BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

PHẠM HOÀNG LINH

ỨNG DỤNG WEB NGỮ NGHĨA XÂY DỰNG HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

Chuyên ngành: **Khoa học máy tính**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: TS. HUỲNH CÔNG PHÁP

Phản biện 1: **TS. NGUYỄN TRẦN QUỐC VINH**

Phản biện 2: **PGS.TS. ĐOÀN VĂN BAN**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 18 tháng 05 năm 2013.

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại Học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Trong ngành giáo dục, bên cạnh việc ứng dụng công nghệ thông tin (CNTT) vào công tác giảng dạy thì việc ứng dụng CNTT vào công tác quản lý cũng đóng vai trò không kém phần quan trọng trong việc nâng cao chất lượng giáo dục. Tối ưu hóa các quy trình quản lý là mục tiêu hàng đầu nhằm tạo điều kiện tốt nhất cho lực lượng giảng viên, giáo viên chuyên tâm nâng cao chất lượng dạy học.

Thực tế hiện nay, lượng văn bản được ban hành ngày càng nhiều và mỗi trường học lại có những văn bản riêng biệt. Mặc dù toàn bộ các công văn đều được lưu trữ dưới dạng các file mềm số hóa nhưng chỉ đơn thuần là lưu trữ bản sao chứ chưa được sắp xếp theo hệ thống cơ sở dữ liệu chuẩn nhất định.

Trong khi đó, các website tìm kiếm và các công cụ hỗ trợ quản lý giáo dục chưa đáp ứng được nhu cầu tìm kiếm một cách chính xác và nhanh chóng. Việc tìm kiếm thông tin hiện nay không theo chủ đề mà chỉ là tìm theo từ khoá đơn thuần. Kết quả trả về sẽ ở dưới dạng những tri thức chứa từ hoặc cụm từ cần tìm mà không được tổng hợp chính xác làm cho khối lượng thông tin rất lớn. Chính vì phương thức quản lý vẫn còn thủ công khiến cho việc xử lý các chính sách, khiếu nại của từng cá nhân trong từng trường hợp cụ thể gặp rất nhiều khó khăn và tốn thời gian.

Nhận thấy rằng, semantic web có thể giúp chúng tôi giải quyết những vấn đề trên. Vì vậy, tôi đã chọn đề tài "Ứng dụng semantic web xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục" cho luận văn tốt nghiệp của mình.

2. Mục đích nghiên cứu

Đề tài hướng đến xây dựng một ontology đầy đủ về văn bản giáo dục trong nước, từ đó xây dựng hệ thống tìm kiếm văn bản thông minh dành riêng cho ngành giáo dục.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là các vấn đề liên quan đến semantic web, vấn đề xử lý ngôn ngữ tự nhiên và văn bản trong ngành giáo dục.

Phạm vi nghiên cứu là các văn bản liên quan đến ngành giáo dục trong cả nước và hệ thống được xuất bản dưới dạng website

4. Phương pháp nghiên cứu

Phương pháp lý thuyết: Tìm hiểu về semantic web; Tìm hiểu về xử lý ngôn ngữ tự nhiên; Tìm hiểu về quá trình xây dựng một công cụ search engine

Phương pháp thực nghiệm: Xây dựng ontology; Xây dựng cơ sở dữ liệu; Xây dựng kho dữ liệu huấn luyện; Triển khai thực tế trên Internet

5. Ý nghĩa khoa học và thực tiễn

Về mặt khoa học đề tài đóng góp một công cụ tìm kiếm theo công nghệ semantic web dành riêng cho ngành giáo dục, phương pháp xây dựng ontology về văn bản nói chung và văn bản cho ngành giáo dục nói riêng, ứng dụng semantic web về mặt tìm kiếm, xử lý Tiếng Việt và vấn đề đa ngữ trong ontology.

Về thực tiễn đề tài mở ra hướng nghiên cứu ứng dụng mới về tìm kiếm văn bản giáo dục, hỗ trợ tìm kiếm văn bản chính xác hơn.

6. Cấu trúc của luận văn

Chương 1 : Tổng quan về semantic web, search engine và hệ hỏi-đáp. Trong chương này, chúng tôi trình bày cơ sở lý thuyết về

semantic web, search engine và hệ hỏi-đáp, đồng thời chúng tôi nêu ra những yếu tố liên quan đến văn bản trong ngành giáo dục.

Chương 2 : Đề xuất hệ thống tìm kiếm văn bản trong ngành giáo dục. Chương này chúng tôi đưa ra nhận xét về những ưu điểm và khuyết điểm của các hệ thống phục vụ cho ngành giáo dục hiện nay. Từ đó, chúng tôi đề xuất xây dựng hệ thống tìm kiếm mới hiệu quả hơn.

Chương 3 : Phân tích, thiết kế và xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục. Dựa trên những đề xuất ở chương 2, chúng tôi tiến hành phân tích, thiết kế, xây dựng ontology và hệ thống tìm kiếm văn bản giáo dục.

Ngoài ra, để đánh giá toàn bộ quá trình nghiên cứu, phần cuối của luận văn có nêu lên kết quả và hướng phát triển cho đề tài.

CHƯƠNG 1
TỔNG QUAN VỀ SEMANTIC WEB, SEARCH ENGINE VÀ HỆ HỎI-ĐÁP

Toàn chương giới thiệu về semantic web, search engine, hệ hỏi-đáp và tìm hiểu về các loại văn bản trong ngành giáo dục. Đây là chương tiền đề để tiến hành xây dựng các chương sau.

1.1 LÝ THUYẾT VỀ SEMANTIC WEB

Chúng tôi sẽ trình bày khái niệm semantic web là gì? Đồng thời đưa ra ví dụ về semantic web, so sánh giữa semantic web và web hiện tại để từ đó rút ra lợi ích của semantic web

1.1.1 Giới thiệu semantic web

1.1.2 Kiến trúc semantic web

1.1.3 Ontology

Khái niệm ontology, vai trò, ứng dụng và các công cụ hỗ trợ xây dựng ontology hiện nay.

1.1.4 Các ngôn ngữ semantic web

1.2 LÝ THUYẾT VỀ SEARCH ENGINE

Chúng ta sẽ biết được search engine là gì? Nguyên tắc hoạt động của search engine trong phần này.

1.2.1 Các bộ phận cấu thành hệ thống search engine

1.2.2 Nguyên lý hoạt động

1.3 LÝ THUYẾT VỀ TRA CỨU HỆ HỎI-ĐÁP

Chúng tôi trình bày chuyên sâu về hệ hỏi-đáp trong mục này. Từ đó có cơ sở xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục.

1.3.1 Lịch sử phát triển

1.3.2 Khái niệm hệ thống hỏi-đáp

1.3.3 Kiến trúc hệ thống hỏi-đáp

1.3.4 Hệ thống hỏi-đáp tiếng Việt

1.4 TÌM HIỂU VỀ CÁC LOẠI VĂN BẢN TRONG NGÀNH GIÁO DỤC

Phần này, chúng tôi trình bày kết quả nghiên cứu những loại văn bản hiện có trong hệ thống giáo dục của nước Việt Nam. Từ đây, làm cơ sở để tiến hành xây dựng ontology văn bản giáo dục.

1.4.1 Những yếu tố chính của một văn bản giáo dục

Theo tìm hiểu của chúng tôi, những yếu tố cơ bản nhất đối với một văn bản giáo dục đó là *:lĩnh vực, loại văn bản, cơ quan, đơn vị, thời gian, cá nhân liên quan và nội dung.*

1.4.2 Nhận xét

Việc chọn lọc ra những yếu tố cơ bản cấu thành nên một văn bản giáo dục có tác dụng rất lớn trong việc tổ chức cơ sở dữ liệu, xây dựng nền tảng để phát triển hệ thống tìm kiếm theo ngữ nghĩa. Với một văn bản được tiếp nhận, chúng ta sẽ dễ dàng phân loại được văn bản đó liên quan đến vấn đề gì, liên quan đến ai... để từ đó việc tiếm kiếm đạt kết quả tối ưu nhất.

CHƯƠNG 2
ĐỀ XUẤT HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

Chương này, chúng tôi sẽ trình bày những hệ thống phục vụ giáo dục trong nước và trên thế giới, phân tích ưu điểm nhược điểm để từ đó định hướng xây dựng hệ thống của chúng tôi. Đồng thời, chúng tôi trình bày những ý tưởng về hệ thống, giải pháp xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục.

2.1 TỔNG QUAN VỀ CÁC HỆ THỐNG TÌM KIẾM VĂN BẢN GIÁO DỤC HIỆN NAY

2.1.1 Giới thiệu chung

2.1.2 Phân loại

Trong quá trình nghiên cứu luận văn " Ứng dụng semantic web xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục ", chúng tôi đã tham khảo rất nhiều website, hệ thống và ứng dụng khác nhau. Và chúng tôi đã tạm phân loại thành 4 phong cách thiết kế website tìm kiếm văn bản phổ biến hiện nay. Dựa trên 4 phong cách được phân loại này, chúng ta sẽ dễ dàng đánh giá được 1 website hoặc hệ thống tìm kiếm văn bản.

Phong cách cổ điển : Phong cách này chỉ đơn thuần tìm theo đoạn văn bản được nhập vào. Cơ chế làm việc sẽ là so sánh đoạn được nhập vào với cơ sở dữ liệu (CSDL) nếu khớp sẽ xuất ra toàn bộ văn bản chứa thông tin cần tìm.

Phong cách bán cổ điển : CSDL trong phong cách bán cổ điển được tổ chức một cách khoa học và rõ ràng theo từng chuyên đề,

chuyên mục… Phần tìm kiếm ngoài đoạn văn bản được nhập vào còn cho phép người dùng chọn chuyên mục muốn tìm, tạo sự thuận tiện cho người dùng. Tuy nhiên, về cơ bản phong cách bán cổ điển vẫn sử dụng cơ chế tìm kiếm của phong cách cổ điển. Phong cách này hiện nay đang được sử dụng rất phổ biến tại các website trong nước và trên thế giới.

Phong cách hiện đại : Với phong cách thiết kế website này, khối lượng CSDL rất lớn, được tổ chức khoa học và rõ ràng. Tuy nhiên, chính vì khối lượng CSDL quá lớn nên cách quản lý và tổ chức gặp nhiều khó khăn. Vì vậy, những website này thường tổ chức theo dạng hệ thống lớn, với mỗi hệ thống sẽ có cách trình bày và quản lý thông tin khác nhau.

Phong cách semantic web : Các website theo phong cách này có chức năng tìm kiếm theo ngữ nghĩa của thông tin cần tìm, rất tiện lợi cho người sử dụng. Website semantic có khả năng tổng hợp nội dung, phân tích đánh giá để đưa ra kết quả chính xác nhất. Tuy nhiên, những website semantic rất hiếm và CSDL được tích hợp cũng chưa được nhiều. Vì vậy việc ứng dụng semantic web vẫn còn là vấn đề của tương lai.

2.1.3 Các hệ thống phục vụ cho giáo dục trên thế giới

Chúng tôi sẽ trình bày về các hệ thống tiêu biểu trên thế giới như : Cổng thông tin Teachingwithdata.org, Thư viện online của trường đại học British Columbia, Website của chương trình đào tạo và tài trợ để thúc đẩy sự phát triển trong nông nghiệp SARE, Website tìm kiếm theo ngữ nghĩa nổi tiếng Wolframalpha

2.1.4 Nhận xét chung về các hệ thống phục vụ cho ngành giáo dục trên thế giới

Chúng tôi nhận thấy rằng đa phần các website nước ngoài đều được thiết kế theo phong cách bán cổ điển và hiện đại. Về mặt thiết kế, các website nước ngoài sở hữu những thiết đơn giản, đẹp và hiệu quả. Về mặt tìm kiếm, CSDL được tổ chức tốt nên việc tìm kiếm nhanh chóng và dễ dàng hơn so với các website trong nước. Tuy nhiên, ngoài các hệ thống tiên tiến thì phần lớn vẫn chỉ tìm kiếm theo đoạn văn bản được nhập vào chứ chưa phân tích và tìm kiếm theo ngữ nghĩa. Người dùng cần phải tự mình chắt lọc các thông tin cần thiết từ rất nhiều các kết quả trả về.

2.1.5 Các website và hệ thống phục vụ cho ngành giáo dục trong nước

Các hệ thống tiêu biểu trong nước có thể kể đến như : Hệ thống tìm kiếm Wada.vn, Cổng thông tin tuyển sinh thidaihoc.org, Website của bộ giáo dục đào tạo Việt Nam, Các trường đại học lớn trên cả nước

2.1.6 Nhận xét về các website, hệ thống phục vụ cho ngành giáo dục trong nước

Nhìn chung, các website trong nước có kho dữ liệu dồi dào và được tổ chức rất tốt. Tuy nhiên, giao diện còn rườm rà rắc rối, chứa quá nhiều thông tin. Các website chưa có tính liên kết và chưa thống nhất với nhau, vì vậy làm cho việc tìm kiếm 1 thông tin cụ thể nào đó rất khó khăn và mất thời gian.

2.1.7 Giới thiệu các công trình nghiên cứu semantic web trong nước

Các công trình nghiên cứu trong nước tiêu biểu về ontology : Ontology for Vietnamese Language, Ontology khoa học công nghệ,

Ứng dụng web ngữ nghĩa xây dựng hệ thống trợ giúp học tập cho học sinh bậc học phổ thông

2.1.8 Nhận xét chung về các công trình nghiên cứu semantic web trong nước

Nhìn chung, các công trình nghiên cứu về web ngữ nghĩa đã đạt được những thành công bước đầu như : xây dựng ontology, xây dựng ứng dụng... Tuy nhiên, những công trình này vẫn chưa được ứng dụng rộng rãi. Có rất ít các công trình web ngữ nghĩa được ứng dụng trong thực tiễn. Đa phần các lĩnh vực áp dụng web ngữ nghĩa đều là những lĩnh vực đang thu hút rất nhiều sự quan tâm và mang lại nhiều lợi nhuận. Vì vậy, hướng phát triển của web ngữ nghĩa trong tương lai chắc chắn sẽ hướng đến phục vụ cho đời sống chứ không mang nặng tính hàn lâm. Chúng tôi hy vọng trong tương lai sẽ có nhiều thêm các công trình hữu ích phục vụ cho xã hội.

2.1.9 Các công trình nghiên cứu semantic web trên thế giới

2.1.10 Nhận xét về các công trình nghiên cứu semantic web trên thế giới

Những công trình nghiên cứu semantic web trên thế giới đã tiến rất gần đến mô hình web ngữ nghĩa hoàn chỉnh. Tuy nhiên, những công trình này hầu như chỉ hỗ trợ cho những ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, hoàn toàn chưa hỗ trợ tiếng Việt.

2.2 Ý TƯỞNG VỀ HỆ THỐNG TÌM KIẾM VĂN BẢN GIÁO DỤC

2.3 PHÂN TÍCH KHẢ NĂNG ỨNG DỤNG CỦA SEMANTIC WEB CHO BÀI TOÀN

2.3.1 Đặt vấn đề

2.3.2 Phân tích vấn đề

2.3.3 Giải pháp

2.4 ĐẶC TẢ HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

2.4.1 Dự kiến chức năng của hệ thống

Chúng tôi kiến sẽ phát triển hệ thống tìm kiếm văn bản trong ngành giáo dục với các chức năng nổi bật sau :

- Hệ thống sẽ có kho dữ liệu về văn bản lớn nhất, đầy đủ nhất trong ngành giáo dục hiện nay.
- Hệ thống sẽ ứng dụng công nghệ web ngữ nghĩa, nhằm tạo điều kiện cho người sử dụng dễ dàng tìm kiếm với lượng thông tin quá lớn.
- Hệ thống sẽ có chế độ học tập thông minh, tự động thu thập thông tin từ nhiều nguồn khác nhau và tự động tổ chức dữ liệu.
- Hệ thống sẽ là diễn đàn tương tác thông minh giữa người dùng với nhau. Người sử dụng có thể thảo luận, trao đổi hoặc cung cấp thông tin qua lại với nhau.

2.4.2 Mô tả hệ thống

Cấu trúc của một máy tìm kiếm theo công nghệ web semantic, về cơ bản cũng có cấu trúc tương tự với một máy tìm kiếm thông thường, bao gồm 2 thành phần chính là giao diện truy vấn và phần kiến trúc bên trong.

a. Giao diện truy vấn

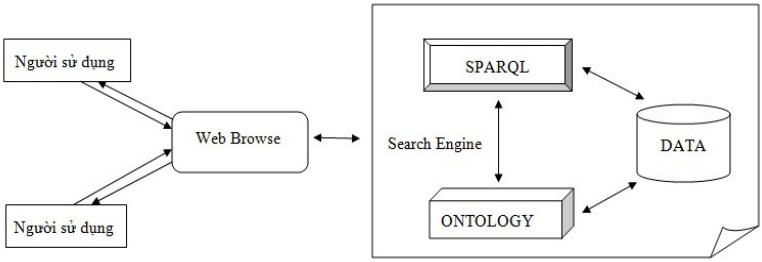
- Cho phép người dùng nhập yêu cầu tìm kiếm.

- Hiển thị kết quả tìm kiếm.

b. Phần kiến trúc bên trong

Đây là phần cốt lõi của máy tìm kiếm bao gồm các thành phần: phân tích yêu cầu, tìm kiếm kết quả cho yêu cầu, dữ liệu tìm kiếm, mạng ngữ nghĩa. Sự khác biệt trong cấu trúc của máy tìm kiếm ngữ nghĩa so với tìm kiếm thông thường nằm ở phần kiến trúc bên trong, cụ thể ở 2 phần: phân tích câu hỏi và tập dữ liệu tìm kiếm.

Mô hình được đề xuất trong luận văn cho ứng dụng tìm kiếm ngữ nghĩa như hình sau :



Hình 2.1. Mô hình đề xuất cho hệ thống tìm kiếm văn bản giáo dục.

c. Cơ sở dữ liệu

Cơ sở dữ liệu nhằm cung cấp cho trang web tìm kiếm được thu thập tự động từ các website phổ biến trên Internet hoặc tự nhập vào bằng tay.

Hệ thống tìm kiếm văn bản trong ngành giáo dục sẽ hướng đến việc cập nhật dữ liệu tự động thông qua các robot tìm kiếm, phân tích dữ liệu thông minh. Nhân tố con người sẽ đóng vai trò kiểm tra và chỉnh sửa dữ liệu trong hệ thống đó.

CHƯƠNG 3
PHÂN TÍCH, THIẾT KẾ VÀ XÂY DỰNG HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

Đây là chương thể hiện cụ thể những gì đã nêu ở 2 chương trước. Chúng tôi sẽ trình bày về ontology văn bản giáo dục, mô hình hoạt động hệ thống, các ngôn ngữ, công cụ hỗ trợ, quy trình xây dựng ứng dụng và kết quả sẽ được trình bày trong chương cuối này.

3.1 PHÂN TÍCH HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

3.1.1 Các giai đoạn xây dựng hệ thống

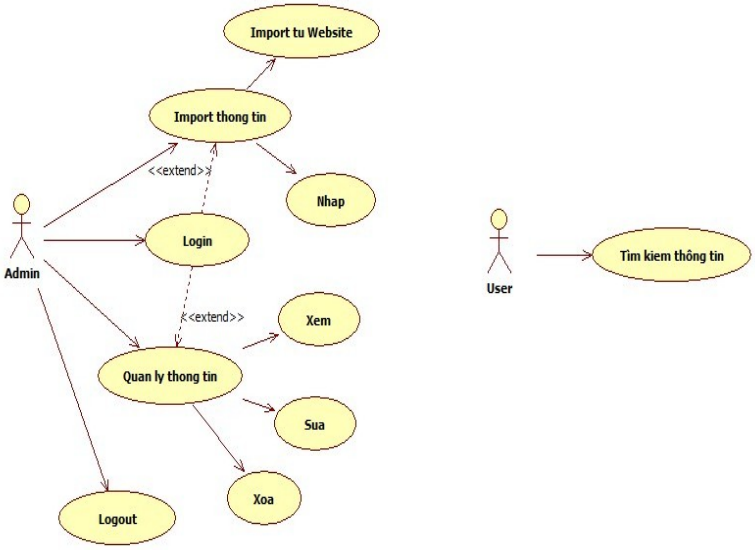
3.1.2 Phân tích chức năng tìm kiếm của hệ thống

Mục này tập trung phân tích 3 chức năng tìm kiếm chính của hệ thống : duyệt theo ngữ nghĩa, tìm kiếm theo từ khóa, tìm kiếm nâng cao

3.2 CÁC QUYỀN CỦA HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

Mục này chúng tôi trình bày về các hành động xảy ra trên hệ thống. Đi sâu phân tích và đưa ra giải pháp với từng hành động cụ thể.

3.3 BIỂU ĐỒ CA SỬ DỤNG



Hình 3.1. Biểu đồ Ca sử dụng

3.4 ĐẶC TẢ CA SỬ DỤNG

3.4.1 Login

3.4.2 Logout

3.4.3 Import thông tin từ website

3.4.4 Import thông tin bằng tay

3.4.5 Xem thông tin

3.4.6 Cập nhật thông tin

3.4.7 Xóa thông tin

3.4.8 Tìm kiếm thông tin

3.5 THIẾT KẾ HỆ THỐNG TÌM KIẾM VĂN BẢN TRONG NGÀNH GIÁO DỤC

Phân tích và thiết kế hệ thống tìm kiếm văn bản trong ngành giáo dục. Với mỗi hệ thống chúng tôi đều đưa ra hình ảnh miêu tả cụ thể.

3.5.1 Biểu đồ Login

3.5.2 Biểu đồ Logout

3.5.3 Biểu đồ Import thông tin từ website

3.5.4 Biểu đồ Import thông tin thủ công bằng tay

3.5.5 Biểu đồ Xem thông tin

3.5.6 Biểu đồ Sửa thông tin

3.5.7 Biểu đồ Xóa thông tin

3.5.8 Biểu đồ Tìm kiếm thông tin

3.6 THIẾT KẾ ONTOLOGY

Từ những kết quả nghiên cứu được ở 2 chương trước, chúng tôi đề xuất mô hình dữ liệu ontology chuẩn có thể áp dụng rộng rãi đối với văn bản giáo dục nói riêng và các loại văn bản nói chung.

3.6.1 Các bước xây dựng ontology

3.6.2 Thiết kế mô hình dữ liệu ontology

Chúng tôi tiến hành xây dựng ontology cho văn bản giáo dục như sau :

Bước 1: Xác định miền quan tâm và phạm vi của ontology

- Miền quan tâm của ontology: *Văn bản trong ngành giáo dục*
- Phục vụ mục đích: *tìm kiếm thông tin văn bản giáo dục*
- Phục vụ đối tượng: *là những người có nhu cầu tìm kiếm thông tin về giáo dục hoặc xây dựng hệ thống thông tin về giáo dục.*

- Phạm vi của ontology: *ngành giáo dục trong cả nước.*

Bước 2: Xem xét việc kế thừa các ontology có sẵn

Đối với ontology văn bản giáo dục, không có sự thừa kế từ các ontology có sẵn.

Bước 3: Liệt kê các thuật ngữ quan trọng trong ontology

Văn bản, lĩnh vực, loại văn bản, cơ quan, thời gian, nội dung văn bản, cá nhân...

Bước 4: Xây dựng các lớp và cấu trúc lớp phân cấp

Với bài toán ta sẽ xây dựng một ontology định nghĩa văn bản trong ngành giáo dục có 7 class chính :

- Linh_vuc : class mô tả về lĩnh vực mà một văn bản cụ thể đề cập đến.

- Loai_van_ban : class mô tả về loại văn bản được đề cập đến.

- Thoi_gian : class mô tả về yếu tố thời gian có liên quan đến văn bản.

- Co_quan : class mô tả về các cơ quan liên quan trong văn bản, đồng thời đây cũng là kho dữ liệu lưu trữ thông tin của các cơ quan.

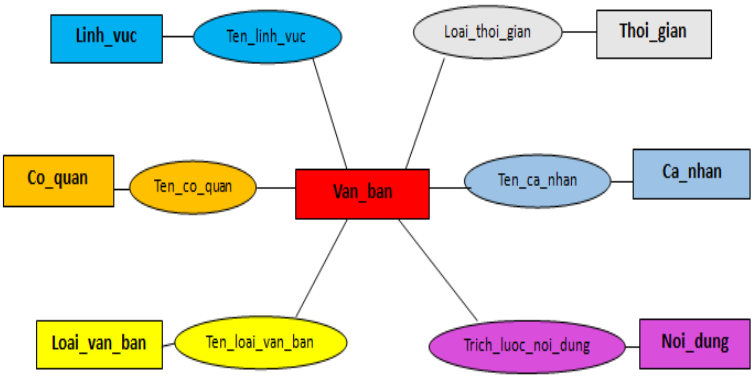
- Ca_nhan : class mô tả về các cá nhân có liên quan trong văn bản, đồng thời đây cũng là kho dữ liệu lưu trữ thông tin của nhiều cá nhân.

- Noi_dung : class mô tả về nội dung của văn bản, nội dung của văn bản sẽ được phân loại nhờ vào class con Kieu_noi_dung.

- Van_ban : class bao quát nhất, miêu tả cụ thể rõ ràng đối tượng chính là văn bản. Class này chứa dữ liệu liên quan đến tất cả các class còn lại.

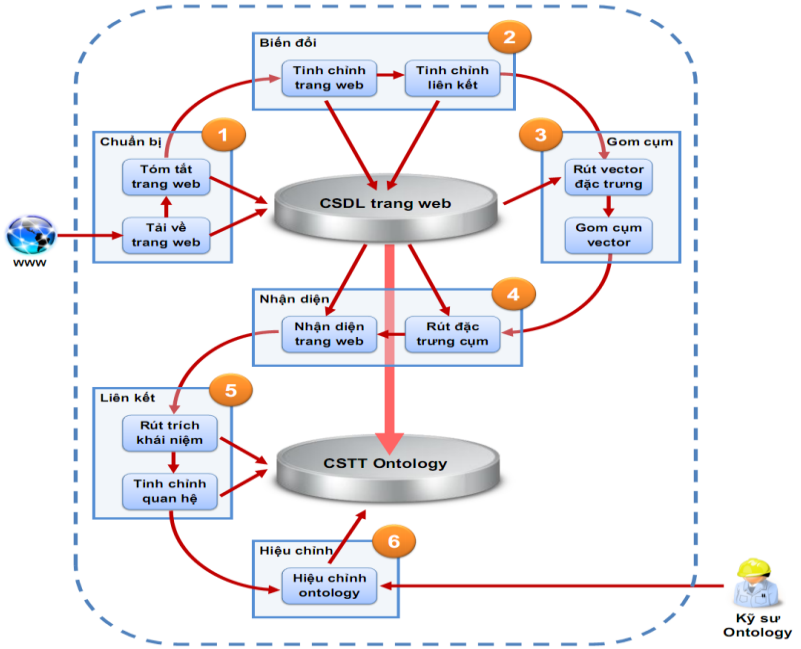
Bước 5 + 6: Định nghĩa các thuộc tính và quan hệ cho lớp, định nghĩa các ràng buộc về thuộc tính và quan hệ của lớp

Để trình bày rõ hơn về các Class cơ bản trong ontology văn bản giáo dục, chúng tôi sẽ miêu tả kèm theo sơ đồ mô phỏng từng Class liên quan.



Hình 3.2. Sơ đồ mô tả tổng quát ontology

3.7 PHƯƠNG PHÁP THU THẬP, TRÍCH RÚT THUỘC TÍNH TỰ ĐỘNG



Hình 3.3. Quy trình rút trích ontology từ WWW [4]

Đầu tiên các trang web thuộc về một website được tải về, loại bỏ những thẻ không phù hợp và được lưu trữ dưới dạng chuẩn hóa (David, Ling, &Calton, 2001) và mô tả tóm tắt sử dụng những từ khoá ở bước chuẩn bị. Bước biến đổi thực hiện việc tinh chỉnh trang web bằng cách loại bỏ các thành phần lặp và xử lý các đường dẫn.

Tiếp theo mỗi trang web được biểu diễn bằng một vector đại diện thể hiện đặc trưng nội dung của trang web đó, gọi là vector đặc trưng. Các trang web này sau đó được gom cụm dựa trên độ tương

đồng giữa các vector đặc trưng của chúng ở bước gom cụm. Mỗi cụm sau đó được nhận diện đặc trưng cụm bằng cách rút ra vector đặc trưng cụm của cụm đó thông qua quá trình nhận diện. Ở bước liên kết, mối quan hệ giữa các cụm được gán và tinh chỉnh dựa trên các đường dẫn giữa các trang web trong cụm. Cuối cùng, việc tinh chỉnh ontology được thực hiện bởi một chuyên gia xử lý ontology ở bước tinh chỉnh.

3.8 CÔNG CỤ, MÔI TRƯỜNG, THƯ VIỆN VÀ NGÔN NGỮ

Chúng tôi tập trung trình bày những thư viện, công cụ, môi trường phát triển hệ thống tìm kiếm văn bản trong ngành giáo dục.

3.8.1 Protégé - Công cụ xây dựng ontology

a. Đặc điểm của Protégé

Đây là phần mềm miễn phí dùng để tạo ra các mô hình và các ứng dụng bằng cách sử dụng các ontology. Protégé được phát triển bởi trường Đại học Stanford và Mark Musen. Chức năng nổi bật nhất của phần mềm này là cho phép người dùng sử dụng tạo ra các ontology để phát triển web ngữ nghĩa theo đúng chuẩn của ngôn ngữ W3C OWL.

Các đối tượng xây dựng chính của Protégé là :

- Classes – tổ chức các quan hệ tham chiếu và các kiểu thực thi
- Axioms – mô hình câu lệnh đúng
- Instances – các thể hiện, các thành phần của đối tượng
- Domain – giới hạn của ontology
- Vocabulary – các lớp và khai báo

b. Protégé sử dụng giao diện đồ họa

c. Protégé phát triển để tích hợp các công cụ

3.8.2 Thư viện SemWeb

SemWeb lần đầu tiên được phát hành vào tháng sáu năm 2005 và đã được thử nghiệm gần đây hơn với những bộ lưu trữ hơn một tỉ bộ ba. Các tính năng cốt lõi như đọc ghi dữ liệu XML với bộ ba RDF, liên tục lưu trữ dữ liệu với nền tảng SQL và các truy vấn SPARQL cơ bản đã được kiểm nghiệm nhiều lần. Các chức năng bên ngoài như hoạt động RDFS hoặc hoạt động của backware-chaining đã làm việc nhưng ít được thử nghiệm và không hoàn chỉnh. Thư viện không có công cụ đặc biệt đối với OWL schema và nó hoạt động ở mức bộ ba của RDF.

a. Giấy phép

b. Đặc điểm của SemWeb

3.8.3 Giao diện lập trình ứng dụng OwlDotNetApi

OwlDotNetApi là một giao diện lập trình ứng dụng với bộ phân tích cú pháp viết bằng C# theo công nghệ .NET dựa trên phân tích cú pháp RDF Drive. Hoàn toàn phù hợp với đặc điểm kỹ thuật của W3C.

a. Phiên bản

b. Chức năng

3.8.4 Hệ truy vấn SPARQL

SPARQL là một ngôn ngữ để truy cập thông tin từ các đồ thị RDF . SPARQL cung cấp các tính năng sau: trích thông tin trong các dạng của URI, các nút trống hay giá trị nguyên thủy hoặc các kiểu được định nghĩa từ các giá trị nguyên thủy, trích thông tin từ các đồ

thị con và xây dựng một đồ thị RDF mới dựa trên thông tin trong đồ thị truy vấn.

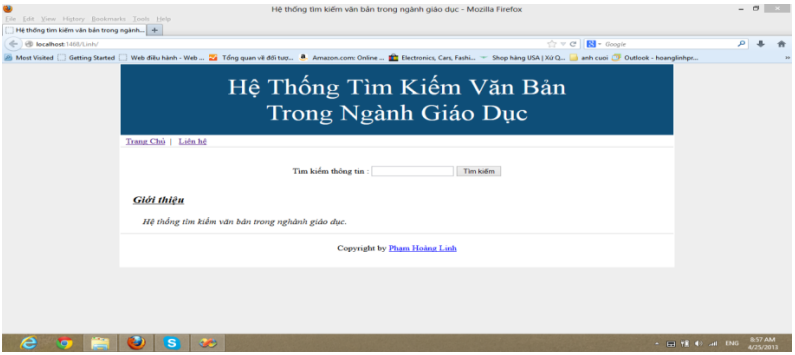
3.8.5 Môi trường và thư viện phát triển ứng ụng cho semantic web

Web ngữ nghĩa được đánh giá và xây dựng mô hình bởi tổ chức W3C. Tổ chức này muốn phát triển công nghệ này theo hướng chia sẻ phát triển. Tất cả những quy tắc phát triển đều được công bố và mọi người có thể sử dụng mà không cần tính bản quyền. Điều này rất phù hợp với định hướng của mã nguồn mở và do vậy có rất nhiều công cụ hỗ trợ để phát triển ứng dụng web ngữ nghĩa.

Với sở trường và thói quen lập trình với ngôn ngữ C#, chúng tôi quyết định phát triển hệ thống tìm kiếm văn bản trong ngành giáo dục trên nền internet, dưới dạng website, sử dụng ngôn ngữ C# và các công cụ cần thiết đã giới thiệu để xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục.

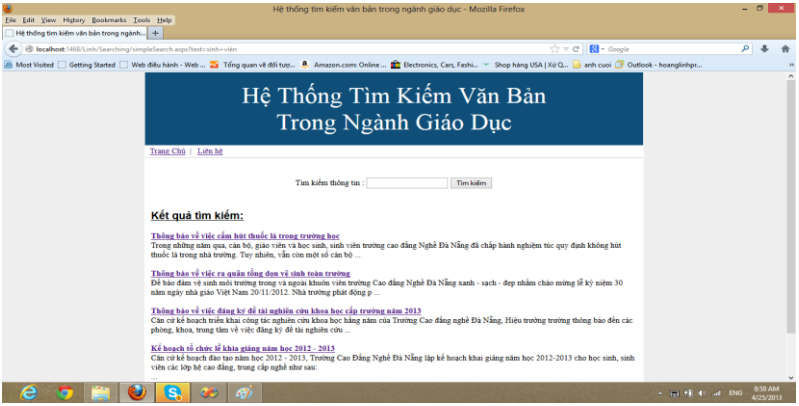
3.9 CHƯƠNG TRÌNH THỬ NGHIỆM, KẾT QUẢ VÀ ĐÁNH GIÁ

3.9.1 Trang chủ hệ thống



Hình 3.4. Trang chủ hệ thống

3.9.2 Màn hình hiển thị tất cả kết quả tìm kiếm



Hình 3.5. Màn hình hiển thị tất cả kết quả tìm kiếm

3.9.3 Màn hình hiển thị văn bản chính xác

Hình 3.6. Màn hình hiển thị văn bản chính xác

3.9.4 Màn hình hiển thị thông tin chi tiết các thành phần có liên quan

3.9.5 Màn hình hiển thị thông tin tác giả

3.9.6 Đánh giá chương trình thử nghiệm

Dựa trên ontology đã xây dựng, website đã có thể cho phép người dùng tìm kiếm văn bản giáo dục chính xác hơn. Sử dụng ngôn ngữ truy vấn SPARQL truy vấn dữ liệu. Việc truy vấn này không tìm theo dữ liệu thuần túy, mà dựa trên dữ liệu có nghĩa, theo các element ðýợc ðịnh nghĩa trong RDF trýớc ðó.

Chúng tôi đã xây dựng được ontology và website tìm kiếm văn bản giáo dục trong trường cao đẳng nghề Đà Nẵng. Tuy nhiên, vì điều kiện thời gian không cho phép nên ontology còn hạn hẹp và website hoạt động chưa thực sự ổn định .

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. KẾT QUẢ ĐẠT ĐƯỢC CỦA LUẬN VĂN

Việc nghiên cứu, ứng dụng semantic web để xây dựng website tìm kiếm văn bản trong ngành giáo dục đã thu được những kết quả ban đầu. đáng khích lệ.

Về mặt lý thuyết, nghiên cứu này đã nêu được những nét đặc trưng, ưu thế của web semantic. Dựa trên việc tìm hiểu những ngôn ngữ, công cụ hỗ trợ lập trình web ngữ nghĩa, luận văn đã đưa ra được một ontology về văn bản giáo dục, cơ bản xây dựng được một website tìm kiếm văn bản theo chuẩn semantic web

Đối với kết quả thực nghiệm với hệ thống tìm kiếm văn bản trong ngành giáo dục, ứng dụng cho phép người dùng có thể tìm kiếm văn bản ở mức cơ bản hoặc tìm kiếm nâng cao hoặc cũng có thể đưa ra những gợi ý cho người dùng khi có nhiều kết quả trùng nhau.

Với việc sử dụng hệ truy vấn SPARQL, việc truy vấn dữ liệu sẽ không tìm theo dữ liệu thuần túy, mà dựa trên dữ liệu có nghĩa, theo các element được định nghĩa trong RDF trước đó.

2. HẠN CHẾ CỦA HỆ THỐNG

Bên cạnh thành công đạt được thì nghiên cứu vẫn còn những hạn chế, đó là ontology chỉ ở mức độ nhỏ, chưa thật sự lớn và phong phú. Ta cần phải có được dữ liệu ontology đầy đủ để đánh giá mức độ xử lý tìm kiếm chính xác cũng như mức độ đáp ứng được bao nhiêu người dùng truy cập ứng dụng cùng một lúc.

Ngoài ra, còn chưa có sự kết nối giữa dữ liệu được trích rút từ WWW và dữ liệu trong ontology. Chức năng trích rút thuộc tính tự động này còn đang được nghiên cứu và có nhiều điểm chưa thống nhất trong các nghiên cứu khác nhau trên thế giới. Bên cạnh đó, việc cài đặt vẫn ở máy local, chưa triển khai lên một server trên Internet.

3. HƯỚNG PHÁT TRIỂN CỦA LUẬN VĂN

Trong tương lai luận văn này có thể tiếp tục phát triển để ứng dụng được vào thực tiễn. Để đạt được mục đích này cần phải xây dựng hệ thống bóc tách thông tin tự động, một chương trình sẽ tự động dò tìm các trang web trên mạng và tiến hành bóc tách theo các thuật toán rẽ nhánh thông minh.

Mở rộng phạm vi của ontology ra cả nước để xây dựng website tìm kiếm văn bản chung cho cả nước. Triển khai ứng dụng thực tế trên internet, phát triển website trở thành một diễn đàn có thể cho phép người sử dụng có thể thêm mới hoặc sửa đổi thông tin văn bản.

Trên đây là toàn bộ nghiên cứu về lý thuyết và ứng dụng semantic web để xây dựng hệ thống tìm kiếm văn bản trong ngành giáo dục.