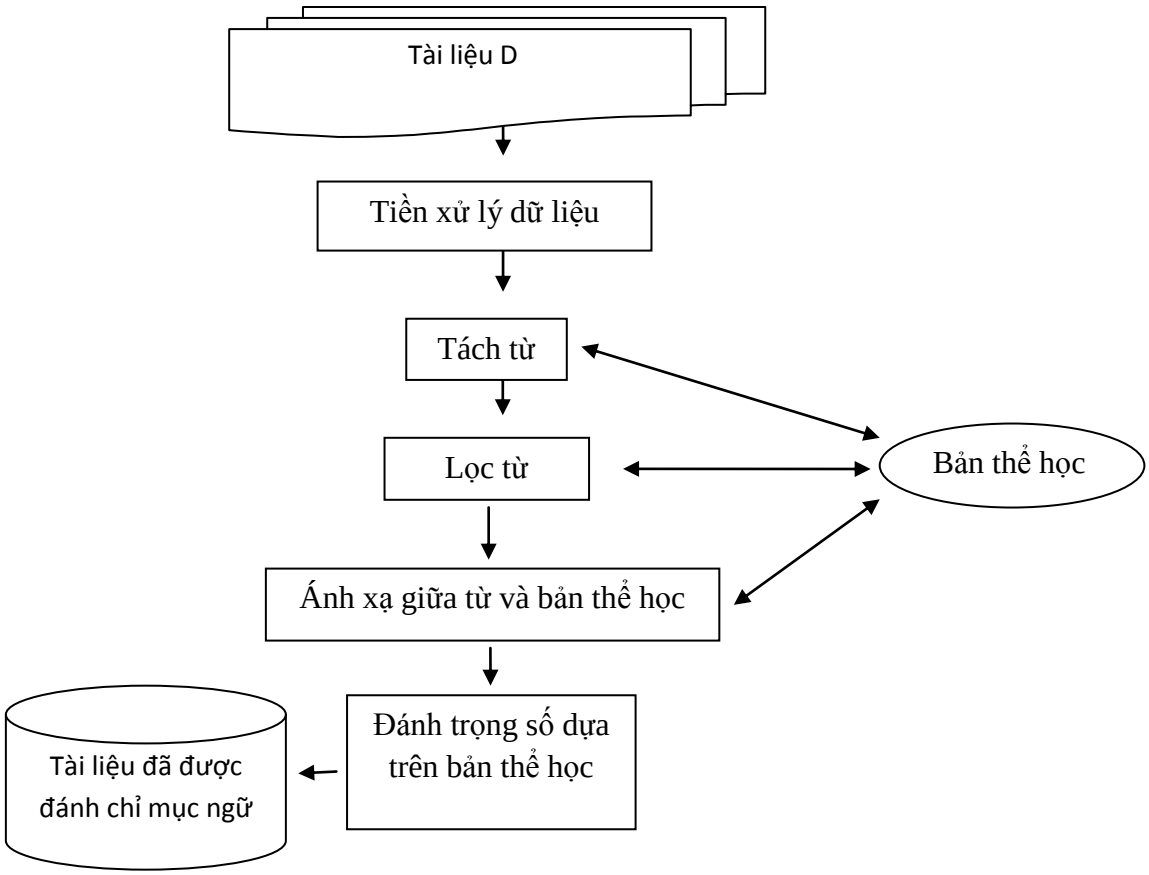
$$TF\text{-}IDF_{t,d} = TF_{t,d} * IDF_t$$

Trọng số của mỗi từ có ý nghĩa:

- ✓ Cao khi xuất hiện nhiều trong một số lượng ít tài liệu.
- ✓ Thấp khi xuất hiện một vài lần trong một số tài liệu.
- ✓ Thấp nhất khi xuất hiện nhiều lần trong tất cả các tài liệu.

2.2.1.2. Đánh chỉ mục ngữ nghĩa dựa trên bản thể học

Như đã trình bày ở phần 2.1. Bản thể học, trong bản thể học chúng ta sẽ định nghĩa các lớp, thuộc tính, thực thể và mối quan hệ giữa các thực thể. Quá trình đánh trọng số dựa trên bản thể học cũng tương đương với đánh trọng số dựa trên từ trong đó thay vì dùng từ thì chúng ta sẽ sử dụng bản thể học. Quá trình đánh chỉ mục ngữ nghĩa dựa trên bản thể học bao gồm các bước sau.



Hình 2.5. Quá trình đánh chỉ mục ngữ nghĩa dựa vào bản thể học[9]

1) **Xử lý dữ liệu**: Trong bước này, dữ liệu được xử lý như loại bỏ các ký tự đặc biệt, các thẻ HTML…

2) **Tách từ**: Trong bước này, dữ liệu sau khi được xử lý sẽ được phân tích, tách và gán nhãn để phân loại từ loại.

3) **Lọc từ**: Trong bước này, dữ liệu sau khi được tách và phân loại sẽ được lọc, dữ liệu được lấy là những từ nằm trong miền của bản thể học.

4) **Ánh xạ giữa từ và bản thể học**: Trong phần này dựa vào từ ta sẽ tìm được miền của nó trong bản thể học.

5) **Đánh trọng số dựa trên bản thể học**: Tính trọng số theo thuật toán TF-IDF, nhưng thay vì dựa vào từ thì ta sẽ dựa vào bản thể học.

2.2.2. Xử lý câu truy vấn và tìm kiếm

2.2.2.1. Xử lý câu truy vấn

Ta có thể coi câu truy vấn như một tài liệu và thực hiện xử lý câu truy vấn giống như xử lý tài liệu trong phần 2.2.1.2.

2.2.2.2. Tìm kiếm và phân hạng kết quả - Mô hình không gian vector.

Mô hình không gian vector là một mô hình đại số thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của từ trong tài liệu – chính là trọng số của từ trong tài liệu. Cách biểu diễn này không quan tâm đến thứ tự xuất hiện của từ mà chỉ quan tâm đến việc từ có xuất hiện hay không mà thôi.

Mỗi văn bản sẽ được biểu diễn bằng một vector một chiều của từ và trọng số. D (d1, d2, …., dn) với di là trọng số của từ thứ i trong văn bản D.

Tương tự ta câu truy vấn cũng được biểu diễn bằng một vector Q(q1, q2, …, qn) với qi là trọng số của từ i trong câu truy vấn.

Độ tương tự giữa văn bản và câu truy vấn được tính bằng độ đo cosin giữa chúng[7]:

$$\cos\theta_j = \frac{d_j^T q}{\|d_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} q_i^2} \sqrt{\sum_{i=1}^{m} d_{ij}^2}}$$

Công thức tính độ tương tự giữa câu truy vấn và tài liệu[7].

Trong tìm kiếm văn bản dựa vào từ khóa, độ tương tự giữa văn bản và câu truy vấn được tính dựa vào trọng số của từ,

Để áp dụng mô hình không gian vector cho tìm kiếm ngữ nghĩa, từ sẽ được thay bằng bản thể học tương ứng với nó, trọng số của từ sẽ được thay thế tương ứng với trọng số của bản thể học, giá trị này đã được tính toán trong phần 2.2.1.2.

Thứ hạng của kết quả trả về sẽ dựa vào giá trị của hàm cosin, giá trị cao sẽ được ưu tiên trả về đầu tiên cho người dùng.

Kết Luận Chương 2

Như vậy trong chương 2 đã trình bày về bản thể học, phương pháp tìm kiếm văn bản dựa vào mô hình không gian vector từ đó mở rộng ra, áp dụng mô hình không gian vector cho việc tìm kiếm ngữ nghĩa dựa vào bản thể học.

CHƯƠNG III: CÀI ĐẶT VÀ THỬ NGHIỆM HỆ THỐNG

Trong chương 3 sẽ trình bày quá trình xây dựng máy tìm kiếm ngữ nghĩa trong tiếng việt bằng cách áp dụng bản thể học, thuật toán TF-IDF và mô hình không gian vector.

3.1. Mô tả ứng dụng

Chương này sẽ xây dựng một ứng dụng tìm kiếm ngữ nghĩa sử dụng bản thể học. Ứng dụng này xây dựng cho việc tìm thông tin về máy tính xách tay. Dữ liệu được lấy từ trên mạng hoặc do người dùng thêm vào cơ sở dữ liệu, dữ liệu được lưu trữ trong cơ sở dữ liệu và được đánh chỉ mục dựa vào bản thể học đây chính là phần xử lý ngữ nghĩa của hệ thống

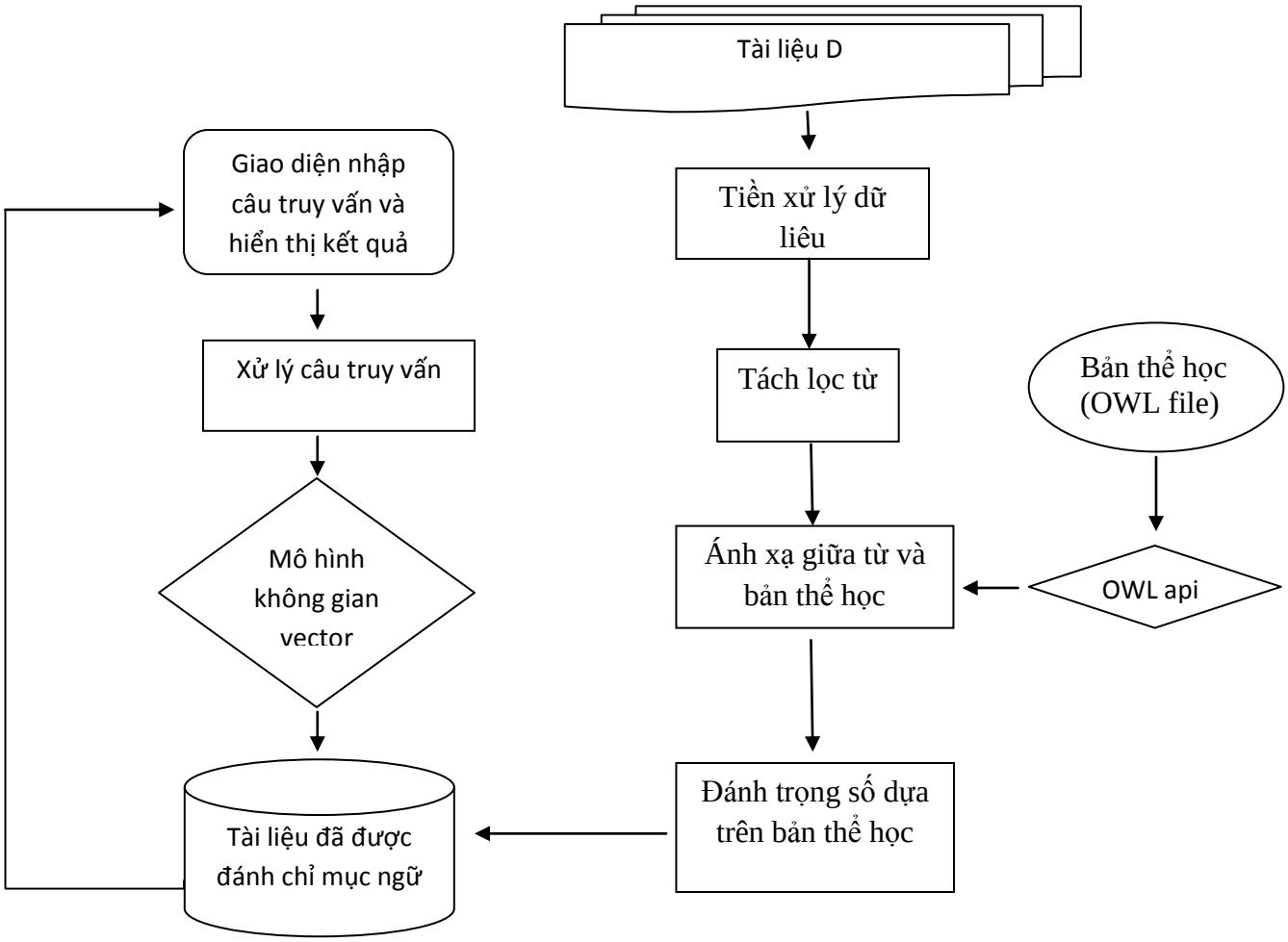
Ứng dụng được viết riêng cho việc tìm kiếm thông tin liên quan đến máy tính xách tay nhưng cũng có thể áp dụng dễ dàng cho nhiều lĩnh vực khác bằng cách xây dựng bản thể học tương ứng cho từng lĩnh vực.

3.2. Phân tích thiết kế hệ thống

3.2.1. Yêu cầu của hệ thống

Hệ thống tìm kiếm thông tin máy tính xách tay được xây dựng trên bản thể học, đáp ứng được tính chính xác cao khi người dùng thực hiện tìm kiếm.

3.2.2. Mô hình kiến trúc của hệ thống



Hình 3.1. Mô hình kiến trúc của hệ thống

Thông tin hoặc dữ liệu về máy tính được lưu trữ trong bộ nhỡ, dữ liệu này được đánh chỉ mục ngữ nghĩa để phục vụ cho việc xây dựng hệ thống tìm kiếm.

Quá trình đánh chỉ mục ngữ nghĩa được mô tả ngắn gọn như sau: dữ liệu tiếng Việt ban đầu sẽ được tách từ bằng cách sử dụng thư viện VietNamTagger [14], sau khi tách từ, tập từ sẽ được ánh xạ với bản thể học thông qua OWL Api để tạo chỉ mục ngữ nghĩa. Dữ liệu sau khi được đánh chỉ mục ngữ nghĩa sẽ được lưu vào trong cơ sở dữ liệu để phục vụ cho quá trình tìm kiếm.

3.2.3. Xây dựng các thành phần của hệ thống

3.2.3.1. Thiết kế bản thể học

Do ứng dụng là tìm kiếm thông tin về máy tính xách tay nên bản thể học sẽ được thiết kế cho lĩnh vực máy tính xáy tay. Đầu tiên để tạo bản thể học là xác định các lớp và mối quan hệ giữa chúng. Một máy tính xách tay sẽ bao gồm một số yếu tố sau: hãng sản xuất, model máy, năm sản xuất, cấu hình phần cứng bao gồm ram, ổ cứng, bộ vi xử lý.

Như vậy ta phải xây dựng được một bản thể học, có thể biểu diễn được máy tính xách tay với các thông tin cơ bản như trên. Thiết kế bản thể học càng chi tiết thì kết quả của quá trình tìm kiếm càng chính xác.

3.2.3.2. Quá trình đánh chỉ mục ngữ nghĩa

Quá trình đánh chỉ mục ngữ nghĩa là quá trình quan trọng nhất, quá trình này sẽ xử lý dữ liệu từ dạng không có cấu trúc sang dạng có cấu trúc. Trong ví dụ này chúng ta sẽ thực hiện lấy dữ liệu từ trang web http://www.pcworld.com.vn/ trong chuyên mục *SẢN PHẨM > Laptop*.

Quá trình đánh chỉ mục ngữ nghĩa cho mỗi tài liệu được mô tả như dưới đây.

1) ***Loại bỏ HTML Tag***: dữ liệu được lấy về từ website http://www.pcworld.com.vn/ dưới dạng HTML cần phải loại bỏ HTML Tag. Trong phần này sẽ sử dụng thư viện HTML parser.

2) ***Tách từ***: Dữ liệu đã được loại bỏ HTML Tag, cần phải được tách từ, như trong phần 1.2.1 đã trình bày về đặc trưng của tiếng việt. Từ trong tiếng Việt không được phân đoạn bằng các khoảng trắng như tiếng Anh hoặc một số

ngôn ngữ khác. Từ tiếng Việt có thể chỉ gồm một tiếng như: ăn, ngủ, nghỉ, nói… bên cạnh đó từ tiếng Việt cũng có thể bao gồm nhiều tiếng như:giúp đỡ, máy tính, nhu cầu, bộ vi xử lý… Do đó việc tách từ trong tiếng Việt gặp rất nhiều khó khăn. Trong phạm vi luận văn sẽ không để cập đến việc làm thế nào để tách từ trogn tiếng việt, ở đây luận văn sẽ sử dụng bộ thư viện vnTagger của tác giả Lê Hồng Phương[14], bộ thư viện đưa ra được kết quả tách từ chính xác trong khoảng 94% - 95%.

3) ***Lọc từ***: quá trình này sẽ loại bỏ những từ quá phổ biến, chung chung và tiến hành lọc từ để đánh chỉ mục, sử dụng thư viện OWL để truy xuất vào bản thể học và xác định những từ được lọc. Những từ được lọc ra là những từ nằm trong bản thể học

4) ***Đánh chỉ mục tìm kiếm với bản thể học***: mỗi từ sẽ tương ứng với một hoặc nhiều thực thể trong bản thể học

Từ	Thực thể trong bản thể học
Máy tính xách tay	E1, E3, E4
Lenovo	E1, E5, E7
…..	….
Core i3	E7, E8

5) ***Tính số lần xuất hiện của thực thể*** trong bản thể học trong mỗi tài liệu

Thực thể trong bản thể học	Tài liệu
E1	D1(2), D4(4)
E3	D1(4)
…..	….
E5	D10(7), D12(9)

6) Sử dụng thuật toán TF-IDF để tính toán trọng số của bản thể học cho mỗi tài liệu

Thực thể trong bản thể học	Tài liệu	Weight
E1	D1	0.74
E3	D1	0.76
….	…	…

3.3. Cài đặt và đánh giá kết quả

3.3.1. Cài đặt hệ thống

3.3.2. Kết quả

TÀI LIỆU THAM KHẢO

[1]. Bonino, D., Corno, F., Farinetti, L., & Bosca, A. (2004). Ontology driven semantic search. *WSEAS Transaction on Information Science and Application*, *1*(6), 1597-1605.

[2]. Horridge, M. (2009). A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition1. 2. *The University Of Manchester*.

[3]. Kassim, J. M., & Rahmany, M. (2009, August). Introduction to semantic search engine. In *Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on* (Vol. 2, pp. 380-386). IEEE.

[4]. Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *Software, IEEE*, *14*(2), 67-75.

[5]. Lukasiewicz, T., Fazzinga, B., Gianforme, G., & Gottlob, G. (2012). Semantic Web Search Based on Ontological Conjunctive Queries. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(4).

[6]. Manh Hung Nguyen and Tan Hiep Nguyen. (2013). *Towards a Semantic Search Mechanism based on Query Expansion*

[7]. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 6). Cambridge: Cambridge university press.

[8]. Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

[9]. Sánchez, M. F. (2009). *Semantically enhanced Information Retrieval: an ontology-based approach* (Doctoral dissertation, Doctoral dissertation. Unitversidad de Autónoma, Madrid).

[10]. Singto, P., & Mingkhwan, A. (2013). Semantic Searching IT Careers Concepts Based on Ontology. *Journal of Advanced Management Science*, *1*(1).

[11]. Swartout, B., Patil, R., Knight, K., & Russ, T. (1996, November). Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*.

[12]. Wei, W., Barnaghi, P. M., & Bargiela, A. (2008). Search with meanings: an overview of semantic search systems. *Int. J. Communications of SIWN*, *3*, 76-82.

[13]. http://www.vietlex.com/ngon-ngu-hoc/11-Dac_diem_tieng_Viet

[14]. http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger