BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

DƯƠNG NGỌC DUY

XÂY DỰNG WEB NGỮ NGHĨA TRỢ GIÚP TRA CỨU TỪ HÁN 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. TS. PHAN HUY KHÁNH**

Phản biện 1 : **PGS.TS. VÕ TRUNG HÙNG**

Phản biện 2 : **TS. TRƯƠNG CÔNG TUẤN**

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 15 tháng 12 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

Từ Hán Việt chiếm tỷ lệ rất lớn trong kho từ vựng tiếng Việt, việc tra cứu thông tin, ý nghĩa từ Hán Việt được nhiều sự quan tâm của nhà nghiên cứu văn hóa, lịch sử, ngôn ngữ cũng như học sinh, sinh viên.

Theo thống kê một cách tương đối của GS. Phan Ngọc Thạch có hơn 7000 từ Hán Việt đang được sử dụng phổ biến hiện nay, chiếm gần 60% số lượng từ của tiếng Việt hiện nay.

Vấn đề sử dụng sai từ Hán Việt hiện nay trong một bộ phận người dân cũng như sinh viên là rất đáng lo ngại.

Trong thời đại ngày nay ngôn ngữ luôn biến đổi, lượng kiến thức từ về các lĩnh vực khoa học công nghệ hay kinh tế từ các nước phương tây nhu nhập về nước ta ngày càng nhiều, chúng ta lại vay mượn từ tiếng Trung Quốc để thể hiện, vậy làm thế nào để quản lý lượng từ Hán Việt mới này.

Việc tra cứu thông tin từ Hán Việt còn gặp nhiều khó khăn, kết quả tìm kiếm không chính xác, vẫn còn nhiều nhập nhằng về nghĩa.

Hiện nay có nhiều công trình nghiên cứu Hán Việt, xây dựng từ điển Hán Việt: Xây dựng công cụ chuyển đổi nhanh giữa văn bản Hán Việt và văn bản chữ, Từ điển Vdict, Từ điển trực tuyến… nhưng những ứng dụng này vẫn còn một số hạn chế như:

- Tất cả ứng dụng trên điều chưa có một kho ngữ vựng dùng chung mang tính chất mở.

- Thiếu định hướng về cấu trúc kho ngữ vựng, tạo khó khăn cho quá trình chia sẽ, tái sử dụng hay kết hợp các kho ngữ vựng Hán Việt lại với nhau.

- Các từ điển hiện nay vẫn còn thiếu nhiều từ Hán Việt gây khó khăn cho người dùng trong việc tra cứu.

Các công cụ tra cứu chỉ hỗ trợ tra nghĩa theo từ khóa nhập vào như từ điển Vdict tuy nhiêu chưa có website cho phép tìm kiếm theo nghĩa của từ khóa, đồng thời hỗ trợ nhiều tùy chọn.

Web ngữ nghĩa có thể giúp chúng ta xây dựng một website giải quyết những khả năng chưa được thực hiện trên. Vì vậy, tôi đã chọn đề tài "Xây dựng Web ngữ nghĩa trợ giúp tra cứu từ Hán Việt" cho luận văn tốt nghiệp của mình.

2. Mục tiêu và nhiệm vụ nghiên cứu

- **Mục tiêu:**

Tìm hiểu được các khái niệm tổng quan về Web ngữ nghĩa, các công cụ, ứng dụng hỗ trợ xây dựng Web ngữ nghĩa. Tìm hiểu từ Hán Việt, về cấu trúc và cách nhận biết các từ Hán Việt.

Xây dựng được một Ontology đầy đủ về từ Hán Việt

Xây dựng được một website thông minh, tìm kiếm và phổ biến thông tin trợ giúp tra nghĩa Hán Việt.

- **Nhiệm vụ:**

Xây dựng Ontology về Hán Việt.

Xây dựng công cụ tìm kiếm nghĩa Hán Việt.

Xây dựng website trợ giúp tra nghĩa Hán Việt đầy đủ và thông minh.

3. Đối tượng và phạm vi nghiên cứu

- **Đối tượng:**

Các vấn đề liên quan đến web ngữ nghĩa.

Xử lý ngôn ngữ tự nhiên

Từ Hán Việt

- **Phạm vi:**

Nghĩa từ Hán Việt

Chương trình dưới dạng Web.

4. Phương pháp nghiên cứu

- **Phương pháp lý thuyết:**

Tìm hiểu về Web ngữ nghĩa.

Tìm hiểu về từ Hán Việt.

Tìm hiểu về xử lý ngôn ngữ tự nhiên.

Tổng hợp từ và nghĩa Hán Việt thu thập được.

- **Phương pháp thực nghiệm**

Xây dựng một Ontology bán tự động

Xây dựng kho dữ liệu Hán Việt có cấu trúc

Xây dựng cơ sở dữ liệu cập nhật tự động và bằng tay

Triển khai thực tế trên Internet.

5. Ý nghĩa khoa học và thực tiễn

- **Ý nghĩa khoa học:**

- Đóng góp một công cụ Search Engine theo công nghệ web ngữ nghĩa trợ giúp người dùng tra cứu nghĩa Hán Việt.

- Phương pháp xây dựng Ontology về từ Hán Việt.

- Ứng dụng sematic web về mặt tìm kiếm.

- Xử lý Tiếng Việt trong Ontology

- **Ý nghĩa thực tiễn:**

- Đây là lĩnh vực chưa được nghiên cứu và phổ biến ở Việt Nam, điều đó mở ra hướng nghiên cứu, ứng dụng mới.

- Đề tài được áp dụng ở Việt Nam, trợ giúp công việc nghiên cứu, học tập và tra cứu của học sinh, sinh viên, các nhà nghiên cứu ngôn ngữ cũng như những người quan tâm đến từ Hán Việt.

- Hỗ trợ tra cứu nghĩa từ Hán Việt chính xác hơn.

- Đem lại ý nghĩa nhân văn.

6. Bố cục luận văn

Luận văn được trình bày bao gồm các nội dung như sau :

Chương 1: Tổng quan về Web Ngữ Nghĩa.

Chương 2: Tìm hiểu từ Hán Việt và giải pháp xây dựng kho từ vựng Hán Việt.

Chương 3: Trình bày giải pháp xây dựng kho từ Hán Việt và web ngữ nghĩa.

CHƯƠNG 1. TỔNG QUAN VỀ WEB NGỮ NGHĨA

1.1. KHÁI NIỆM WEB NGỮ NGHĨA

Theo thống kê của tổ chức W3C, hiện nay thông tin dưới dạng website chiếm gần 70% lượng thông tin giao tiếp trên toàn thế giới và ngày càng không ngừng tăng cao. Với một lượng quá lớn những thông tin như vậy dẫn đến việc quản lý và chia sẻ những thông tin này không còn hiệu quả như mong đợi.

Như vậy, vấn đề đặt ra là những thách thức về việc làm thế nào để web 2.0 có thể chuyển hóa những thông tin văn bản thành những dữ liệu có định dạng đúng với nội dung, nhằm quản lý và sử dụng hiệu quả hơn. Đó là vấn những yêu cầu mà chúng ta cần phải giải quyết.

Web ngữ nghĩa ra đời đáp ứng những yêu cầu tìm kiếm và xử lý thông tin một cách hiệu quả nhất.

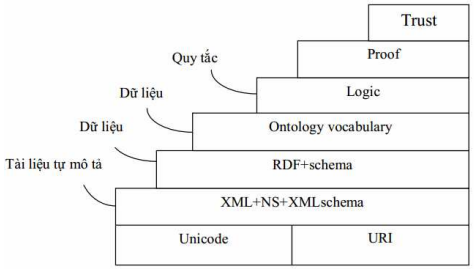
Web ngữ nghĩa không phải là một định dạng web mới riêng biệt. Nó là sự kết hợp giữa web 2.0 hiện tại với những định nghĩa dữ liệu thông minh để nâng cao tính giao tiếp giữa người và máy.

Web ngữ nghĩa được phát triển bởi Tim- Berners Lee, cha đẻ của WWW, URIs, HTTP và HTML.

Hiện nay có các công nghệ hỗ trợ phát triển Web ngữ nghĩa điển hình như theo công nghệ của java có jena, theo công nghệ Microsoft có Semweb, OwlDotNetApi…

Ở Việt Nam, trong khoảng vài năm trở lại đây đã có những nghiên cứu về vấn đề này nhưng chúng ta chỉ tập trung xây dựng các ứng dụng hoặc minh họa cho lý thuyết nghiên cứu.

Mô hình chung của Web ngữ nghĩa:



Hình 1.1 Mô hình các tầng của Web ngữ nghĩa

Mô hình trên có tất cả 7 lớp, trong đó có một số tầng còn đang trong quá trình hoàn thiện. Nội dung các tầng như sau:

Lớp URI, Unicode : đây là tầng cơ bản định nghĩa định dạng xử lý nhằm chuẩn hoá dữ liệu xử lý.

Lớp XML : là ngôn ngữ đánh dấu mở rộng, dùng để lưu trữ dữ liệu, cho phép người dùng có thể tùy ý thêm vào những thẻ theo yêu cầu của mình.

Lớp RDF : khung mô tả tài nguyên RDF - được phát triển dựa trên kỹ thuật lưu trữ dữ liệu của XML và kiểu cấu trúc dữ liệu thông minh để tạo và thay đổi sử dụng các chú thích trong Web ngữ nghĩa.

Lớp Ontology : Ontology là cấu trúc dữ liệu biểu diễn ngữ nghĩa nâng cao. Được phát triển trên nền tảng RDF có phát triển thêm những định nghĩa về từ vựng ngữ nghĩa bổ sung những ràng buộc dữ liệu.

Lớp Logic: Việc biểu diễn các tài nguyên dưới dạng các bộ từ vựng ontology có mục đích là để máy có thể lập luận được trong khi cơ sở lập luận chủ yếu dựa vào logic.

Lớp Proof: Tầng này đưa ra các luật để suy luận. Cụ thể từ các thông tin đã có ta có thể suy ra các thông tin mới.

Lớp Trust: Để đảm bảo tính tin cậy của các ứng dụng trên Web ngữ nghĩa.

1.2. VAI TRÒ CÁC LỚP TRONG KIẾN TRÚC WEB NGHỮ NGHĨA

1.2.1. Vai trò Lớp định danh tài nguyên-URI và Unicode

URI : URI đơn giản chỉ là một định danh Web giống như các chuỗi bắt đầu bằng “http” hay “ftp”.

Một dạng thức quen thuộc của URI là URL - Uniform Resource Locator, URL là một địa chỉ cho phép chúng ta thăm một trang Web.

URI là nền tảng của Web ngữ nghĩa. Trong khi mọi thành phần khác của Web gần như có thể được thay thế nhưng URI thì không.

Unicode: là chuẩn biểu diễn ký tự nhằm mục đích hỗ trợ đa ngôn ngữ. Giúp các trang web ngữ nghĩa thể hiện được trên nhiều ngôn ngữ khác nhau.

1.2.2. Vai trò Lớp XML và XML Schema

XML – (eXtensible Markup Language)là ngôn ngữ đánh dấu mở rộng, cho phép người dùng có thể tùy ý thêm vào những thẻ theo yêu cầu của mình. XML được sử dụng trong web ngữ nghĩa với vai trò định nghĩa cú pháp và cấu trúc của một tài liệu web ngữ nghĩa.

1.2.3. Vai trò Lớp RDF - RDF Schema

RDF 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.

Cấu trúc căn bản của một RDF statement gồm 3 thành phần:

- Tài nguyên (Subject) - là cái mà chúng ta đề cập, thường được nhận diện bởi một URI.
- Vị ngữ (Predicate), có kiểu metadata (ví dụ như tiêu đề, tác giả,...), cũng có thể được xác định bởi một URI.
- Bổ ngữ (Object) ví dụ: một người có tên Eric Miller. Tập hợp các RDF statement được lưu dưới dạng cú pháp của XML, còn được gọi là RDF/XML.

1.2.4. Vai trò Lớp Ontology

Định nghĩa : 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.

Đây cũng là một hướng tiếp cận để xây dựng Web ngữ nghĩa. Tổ chức W3C cũng đã đề ra một ngôn ngữ ontology trên Web (OWL) để xây dựng Sematic Web dựa trên nền tảng của ontology.

Một số lý do cần phát triển một Ontology :

- Để chia sẻ những hiểu biết chung về cấu trúc thông tin giữa con người và các software agent.
- Để cho phép tái sử dụng lĩnh vực tri thức (domain knowledge).
- Để làm cho các giả thuyết về lĩnh vực được tường minh.
- Để tách biệt tri thức lĩnh vực (domain knowledge) ra khỏi tri thức thao tác(operational knowledge).

1.3. CÔNG CỤ XÂY DỰNG ONTOLOGY PROTÉGÉ

1.3.1. 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, protégé có hai phiên bản OWL và API.

Protégé-OWL được phát triển dựa trên hai yêu cầu chính : định nghĩa các đối tượng và quan hệ tồn tại giữa chúng.

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

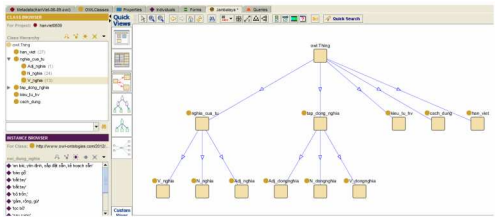
1.3.2. Protégé sử dụng giao diện đồ họa

Ngay từ phiên bản Protégé API, thì phần mềm đã không chỉ cho phép tạo mô hình bằng cách thủ mà nó còn cho phép người sử dụng giao diện đồ họa để phát triển.

Hình 1.2 Giao tiếp bằng đồ họa của Protégé

1.3.3. Protégé phát triển để tích hợp các công cụ

Protégé cung cấp một số điểm mở rộng nơi các nhà phát triển có thể chủ động thêm các thành phần mà ta thường gọi là plug-ins.



Hình 1.3 Protégé tích hợp công cụ Jabalaya

1.4. THƯ VIỆN PHÁT TRIỂN ỨNG DỤNG WEB NGỮ NGHĨA

1.4.1. SemWeb

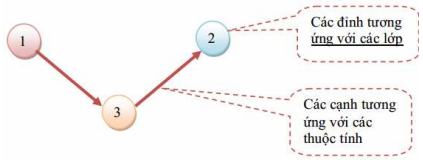
SemWeb lần đầu tiên được phát hành vào tháng sáu năm 2005 và đã được thử nghiệm gần đây hơn với những bộ lưu trữ hơn một tỉ bộ ba. Các tính năng cốt lõi như đọc/ghi dữ liệu XML với bộ ba RDF, liên tục lưu trữ dữ liệu với nền tảng SQL và các truy vấn SPARQL cơ bản đã được kiểm nghiệm nhiều lần. Thư viện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

1.4.2. OwlDotNetApi

OwlDotNetApi là một OWL API với bộ phân tích cú pháp viết bằng C# theo công nghệ .NET dựa trên phân tích cú pháp RDF Drive.

- **Phiên bản**
- **Chức năng**

Mục tiêu của OwlDotNetApi là đọc/ghi dữ liệu của XML dựa trên đồ thị với các cạnh tương ứng với thuộc tính liên kết và các đỉnh tương ứng với các nút hay còn gọi là các lớp.



Hình 1.4 Mô hình quan hệ giữa các nút và các cạnh

Xuất phát từ việc đồ thị hoá nội dung của dữ liệu nên OwlDotNetApi đáp ứng được hầu hết tất cả các chuẩn mà W3C đưa

ra. Tuy nhiên việc truy cập dữ liệu không thông qua câu lệnh truy vấn nên việc lập trình với thư viện này chưa thuận lợi về thời gian xử lý.

CHƯƠNG 2. TÌM HIỂU TỪ HÁN VIỆT VÀ GIẢI PHÁP XÂY DỰNG KHO TỪ VỰNG HÁN VIỆT

2.1. TÌM HIỂU VỀ TỪ HÁN VIỆT

2.1.1. Nguồn gốc từ Hán Việt

Chữ Hán hay còn được gọi là chữ Nho được người Hán sáng tạo cách đây khoản hơn 3000 năm.

Ở nước ta, trước khi sử dụng văn tự Hán cách đây 3000 năm, người Việt đã có ngôn ngữ riêng của mình, đó là ngôn ngữ cổ Việt Mường.

Vào thế kỷ thứ nhất trước Công Nguyên cùng với việc phong kiến phương Bắc xâm lược Việt Nam, cũng do đặc điểm địa lý, có sự giao lưu giữa cư nhân hai thì ngôn ngữ văn tự Hán cũng được đưa vào Việt Nam.

Người Việt dùng các từ ngữ gốc Hán ghép với nhau theo cách riêng của mình để tạo ra từ Hán Việt.

Về sau, người Việt dùng văn tự này để ghi lại tiếng nói của mình (tức là chử nôm).

2.1.2. Các đặc điểm của từ Hán Việt

Theo các nhà nghiên cứu ngôn ngữ thì ước chừng có khoản 60% số từ Hán Việt trong ngôn ngữ hiện nay của chúng ta.

Việc sử dụng Hán Việt rất khó khăn. Có nhiều sự hiểu sai từ Hán Việc dẫn đến cách dùng từ Hán Việt sai lệch trong văn bản và lời nói.

Về năng lực hoạt động, khả năng nhập hệ của các từ gốc Hán trong tiếng Việt, rất không đồng đều.

Đôi khi trong những tổ hợp vay mượn nguyên khối từ gốc Hán, nói mới lưu giữ ý nghĩa .

Với cách nhập lẻ tẻ, các từ đơn tiết Hán Việt xuất hiện với vai trò lấp đầy, bổ sung những khái niệm mới cho các trường từ vựng.

Sự xuất hiện theo trường từ vựng của các từ Hán- Việt mới trong Tiếng Việt một mặt thể hiện ảnh hưởng của văn hóa văn minh Trung Hoa đối với châu Á nói chung và Việt Nam nói riêng.

2.1.3. Cấu trúc từ Hán Việt

a. Từ đơn Hán Việt

➢ Từ đơn Hán Việt nhìn theo tiêu chí ngữ âm

- Từ đơn thuần âm Hán Việt
- Từ đơn biến âm Hán Việt

➢ Từ đơn Hán Việt nhìn từ tiêu chí ngữ nghĩa

Nghĩa của từ đơn Hán Việt ở đây có thể phân ra hai loại :

- Từ đơn Hán Việt theo nghĩa
- Từ đơn Hán Việt biến

➢ Từ đơn Hán Việt nhìn theo tiêu chí ngữ pháp

- Từ đơn Hán Việt là danh từ
- Từ đơn Hán Việt là động từ
- Từ đơn Hán Việt là tính từ

b. Từ ghép Hán Việt

Từ ghép Hán Việt là những từ do hai yếu tố Hán Việt có nghĩa ghép lại với nhau mà thành.

➢ Từ ghép Hán Việt nhìn theo tiêu chí ngữ âm

- Từ ghép thuần âm Hán Việt
- Từ ghép biến âm Hán Việt

➢ Từ ghép Hán Việt nhìn từ tiêu chí ngữ nghĩa

- Từ ghép nguyên nghĩa Hán Việt
- Từ ghép Hán Việt biến nghĩa

➢ Từ ghép Hán Việt nhìn từ tiêu chí ngữ pháp

- Từ ghép Hán Việt đẳng lập
- Từ ghép chính phụ Hán Việt

2.1.4. Các luật nhận biết từ Hán Việt

Chúng ta sẽ sử dụng các mẹo tên để nhận biết từ Hán Việt để có được kho từ Hán Việt chính xác trong giai đoạn xây dựng kho từ thô Hán Việt.

2.2. HIỆN TRẠNG VÀ NHU CẦU TRA CỨU TỪ HÁN VIỆT HIỆN NAY

2.2.1. Nhu cầu tra cứu từ Hán Việt

2.2.2. Hiện trạng tra cứu từ Hán Việt

Hiện nay đối với học sinh, sinh viên vấn đề sử dụng đúng ngôn ngữ tiếng Việt cũng là một vấn đề hết sức khó khăn. Có thể kể ra đây một số lỗi thường gặp như :

- Dùng từ sai phong cách
- Viết sai chính tả
- Sử dụng từ không đúng

Những trường hợp trên đây xuất phát từ một thực trạng là học sinh không hiểu được nghĩa cũng như phạm vi sử dụng của từ Hán Việt.

Các từ điển hiện nay vẫn còn thiết nhiều từ gây khó khăn cho người dùng.

Trong tiếng Việt, từ Hán Việt chiếm số lượng tương đối cao - trên 60%, gây khó khăn cho người tiếp nhận và sử dụng.

Trên thực tế, trước nay đã có nhiều công trình nghiên cứu, chuyên luận bàn ở nhiều khía cạnh khác nhau và hỗ trợ khả năng sử dụng từ Hán Việt cho các đối tượng người dùng như: “Mẹo giải nghĩa từ Hán Việt và chữa lỗi chính tả” của tác giả Phan Ngọc, từ điển Hán Việt.

2.2.3. Tìm hiểu từ điển

Từ điển là cách tra cứu tập hợp các đơn vị ngôn ngữ (thường là đơn vị từ vựng) và sắp xếp theo một tật tự nhất định, cung cấp một số kiến thức cần thiết đối với từng đơn vị.

Các loại từ điển hiện nay

- Từ điển giấy
- Từ điển điện tử
- Từ điển máy tính

2.3. GIẢI PHÁP XÂY DỰNG KHO TỪ HÁN VIỆT

Khi xây dựng kho từ phục vụ cho quá trình làm ontology chúng ta gặp phải vấn đề là dữ liệu từ đâu ra và tập hợp chúng như thế nào? Làm thế nào để có được dữ liệu chính xác nhất là vấn đề rất được tôi quan tâm. Trong phạm vi luận văn tôi sẽ sử dụng một số nghiên cứu của các tác giả khác với kết quả thực nghiệm đã được công nhận trong thực tế . Nguồn dữ liệu để xây dựng kho từ sẽ được lấy chủ yếu ở trong các từ điển Hán Việt, từ điển Hán Việt online …

2.3.1. Vấn đề xử lý ngôn ngữ tự nhiên

2.3.2. Sơ lược bài toán tách từ

Sau đây tôi xin giới thiệu một số vấn đề liên quan đến bài toán tách từ trong tiếng Việt để làm giàu ontology từ nguồn dữ liệu lấy từ internet.

Các hướng tiếp cận cho bài toán tách từ :

- Hướng tiếp cận dựa trên từ
- Hướng tiếp cận dựa trên ký tự

2.3.3. Công cụ vnTokenize

vnTokenizer là công cụ tách từ tiếng Việt được nhóm tác giả Nguyễn Thị Minh Huyền, Vũ Xuân Lương và Lê Hồng Phương phát triển dựa trên phương pháp so khớp tối đa (Maximum Matching) với

tập dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt.

2.3.4. Xây dựng kho từ Hán Việt

a. Quy mô

- Xây dựng cấu trúc kho
- Thu thập nguồn dữ liệu
- Giải thích từ vựng: chúng ta sẽ dùng xây dựng thủ công và tự động.

b. Chọn lọc dữ liệu đưa vào kho

Là dữ liệu đưa vào kho ngữ vựng, các nguồn dữ liệu :

- Kho từ đơn và kho từ .
- Kho dữ liệu trung gian .
- Kho dữ liệu thô .

c. Đề xuất cấu trúc lưu trữ kho

Chúng ta tổ chức kho dữ liệu theo cấu trúc Alphabet tức là ta tổ chức các mục từ theo thứ tự ABC và lưu theo kiểu file XML.

2.4. GIẢI PHÁP XÂY DỰNG ONTOLOGY HÁN VIỆT

Mô hình ontology tôi xây dựng sẽ dựa theo mô hình ontology hiện có trong Wordnet.

2.4.1. Giới thiệu Wordnet

Năm 1980, Miller và cộng sự tại trường Đại học Princeton (Mỹ) đã xây dựng WordNet, là một cơ sở dữ liệu tri thức ngữ nghĩa từ vựng bằng tiếng Anh.

a. Mô hình Wordnet

WordNet là một loại từ điển tương tự từ điển đồng nghĩa. WordNet phân chia từ vựng thành 5 loại : noun, verb, adjective, adverb và funtion words, nhưng thực tế nó chỉ chứa noun, verb, adjective, adverb.

b. Các quan hệ trong WordNet

- Quan hệ đồng nghĩa (synonymy)
- Quan hệ trái nghĩa (antonymy)
- Quan hệ hạ danh (thuộc cấp hyponym) và quan hệ thượng danh (bao hàm, hypernym)
- Quan hệ bộ phận (meronymy/ holonymy)
- Quan hệ kéo theo (entailment)
- Quan hệ cách thức đặc biệt (troponymy)

2.4.2. Thiết kế mô hình dữ liệu Ontology

- Trong ontology sẽ xây dựng gồm 5 class lớn là :

- Han_viet
- Nghia_cua_tu : Đây là class chứa các class con n_nghia, v_nghia, adj_nghia.
- Tap_dong_nghia : chứa các class con n_dongnghia, v_dongnghia, adj_dongnghia.
- Kieu_tu_hv : là class dùng để chỉ kiểu từ Hán Việt.
- Cach_dung : là class dùng để thể hiện các sử dụng từ Hán Việt.

- Thuộc tính :

Đối tượng từ Hán Việt (han_viet): Trong class này ta sẽ định nghĩa thuộc tính cơ bản của từ đó là tên, id từ, kiểu từ và có một property thể hiện nghĩa của từ (co_nghia) .

Đối tượng nghĩa của từ (nghia_cua_tu) : Các lớp con là n_nghia, v_nghia, adj_nghia gồm có: id_nghia , noi_dung_nghia , co_tap_dong_nghia, trai_nghia, co_tu_hanviet.

Đối tượng tập đồng nghĩa (tap_dong_nghia) : các lớp tương ứng là n_dongnghia, v_dongnghia, adj_dongnghia gồm : id_dongnghia, mo_ta, vi_du.

Đối tượng kiểu từ (kieu_tu) : Trong class này sẽ có thuộc tính kieu_tu để định nghĩa kiểu từ.

Đối tượng các dùng (cach_dung)

Doi_tuong : thể hiện đối tượng của từ Hán Việt.

Hoan_canh : thể hiện hoàn cảnh sử dụng.

Ngu_phap : thể hiện vị trí đặt từ.

CHƯƠNG 3. PHÁT TRIỂN ỨNG DỤNG

3.1. PHÂN TÍCH BÀI TOÁN

3.1.1. Xác định đối tượng sử dụng

Trong giới hạn luận văn tôi sẽ nghiên cứu và phát triển ứng dụng phục vụ cho đối tượng là học sinh, sinh viên.

3.1.2. Yêu cầu bài toán

Bài toán đặt ra yêu cầu xây dựng một trang web giúp người dùng tra cứu và sử dụng từ Hán Việt với những yêu cầu chức năng như :

- Thu thập từ Hán Việt từ Internet, sách báo, từ điển tạo kho từ Hán Việt dựa nghĩa.
- Quản lý các từ mới tìm được, chỉnh sửa các thông tin.
- Cho phép người dùng tìm kiếm, tra cứu từ Hán Việt.
- Website lưu trữ đầy đủ thông tin về từ Hán Việt .

3.1.3. Phân tích hệ thống

a. Hướng tiếp cận

Chương trình được xây dựng là một Semantic Web. Công nghệ Web Semantic sử dụng mô hình dữ liệu thông minh.

Chương trình hỗ trợ tra cứu từ Hán Việt sẽ được xây dựng dựa trên đối tượng chính là từ Hán Việt, cụ thể ở đây chúng ta có từ đơn và từ ghép.

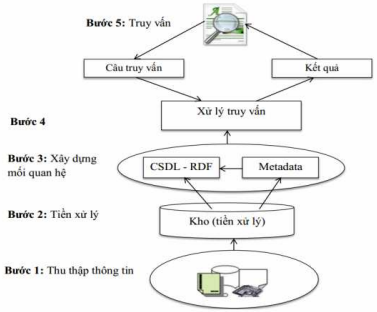
Với công việc xác định là phát triển một trang web semantic ta cần xây dựng ứng dụng gồm 2 phần chính :

Ontology : Trong phần này chúng ta sẽ tiến hành xây dựng các lớp, các thuộc tính và tạo ra các mối quan hệ đồng cấp, phân cấp theo W3C và tất cả các định nghĩa mới đã xác định cho ontology.

Trình duyệt web : Phần trình duyệt ta không xây dựng mới hoàn toàn đáp ứng đáp ứng đầy đủ các yêu cầu truy cập dữ liệu ở bất kỳ ontology nào mà ta xây dựng trình duyệt tương tự các ứng dụng web hiển thị nội dung cơ sở dữ liệu đã xây dựng.

b. Mô hình hóa

Đây là bài toán dựa trên cơ sở dữ liệu được lưu trữ và đưa thông tin một cách thông minh về phía người dùng. Trước khi có thiết kế chi tiết ta cần phân chia chương trình làm 5 hạn mục chính bao gồm các phần ta có thể tóm lại các mục của mô hình bằng hình vẽ bên dưới.



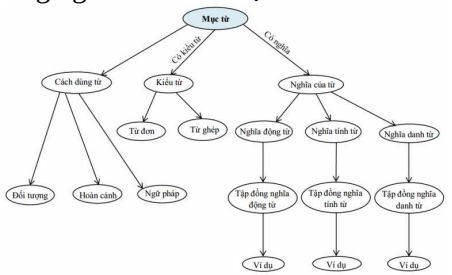
Hình 3.1 Mô hình tổng quát hệ thống.

3.2. XÂY DỰNG ONTOLOGY

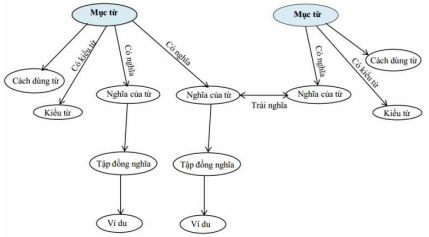
Đối với vấn đề tìm kiếm dữ liệu ngữ nghĩa trong bài toán này là xác định các thông tin mà ta cần tìm kiếm, ở đây các thông tin cần tìm kiếm cho một từ Hán Việt là ngữ nghĩa, loại từ của từ đó. Vì vậy ý nghĩa và các dữ liệu liên quan phải được lưu trữ trong lớp và đây là những lớp quan trọng của bài toán cần xây dựng.

Dữ liệu liên quan đến từ cần tìm kiếm gồm có: nghĩa của từ, loại từ , từ đồng nghĩa, từ phản nghĩa.

Trong cấu trúc được xây dựng chúng ta sẽ quản lý các từ, nghĩa của từ, các tập đồng nghĩa và các thuộc tính đi kèm của từ.



Hình 3.2 Từ trong ontology Hán Việt



Hình 3.3 Mối quan hệ trong ontology Hán Việt

3.2.1. Công cụ xây dựng ontology

Ontology Hán Việt được xây dựng dùng công cụ soạn thảo Protégé.

3.2.2. Các bước xây dựng ontology

Dựa trên các bước xây dựng ontology của Noy và McGuinness ta có sự tinh gọn công việc trong mỗi bước như sau:

- Bước 1. Xác định mục đích phát triển ontology.

Chúng ta đã thấy được các kho từ Hán Việt hiện nay vẫn còn nhiều hạn chế về tính mở cũng như cấu trúc đã được nêu ra ở chương 2.

Xây dựng ontology Hán Việt giúp mô tả mối quan hệ giữa các từ được tường minh và dễ truy vấn hơn.

Người dùng có thể sử dụng hay kế ontology Hán Việt để phát triển các chức năng như người dùng mong muốn.

➢ Bước 2. Nắm bắt kỹ thuật xây dựng ontology :

Bước này gồm ba giai đoạn như sau :

- Xác định phạm vi của ontology : gồm kiểu từ là từ đơn và từ ghép Hán Việt, các loại từ chính gồm có danh từ, động từ và tính từ. Các mối quan hệ quan trong gồm: quan hệ về nghĩa là mối quan hệ đồng nghĩa phản nghĩa, phương pháp sử dụng hợp lý từ Hán Việt.

- Chọn phương thức nắm bắt ontology : phân tích hướng đối tượng tập trung vào các phương thức trong lớp.

- Định nghĩa các khái niệm trong ontology: Chúng ta tiến hành định nghĩa các khái niêm cho ontology gồm : Từ Hán Việt, nghĩa của từ, tập đồng, kiểu từ và sử dụng.

➢ Bước 3. Xem xét sử dụng lại các ontology đang tồn tại.

Hiện nay có ontology Wordnet có cấu trúc khá phù hợp với yêu cầu đặt ra của bài toán là xây dựng một ontology Hán Việt.

➢ Bước 4. Mã hoá ontology

Luận văn sử dụng công cụ Protégé để mã hoá ontology. Việc mã hóa liên quan đến biểu diễn ontology trong một ngôn ngữ hình thức. Lớp trong ontology mô tả các khái niệm cùng các thuộc tính và quan hệ. Mã hóa ontology là tiến trình lặp, gồm các bước con sau:

- Định nghĩa lớp : Để tiện việc phân biệt các lớp "thông tin liên quan" với các lớp con của các lớp này, ta gọi các lớp ngoài cùng là siêu lớp. Các lớp con bên trong ta vẫn gọi bình thường là lớp. Như vậy quan hệ giữa cá từ và các lớp bên trong.

Hình 3.5 Class trong ontology

Tập đồng nghĩa <owl:Class rdf:about="#tap_dong_nghia">: Nó là một tập hợp các từ đồng nghĩa, các lớp con ở mức thấp hơn : Tính từ đồng nghĩa, động từ đồng, danh từ đồng nghĩa.

Nghĩa của từ Hán Việt <owl:Class rdf:about="#nghia_cua_tu"> : gồm các lớp con như sau : nghĩa của tính từ, nghĩa của danh từ, nghĩa của động .

Từ Hán Việt <owl:Class rdf:about="#han_viet">: lớp chứa các từ Hán Việt.

Kiểu từ: <owl:Class rdf:about="#kieu_tu">: lớp chứa các kiểu Hán Việt.

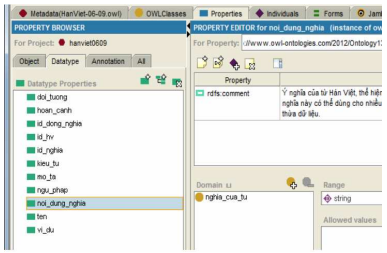
Cách dùng từ: <owl:Class rdf:about="#cach_dung">: lớp chứa các kiểu Hán Việt.

- Mô tả thuộc tính: Các thuộc tính thể hiện mối quan hệ giữa các đối tượng dữ liệu individual) với nhau hoặc quan hệ với dữ liệu Ngôn ngữ :

Dưới đây là một số thuộc tính dữ liệu cơ bản có trong chương trình:

Thuộc tính dữ liệu: Thuộc tính id_hv, id_dong_nghia, id_nghia, ten, kieu_tu, mo_ta, noi_dung_nghia, doi_tuong.

Thuộc tính quan hệ : Thuộc tính co_nghia, co_tu_hanviet, co_Tap_dong_nghia, trai_nghia, co_kieu, co_cach_dung .



Hình 3.6 Thuộc tính datatype trong ontology

- Bước 5. Cải tiến ontology

Bao gồm hai giai đoạn :

Cải tiến mã hóa bên trong (intra-coding)

Cải tiến mã hóa bên ngoài (extra-coding)

- Bước 6 : Kiểm thử

Phát hiện nhược điểm của ontology. Bước này được thực hiện trong tất cả các giai đoạn phát triển. Ngay khi tạo cơ sở tri thức, cần tiến hành kiểm thử để phát hiện lỗi trong ontology và công cụ thu nhận tri thức, và sửa đổi ontology hợp lý.

- Bước 7 : Duy trì

Thực hiện các việc hiệu chỉnh, thích ứng hoặc hoàn tất ontology Hán Việt.

3.2.3. Kết quả Ontology

Sau khi đã đĩnh nghĩa các class cũng như các đối tượng trong luận văn thông quá công cụ protégé ta sẽ save lại thành một file có định dạng theo đuôi chuẩn chung là ".owl".

3.3. XÂY DỰNG WEBSITE TRA TỪ HÁN VIỆT

3.3.1. Giải pháp xây dựng

Khai thác thư viện mã nguồn mở OwlDotNetApi.

Truy xuất dữ liệu ontology sang giao diện web

Thuật toán này dùng để điền đầy các quan hệ của ứng dụng và tạo cho ứng dụng có thông tin hai chiều.

Đối với vấn đề này luận văn sẽ xây dựng thuật toán như sau :

Mở tệp tin chứa ontology

Đọc tất cả các Properties có khai báo đưa vào danh sách đối chiếu.

Duyệt qua tất cả các đỉnh của ontology

Nếu một đỉnh có chứa quan hệ cần điền đầy theo danh sách đối chiếu ở trên (B1)

Điền thông tin quan hệ ngược lại

Quay lại xét cho đỉnh vừa điền như B1

Ngược lại bỏ qua bước này

Đóng truy cập vào ontology

Duyệt ngữ nghĩa từ ontology

3.3.2. Xây dựng giao diện

Website được phát triển trên nền.Net, với ngôn ngữ C# và ASP.Net. Công cụ dùng để triển khai là Visual Studio 2008 sử dụng thư viện OwlDotNetApi.

Chương trình có một số chức năng cơ bản như sau :

a. Trang chính của hệ thống : Đây là trang chứa menu với chức năng là thực hiện đọc dữ liệu từ nội dung ontology, lấy các siêu lớp .

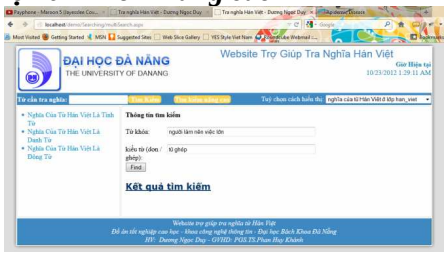
b. Các thuật toán bổ trợ cho việc xây dựng các thuật toán tìm kiếm

c. Trang thực hiện tìm kiếm đơn giản

Chức năng tìm kiếm đơn giản dựa theo từ khóa nhập vào bàn phím để tìm kiếm nghĩa của từ Hán Việt cần tra. Việc tìm kiếm sẽ dựa trên sự đối chiếu, so khớp thông tin từ các từ khoá nhập vào của người dùng.

Hình 3.11 khung tìm kiếm đơn giản

d. Trang thực hiện tìm kiếm nâng cao



Hình 3.12 Hình ảnh tìm kiếm nâng cao

Khi người dùng sử dụng chức năng tìm kiếm đơn giản thì kết quả trả về thường nhiều vì người dùng thường nhập vào từ khóa đơn giản là từ muốn tìm. Vì vậy để kết quả chính xác hơn thì việc cung cấp thông tin ngữ nghĩa cho quá trình tìm kiếm là điều rất được quan tâm.

e. Trang chi tiết

3.3.3. Thống kê và đánh giá kết quả

Trong quá trình nghiên cứu xây dựng web ngữ nghĩa trợ giúp tra cứu từ Hán Việt cho đến nay đã đạt được những kết quả sau :

- Xây dựng ứng dụng web ngữ nghĩa hỗ trợ tra cứu từ Hán Việt với những chức năng tra cứu nghĩa đơn giản và nâng cao.
- Đã tạo được ontology Hán Việt khoảng 500 từ đơn và từ ghép Hán Việt. Trong thời gian đến ontology Hán Việt sẽ tiếp tục được cập nhật dữ liệu.

KẾT LUẬN

1. Kết quả đạt được

Về mặt lý thuyết

- Nắm được các kiến thức về web ngữ nghĩa, cách xây dựng ontology và ứng dụng web ngữ nghĩa .
- Tìm hiểu được cấu trúc nghĩa từ Hán Việt từ đó áp dụng xây dựng được kho từ Hán Việt cơ bản và ontology Hán Việt.

Về mặt thực tiễn

- Xây dựng được kho từ Hán Việt.
- Xây dựng ontology Hán Việt và web ngữ nghĩa hỗ trợ tra nghĩa từ Hán Việt.
- Góp phần giúp cho mọi người có một công cụ tra cứu nghĩa của từ Hán Việt phục vụ nhu cầu học tập nghiên cứu của học sinh – sinh viên, những người có nhu cầu tìm hiểu, tra nghĩa từ Hán Việt.

2. Hướng phát triển của đề tài

- Trong luận văn tôi đã tái sử dụng lại một phần cấu trúc ontology Wordnet để xây dựng ontology Hán Việt và vẫn chưa khai thác hết thế mạnh của bộ ontology này.
- Với vốn kiến thức về từ Hán Việt khá hạn chế, tôi hy vọng trong tương lai sẽ có sự góp mặt của các chuyên gia ngôn ngữ để dữ liệu được chính xác hơn.
- Phát triển bài toán có thể thêm các ký tự tiếng trung vào ontology giúp hoàn thiện hơn chức năng tra hỗ trợ tiếng trung.