

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

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành tới thầy giáo, TS Nguyễn Mạnh Hùng, người đã tận tình hướng dẫn, động viên và giúp đỡ em thực hiện luận văn ngay từ những bước đầu tiên đến khi hoàn thành. Em xin trân trọng cảm ơn các thầy cô giáo trong khoa Quốc tế & Đào tạo sau đại học, Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện học tập thuận lợi nhất cho em trong suốt 2 năm qua.

Em xin gửi lời cảm ơn tới gia đình và bạn bè, những người luôn ở bên cạnh động viên, ủng hộ, khích lệ mỗi khi em gặp khó khăn.

Do thời gian có hạn và vốn kiến thức còn ít ỏi, luận văn chắc chắn không thể tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp của thầy cô và các bạn để luận văn này được hoàn thiện hơn.

Hà Nội, tháng 11 năm 2014

Học viên thực hiện

Hoàng Anh

LỜI CAM ĐOAN

Tôi xin cam đoan:

Những nội dung trong luận văn này là do tôi thực hiện. Mọi tham khảo dùng trong luận văn đều được trích dẫn rõ ràng và trung thực về tên tác giả, tên công trình, thời gian và địa điểm công bố.

Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo, hay gian trá, tôi xin chịu hoàn toàn trách nhiệm.

Tác giả luận văn ký và ghi rõ họ tên

Hoàng Anh

MỤC LỤC

LỜI CẢM ƠN.....	i
LỜI CAM ĐOAN.....	ii
MỤC LỤC	iii
DANH MỤC CHỮ VIẾT TẮT	vi
DANH MỤC CÁC BẢNG.....	vii
DANH MỤC HÌNH VẼ	vii
MỞ ĐẦU	1
CHƯƠNG I: GIỚI THIỆU BÀI TOÁN MÁY TÌM KIẾM NGŨ NGHĨA CHO TIẾNG VIỆT	3
1.1. Máy tìm kiếm và máy tìm kiếm ngữ nghĩa.....	3
1.1.1. Máy tìm kiếm	3
1.1.2. Máy tìm kiếm ngữ nghĩa	6
1.2. Tìm kiếm ngữ nghĩa trong tiếng Việt	8
1.2.1. Đặc trưng của tiếng Việt	8
1.3. Một số phương pháp xử lý ngữ nghĩa trong máy tìm kiếm	10
1.4. Giải pháp đề xuất của luận văn	11
1.5. Kết luận	11
CHƯƠNG II: XỬ LÝ NGŨ NGHĨA TRONG MÁY TÌM KIẾM	13
2.1. Bản thể học.....	13
2.1.1. Định nghĩa bản thể học	13
2.1.2. Các thành phần của bản thể học	13
2.1.3. Ngôn ngữ biểu diễn bản thể học.....	15

2.1.4. Cách xây dựng Bản thể học.....	19
2.1.5. Công cụ phát triển bản thể học.....	22
2.2. Tìm kiếm ngữ nghĩa dựa trên bản thể học	23
2.2.1. Đánh chỉ mục ngữ nghĩa dựa trên bản thể học.....	23
2.2.2. Xử lý câu truy vấn và tìm kiếm.....	27
2.3. Kết Luận.....	28
CHƯƠNG III: CÀI ĐẶT VÀ THỬ NGHIỆM HỆ THỐNG.....	29
3.1. Mô tả ứng dụng	29
3.2. Phân tích hệ thống.....	29
3.2.1. Yêu cầu của hệ thống.....	29
3.2.2. Quá trình đánh chỉ mục ngữ nghĩa và tìm kiếm	30
3.3. Xây dựng các thành phần của hệ thống	31
3.3.1. Thiết kế bản thể học.....	31
3.3.2. Xây dựng tập dữ liệu	34
3.3.2. Quá trình đánh chỉ mục ngữ nghĩa	35
3.3. Thiết kế hệ thống.....	38
3.3.1. Kiến trúc hệ thống.....	38
3.3.2. Các module của hệ thống.....	39
3.3.3. Công cụ phát triển.....	41
3.3.4. Giao diện chương trình	41
3.4. Kết quả	45
3.5. Kết luận	47
KẾT LUẬN	48

TÀI LIỆU THAM KHẢO	49
--------------------------	----

DANH MỤC CHỮ VIẾT TẮT

Chữ viết tắt	Ý nghĩa
TF-IDF: term frequency-inverse document frequency	Trọng số của từ trong tài liệu

DANH MỤC CÁC BẢNG

Bảng 1. 2 Danh sách các công cụ phát triển <i>bản thể học</i>	23
---	----

DANH MỤC HÌNH VẼ

Hình 1. 1 Quá trình tìm kiếm tài liệu [7]	4
Hình 1. 2 Quá trình xử lý ngữ nghĩa và tìm kiếm [7]	7
 Hình 2. 1 Thể hiện mối quan hệ cha con trong bản thể học	15
Hình 2. 2 Mô tả lớp trong bản thể học (bao gồm cả thực thể) [2]	16
Hình 2. 3 Mối quan hệ giữa các thực thể [2]	16
Hình 2. 4 Các thực thể trong bản thể học [2]	19
Hình 2. 5 Quá trình đánh chỉ mục ngữ nghĩa dựa vào bản thể học [7]	26
 Hình 3. 1 Quá trình đánh chỉ mục nghĩa và tìm kiếm	30
Hình 3. 2 Bản thể học trong miền máy tính xách tay	34
Hình 3. 3 Bảng Document - lưu tài liệu trong cơ sở dữ liệu.	35
Hình 3. 4 Mô hình kiến trúc của hệ thống	38

MỞ ĐẦU

Hiện nay, cùng với sự phát triển của mạng máy tính, nguồn thông tin chia sẻ trên mạng ngày càng phong phú và đa dạng cùng với đó là nhu cầu tìm kiếm thông tin người dùng ngày càng tăng.

Đối với các hệ thống tìm kiếm theo từ khóa truyền thống, khi người dùng nhập từ khóa tìm kiếm, kết quả trả về đơn thuần chỉ là những nội dung chứa từ khóa trong câu truy vấn. Do máy tìm kiếm không hiểu được người dùng nên kết quả trả về từ máy tìm kiếm thường có thêm những thông tin không liên quan đến thông tin mà người dùng mong muốn. Từ đó xuất hiện nhu cầu xây dựng một máy tìm kiếm hiểu được từ khóa của người dùng nhập vào và đưa ra kết quả chính xác cho người dùng.

Xử lý ngữ nghĩa trong quá trình tìm kiếm đã ra đời nhằm cải thiện tìm kiếm theo phương pháp tìm kiếm truyền thống bằng cách dựa vào từ khóa của người dùng nhập vào và từ đó phân tích và hiểu vấn đề của người dùng đang tìm kiếm, từ đó đưa ra kết quả chính xác cho người dùng.

Luận văn tập trung nghiên cứu về xử lý ngữ nghĩa trong máy tìm kiếm bằng cách tìm hiểu về kiến trúc tổng quan của một máy tìm kiếm ngữ nghĩa, quá trình xử lý ngữ nghĩa trong máy tìm kiếm từ đó xây dựng một ứng dụng tìm kiếm ngữ nghĩa.

Nội dung luận văn bao gồm 3 chương:

Chương I: Giới thiệu bài toán máy tìm kiếm ngữ nghĩa

Trong chương này, luận văn sẽ đi tìm hiểu về máy tìm kiếm, máy tìm kiếm ngữ nghĩa và quá trình xử lý ngữ nghĩa của máy tìm kiếm ngữ nghĩa, đặc điểm của tiếng Việt và tìm kiếm ngữ nghĩa trong tiếng Việt.

Chương II: Xử lý ngữ nghĩa trong máy tìm kiếm

Chương này nghiên cứu phương pháp tìm kiếm ngữ nghĩa dựa trên bản thể học bằng cách đi tìm hiểu về bản thể học, các thành phần, ngôn ngữ biểu diễn và cách xây dựng một bản thể học. Xử lý ngữ nghĩa trong máy tìm kiếm ngữ nghĩa có hai quá trình không thể thiếu đó là đánh chỉ mục ngữ nghĩa, xử lý câu truy vấn và đưa ra kết quả tìm kiếm. Trong chương này luận văn sẽ trình bày quá trình đánh chỉ mục ngữ nghĩa dựa trên bản thể học bằng các sử dụng mô hình không gian Vector.

Chương III: Cài đặt và thử nghiệm hệ thống

Cuối cùng, sau khi tìm hiểu về: quá trình xử lý ngữ nghĩa, vấn đề khi xử lý ngữ nghĩa trong tiếng Việt, cách xây dựng bản thể học, cách đánh chỉ mục ngữ nghĩa, xử lý câu truy vấn và tìm kiếm trong chương I và II. Chương III sẽ đi xây dựng và cài đặt thử nghiệm hệ thống.

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

Trong chương I, luận văn sẽ trình bày tổng quan về máy tìm kiếm, máy tìm kiếm ngữ nghĩa, một số vấn đề của tìm kiếm ngữ nghĩa trong Tiếng Việt, các vấn đề liên quan và giải pháp đề xuất của luận văn.

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.

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 [3].

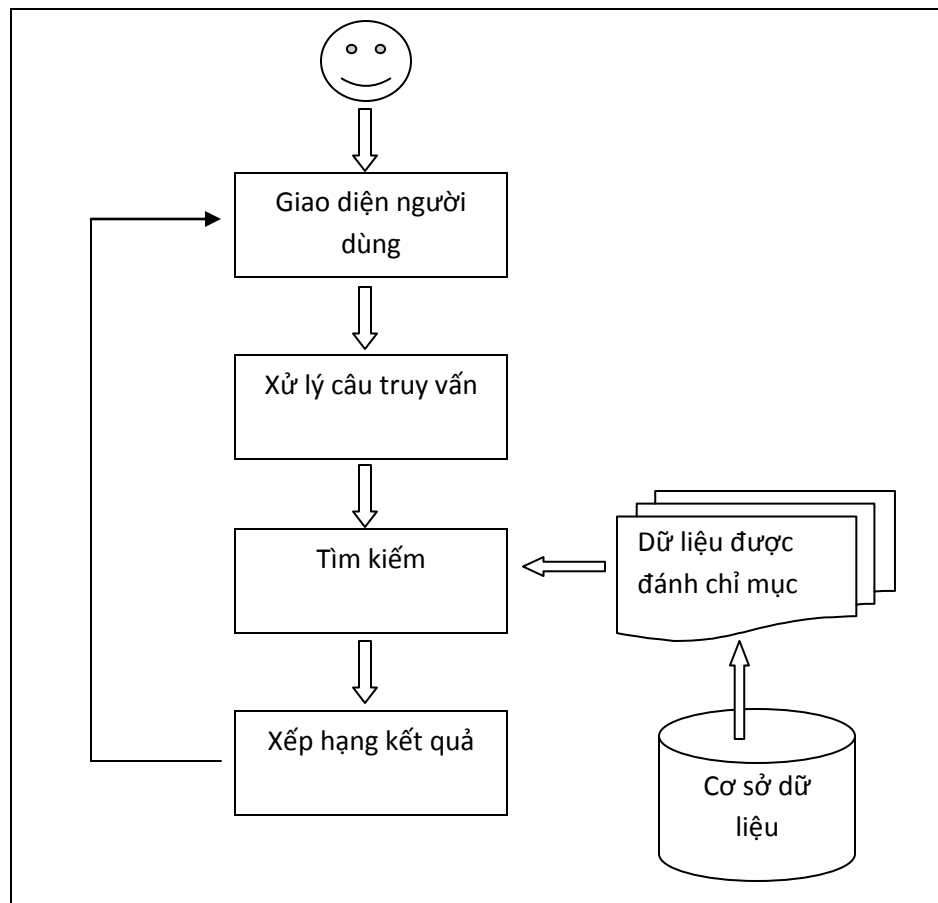
- ✓ Một cơ sở dữ liệu chứa các tài liệu trên Internet.
- ✓ Một máy tìm kiếm hoạt động trên cơ sở dữ liệu đó.
- ✓ Một tập các chương trình có nhiệm vụ xác định làm thế nào để kết quả tìm kiếm được hiển thị.

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 yếu tố:

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

Quá trình xử lý và tìm kiếm thông tin được mô tả như hình dưới đây.



Hình 1. 1 Quá trình tìm kiếm tài liệu [7]

Dữ liệu đầu vào của quá trình tìm kiếm là thông tin mà người dùng nhập vào từ giao diện dưới dạng ký tự, trong một số trường hợp máy tìm kiếm hỗ trợ thông tin dạng ảnh, audio, video...

Dữ liệu đầu ra là những tài liệu mà hệ thống tìm được thích hợp với thông tin mà người dùng nhập vào từ giao diện. Những tài liệu trả ra đã được sắp xếp dựa trên sự phù hợp giữa tài liệu và thông tin truy vấn của người dùng nhập vào.

Quá trình xử lý bao gồm 3 phần chính: trích chọn những thông tin quan trọng trong tài liệu (*đánh chỉ mục*), xử lý thông tin người dùng nhập vào từ giao diện và biểu diễn theo một cách nào đây để phục vụ quá trình tìm kiếm (*xử lý câu truy vấn*), so sánh giữa câu truy vấn và tài liệu được đánh chỉ mục từ đó tìm những tài liệu thích hợp và trả về cho người dùng (*tìm kiếm và xếp hạng kết quả*).

Đánh chỉ mục: Không phải tất cả thông tin trong mỗi tài liệu đều hữu ích như nhau. Trong văn viết có một vài từ sẽ mang nghĩa hơn những từ khác, bởi vậy những từ này thường được xem xét và lựa chọn để đánh chỉ mục cho tài liệu. Đánh chỉ mục tài liệu nhằm mục đích tăng tốc độ cho quá trình tìm kiếm.

Xử lý câu truy vấn: Câu truy vấn của người dùng được xử lý trước khi tiến hành tìm kiếm. Trong trường hợp là tìm kiếm dữ liệu kiểu text, câu truy vấn thường được xử lý giống như tài liệu tách và chọn những từ mang nghĩa hơn những từ khác. Thêm vào đó có thể mở rộng câu truy bằng các sử dụng từ điển để tìm từ đồng nghĩa.

Tìm kiếm: Câu truy vấn của người dùng được đem so sánh với các tài liệu, từ đó một tập các tài liệu thích hợp với câu truy vấn sẽ được trả về.

Xếp hạng kết quả: Quá trình tìm kiếm trả về một tập các tài liệu, nhưng không phải tất cả các tài liệu đều thích hợp với người dùng. Quá trình xếp hạng kết quả sẽ đánh giá một tài liệu như thế nào là thích hợp với câu truy vấn của người dùng hơn những tài liệu khác. Từ đó tài liệu trả về cho người dùng được sắp xếp theo thứ tự giảm dần độ thích hợp giữa câu truy vấn và tài liệu.

Như vậy quá trình tìm kiếm và xếp hạng kết quả có thể được xem như là phần chính của một máy tìm kiếm, quá trình này quyết định đến hiệu năng của hệ thống.

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 thường bao gồm các quá trình sau:

- ✓ 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.

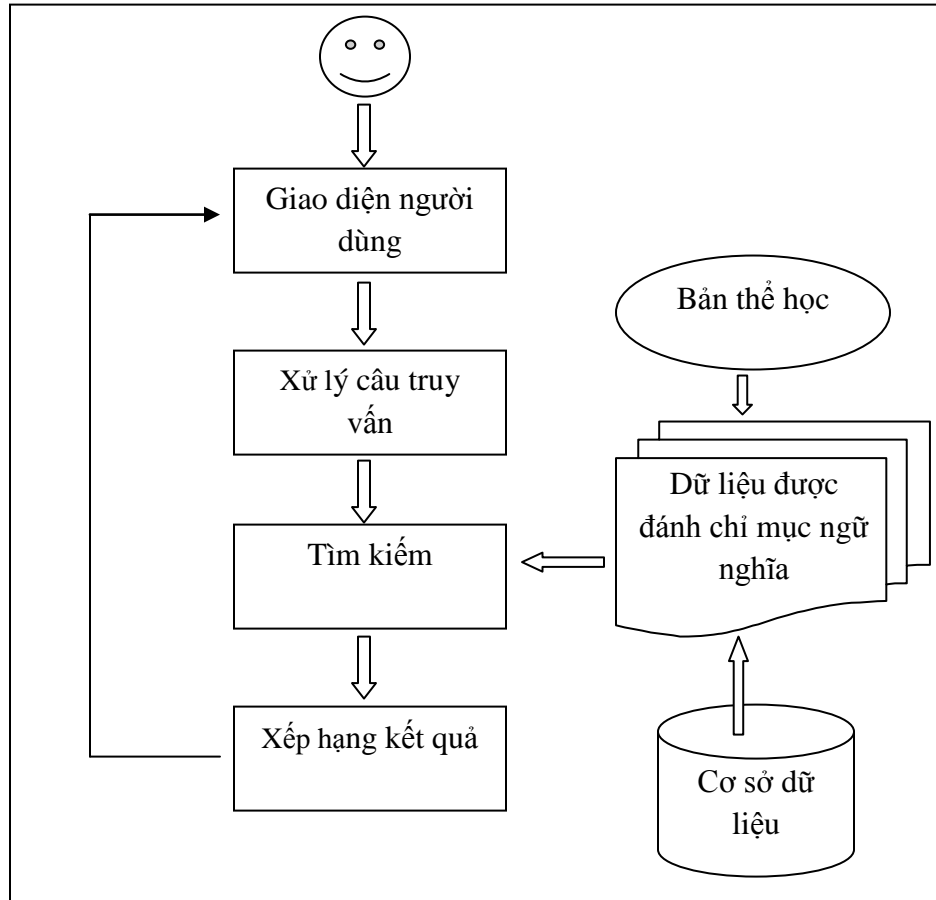
Một máy tìm kiếm ngữ nghĩa về cơ bản cũng có cấu trúc giống như máy tìm kiếm thông thường bao gồm hai phần chính

❖ **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 tả quá trình xử lý ngữ nghĩa và tìm kiếm



Hình 1. 2 Quá trình xử lý ngữ nghĩa và tìm kiếm [7]

Quá trình xử lý ngữ nghĩa và tìm kiếm trong máy tìm kiếm ngữ nghĩa bao gồm 4 bước:

- ✓ 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, quá trình xử lý câu truy vấn và đánh chỉ mục đều dựa vào bản thể học. Về chi tiết luận văn sẽ trình bày 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 [9] tiếng Việt có một số đặc trưng về ngữ âm, từ vựng, ngữ pháp như sau:

a) 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.

b) 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 lấy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng, chẳng hạn: chòm chia, chông chơ, đồng đa đồng đánh, thơ thân, 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 - kỹ thuật, đặc biệt là công nghệ thông tin, thì tiềm năng đó còn được phát huy mạnh mẽ hơn.

c) 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 dựa bản thể học bằng cách đánh chỉ mục ngữ nghĩa dựa vào độ tương tự của từ. Máy tìm kiếm sẽ xem xét và tự động tìm kiếm những từ thích hợp bằng cách dựa trên cây khái niệm được xây dựng trong bản thể học. Trong nghiên cứu này tác giả cũng đề cập đến việc đánh chỉ mục ngữ nghĩa dựa vào bản thể học, bản thể học sẽ xây dựng một cây, đưa ra các quan hệ giữa các lớp với nhau. Từ đó hệ thống có thể đánh chỉ mục và có thể tìm kiếm được.

Trong nghiên cứu [4], tác giả đã trình bày một công cụ tìm kiếm ngữ nghĩa, công cụ này được dựa trên các công cụ tìm kiếm thông thường và thêm vào một tầng xử lý ngữ nghĩa. Tầng xử lý ngữ nghĩa này có nhiệm vụ phân tích lại câu truy vấn và đánh giá lại kết quả trả về cho người dùng. Đầu tiên khi người dùng nhập từ khóa tìm kiếm, hệ thống

sẽ tìm ra một tập các từ khóa có quan hệ ngữ nghĩa với từ khóa ban đầu dựa trên các quan hệ của chúng được định nghĩa trong WordNet. Tập từ khóa này sẽ được cho vào một số công cụ tìm kiếm như Google, Yahoo, ... và sau đó xếp hạng lại kết quả trả về từ những công cụ tìm kiếm thông thường dựa trên mối quan hệ của tài liệu với từ khóa. Như vậy trong nghiên cứu [4], bằng cách thêm vào tầng xử lý ngữ nghĩa tác giả đã hạn chế được những kết quả tìm kiếm không liên quan với từ khóa đầu vào của hệ thống tìm kiếm hiện tại – hệ thống tìm kiếm dựa vào sự phù hợp giữa câu truy vấn và tài liệu

1.4. Giải pháp đề xuất của luận văn

Nghiên cứu [1] dựa vào bản thể học để đánh chỉ mục ngữ nghĩa, quá trình đánh chỉ mục dựa vào độ tương tự giữa bản thể học, bản thể học ở đây sẽ định nghĩa các khái niệm. Để xây dựng được một bản thể học có thể đưa ra được mức độ tương tự của từ là khá phức tạp, khó có thể áp dụng được trong lĩnh vực lớn.

Trong nghiên cứu [4] tác giả đã đưa ra phương pháp xử lý ngữ nghĩa cho máy tìm kiếm bằng cách thêm vào một tầng xử lý ngữ nghĩa và đánh giá lại kết quả tìm kiếm từ một số máy tìm kiếm phổ biến hiện nay, giới hạn của phương pháp này là phụ thuộc vào công cụ tìm kiếm hiện tại.

Luận văn sẽ tìm hiểu về bản thể học, phương pháp đánh chỉ mục ngữ nghĩa dựa vào bản thể học từ đó xây dựng một máy tìm kiếm dựa trên bản thể học.

1.5. Kết luận

Như vậy, trong **Chương I** luận văn đã đi tìm hiểu về máy tìm kiếm. Một máy tìm kiếm thông thường bao gồm ba phần chính: đánh chỉ mục, xử lý câu truy vấn, tìm kiếm và xếp hạng kết quả. Máy tìm kiếm ngữ nghĩa về cơ bản cũng giống như máy tìm kiếm thông thường nhưng được thêm vào phần xử lý ngữ nghĩa cho tài liệu và cho câu truy vấn của người dùng trước khi tìm kiếm.

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. Đây chính là vấn đề khó khăn trong quá trình lọc từ để đánh chỉ mục cho tài liệu tiếng Việt, vấn đề này sẽ được giải quyết trong **Chương II**.

CHƯƠNG II: XỬ LÝ NGỮ NGHĨA TRONG MÁY TÌM KIẾM

Trong chương này luận văn đi vào tìm hiểu về bản thể học, cách xây dựng bản thể học, phương pháp chỉ mục ngữ nghĩa dựa trên bản thể học và mô hình không gian Vector.

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 [8], các thuật ngữ giữ vai trò quan trọng trong bản thể học: “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 đều có tên và giá trị của thuộc tính đó. Các thuộc tính được sử dụng để lưu trữ các thông tin mà đối tượng

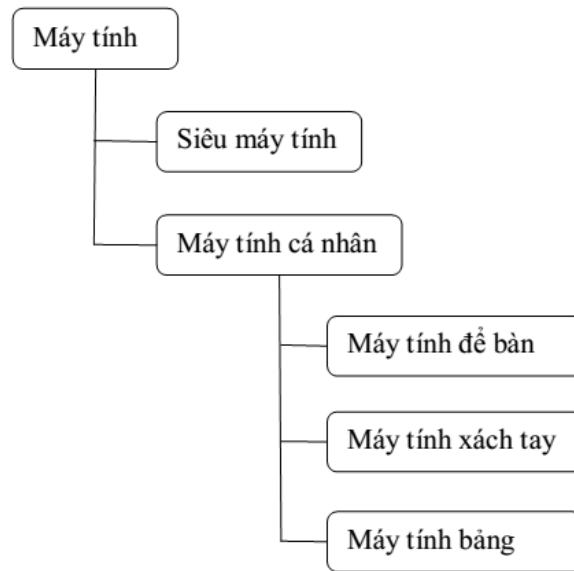
có thể có. Ví dụ, một cá nhân có thể có các thuộc tính như họ tên, ngày sinh, quê quán, số CMND... Giá trị của một thuộc tính có thể là kiểu dữ liệu phức tạp.

❖ *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.



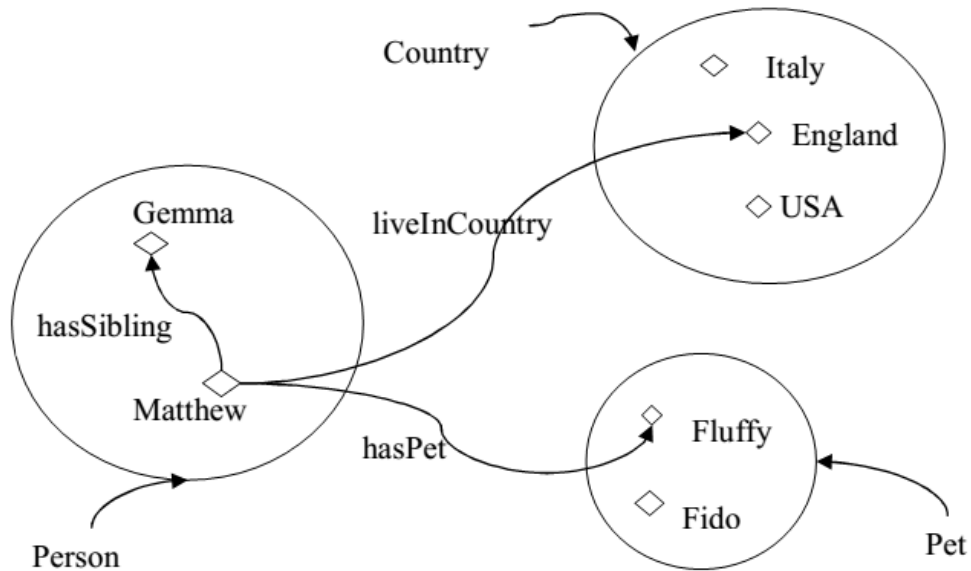
Hình 2. 1 Thể hiện mối quan hệ cha con trong bản thể học

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.

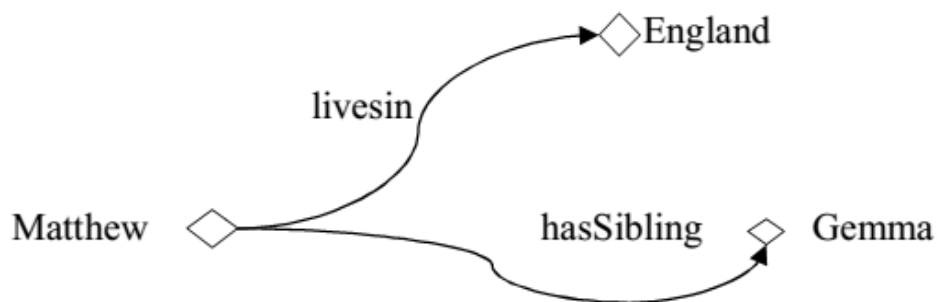
owl: Class - là một nhóm các thực thể có liên quan. Lớp có thể được xây dựng trong một hệ thống phân cấp bằng các sử dụng `subClassOf`. Thing là lớp của tất cả các thực thể và là lớp cha của tất cả các lớp OWL.



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.

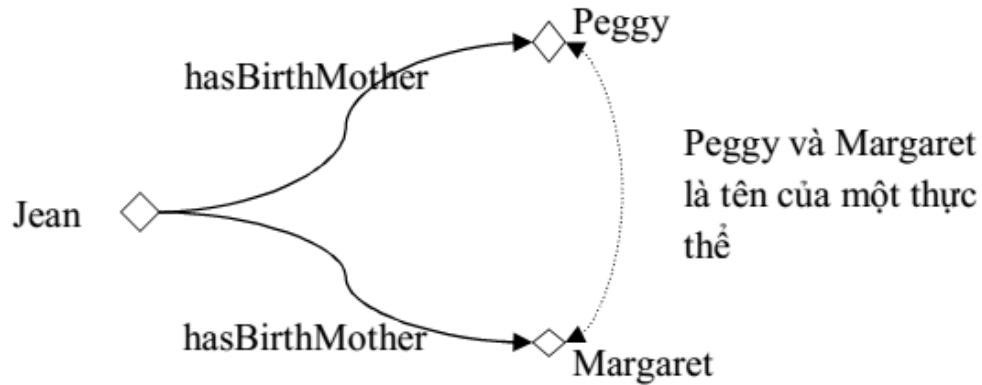
owl: ObjectProperty - là thuộc tính mô tả mối quan hệ giữa hai thực thể. Ví dụ, thuộc tính *hasSibling* liên kết thực thể Matthew với thực thể Gemma.



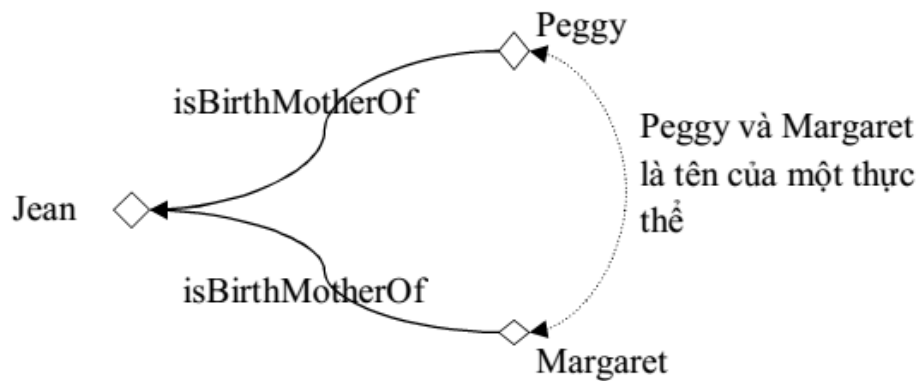
Hình 2. 3 Mối quan hệ giữa các thực thể [2]

ObjectProperty có 4 tính chất sau:

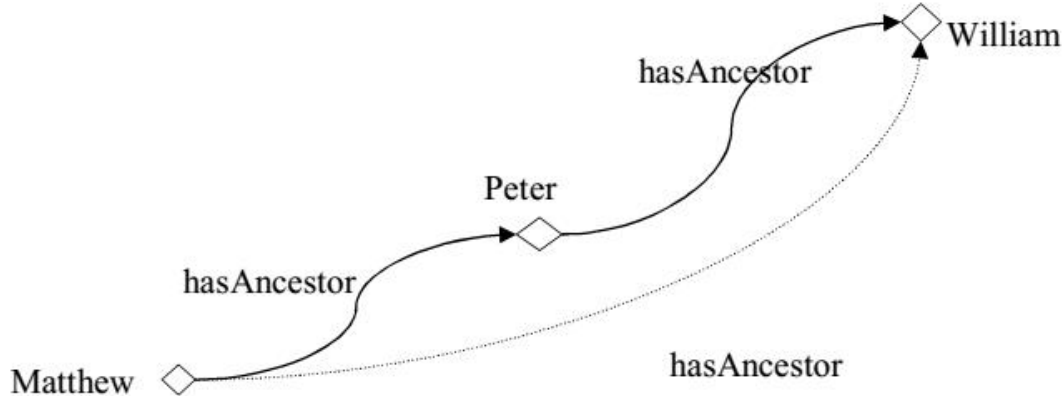
- ✓ Functional: một thực thể chỉ liên quan nhiều nhất đến một thực thể khác



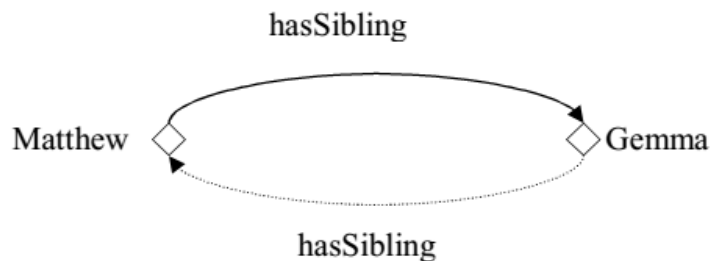
- ✓ Inverse Functional: là thuộc tính đảo ngược của Functional.



- ✓ Transitive: thực thể a quan hệ với thực thể b, thực thể b quan hệ với thực thể c \rightarrow thực thể a quan hệ với thực thể c.



- ✓ Symmetric: thực thể a quan hệ với thực thể b \rightarrow thực thể b quan hệ với thực thể a



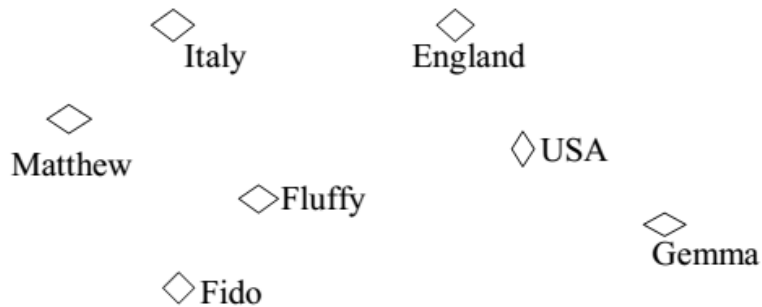
owl: DatatypeProperty - mô tả mối quan hệ giữa thực thể và giá trị của nó. Ví dụ, thuộc tính *hasAge* có thể được sử dụng để chỉ quan hệ một thể hiện của lớp người với một thể hiện của kiểu dữ liệu số nguyên.

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:

a) Xác định miền quan tâm và phạm vi của *bản thể học*

Trong giai đoạn này cần xác định mục đích của việc xây dựng bản thể học là gì? Phục vụ đối tượng nào? Bản thể học sắp xây dựng cần có đặc điểm gì, liên quan đến lĩnh vực, phạm vi nào. Quá trình khai thác, quản lý và bảo trì bản thể học được thực hiện ra sao?

b) Xem xét việc kế thừa các *bản thể học* có sẵn

Cấu trúc của một Bản thể học bao gồm 3 tầng: tầng trừu tượng (Abstract), tầng miền xác định (Domain) và tầng mở rộng (Extension). Trong đó tầng trừu tượng có tính tái sử dụng rất cao, tầng miền xác định có thể tái sử dụng trong một lĩnh vực nhất định. Hiện nay có rất nhiều bản thể học đã được tạo ra, với tâm huyết của nhiều chuyên gia. Do đó trước khi bắt đầu xây dựng bản thể học, cần xét đến khả năng sử dụng lại các bản thể học đã có. Nếu có thể sử dụng lại một phần các bản thể học đã có, chi phí bỏ ra cho quá trình xây dựng bản thể học sẽ giảm đi rất nhiều.

c) Liệt kê các thuật ngữ quan trọng trong *bản thể học*

Bản thể học được xây dựng trên cơ sở các khái niệm trong một lĩnh vực cụ thể, vì vậy khi xây dựng bản thể học cần bắt đầu từ các thuật ngữ chuyên ngành để xây dựng thành các lớp trong bản thể học tương ứng. Tất nhiên không phải thuật ngữ nào cũng đưa vào bản thể học, vì chưa chắc đã định vị được cho thuật ngữ đó. Do đó cần phải liệt kê các thuật ngữ, để xác định ngữ nghĩa cho các thuật ngữ đó, cũng như cân nhắc về phạm vi của bản thể học. Việc liệt kê các thuật ngữ còn cho thấy được phần nào tổng quan về các khái niệm trong lĩnh vực đó, giúp cho các bước tiếp theo được thuận lợi.

d) Xây dựng các lớp và cấu trúc lớp phân cấp

Công việc xác định các lớp không chỉ đơn giản là tiến hành tìm hiểu về ngữ nghĩa của các thuật ngữ đã có để có được các mô tả cho thuật ngữ đó, mà còn phải định vị cho các lớp mới, loại bỏ ra khỏi bản thể học nếu nằm ngoài phạm vi của bản thể học hay hợp nhất với các lớp đã có nếu có nhiều thuật ngữ có ngữ nghĩa như nhau (đồng nghĩa, hay đa ngôn ngữ). Ngoài ra không phải thuật ngữ nào cũng mang tính chất như một lớp.

Một công việc cần phải tiến hành song song với việc xác định các lớp là xác định phân cấp của các lớp đó. Việc này giúp định vị các lớp dễ dàng hơn.

Có một số phương pháp tiếp cận trong việc xác định phân cấp của các lớp:

- ✓ *Phương pháp từ trên xuống (top-down)*: bắt đầu với định nghĩa của các lớp tổng quát nhất trong lĩnh vực và sau đó chuyên biệt hóa các khái niệm đó. Ví dụ: Trong bản thể học về quản lý nhân sự, ta bắt đầu với lớp Người, sau đó chuyên biệt hóa lớp Người đó bằng cách tạo ra các lớp con của lớp Người như : Kỹ sư, Công nhân, Bác sỹ,... Lớp Kỹ sư cũng có thể chuyên biệt hóa bằng cách tạo ra các lớp con như Kỹ sư CNTT, Kỹ sư điện, Kỹ sư cơ khí, ...
- ✓ *Phương pháp từ dưới lên (bottom-up)*: bắt đầu với định nghĩa của các lớp cụ thể nhất, như các lá trong cây phân cấp. Sau đó gộp các lớp đó lại thành các khái niệm tổng quát hơn. Ví dụ: ta bắt đầu với việc định nghĩa các lớp như: nhân viên lễ tân, nhân viên vệ sinh, nhân viên kỹ thuật. Sau đó tạo ra một lớp chung hơn cho các lớp đó là lớp nhân viên.
- ✓ *Phương pháp kết hợp*: kết hợp giữa phương pháp từ trên xuống và từ dưới lên: bắt đầu từ định nghĩa các lớp dễ thấy trước và sau đó tổng quát hóa và chuyên biệt hóa các lớp đó một cách thích hợp. Ví dụ ta bắt đầu với lớp nhân viên trước, là thuật ngữ hay gặp nhất trong quản lý nhân sự. Sau đó chúng ta có thể chuyên biệt hóa thành các lớp con: nhân viên lễ tân, nhân viên vệ sinh,... hoặc tổng quát hóa lên thành lớp Người.

e) Định nghĩa các thuộc tính và quan hệ cho lớp

Để xác định thuộc tính cho các lớp, ta quay trở lại danh sách các thuật ngữ đã liệt kê được. Hầu hết các thuật ngữ còn lại (sau khi đã xác định lớp) là thuộc tính của các lớp đó. Với mỗi thuộc tính tìm được, ta phải xác định xem nó mô tả cho lớp nào. Các thuộc tính đó sẽ trở thành thuộc tính của các lớp xác định. Ví dụ lớp Người có các thuộc tính sau: Họ, Tên, Ngày sinh, Giới tính, Nghề nghiệp, Địa chỉ, Điện thoại,...

f) Định nghĩa các ràng buộc về thuộc tính và quan hệ của lớp

Các thuộc tính có thể có nhiều khía cạnh khác nhau: như kiểu giá trị, các giá trị cho phép, số các thuộc tính (lực lượng), và các đặc trưng khác mà giá trị của thuộc tính có thể nhận. Ví dụ: “Năm sinh” của một “nhân viên” chỉ có duy nhất và là số nguyên, có thể

nhận giá trị từ 1948 đến 1990. Cần phải xác định các ràng buộc cho một thuộc tính càng chặt chẽ càng tốt, để tránh trường hợp nhập dữ liệu sai, dẫn đến đổ vỡ của các ứng dụng sử dụng bản thể học này.

g) Tạo các thực thể cho lớp

Bước cuối cùng là tạo ra các thể hiện của các lớp trong sự phân cấp. Việc tạo thể hiện cho một lớp là quá trình điền các thông tin vào các thuộc tính của lớp đó.

2.1.5. Công cụ phát triển bản thể học

Một số công cụ phát triển và hiệu chỉnh có giá trị trong việc làm giảm độ phức tạp và thời gian dùng cho nhiệm vụ xây dựng ontology. Các công cụ như Kaon, OileEd và Protégé cung cấp các giao diện nhằm giúp đỡ người sử dụng thực hiện các hoạt động chính yếu trong quá trình phát triển một ontology. Việc lựa chọn một công cụ hiệu chỉnh phù hợp nhất có nhiều khó khăn vì mỗi kiểu ontology có các yêu cầu về kinh phí, thời gian, tài nguyên khác nhau. Để giúp cho việc giải quyết vướng mắc này, Singh & Murshed (2005) đưa ra các tiêu chuẩn đánh giá công cụ tạo ontology. Tiêu chuẩn bao gồm tính năng, khả năng sử dụng lại, lưu trữ dữ liệu, mức độ phức tạp, quan hệ, tính lâu bền, độ an toàn, độ chắc chắn, khả năng học, tính khả dụng, hiệu lực, và tính rõ ràng. Protégé và OntoEditFree được phát triển bởi Singh & Murshed sử dụng các tiêu chuẩn này.

Một số công cụ phát triển *bản thể học* phổ biến hiện nay:

Developer	Product	Availability	Language Support
FZI – AIFB http://kaon.semanticWeb.org/frontpage	KAON 1.2.7	Open source	KAON RDF(S)
IMG (University of Manchester) http://oiled.man.ac.uk/index.shtml	OilEd 3.5	Open source	RDF(S) OIL DAML+OIL OWL SHIQ
Ontoprise http://www.ontoprise.de/content/e3/e43/index_eng.html	Ontostudio 1.4	Freeware Licenses	RDF(S) OWL F-Logic OXML
SMI (Stanford University) http://protege.stanford.edu/	Protégé 3.2	Open source	XML RDF(S) XML Schema OWL
KMI (Open University) http://kmi.open.ac.uk/projects/Webonto/	WebOnto 2.3	Free access	OCML RDF(S)
Mindswap http://www.mindswap.org/2004/SWOOP/	Swoop 2.3	Open Source	RDF(S) OWL

Bảng 1. 1 Danh sách các công cụ phát triển *bản thể học*

Protégé hỗ trợ OWL, là công cụ được sử dụng rộng rãi và lâu nhất hiện nay. Nó cho phép người sử dụng định nghĩa và chỉnh sửa các lớp ontology, các thuộc tính và quan hệ và các thể hiện sử dụng cấu trúc cây. Các ontology có thể được đưa ra theo các định dạng RDF(S), XML Schema. Platform protégé cung cấp hai cách chính mô hình hóa ontology thông qua Protégé - Frame và Protégé – OWL, ngoài ra có thể có nhiều plugin. Chúng ta có thể quan sát một cách trực quan thông qua OWL Viz, nó cho phép quan sát ontology dưới dạng đồ họa và đưa file ảnh JPEG. Ngoài ra còn hỗ trợ truy vấn SPARQL.

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

mối quan hệ của từ đó với những từ khác. 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.

a) **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 [5].

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

- ✓ D_1 : “lập trình viên Java”
- ✓ D_2 : “lập trình viên PHP”

Chúng ta tính toán trọng số cho mỗi từ trong D_1, D_2

	lập	trình	viên	Java	PHP
D_1	1	1	1	1	0
D_2	1	1	1	0	1

Tính trọng số dựa vào tần suất 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 [5]:

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

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 [5]:

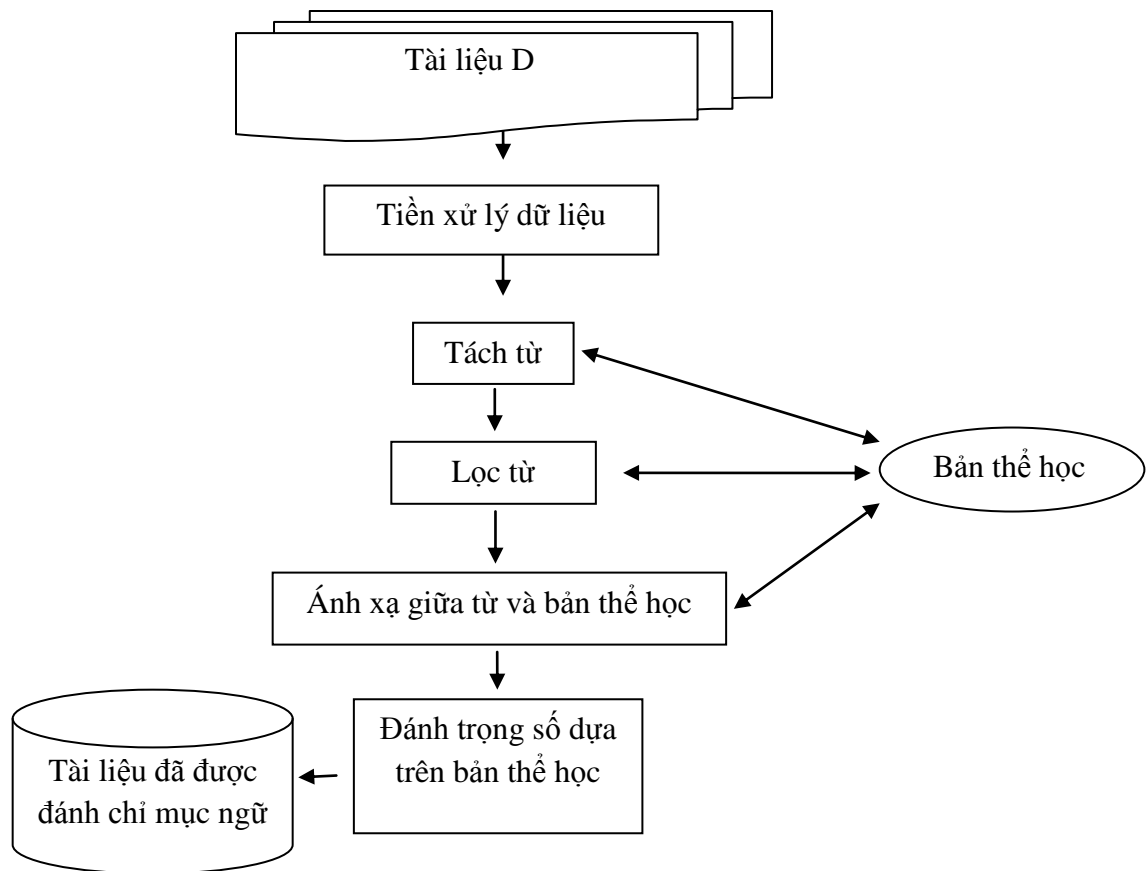
$$TF-IDF_{t,d} = TF_{t,d} * IDF_t$$

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.

b) 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 [7]

- 1) **Tiền 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

a) 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.b.

b) 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. $D (d_1, d_2, \dots, d_n)$ với d_i là trọng số của từ thứ i trong văn bản D .

Tương tự ta câu truy vấn cũng được biểu diễn bằng một vector $Q(q_1, q_2, \dots, q_n)$ với q_i 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 [5]:

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

Công thức tính độ tương tự giữa câu truy vấn và tài liệu [5].

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.b.

Thứ hạng của kết quả trả về sẽ dựa vào giá trị của hàm cosin, giá trị cao có nghĩa là mức độ tương tự giữa câu truy vấn và tài liệu lớn. Kết quả trả về sẽ được sắp xếp giảm dần dựa vào giá trị hàm cosin.

2.3. Kết Luận

Như vậy trong **chương II** đã trình bày về bản thể học, các thành phần và phương pháp xây dựng bản thể học. Dựa vào bản thể học để đánh trọng số cho tài liệu và 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 III** sẽ trình bày quá trình xây dựng máy tìm kiếm ngữ nghĩa trong tiếng Việt bằng cách áp dụng bản thể học, thuật toán đánh trọng số TF-IDF và mô hình không gian vector.

3.1. Mô tả ứng dụng

Hiện nay, nhu cầu tìm kiếm thông tin của người dùng Internet là vô cùng lớn, nếu không có công cụ tìm kiếm thì người dùng Internet sẽ không biết làm thế nào để tìm được thông tin cần thiết trong hàng triệu, hàng tỷ thông tin trên mạng Internet.

Trong chương này luận văn sẽ xây dựng một ứng dụng tìm kiếm ngữ nghĩa sử dụng bản thể học. Ứng dụng phục vụ cho việc tìm thông tin về máy tính xách tay trên cơ sở dữ liệu cục bộ của hệ thống. Dữ liệu sẽ được lấy từ trên mạng Internet hoặc do người dùng thêm vào trong cơ sở dữ liệu. Dữ liệu được lưu trong cơ sở dữ liệu sẽ qua quá trình xử lý và được đánh chỉ mục ngữ nghĩa dựa vào bản thể học. Đây chính là phần xử lý ngữ nghĩa của hệ thống.

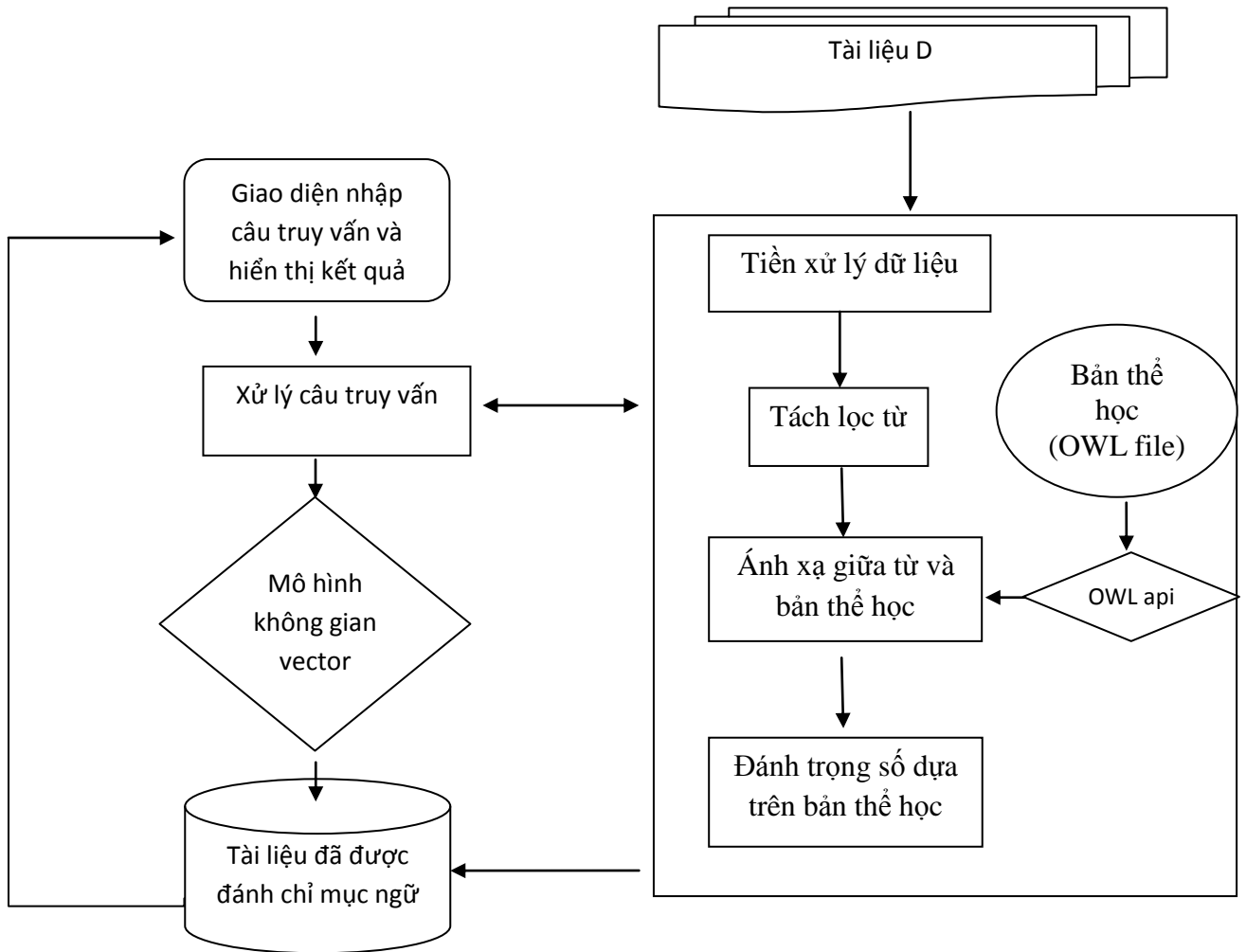
Ứng dụng được viết riêng cho việc tìm kiếm thông tin liên quan đến máy tính xách tay nhưng cũng có thể áp dụng dễ dàng cho nhiều lĩnh vực khác bằng cách xây dựng bản thể học tương ứng cho từng lĩnh vực.

3.2. Phân tích hệ thống

3.2.1. Yêu cầu của hệ thống

Hệ thống tìm kiếm thông tin máy tính xách tay được cho phép người dùng tìm kiếm thông tin về máy tính xách tay trong hệ thống và hệ thống phải đáp ứng được tính chính xác cao khi người dùng thực hiện tìm kiếm.

3.2.2. Quá trình đánh chỉ mục ngữ nghĩa và tìm kiếm



Hình 3. 1 Quá trình đánh chỉ mục ngữ nghĩa và tìm kiếm

Thông tin hoặc dữ liệu về máy tính xách tay được lưu trữ trong bộ nhớ, 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 [10], sau khi tách từ, tập từ sẽ được ánh xạ với bản thể học thông qua OWL API để lấy những bản thể học có quan hệ với từ, từ đó tạo chỉ mục ngữ nghĩa dựa trên bản thể học. Dữ liệu sau khi được đánh chỉ mục ngữ nghĩa sẽ được lưu vào bộ nhớ để phục vụ cho quá trình tìm kiếm.

3.3. Xây dựng các thành phần của hệ thống

3.3.1. 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 xách 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.

Bản thể học cho miền máy tính xách tay là một hệ thống phân cấp ngữ nghĩa các thuật ngữ trong lĩnh vực máy tính xách tay cùng với những mô tả về các mối quan hệ giữa các thực thể hoặc các lớp. Để xây dựng được bản thể học cho miền máy tính xách tay ta cần phải đi tìm hiểu xem một đối tượng máy tính xách tay có những đặc điểm gì, có những loại máy tính xách tay nào, thành phần của nó ra sao và mối quan hệ giữa các thành phần.

Cách thức xây dựng bản thể học cho miền máy tính xách tay:

- Bước 1: Xác định những thuật ngữ thường được dùng trong lĩnh vực máy tính xách tay như “Máy tính xách tay”, “Phần cứng”, “Màn hình”, “Phần mềm”, “Bộ xử lý”, “Hệ điều hành”,... mỗi thuật ngữ này sẽ tương ứng với một lớp trong bản thể học.

- Bước 2: Xây dựng hệ thống cây phân cấp, xác định lớp nào là lớp con, lớp nào là lớp cha. Ví dụ như lớp “Máy tính” là cha của lớp “Máy tính cá nhân”, “Máy tính cá nhân” là cha của lớp “Máy tính để bàn”, “Máy tính xách tay”.

- Bước 3: Tạo thực thể cho lớp máy tính xách tay như sản phẩm “Macbook Pro” thuộc lớp “Máy tính xách tay”.

- Bước 4: Xác định mối quan hệ giữa các thực thể và các lớp với nhau ví dụ “Macbook Pro” là sản phẩm của “Apple”. Như vậy thực thể “Macbook Pro” có quan hệ is_product_of với thực thể “Apple”.

Về lý thuyết, người xây dựng và phát triển bản thể học có thể không cần các công cụ hỗ trợ, có thể thực hiện trực tiếp bằng các ngôn ngữ. Tuy nhiên, sẽ không khả thi khi bản thể học có kích thước lớn và cấu trúc phức tạp. Thêm vào đó, việc xây dựng bản thể học không chỉ đòi hỏi việc tạo cấu trúc lớp phân cấp, định nghĩa các thuộc tính, ràng buộc... mà còn bao hàm việc giải quyết các bài toán liên quan đến nó như:

- ✓ Kiểm tra tính đúng đắn và đầy đủ
- ✓ Suy luận trên bản thể học
- ✓ Xóa, sửa và tinh chỉnh các thành phần bên trong bản thể học
- ✓ Tách biệt bản thể học với ngôn ngữ sử dụng (DAML, OWL...).

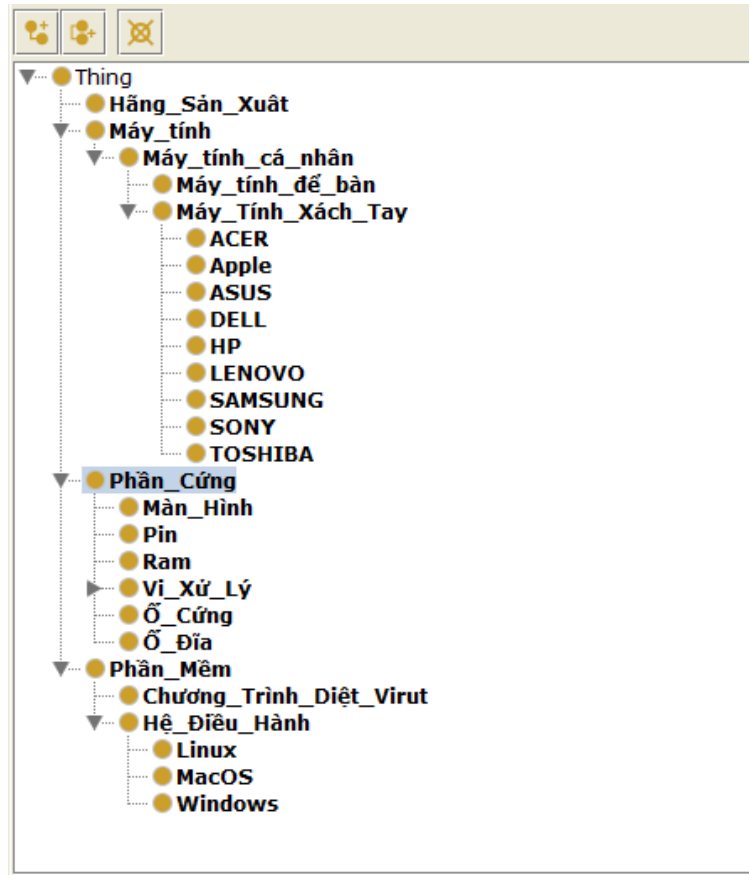
Những khó khăn trên đã khiến các công cụ trở thành một phần không thể thiếu và là yếu tố quyết định đến chất lượng của một hệ thống bản thể học. Chính vì điều này, tác giả khi xây dựng bản thể học chủ đề máy tính sẽ sử dụng công cụ Protégé. Protégé là bộ phần mềm mã nguồn mở Java nổi tiếng. Protégé được nghiên cứu và phát triển bởi nhóm nghiên cứu của Mark Musen, ĐH. Stanford.

Các ưu điểm của Protégé là:

- ✓ Hỗ trợ đầy đủ ba phiên bản của ngôn ngữ OWL là OWL-Full, OWL-Lite và OWL-DL.

- ✓ Nhờ sử dụng mô hình hướng đối tượng của ngôn ngữ Java, Protégé tỏ ra hiệu quả trong việc mô hình các lớp, thực thể, quan hệ...
- ✓ Giao diện thiết kế trực quan có tính tương tác cao. Người sử dụng có thể định nghĩa các thành phần của bản thể học trực tiếp từ các form.
- ✓ Cho phép biểu diễn trực quan bản thể học dưới dạng các sơ đồ.
- ✓ Cho phép xây dựng bản thể học từ nhiều nguồn khác nhau.
- ✓ Cung cấp chức năng tìm kiếm lỗi, kiểm tra tính nhất quán và đầy đủ của bản thể học.
- ✓ Hỗ trợ khả năng suy luận trực tiếp trên bản thể học.
- ✓ Hỗ trợ sinh mã tự động. Protégé cho phép chuyển bản thể học thành mã nguồn RDF/XML, OWL, DIG...

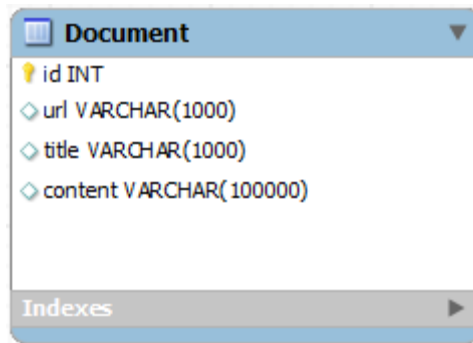
Kết quả xây dựng bản thể học trong miền máy tính xách tay



Hình 3. 2 Bản thể học trong miền máy tính xách tay

3.3.2. Xây dựng tập dữ liệu

Như đã trình bày ở phần 3.1. *Mô tả ứng dụng*, hệ thống chỉ cho phép tìm kiếm những thông tin máy tính xách tay, dữ liệu được lấy từ những bài viết về máy tính xách tay trong websites <http://www.pcworld.com.vn/> và <http://vnexpress.net/> sẽ được lưu vào cơ sở dữ liệu cục bộ để phục vụ cho quá trình đánh chỉ mục ngữ nghĩa. Dữ liệu được lấy về dựa vào thư viện mã nguồn mở HTML parser [11] của Java và được lưu vào cơ sở dữ liệu.



Hình 3. 3 Bảng Document - lưu tài liệu trong cơ sở dữ liệu.

Dữ liệu được lấy về từ hai website <http://www.pcworld.com.vn/> và <http://vnexpress.net/> có khoảng 2000 tài liệu. Những tài liệu này sẽ được dùng cho quá trình đánh chỉ mục ngữ nghĩa để phục vụ cho quá trình tìm kiếm.

3.3.2. 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 để cho phép hệ thống có thể tìm kiếm được trên tập dữ liệu.

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[15]. HTML parser là thư viện mã nguồn mở của Java cho phép phân tích trang HTML và lấy dữ liệu cần thiết trong các thẻ HTML.

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 đề cập đến việc làm thế nào để tách từ trong tiếng Việt, ở đây luận văn sẽ sử dụng bộ thư viện vnTagger của tác giả Lê Hồng Phương[10], bộ thư viện đưa ra được kết quả tách từ chính xác trong khoảng 94% - 95%.

Lọc từ: quá trình này sẽ loại bỏ những từ quá phổ biến, chung chung, không mang nghĩa và tiến hành lọc từ để đánh chỉ mục, sử dụng thư viện OWL[12] để truy xuất vào bản thể học để lấy những thực thể tương ứng trong bản thể học. Những từ được lọc ra là những từ nằm trong bản thể học.

Trong quá trình tìm hiểu về dữ liệu máy tính xách tay, các tài liệu có sử dụng một số thuật ngữ viết tắt hoặc sử dụng từ nguyên gốc tiếng anh. Khi gặp từ viết tắt hoặc từ tiếng anh hệ thống không thể xác định được bản thể học tương ứng. Do vậy cần phải xây dựng bộ từ điển cho phép lấy từ nguyên gốc từ từ viết tắt hoặc từ tiếng anh. Từ điển được xây dựng dựa trên định dạng XML có dạng như sau:

```
<dictionary>
  <word word="MTXT">máy tính xách tay</word>
  <word word="Laptop">máy tính xách tay</word>
</dictionary>
```

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	E ₁ , E ₃ , E ₄
Lenovo	E ₁ , E ₅ , E ₇
.....
CoreI3	E ₇ , E ₈

Tính số lần xuất hiện của thực thể trong mỗi tài liệu

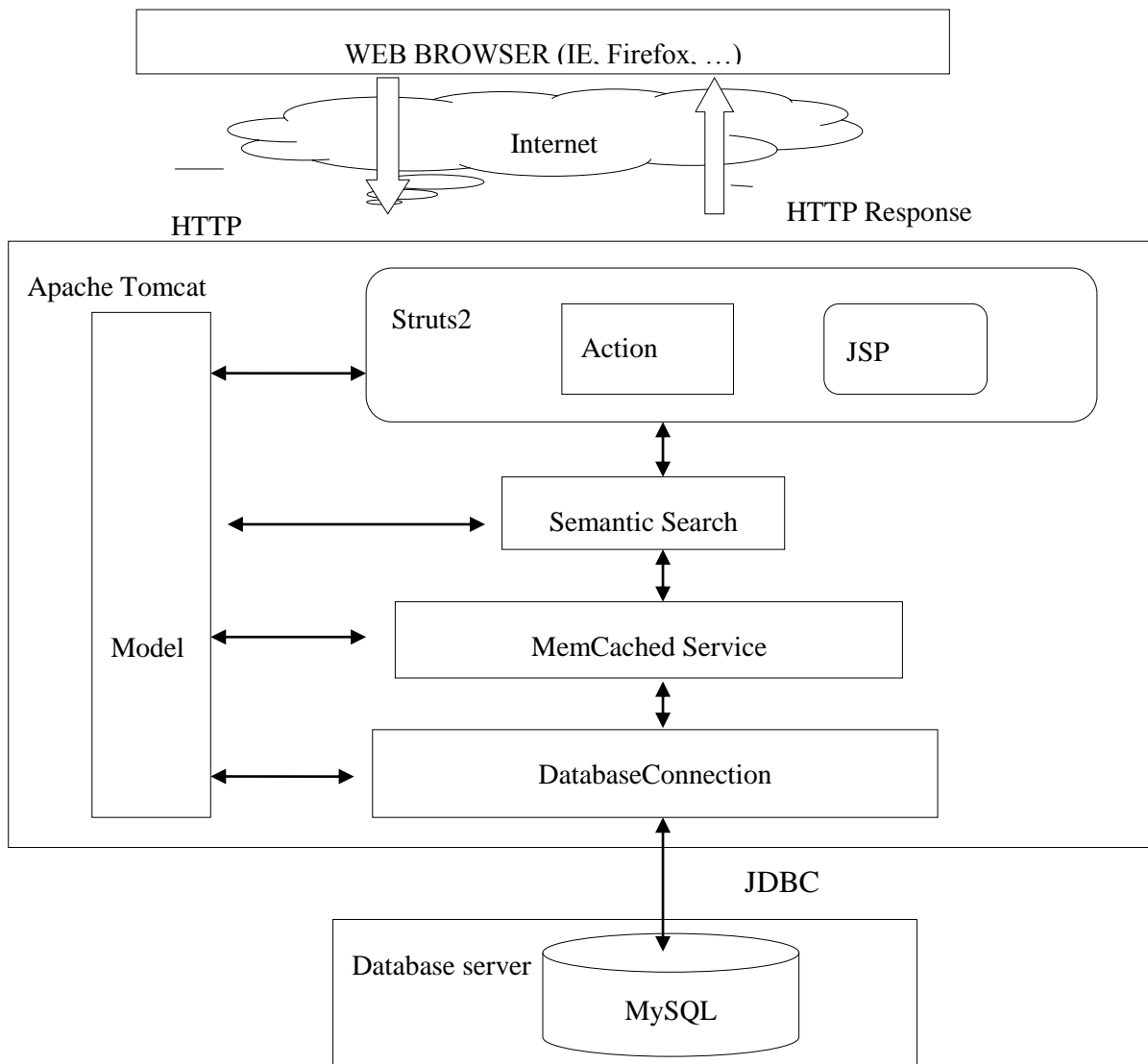
Thực thể trong bản thể học	Tài liệu
E_1	$D_1(2), D_4(4)$
E_3	$D_1(4)$
.....
E_5	$D_{10}(7), D_{12}(9)$

3) **Tính trọng số của thực thể trong mỗi tài liệu** bằng các sử dụng thuật toán TF-IDF

Thực thể trong bản thể học	Tài liệu	Trọng số
E_1	D_1	0.74
E_3	D_2	0.76
....

3.3. Thiết kế hệ thống

3.3.1. Kiến trúc hệ thống



Hình 3. 4 Mô hình kiến trúc của hệ thống

Mô tả các thành phần hệ thống:

Database server: là cơ sở dữ liệu của hệ thống, chứa tập các tài liệu, đây là tầng lưu dữ liệu chưa qua xử lý.

DatabaseConnection: Phần này chịu trách nhiệm kết nối giữa ứng dụng và cơ sở dữ liệu. Định nghĩa một số phương thức cho phép ứng dụng truy xuất dữ liệu.

Memcached Service: Phần này lưu trữ tài liệu đã được đánh chỉ mục, để phục vụ cho quá trình tìm kiếm. Quá trình ứng dụng mở kết nối đến cơ sở dữ liệu để lấy tài liệu có thể chiếm khá nhiều thời gian, lưu tài liệu đã được đánh chỉ mục vào bộ nhớ memcached sẽ đẩy nhanh quá trình tìm kiếm.

Semantic Search: bao gồm các lớp và phương thức được cài đặt để đánh trọng số của tài liệu dựa vào bản thể học, tìm kiếm tài liệu dựa vào bản thể học và mô hình không gian Vector.

Action: định nghĩa các phương thức điều khiển và xử lý các hoạt động trên trang **JSP**.

Model: bao gồm 2 lớp là Word để định nghĩa từ, và Document để định nghĩa tài liệu.

3.3.2. Các module của hệ thống

- ❖ Module lấy dữ liệu từ website <http://www.pcworld.com.vn/> và <http://vnexpress.net/> bao gồm 2 packages.

Package: com.search.crawler

- ✓ PCWorld.java
- ✓ VNExpress.java

Package: com.search.crawler.utils

- ✓ Entry.java
- ✓ HTMLNode.java
- ✓ HTMLNodeFilter.java
- ✓ HTMLParser.java
- ✓ HTMLParserUtil.java
- ✓ HTMLUtil.java
- ✓ Image.java
- ✓ ImageUtil.java

- ✓ SimpleHtmlOption.java

- ✓ StringUtility.java

- ✓ URLLoader.java

- ❖ Module truy xuất cơ sở dữ liệu: module này cung cấp các phương thức cho phép truy xuất cơ sở dữ liệu.

Package: com.search.dao

- ✓ DataAccessManager.java

- ❖ Module truy xuất bản thể học: module này sử dụng thư viện OWL để cung cấp các API cho phép truy xuất bản thể học.

Package: com.search.ontology

- ✓ OntologyManager.java

- ❖ Các Model chứa thông tin về tài liệu và từ

Package: com.search.model

- ✓ Document.java

- ✓ Word.java

- ❖ Module tính trọng số của từ trong tài liệu và tìm kiếm tài liệu

Package: com.search.tfidf

- ✓ TFIDF.java

- ✓ SearchDocument.java

- ❖ Module chứa dữ liệu sau khi được đánh chỉ mục ngữ nghĩa,

Package: com.search.cached

- ✓ CachedService.java

- ✓ SemantichSearchCached.java

- ❖ Module giao diện, cho phép nhập câu truy vấn và hiển thị kết quả trả về

Package: WebContent.WEB-INF.jsp

- ✓ normalDetails.jsp

- ✓ normalSearch.jsp

- ✓ notFound.jsp

- ✓ semanticDetails.jsp
- ✓ semanticSearch.jsp
- ❖ Module điều khiển các thao tác trên giao diện,
 - Package: com.search.action*
 - ✓ BaseAction.java
 - ✓ SearchAction.java

Như vậy các thành phần trong mô hình kiến trúc sẽ tương ứng với các module như sau:

- ✓ **DatabaseConnection** là module truy xuất cơ sở dữ liệu.
- ✓ **Memcached Service** là module chứa dữ liệu sau khi được đánh chỉ mục ngữ nghĩa
- ✓ **Semantic Search** bao gồm ba module: module truy xuất bản thể học, module tính trọng số của từ trong tài liệu và module tìm kiếm tài liệu.
- ✓ **Action và JSP** bao gồm hai module: module giao diện và module điều khiển các thao tác trên giao diện.

3.3.3. Công cụ phát triển

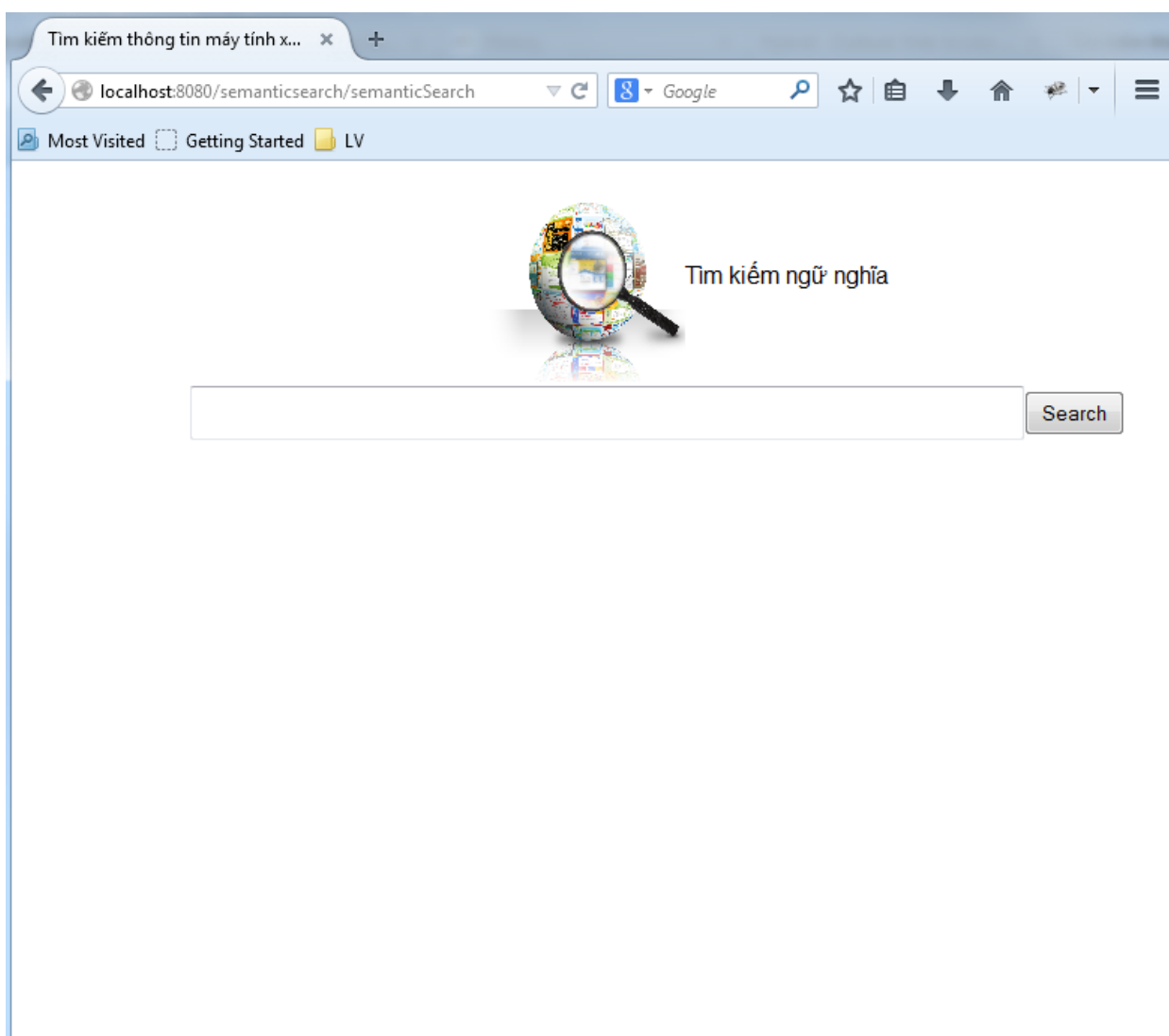
Hệ thống được phát triển dựa trên

- ✓ Hệ điều hành: *Môi trường Windows 7 - 64bit*
- ✓ Công cụ lập trình: *Eclipse Java EE IDE for Web Developers.*
- ✓ Ngôn ngữ lập trình: *Java*
- ✓ Hệ quản trị cơ sở dữ liệu: *MySQL.*
- ✓ Server: *Apache-tomcat-7.0.52*
- ✓ Bộ nhớ cached: *Memcached*

3.3.4. Giao diện chương trình

Máy tìm kiếm ngữ nghĩa bao gồm 3 màn hình chính

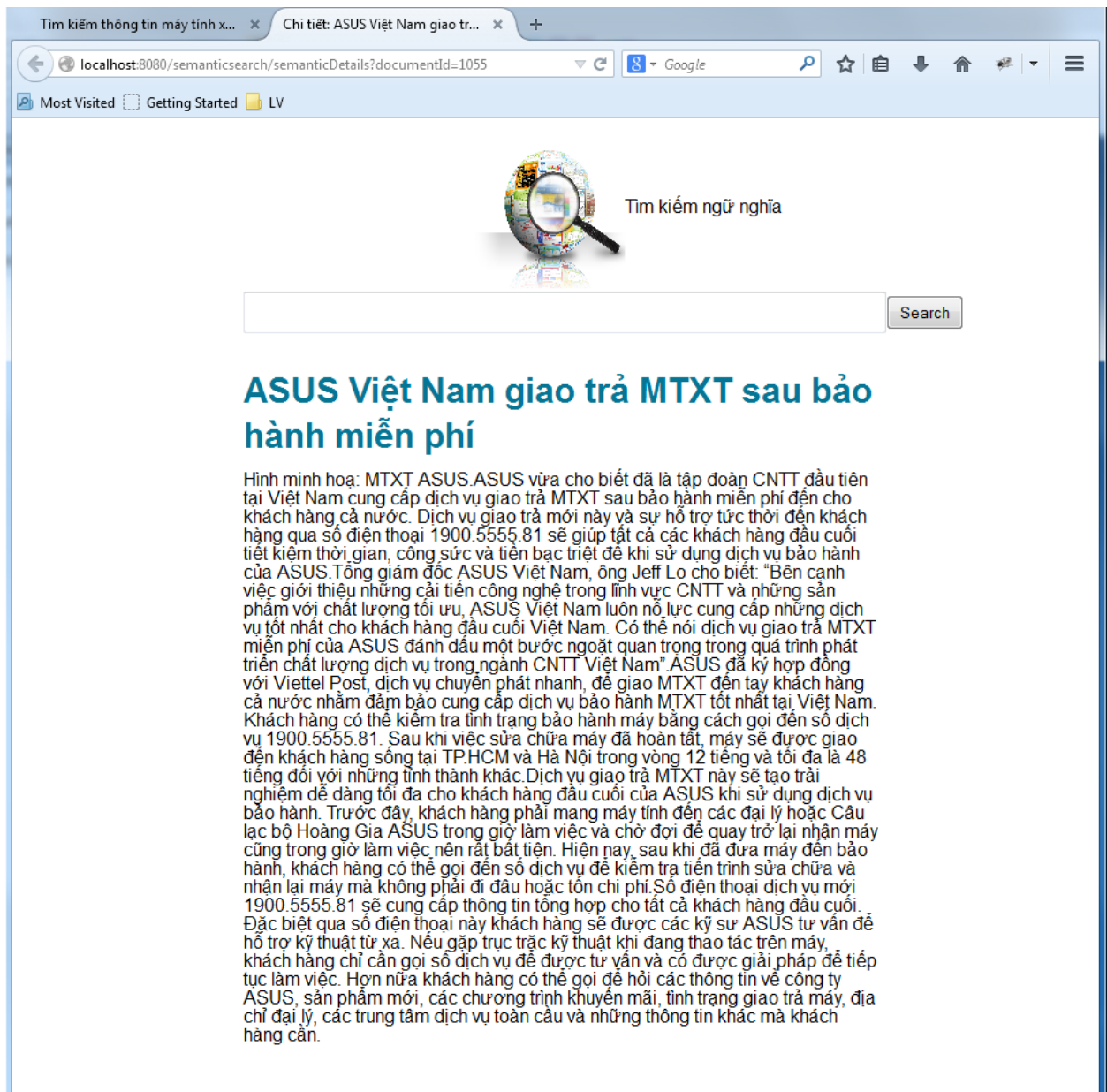
- ✓ Giao diện nhập câu truy vấn



- ✓ Màn hình hiển thị kết quả trả về: kết quả trả về sẽ được phân trang và sắp xếp giảm dần dựa trên giá trị hàm cosin.



✓ Màn hình chi tiết từng tài liệu



The screenshot shows a web browser window with the address bar displaying 'localhost:8080/semanticsearch/semanticDetails?documentId=1055'. The page features a search bar with the text 'Tìm kiếm ngữ nghĩa' and a 'Search' button. Below the search bar, the main heading reads 'ASUS Việt Nam giao trả MTXT sau bảo hành miễn phí'. The content area contains a detailed paragraph about ASUS's service, mentioning that they offer free MTXT (Mobile Text) return after warranty for customers in Vietnam. The text describes the benefits of this service, such as saving time and money, and mentions that ASUS has signed a partnership with Viettel Post for faster delivery. It also provides the service hotline number 1900.5555.81 and states that the service is available for customers in HCMC and Hanoi within 12 hours, and for other cities within 48 hours. The paragraph concludes by stating that ASUS aims to provide the best service to its customers and that this service is a significant step in improving their customer experience.

3.4. Kết quả

Để đánh giá kết quả của hệ thống tìm kiếm ngữ nghĩa cho thông tin về máy tính xách tay, luận văn đã tiến hành cài đặt thêm một hệ thống tìm kiếm thông tin máy tính xách tay theo từ khóa dựa trên thuật toán TF-IDF và mô hình không gian vector. Hệ thống tìm kiếm dựa vào từ khóa sẽ chỉ dựa vào từ để tìm kiếm tài liệu. Dưới đây là một số so sánh về kết quả trả về giữa hệ thống tìm kiếm ngữ nghĩa và hệ thống tìm kiếm dựa vào từ khóa khi thực hiện một số câu truy vấn trên

❖ *Câu truy vấn 1: Máy tính xách tay*

- ✓ Kết quả tìm kiếm từ máy tìm kiếm không có xử lý ngữ nghĩa: trả về 389 kết quả.
- ✓ Kết quả tìm kiếm từ máy tìm kiếm ngữ nghĩa: trả về 389 kết quả.
- ✓ Giải thích: từ khóa “máy tính xách tay” đều được đánh chỉ mục ở 389 tài liệu trong cả máy tìm kiếm có xử lý ngữ nghĩa và máy tìm kiếm không có xử lý ngữ nghĩa.

❖ *Câu truy vấn 2: Máy tính xách tay Aspire*

- ✓ Kết quả tìm kiếm từ máy tìm kiếm không có xử lý ngữ nghĩa: trả về 396 kết quả - kết quả trả về nhiều hơn so với câu truy vấn 1 vì ngoài từ khóa “máy tính xách tay” còn có kết quả trả về liên quan đến từ khóa “Aspire”.
- ✓ Kết quả tìm kiếm từ máy tìm kiếm ngữ nghĩa: trả về 417 kết quả - kết quả trả về nhiều hơn so với câu truy vấn 1.
- ✓ Giải thích: Trong máy tìm kiếm thông thường trả về 396 kết quả, máy tìm kiếm ngữ nghĩa trả về 417 kết quả do ngoài tìm kiếm những tài liệu liên quan đến “máy tính xách tay”, “Aspire” hệ thống tìm kiếm ngữ nghĩa còn tìm kiếm những tài liệu liên quan đến “Acer” vì “Aspire” là một sản phẩm của “Acer”. Những kết quả về liên quan đến “Aspire” vẫn được xếp đầu tiên và ở trong máy tìm kiếm ngữ nghĩa, giá trị hàm cosin liên quan đến

“*Aspire*” trong 20 kết quả đầu tiên luôn cao hơn nhiều so với máy tìm kiếm thông thường.

❖ *Câu truy vấn 3: Aspire*

- ✓ Kết quả tìm kiếm từ máy tìm kiếm không có xử lý ngữ nghĩa: 26 kết quả.
- ✓ Kết quả tìm kiếm từ máy tìm kiếm ngữ nghĩa: 105 kết quả.
- ✓ Giải thích: Máy tìm kiếm không có xử lý ngữ nghĩa chỉ trả về những tài liệu có từ khóa “*Aspire*”. Trong trường hợp này, trong hệ thống có 26 tài liệu chứa từ khóa “*Aspire*” nên máy tìm kiếm trả về 26 tài liệu liên quan, nếu hệ thống không có tài liệu chứa từ khóa “*Aspire*” thì máy tìm kiếm sẽ không thể trả về kết quả. Máy tìm kiếm ngữ nghĩa trả về 105 kết quả, ngoài những tài liệu có chứa từ khóa “*Aspire*” được trả về đầu tiên thì những tài liệu liên quan đến “*Aspire*” là “*Acer*” cũng được trả về. Trong trường hợp trong cơ sở dữ liệu không có tài liệu nào chứa từ khóa “*Aspire*” thì máy tìm kiếm ngữ nghĩa vẫn trả về kết quả liên quan đến “*Aspire*” là các sản phẩm của “*Acer*” nhưng giá trị của hàm cosin sẽ thấp.

Như vậy, khác với hệ thống tìm kiếm theo từ khóa, chỉ trả về những tài liệu chứa từ khóa, hệ thống tìm kiếm ngữ nghĩa ngoài trả tài liệu có chứa từ khóa còn trả về những tài liệu liên quan đến từ khóa.

3.5. Kết luận

Như vậy, trong chương III, luận văn đã xây dựng thành công một hệ thống tìm kiếm ngữ nghĩa cho máy tính xách tay trên cơ sở dữ liệu cục bộ. Hệ thống có thể áp dụng cho nhiều lĩnh vực khác nhau bằng cách mở rộng bản thể học và cơ sở dữ liệu tương ứng.

KẾT LUẬN

Nội dung luận văn đã đạt được một số kết quả sau đây:

- ***Tìm hiểu về máy tìm kiếm và máy tìm kiếm ngữ nghĩa:*** Trình bày về máy tìm kiếm, quá trình xử lý và tìm kiếm thông tin của một máy tìm kiếm, quá trình xử lý và tìm kiếm thông tin của máy tìm kiếm ngữ nghĩa. Vấn đề khó khăn khi tìm kiếm trong tiếng Việt. Tìm hiểu một số nghiên cứu liên quan đến xử lý ngữ nghĩa trong máy tìm kiếm. Từ đó đưa ra giải pháp xây dựng máy tìm kiếm ngữ nghĩa dựa vào bản thể học.
- ***Trình bày về cách xây dựng bản thể học và quá trình đánh chỉ mục ngữ nghĩa dựa trên bản thể học:*** Trình bày về bản thể học, thành phần và cách xây dựng. Tìm hiểu thuật toán tính trọng số của từ trong tài liệu TF-IDF, từ đó mở rộng để đánh chỉ mục ngữ nghĩa cho tài liệu dựa vào bản thể học và áp dụng mô hình không gian Vector để tìm kiếm tài liệu.
- ***Xây dựng thành công máy tìm kiếm ngữ nghĩa dựa trên bản thể học:*** thực hiện cài đặt hệ thống tìm kiếm ngữ nghĩa dựa trên bản thể học cho máy tính xách tay

Hướng nghiên cứu tiếp theo

- Xây dựng bản thể học một cách đầy đủ để có thể áp dụng cho một số lĩnh vực khác.
- Nghiên cứu và xây dựng module tách từ trong tiếng Việt để có thể đánh trọng số được chính xác, từ đó tìm kiếm chính xác hơn.

TÀI LIỆU THAM KHẢO

1. Tài liệu tiếng Anh

- [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]. Manh Hung Nguyen and Tan Hiep Nguyen. (2013). *Towards a Semantic Search Mechanism based on Query Expansion*
- [5]. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 6). Cambridge: Cambridge university press.
- [6]. Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.
- [7]. Sánchez, M. F. (2009). *Semantically enhanced Information Retrieval: an ontology-based approach* (Doctoral dissertation, Doctoral dissertation. Unitversidad de Autónoma, Madrid).
- [8]. 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*.

2. WebSites

- [9]. http://www.vietlex.com/ngon-ngu-hoc/11-Dac_diem_tiang_Viet
- [10]. <http://mim.hus.vnu.edu.vn/phuonglh/software/vnTagger>
- [11]. <http://htmlparser.sourceforge.net/>
- [12]. <https://github.com/owlcs/owlapi/wiki>
- [13]. <http://www.pcworld.com.vn/>
- [14]. <http://vnexpress.net/>