BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRẦN NGỌC ĐỨC

TÌM HIỂU WEB NGỮ NGHĨA, XÂY DỰNG ỨNG DỤNG TÌM KIẾM TÀI LIỆU TIẾNG VIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TSKH TRẦN QUỐC CHIẾN**

Phản biện 1: TS. Nguyễn Trần Quốc Vinh

Phản biện 2: PGS.TS. Lê Mạnh Thạnh

Luận văn được bảo vệ tại Hội đồng chấm
Luận văn tốt nghiệp Thạc sĩ kỹ thuật họp tại
Đại học Đà Nẵng vào ngày 03 tháng 03 năm 2012

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

1. Lý do chọn đề tài

Sự phát triển nhanh chóng của khoa học, công nghệ làm cho kho kiến thức của con người ngày càng mở rộng. Ngày nay, dữ liệu của con người một phần lớn được lưu giữ dưới dạng tài liệu điện tử và được lưu giữ trong các thiết bị lưu trữ. Với lượng dữ liệu đồ sộ như vậy, việc tìm kiếm và nắm bắt thông đã trở thành một nhu cầu không thể thiếu đối với mỗi con người.

Trong các cơ quan, doanh nghiệp, thậm chí là các thư viện hầu hết các văn bản, tài liệu, sách đều được lưu trữ dưới dạng tài liệu điện tử. Hiện nay các công cụ hỗ trợ cho việc tìm kiếm các tài liệu trong phạm vi một cơ quan, doanh nghiệp thường rất hạn chế về mặt chức năng cũng như khả năng xử lý tiếng Việt và văn bản tiếng Việt.

Do đặc thù của chữ viết tiếng Việt và sự phát triển của nền tin học Việt Nam, các văn bản tiếng Việt được lưu trữ với nhiều bảng mã khác nhau làm cho việc tìm kiếm trở nên rất khó khăn. Các hệ thống tìm kiếm hiện nay đều chưa chuẩn hóa bảng mã trong tài liệu, làm cho kết quả tìm kiếm có thể bị sai lệch. Các hệ thống tìm kiếm hiện nay hầu hết đều tìm theo từ khóa, không hỗ trợ việc tìm kiếm theo ngữ nghĩa điều này làm hạn chể khả năng tìm kiếm cũng như khả năng hỗ trợ người sử dụng trong quá trình tìm kiếm trên hệ thống tìm kiếm.

Từ thực tế đó, việc xây dựng một hệ thống tìm kiếm có thể dễ dàng triển khai trong môi trường cơ quan, doanh nghiệp và có khả năng "hiểu" ngữ nghĩa tiếng Việt, xử lý văn bản tiếng Việt là cần thiết. Vì vậy tôi thực hiện đề tài*"Tìm hiểu web ngữ nghĩa xây dựng ứng dụng tìm kiếm tài liệu tiếng Việt"***.**

2. Mục đích nghiên cứu

- Tìm hiểu về công nghệ, phương pháp xây dựng Web ngữ nghĩa và các vấn đề có liên quan.
- Tìm hiểu các phương pháp bóc tách dữ liệu tự động bằng cách sử dụng các công cụ xử lý ngôn ngữ thông dụng.
- Đề xuất giải pháp xây dựng và tiến hành xây dựng thử nghiệm hệ thống tìm kiếm thông tin tài liệu tiếng Việt dựa trên công nghệ Web ngữ nghĩa.
- Đưa ra một số nhận định, đánh giá về phương pháp đã lựa chọn để thử nghiệm và khả năng phát triển ứng dụng vào thực tế.

3. Đối tượng và phạm vi nghiên cứu

- Dữ liệu, tài liệu, thông tin văn bản được lưu trữ, truy cập thông qua máy tính và môi trường mạng máy tính.
- Các công cụ mã nguồn mở được sử dụng để thao tác, xử lý ngôn ngữ tự nhiên trên các văn bản được lưu trữ trong máy tính.
- Ứng dụng bóc tách và khai thác dữ liệu, phục vụ tìm kiếm theo ngữ nghĩa cho văn bản tiếng Việt.

4. Phương pháp nghiên cứu

Luận văn sử dụng các phương pháp nghiên cứu như sau:

- Thứ nhất, tìm hiểu và đánh giá các kết quả nghiên cứu về các phương pháp xử lý ngôn ngữ tự nhiên, công nghệ Web ngữ nghĩa đang được phát triển hiện nay.

- Thứ hai, từ kết quả thu được của bước thứ nhất, lựa chọn phương pháp xây dựng ứng dụng.
- Thứ ba, từ phương pháp đã lựa chọn, tìm kiếm công cụ thích hợp để xây dựng ứng dụng.

Từ giải pháp và công cụ đã lựa chọn được, tiến hành xây dựng ứng dụng tìm kiếm tài liệu tiếng Việt.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Về mặt khoa học, đề tài tiếp cận vấn đề xử lý ngôn ngữ tự nhiên một cách tự động dựa trên công nghệ Web ngữ nghĩa. Điều này góp phần làm cho việc tìm kiếm trở nên chính xác và hiệu quả hơn. Phục vụ cho việc giải quyết bài toán bóc tách dữ liệu từ văn bản.

Về mặt thực tiễn, đề tài đưa ra được phương pháp xây dựng một ứng dụng xử lý ngôn ngữ dựa trên những công cụ xử lý ngôn ngữ tự nhiên có sẵn và bước đầu xây dựng ứng dụng minh họa.

6. Giải pháp

Để xây dựng được ứng dụng tìm kiếm tài liệu tiếng Việt, đề tài có thể có giải pháp như sau:

- Xây dựng Ontology tiếng Việt cho một số lĩnh vực nhằm minh họa cho ứng dụng.
- Lựa chọn công cụ để xây dựng chú giải cho các văn bản tiếng Việt dựa trên Ontology đã có.
- Xây dựng ứng dụng tìm kiếm ngữ nghĩa dựa trên chú giải đã gán cho các văn bản tiếng Việt.

7. Cấu trúc của luận văn

Sau phần mở đầu, luận văn gồm có 3 chương và phần kết luận. Các chương của luận văn bao gồm:

- Chương 1, “**Tổng quan về Web ngữ nghĩa**”. Chương này cung cấp cho chúng ta cái nhìn tổng quan về công nghệ Web hiện tại và Web ngữ nghĩa. Phân biệt những điểm khác nhau cơ bản giữa Web và Web ngữ nghĩa cũng như trình bày một số ngôn ngữ, công cụ và công nghệ hiện có để xây dựng ứng dụng Web ngữ nghĩa.
- Chương 2, “**Ontology và phương pháp xây dựng Ontology**”. Chương này sẽ trình bày khái niệm, các thành phần, ngôn ngữ, phương pháp và công cụ để xây dựng Ontology .
- Chương 3, “**Xây dựng ứng dụng tìm kiếm tài liệu tiếng Việt**”. Chương này sẽ mô tả các bước xây dựng ứng dụng tìm kiếm tài liệu tiếng Việt và các kết quả chạy thử nghiệm.

Phần kết luận, tổng hợp các kết quả nghiên cứu của luận văn. Các kết quả đạt được, hạn chế của luận văn. Thông qua các kết quả đạt được của luận văn, đề xuất hướng phát triển tiếp theo cho đề tài.

Chương 1 - TỔNG QUAN VỀ WEB NGỮ NGHĨA

1.1. Công nghệ Web hiện tại và những hạn chế

Khối lượng khổng lồ các tài nguyên trên Web làm nảy sinh vấn đề nghiêm trọng là làm thế nào để tìm kiếm chính xác tài nguyên mình mong muốn. Dữ liệu trong các file HTML – ngôn ngữ trình bày dữ liệu của công nghệ Web hiện tại- hữu ích trong một vài ngữ cảnh nhưng vô nghĩa đối với những ngữ cảnh khác. Thêm vào đó HTML không thể mô tả về dữ liệu đóng gói trong nó. Hiện nay, hầu hết các công cụ tìm kiếm tài liệu trên Web được coi là tìm kiếm hiệu quả cũng chủ yếu tìm kiếm được trên bề nổi của Web . Trong khi ở tầng sâu của Web chứa một khối lượng thông tin khổng lồ và thường rất có giá trị cho các nhà nghiên cứu, các học giả hay đơn thuần là những người thích tìm hiểu. Bên cạch đó, các trang Web hiện nay có rất ít đường liên kết với các trang Web khác nên việc tìm kiếm là khó khăn. Ngoài ra, thông tin tìm kiếm được không theo chủ đề mà chỉ là vấn đề tìm thoả theo từ khoá đơn thuần, kết quả tìm kiếm phải do con người chọn lại theo chủ đề mong muốn.

Ví dụ, khi chúng ta biết tên một quốc gia và muốn tìm tên thủ đô của quốc gia đó. Vì mỗi quốc gia có một thủ đô khác nhau và Web không biểu diễn được mối liên hệ này, nên chúng ta không nhận được điều chúng ta mong đợi. Trái lại, đối với Semantic Web, chúng ta có thể chỉ ra kiểu của mối liên hệ này; ví dụ, tên quốc gia có tên thủ đô tương ứng.

Vì vậy, nếu như các thành phần chính yếu của dữ liệu trong Web trình bày theo dạng thức thông thường, thì rất khó sử dụng dữ liệu này một cách phổ biến để có thể mô tả được mối quan hệ như tương tự trên. Một thiếu sót của Web hiện nay là thiếu cơ cấu hiệu quả để chia sẻ dữ liệu khi ứng dụng được phát triển một cách độc lập. Do dó cần phải mở rộng Web để máy có thể hiểu, tích hợp dữ liệu, cũng như tái sử dụng dữ liệu thông qua các ứng dụng khác nhau.

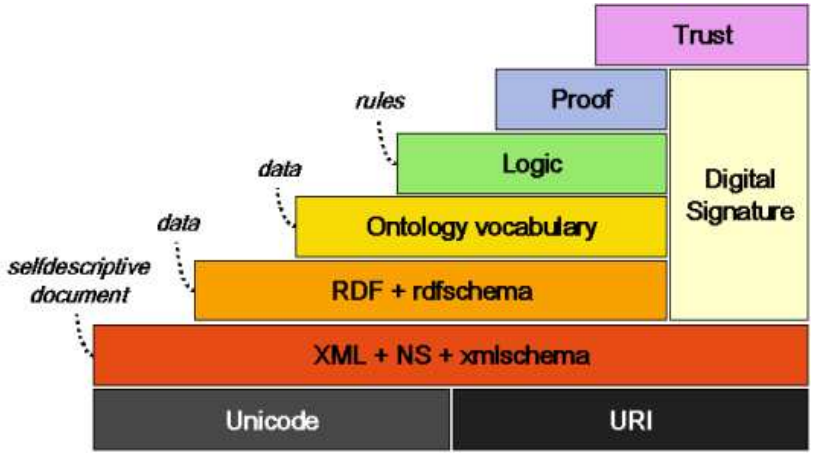
1.2. Web có ngữ nghĩa

Từ những hạn chế, vấn đề về mặt khai thác dữ liệu của công nghệ Web hiện tại đã thúc đẩy sự ra đời của ý tưởng Web ngữ nghĩa (Semantic Web), một thế hệ mới của Web , mà chính cha đẻ của World Wide Web là Tim Berners-Lee đề xuất vào năm 1998. Web ngữ nghĩa là sự mở rộng của Web hiện tại mà trong đó thông tin được định nghĩa rõ ràng sao cho con người và máy tính có thể cùng làm việc với nhau một cách hiệu quả hơn. Mục tiêu của Web có ngữ nghĩa là để phát triển các chuẩn chung và công nghệ cho phép máy tính có thể hiểu được nhiều hơn thông tin trên Web , sao cho chúng có thể hỗ trợ tốt hơn việc khám phá thông tin (thông tin được tìm kiếm nhanh chóng và chính xác hơn), tích hợp dữ liệu (dữ liệu liên kết động), và tự động hóa các công việc.

1.3. Kiến trúc của Web ngữ nghĩa

Web ngữ nghĩa là một tập hợp các ngôn ngữ. Tất cả các lớp của Web ngữ nghĩa được sử dụng để đảm bảo độ an toàn và khai thác thông tin một cách tốt nhất.

Web ngữ nghĩa được xây dựng trên nền hệ thống web hiện tại. Web ngữ nghĩa được coi là sự mở rộng của Web hiện tại có bổ sung thêm ngữ nghĩa vào dữ liệu trên web. Hình 3 chỉ ra sơ đồ kiến trúc của Web ngữ nghĩa.



Hình 1.1: Kiến trúc của web ngữ nghĩa

1.4. Vai trò của các tầng trong Web ngữ nghĩa

1.4.1. Tầng định danh tài nguyên-URI

URI - Uniform Resource Identifier, URI đơn giản chỉ là một định danh Web giống như các chuỗi bắt đầu bằng "http" hay "ftp" mà bạn thường xuyên thấy trên mạng. Bất kỳ ai cũng có thể tạo một URI, và có quyền sở hữu chúng.

1.4.2. Tầng XML và XML Schema

XML là một mở rộng của ngôn ngữ đánh dấu cho các các cấu trúc tài liệu bất kỳ.

1.4.3. Tầng RDF - RDF Schema

RDF (Resource Description Framework) là nền tảng của Web ngữ nghĩa và xử lý *metadata*, được định nghĩa bởi tổ chức W3C. RDF cho phép trao đổi thông tin giữa các ứng dụng trên Web mà máy có thể hiểu được.

1.4.4. Tầng Ontology

Ontology là một tập các khái niệm và quan hệ giữa các khái niệm được định nghĩa cho một lĩnh vực nào đó nhằm vào việc biểu diễn và trao đổi thông tin.

1.4.5. Tầng logic

Khai báo các nguyên tắc logic và cho phép máy tính suy diễn (bằng cách suy luận) bằng cách dùng những nguyên tắc này.

1.4.6. Tầng Proof

Chúng ta sẽ xây dựng các hệ hiểu logic và dùng chúng để chứng minh. Mọi người trên thế giới có thể viết các khai báo logic. Sau đó máy tính có thể theo những liên kết ngữ nghĩa này để kiểm chứng.

1.4.7. Tầng Trust

Tầng này nhằm đảm bảo tính tin cậy của các ứng dụng trên Web ngữ nghĩa.

1.5. Các ngôn ngữ được sử dụng trong Web ngữ nghĩa

1.5.1. XML–Ngôn ngữ đánh dấu mở rộng

XML là một đặc tả cho các tài liệu mà máy tính đọc được. Đánh dấu có nghĩa là các chuỗi ký tự nào đó trong tài liệu có chứa thông tin chỉ ra vai trò nội dung của tài liệu. đánh dấu mô tả sơ đồ dữ liệu của tài liệu và cấu trúc logic. Các đánh dấu này làm thông tin tự mô tả tùy vào cảm nhận. Các đánh dấu này được mô tả dưới dạng các từ trong dấu ngoặc nhọn hay còn gọi là tag.

1.5.2. RDF - Biểu diễn dữ liệu về dữ liệu

XML cung cấp cú pháp để mã hóa dữ liệu, RDF là một cơ cấu chỉ ra điều gì đó về dữ liệu. Như tên gọi, RDF là một mô hình để biểu diễn dữ liệu về "mọi thứ trên Web".

1.5.2.1. Các khái niệm cơ bản

Namespace và cách khai báo

Qualified name (QName) và cách sử dụng

Mô hình RDF

Bộ ba RDF (RDF Tripple)

Đồ thị RDF

Dữ liệu nguyên thủy(Literal)

1.5.2.2. Cấu trúc RDF/XML

Cú pháp RDF/XML cơ bản

RDF Container

RDF Collection

1.5.2.3. Lược đồ RDF- RDF Schema

- Định nghĩa class (lớp)

Các tài nguyên trên Web có thể chia thành các nhóm gọi là class. Các thành viên (member) của nhóm được xem như là thể hiện của lớp đó. Class cũng chính là tài nguyên. Nó được nhận ra thông qua các định danh URI và có thể được mô tả bằng cách sử dụng các RDF properties.

- Định nghĩa thuộc tính (property)

RDF Schema cũng cung cấp một bộ từ vựng để mô tả làm thế nào mà các thuộc tính (property) và lớp (class) có thể được sử dụng cùng với nhau trong dữ liệu RDF.

1.5.2.4. Truy vấn dữ liệu trong rdf

SPARQL là một ngôn ngữ để truy cập thông tin từ các đồ thị RDF. Nó cung cấp các tính năng sau:

- Trích thông tin trong các dạng của URI, các nút rỗng và các dữ liệu nguyên thủy hay giá trị được định nghĩa từ dữ liệu nguyên thủy.
- Trích thông tin từ các đồ thị con.
- Xây dựng một đồ thị RDF mới dựa trên thông tin trong đồ thị truy vấn.

Chương 2 - ONTOLOGY VÀ PHƯƠNG PHÁP XÂY DỰNG ONTOLOGY

2.1. Giới thiệu Ontology

2.1.1. Khái niệm Ontology

Trong những năm gần đây, thuật ngữ "Ontology" không chỉ được sử dụng ở trong các phòng thì nghiệm trên lĩnh vực trí tuệ nhân tạo mà đã trở nên phổ biến đối với nhiều miền lĩnh vực trong đời sống . Đứng trên quan điểm của ngành trí tuệ nhân tạo, một Ontology là sự môt tả về những khái niệm và những quan hệ của các khái niệm đó nhằm mục đích thể hiện một góc nhìn về thế giới. Trên miền ứng dụng khác của khoa học, một Ontology bao gồm tập các từ vựng cơ bản hay một tài nguyên trên một miền lĩnh vực cụ thể, nhờ đó những nhà nghiên cứu có thể lưu trữ, quản lý và trao đổi tri thức cho nhau theo một cách tiện lợi nhất.

Hiện nay tồn tại nhiều khái niệm về Ontology, trong đó có nhiều khái niệm mâu thuẫn với các khác niệm khác, khóa luận này chỉ giới thiệu một định nghĩa mang tính khái quát và được sử dụng khá phổ biến được Kincho H. Law đưa ra: "Ontology là biểu hiện một tập các khái niệm (đối tượng), trong một miền cụ thể và những mối quan hệ giữa các khái niệm này". Ontology chính là sự tổng hợp của một tập từ vựng chia sẻ và các miêu tả ý nghĩa của từ đó theo cách mà máy tính hiểu được.

2.1.2. Các thành phần của Ontology

Lớp (class) là một bộ những thực thể, các thực thể được mô tả logic đề định nghĩa các đối tượng của lớp; lớp được xây dựng theo cấu trúc phân cấp cha con như là một sự phân loại các đối tượng. Thực thể được xem là thể hiện của một lớp, làm rõ hơn về lớp đó và có thể được hiểu là một đối tương nào đó trong tự nhiên (England, Manchester United, bệnh sởi, thủy đậu…).

Thuộc tính (Property) thể hiện quan hệ nhị phân của các thực thể (quan hệ giữa hai thực thể) như liên kết hai thực thể với nhau. Ví dụ thuộc tính "làm cho" liên kết hai thực thể "người" và "công ty" với nhau.

Thuộc tính (property) có 4 loại (1) Functional: Một thực thể chỉ liên quan nhiều nhất đến một thực thể khác, ví dụ thuộc tính "có hương vị" đối với các thực thể lớp "thức_ăn"; (2) Inverse Functional: Thuộc tính đảo ngược của Functional, thuộc tính "là hương vị của"; (3) Transitive: Thực thể a quan hệ với thực thể b, thực thể b quan hệ với thực thể c thì thực thể a quan hệ với thực thể c; (4) Symmetric: Thực thể a quan hệ với thực thể b thì thực thể b quan hệ với thực thể a.

Thuộc tính có 3 kiểu thể hiện:

- Object Property: Liên kết thực thể này với thực thể khác
- DataType Property: Liên kết thực thể với kiểu dữ liệu XML Schema, RDF literal
- Annotation Property: Thêm các thông tin metadata về lớp, thuộc tính hay thực thể khác thuộc 2 kiểu trên.

2.1.3. Một số công trình liên quan tới xây dựng Ontology

Ngày nay, Ontology được sử dụng rất nhiều trong các lĩnh vực liên quan đến ngữ nghĩa như trí tuệ nhân tạo (AI), semantic web, kĩ nghệ phần mềm, v.v… Vì những ứng dụng của Ontology nên không chỉ riêng Việt Nam, trên thế giới đã có nhiều dự án tập trung xây dựng Ontology đối với từng miền dữ liệu khác nhau và phục vụ cho nhiều mục đích đa dạng khác nhau. Đối với miền dữ liệu y tế có thể kể tới rất nhiều Ontology trong lĩnh vực y tế, sinh học đã được đưa ra bởi tổ chức The National Center for Biomedical Ontology. Dự án này đã đưa ra được rất nhiều Ontology trong y tế cũng như trong sinh học, ví dụ như Ontology về cell type, Gene, FMA, Human disease…danh sách các Ontology đưa ra được hiển thị trong.

Ngoài ra có thể kể tới Disease Ontology là một tập từ về y khoa được phát triển tại Bioinformatics Core Facility cùng với sự cộng tác của dự án NuGene Project tại trung tâm Center for Genetic Medicine. Ontology này được thiết kế với mục đích sắp xếp các bệnh và các điều kiện tương ứng đối với những code về y tế cụ thể như là ICD9CM, SNOMED và những cái khác….Disease Ontology cũng được sử dụng để liên kết những kiểu hình sinh vật mẫu đối với các bệnh của con người cũng như trong việc khai phá dữ liệu y học. Disease Ontology được thực hiện như là một đồ thị xoắn có hướng và sử dụng UMLS (Unified Medical Language System) là tập từ vựng để truy cập các Ontology về y tế khác như ICD9CM.

Một ontology tiếng Anh được đề cập rất nhiều trong lĩnh vực y tế trong thời gian gần đây đó là GENIA. Mục đích chính mà ontology này hướng tới đó là sự phản ứng lại của tế bào trong não người.

Ontology này chủ yếu tập trung trong các lĩnh vực y tế và cũng được sử dụng trong các bài toán xử lý ngôn ngữ tự nhiên: truy hồi thông tin (Information Retrieval – IR), trích chọn thông tin, phân lớp và tóm tắt văn bản.

DBpedia Ontology là một ontology tổng quát, bao trùm nhiều lĩnh vực. Ontology này được tạo ra bằng cách lấy thông tin phổ biến trên Wikipedia và xây dựng lại một cách thủ công. Hiện nay, DBpedia đã có hơn 320 lớp phân cấp bao gồm nhiều lĩnh vực được mô tả bởi hơn 1650 thuộc tính khác nhau.

2.2. Phương pháp xây dựng Ontology

2.2.1. Xây dựng Ontology

Ngày nay, việc nghiên cứu quá trình xây dựng ontology ngày càng được quan tâm nhiều hơn. Có rất nhiều nhóm sau quá trình nghiên cứu đã đưa ra các phương pháp khác nhau nhằm xây dựng Ontology.

Nội dung chương này sẽ đề cập đến một số nguyên tắc cơ bản của việc xây dựng Ontology qua các các công đoạn cụ thể sau đây:

Các bước cụ thể như sau:

- Bước 1, xác định miền quan tâm và phạm vi của Ontology
- Bước 2, xem xét việc kế thừa các Ontology có sẵn
- Bước 3, liệt kê các thuật ngữ quan trọng trong Ontology
- Bước 4, xây dựng các lớp và cấu trúc lớp phân cấp

- Bước 5, định nghĩa các thuộc tính và quan hệ cho lớp
- Bước 6, định nghĩa các ràng buộc về thuộc tính và quan hệ của lớp
- Bước 7, tạo các thực thể cho lớp

2.2.2. Ngôn ngữ xây dựng Ontology

Hiện tại, các ngôn ngữ xây dựng ontology (ngôn ngữ ontology) điển hình bao gồm LOOM, LISP, Ontolingua, XML, SHOE, OIL, DAML+OIL và OWL.

2.2.2.1. RDFS (RDF-Schema)

RDFS là một ngôn ngữ Ontology cơ bản. Nó được phát triển ở tầng trên của RDF cho nên bản thân RDF-Schema cũng chính là RDF, nó được mở rộng từ RDF và bổ sung thêm các tập từ vựng để hỗ trợ cho việc xây dựng các Ontology được dễ dàng.

2.2.2.2. OWL (Ontology Web Language)

OWL là ngôn ngữ ontology khá mạnh, nó ra đời sau RDFS nên biết kế thừa những lợi thế của ngôn ngữ này đồng thời bổ sung thêm nhiều yếu tố giúp khắc phục được những hạn chế của RDFS. OWL giúp tăng thêm yếu tố logic cho thông tin và khả năng phân loại.

2.2.2.3. DAML + OIL

DAML+ OIL ra đời nhằm khắc phục những hạn chế về kiểu dữ liệu trong các ngôn ngữ Ontology trước đó là RDF, RDFS. DAML + OIL (gọi tắt là DAML) là ngôn ngữ đánh dấu cho các tài nguyên trên Web, có hỗ trợ suy luận.

2.2.3. Công cụ xây dựng Ontology

Về mặt lý thuyết, người xây dựng và quản trị Ontology có thể không cần các công cụ hỗ trợ, thay vào đó có thể thực hiện trực tiếp bằng các ngôn ngữ. Tuy nhiên, cách thứ hai sẽ không khả thi khi Ontology có kích thước lớn và cấu trúc phức tạp. Thêm vào đó, việc xây dựng và quản trị Ontology 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 trên nó. Có rất nhiều bài toán liên quan đến một hệ thống Ontology như:

- Trộn hai hay nhiều Ontology.
- Chuẩn đoán và phát hiện lỗi.
- Kiểm tra tính đúng đắn và đầy đủ.
- Ánh xạ qua lại giữa các Ontology.
- Suy luận trên Ontology.
- Sao lưu và phục hồi một Ontology.
- Xóa, sửa và tinh chỉnh các thành bên trong Ontology.
- Tách biệt Ontology với ngôn ngữ sử dụng (DAML, OWL,..).

Những khó khăn trên đã khiến các công cụ trở thành một thành phần không thể thiếu, quyết định đến chất lượng của một hệ thống Ontology. Hiện có rất nhiều công cụ có khả năng hỗ trợ người thiết kế giải quyết những bài toán liên quan. Có thể kể ra một số như: Sesame, Protégé, Ontolingua, Chimaera, OntoEdit, OidEd..

Nội dung phần này sẽ đề cập giới thiệu sơ lược một số công cụ xây dựng và quản trị Ontology và sẽ trình bày chi tiết hai công cụ là Protégé và Chimaera.

2.2.3.1. Protégé

Protégé là bộ phần mềm mã nguồn mở Java nổi tiếng. Protégé được nghiên cứu và phát triển từ năm 1998 bởi nhóm nghiên cứu của Mark Musen, ĐH. Stanford nhằm quản lý các thông tin trong lĩnh vực sinh y học. Đây là dự án được nhận được sự quan tâm và tài trợ từ rất nhiều tổ chức, trong đó có Bộ Quốc Phòng Mỹ.

2.2.3.2. Chimaera

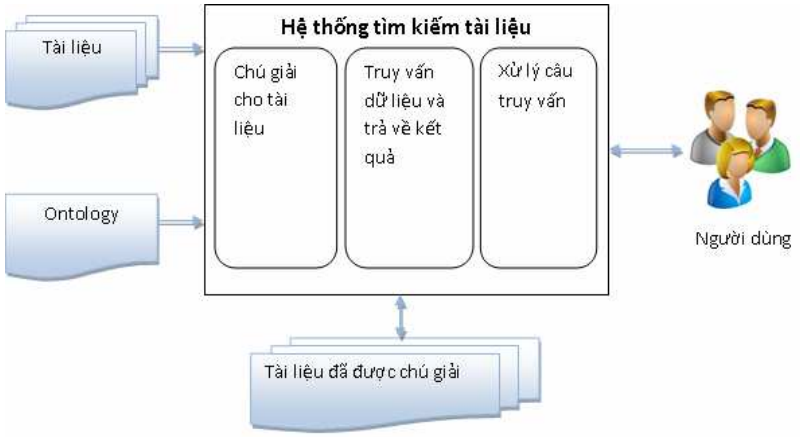
Chimaera cũng là một ứng dụng khác được phát triển bởi đại học Stanford, với mục đích ban đầu nhằm giải quyết hai vấn đề là: trộn các Ontology và chuẩn đoán lỗi, phân tích tính nhất quán giữa các Ontology phân tán.

Chương 3 - XÂY DỰNG ỨNG DỤNG TÌM KIẾM TÀI LIỆU TIẾNG VIỆT

3.1. Mô tả ứng dụng

Ứng dụng có thể thực hiện tìm kiếm trong kho dữ liệu của mình bao gồm việc tìm kiếm trong các tài liệu và trong Ontology đã được xây dựng sẵn. Các tài liệu bao gồm các tập tin dạng văn bản như: file text, một trang Web,...Ontology đóng vai trò xử lý gán chú giải ngữ nghĩa cho các tài liệu cũng như xử lý câu truy vấn do người dùng nhập vào.

Kết quả trả về là một hoặc nhiều tài liệu trong kho dữ liệu của ứng dụng.



Hình 3.1: Mô hình hệ thống ứng dụng tìm kiếm tài liệu tiếng Việt

3.2. Xây dựng ứng dụng

Dựa vào mô tả trên của ứng dụng cần xây dựng, các bước để xây dựng ứng dụng bao gồm:

- Xây dựng Ontology cho ứng dụng.
- Xây dựng chức năng tạo chú giải cho tài liệu dựa trên Ontology đã xây dựng
- Xây dựng chức năng xử lý câu truy vấn và truy vấn dữ liệu dựa trên yêu cầu truy vấn của người dùng.

3.2.1. Công cụ và ngôn ngữ lập trình

Trong luận văn này, tôi tích hợp các tiện ích trong các bộ công cụ Protégé, Gate (General Architecture for Text Mining) để xây dựng ontology, chú thích dữ liệu và nhận dạng thực thể tiếng Việt.

Gate là một kiến trúc phần mềm để phát triển và triển khai các bộ phận phần mềm phục vụ công việc xử lý ngôn ngữ của con người.

3.2.2. Xây dựng Ontology

Để xây dựng Ontology cho ứng dụng ta dựa vào phương pháp xây dựng Ontology đã được trình bày ở trên cùng với công cụ là phần mềm Protégé.

Việc xây dựng Ontology dựa trên Ontology có sẵn là PROTON.

3.2.3. Chú giải cho tài liệu

Chú giải ngữ nghĩa là quá trình chèn những nhãn trong một tài liệu để gán ngữ nghĩa cho những đoạn văn bản cho phép để tạo ra những tài liệu có thể xử lý được bằng những tác nhân tự động.

Luận văn tích hợp Ontology đã xây dựng vào công cụ Gate để chú thích dữ liệu.

3.2.4. Xử lý truy vấn

Để xử lý một truy vấn dữ liệu ta cần qua hai bước: xử lý truy vấn trong Ontology và xử lý truy vấn trong kho dữ liệu đã chú giải.

Xử lý truy vấn trong Ontology ta cần dùng Framework Jena, nó cung cấp đầy đủ các phương thức để truy cập, thao tác trên Ontology đã xây dựng thông qua việc truy vấn dựa trên cú pháp của ngôn ngữ truy vấn SPARQL.

3.3. Cài đặt và thử nghiệm ứng dụng

Dựa vào các công cụ, phương pháp thực hiện ở trên ta tiến hành việc cài đặt ứng dụng.

3.3.1. Cài đặt ứng dụng

3.3.1.1. Môi trường cài đặt

Môi trường cài đặt ứng dụng, bao gồm các môi trường phần cứng, phần mềm.

3.3.1.2. Các bước thực hiện

Qui trình thiết kế, xây dựng ứng dụng theo trình tự dựa trên môi trường cài đặt thử nghiệm như đã lựa chọn.

3.3.2. Chạy thử nghiệm và kết quả đạt được

3.3.2.1. Dữ liệu thử nghiệm

Mô tả dữ liệu thử nghiệm được sử dụng của chương trình thử nghiệm để tiến hành chạy thử.

3.3.2.2. Kết quả

Kết quả thực hiện chương trình như sau:

- Yêu cầu 1

Thực hiện truy vấn với yêu cầu: *"tìm tất cả các tài liệu có chứa thông tin của ít nhất một địa danh"*

- Yêu cầu 2

Thực hiện truy vấn với yêu cầu: *"tìm tất cả các tài liệu chứa thông tin về địa danh có chứa thông tin là Đà Nẵng"*

3.3.3. Đánh giá

Ứng dụng minh họa đã cài đặt thành công trên máy chủ Web Tomcat, thực hiện được yêu cầu đặt ra. Thực hiện truy vấn và trả về kết quả phù hợp với yêu cầu của chương trình đã trình bày ở trên.

Kết quả trả về của ứng dụng chưa được sắp xếp một cách hợp lý. Các tài liệu có thể bị trùng lắp trong danh sách kết quả trả về, thứ tự các tài liệu không được sắp xếp mà trình bày một cách ngẫu nhiên.

KẾT LUẬN

1. Kết luận

Kết quả nghiên cứu đề tài gói gọn trong phạm vi về Web ngữ nghĩa và xây dựng một ứng dụng tìm kiếm nhằm minh họa cho những kiến thức đã đạt được.

Đề tài đã nghiên cứu, tiếp cận công nghệ Web ngữ nghĩa, các vấn đề cơ bản và tổng quát về Web ngữ nghĩa và đã được một số kết quả nhất định.

Nắm được công nghệ về Web ngữ nghĩa, điểm khác biệt giữa công nghệ Web ngữ nghĩa và Web truyền thống. Những điểm mạnh của Web ngữ nghĩa so với công nghệ Web hiện tại cũng như những hạn chế của công nghệ Web mà chúng ta đang sử dụng.

Tìm hiểu được kiến trúc của Web ngữ nghĩa, các thành phần của Web ngữ nghĩa cũng như vai trò của các thành phần của nó.

Tìm hiểu được RDF, là một nền tảng đóng vai trò quan trọng trong kiến trúc của Web ngữ nghĩa. Các khái niệm, thành phần, công cụ cũng như các ngôn ngữ đặt tả được sử dụng để xây dựng mô tả về RDF.

Cách thức truy vấn thông tin trong RDF bằng ngôn ngữ SPARQL. Nghiên cứu được cú pháp, cách xây dựng truy vấn cũng như cách xử lý dữ liệu trong ngôn ngữ truy vấn dữ liệu bằng ngôn ngữ SPARQL.

Tìm hiểu cấu trúc, phương pháp biểu diễn ngôn ngữ suy diễn OWL nhằm xây dựng Ontology.

Áp dụng những lý thuyết đã tìm hiểu được ở trên, đề tài đã xây dựng được ứng dụng minh họa nhằm ứng dụng công nghệ Web ngữ nghĩa. Ứng dụng xây dựng được cho phép người sử dụng có thể tìm kiếm tài liệu mình cần theo ngữ nghĩa. Người dùng có thể nhập dữ liệu và tìm kiếm theo ngữ nghĩa thông qua giao diện người dùng là 1 Website.

Cập nhật dữ liệu về Ontology, các thực thể trong Ontology thông qua ứng dựng Gate.

Cập nhật kho dữ liệu tìm kiếm.

Thông qua việc xây dựng ứng dụng, tìm hiểu được một số công cụ hỗ trợ cho việc phát triển Web ngữ nghĩa như: Protégé, Gate, KIM, Jena và ngôn ngữ lập trình Java.

Đây là cách xử lý dữ liệu dựa trên các công cụ mã nguồn mở cũng là xu hướng nghiên cứu mở rộng các ứng dụng xử lý ngôn ngữ tự nhiên của hiện tại và tương lai.

2. Nhận xét và hướng phát triển

2.1. Nhận xét

Đề tài đã trình bày một cách ngắn gọn và đầy đủ về công nghệ Web ngữ nghĩa.

Xây dựng được một ứng dụng hoàn chỉnh nhằm minh họa cho lý thuyết đã tìm hiểu được.

Do Ontology của ứng dụng của còn hạn chế nên việc tìm kiếm chưa thể mang lại kết quả chính xác và đầy đủ

Việc xử lý tiếng Việt và câu tiếng Việt còn hạn chế. Ứng dụng sử dụng bộ tách từ mặc định của công cụ Gate nên chỉ có thể chú giải cho các thực thể có tên nằm trong Ontology. Ứng dụng không có khả năng chú giải cho câu tiếng Việt, cũng như việc tách từ tiếng Việt và phân tích cú pháp câu theo ngữ pháp tiếng Việt.

2.2. Hướng phát triển

Để đề tài có thể trở thành một ứng dụng có thể sử dụng được trong thực tế ta cần phát triển thêm một số khía cạnh sau về mặt công nghệ và xây dựng thêm Ontology cho ứng dụng.

Tiếp tục nghiên cứu và tiếp cận các nghiên cứu mới nhất về công nghệ Web ngữ nghĩa. Việc này giúp ta có thể có được những phương pháp tiếp cận mới, sử dụng các công cụ hiệu quả hơn giúp ta có thể cải tiến các phương pháp tiến đến áp dụng cho chính mình.

Tìm hiểu và phát triển bộ công cụ tách từ trong tiếng Việt nhằm áp dụng thay thế cho công cụ tách từ của Gate.

Tìm hiểu và xây dựng công cụ có thể nhận dạng và hiểu được ngữ pháp tiếng Việt để nâng cao sự chính xác trong việc xây dựng chú giải ngữ nghĩa cho tài liệu tiếng Việt.

Mở rộng và làm giàu Ontology của ứng dụng.