BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

PHẠM AN BÌNH

TÌM HIỂU CÔNG NGHỆ KIM XÂY DỰNG ỨNG DỤNG CHÚ GIẢI NGỮ NGHĨA TỰ ĐỘNG

Chuyên ngành : Khoa học máy tính

Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - 2010

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS. TS. Phan Huy Khánh**

Phản biện 1 : **TS. Nguyễn Mậu Hân**

Phản biện 2 : **TS. Tăng Tấn Chiến**

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 14 tháng 10 năm 2010.

** 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

Với nhiều tỷ trang web phân bố trên hầu hết các quốc gia, World Wide Web (WWW) là môi trường tốt cho việc biểu diễn và truy cập thông tin dạng số. Tuy nhiên, lượng thông tin khổng lồ đó cũng tạo ra những khó khăn to lớn trong việc tìm kiếm, chia sẻ thông tin trên WWW. Hiện nay thông tin trên WWW được biểu diễn chủ yếu dưới dạng ngôn ngữ tự nhiên. Cách biểu diễn đó phù hợp với con người nhưng gây ra nhiều khó khăn cho các chương trình hỗ trợ tìm kiếm, chia sẻ và trao đổi thông tin. Máy tính không "hiểu" được thông tin và dữ liệu biểu diễn dưới dạng thích hợp với con người.

Để giải quyết vấn đề này, nhiều tổ chức nghiên cứu và kinh doanh đã phối hợp nghiên cứu và phát triển Web có ngữ nghĩa. Theo Tim Berner Lee giám đốc tổ chức World Wide Web Consortium, đồng thời là cha đẻ của WWW, Web có ngữ nghĩa là sự mở rộng của web hiện tại bằng cách thêm vào các mô tả ý nghĩa cho nội dung của trang web dưới dạng mà máy tính có thể hiểu được, do đó có thể xử lý thông tin hiệu quả hơn. Như vậy web có ngữ nghĩa sẽ bao gồm các thông tin được biểu diễn theo cách truyền thống cùng với ngữ nghĩa của các thông tin này được biểu diễn một cách tường minh. Việc thêm phần ngữ nghĩa cung cấp thêm tri thức cho các chương trình, giúp nâng cao chất lượng phân loại, tìm kiếm và trao đổi thông tin.

Sự ra đời của web ngữ nghĩa là một bước tiến vượt bậc so với kỹ thuật web thông thường và hứa hẹn một thế hệ web tương lai. Các phát triển gần đây của công nghệ thông tin và truyền thông đã tạo ra những khả năng để thu thập một lượng lớn dữ liệu mà chúng có liên quan với nhau về mặt khái niệm. Tuy nhiên, đa số những mối quan hệ này được con người "nhớ" chứ không được lưu trữ theo một

cách mà giúp cho máy tính có thể hiểu để xử lý.Thách thức này tạo ra một hướng nghiên cứu đó là tạo ra khả năng cho phép con người tạo, lưu giữ, sắp xếp, ghi phụ chú và truy xuất kho dữ liệu cá nhân rất lớn của mỗi người trong quá khứ theo hình thức như một nhật ký cuộc sống được cá thể hóa và trợ giúp cho bộ nhớ của con người.

Hiện nay, có nhiều hướng nghiên cứu khác nhau về web ngữ nghĩa, như chuẩn hóa ngôn ngữ biểu diễn dữ liệu và siêu dữ liệu trên web, chuẩn hóa ngôn ngữ biểu diễn ontology và phát triển ngữ nghĩa cho web. Đối với hướng nghiên cứu phát triển ngữ nghĩa cho web, người ta tìm cách bổ sung ngữ nghĩa vào các trang web, trong khi có hàng tỷ trang web như vậy trên toàn cầu. Do đó, việc xây dựng các hệ thống tự động chuyển đổi các trang web truyền thống sang các trang web có ngữ nghĩa là vô cùng cần thiết, mang lại nhiều lợi ích và ý nghĩa to lớn. Để thực hiện điều này, chúng ta cần phân tích và trích lọc các ngữ nghĩa và ghi tự động xuống các trang web dưới dạng các chú giải. Đó là lý do tôi chọn đề tài:

“ Tìm hiểu công nghệ KIM

Xây dựng ứng dụng chú giải ngữ nghĩa tự động”

2. MỤC TIÊU VÀ NHIỆM VỤ

Luận văn tập trung vào nghiên cứu những nội dung sau đây:

Thứ nhất, nghiên cứu các nội dung lý thuyết liên về web ngữ nghĩa, chú giải ngữ nghĩa cho trang web.

Thứ hai, nghiên cứu tìm hiểu hệ thống quản lý thông tin và tri thức KIM.

Từ những lý thuyết, kiến thức thu được sau khi nghiên cứu những nội dung trên, luận văn tập trung “xây dựng ứng dụng chú giải

ngữ nghĩa tự động” và đưa ra một số nhận định, kết quả thực hiện đồng thời đề xuất các hướng phát triển của luận văn trong tương lai.

3. ĐỐI TƯỢNG VÀ PHẠM VI NGHÊN CỨU

Đối tượng nghiên cứu của luận văn là dữ liệu dạng văn bản được biểu diễn trên môi trường www. Luận văn tập trung vào nghiên cứu hệ thống quản lý thông tin và tri thức KIM, sau đó xây dựng ứng dụng chú giải ngữ nghĩa tự động.

4. PHƯƠNG PHÁP NGHIÊN CỨU

Luận văn sử dụng các phương pháp nghiên cứu sau :

Thứ nhất, tổng hợp các kết quả nghiên cứu từ các tư liệu liên quan về web ngữ nghĩa, chú giải ngữ nghĩa, KIM.

Thứ hai, phân tích đánh giá các phương pháp và đề xuất các giải pháp lựa chọn để xây dựng ứng dụng có hiệu quả nhất.

Từ những giải pháp lựa chọn đã đề xuất, chọn ra một phương pháp hiệu quả để áp dụng cho việc xây dựng ứng dụng chú giải ngữ nghĩa tự động.

5. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN CỦA ĐỀ TÀI

Đề tài tập trung nghiên cứu, tìm hiểu về công nghệ KIM và tìm hiểu khả năng ứng dụng công nghệ KIM. KIM là một công nghệ còn khá mới mẻ không những trên thế giới mà còn cả ở Việt Nam.

Đề tài đề xuất một hướng tiếp cận mới trong tăng cường ngữ cảnh vào các trang Web bằng cách bổ sung các chú giải tự động vào các trang web, nhằm tăng thêm hiệu quả tìm kiếm, trích lọc, chia sẻ, ... thông tin trên web.

Đề tài cũng góp phần nâng cao khả năng tổ chức và triển khai thành công hệ thống web ngữ nghĩa trong thực tế, giúp người sử dụng hệ thống dễ dàng tìm kiếm được các thông tin mong muốn chính xác hơn và hiệu quả hơn.

6. BỐ CỤC CỦA LUẬN VĂN

Luận văn gồm 3 chương, sau phần mở đầu giới thiệu về lý do chọn đề tài, mục tiêu và nhiệm vụ, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu , ý nghĩa khoa học và thực tiễn của đề tài là:

Chương 1, "**Tìm hiểu web ngữ nghĩa và hệ thống chú giải ngữ nghĩa**" giới thiệu sơ bộ những nội dung tổng quan nhất về sự ra đời của WEB ngữ nghĩa, kiến trúc, ngôn ngữ của WEB ngữ nghĩa. Trong phần này cũng trình bày tổng quan về phương pháp truy vấn dữ liệu trong RDF.

Bên cạnh đó, chương này cũng tập trung trình bày về chú giải ngữ nghĩa, mô hình tổng quát cho hệ thống chú giải ngữ nghĩa tự động, các phương pháp tách từ.

Chương 2, "**Tìm hiểu hệ thống quản lý thông tin và tri thức KIM**" . Trong chương này, luận văn giới thiệu về hệ thông quản lý thông tin và tri thức KIM, đi sâu vào nền tảng, cấu hình, kiến trúc của KIM. Quá trình trích lọc thông tin ngữ nghĩa, chú giải và khôi phục cũng như tính khả thi và giá trị to lớn của KIM.

Chương 3, "**Xây dựng ứng dụng chú giải ngữ nghĩa tự động**". Trong chương này tập trung nghiên cứu phân tích xây dựng kiến trúc tổng thể của hệ thống gồm các thành phần liên quan, cách vận hành của hệ thống, từ kiến trúc tổng thể đã xây dựng tiếp tục triển khai thiết kế các thành phần đã phân tích, xây dựng cơ sở dữ liệu, ứng dụng chú giải ngữ nghĩa tự động.

Phần kết luận, tổng hợp những kết quả nghiên cứu chính của luận văn, chỉ ra một số hạn chế chưa hoàn thiện cài đặt. Đồng thời, luận văn cũng đề xuất một số hướng nghiên cứu cụ thể tiếp theo của tác giả luận văn.

CHƯƠNG 1 - WEB NGỮ NGHĨA VÀ HỆ THỐNG CHÚ GIẢI NGỮ NGHĨA

1.1. CÁC VẤN ĐỀ LIÊN QUAN ĐẾN WEB NGỮ NGHĨA

1.1.1. Sự hạn chế ở World Wide Web

1.1.2. Sự ra đời của Web ngữ nghĩa

1.1.2.1. Web ngữ nghĩa

Theo Tim- Berners Lee, ‘‘ *Web ngữ nghĩa là sự mở rộng của Web hiện tại, cho phép người dùng có thể truy tìm, phối hợp, sử dụng lại và trích lọc thông tin một cách dễ dàng và chính xác* ’’.

1.1.2.2. Một số khái niệm liên quan

Phần này trình bày về Meta data và ontology.

1.1.3. Kiến trúc của Web ngữ nghĩa

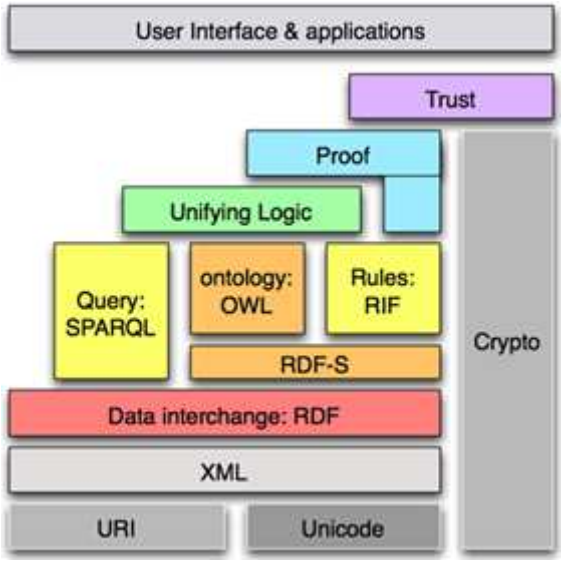
Web ngữ nghĩa là sự mở rộng của web hiện tại có bổ sung thêm ngữ nghĩa vào dữ liệu trên web.

Từ sơ đồ kiến trúc của web ngữ nghĩa ở trên ta thấy có bảy tầng kiến trúc. Với hệ thống web hiện tại là đang ở tầng thứ hai.

1.1.3.1. Unicode: là bảng mã chuẩn chung chứa đầy đủ các ký tự nhằm đáp ứng tính nhất quán toàn cầu của web.

1.1.3.2 URI (Uniform Resource Identifier):là một chuỗi theo hình thức chuẩn cho phép nhận diện các tài nguyên duy nhất.

1.1.3.2. XML: chứa các định nghĩa về XML namespace và XML Schema nhằm có một cú pháp chung được sử dụng trong web ngữ nghĩa. XML là ngôn ngữ đánh dấu tài liệu chứa các thông tin có cấu trúc. Một tài liệu XML chứa các element, các element này có thể lồng nhau và có thể có các thuộc tính và nội dung. XML namespace cho phép chỉ định sự khác nhau của các từ vựng đánh dấu trong một tài liệu XML.



Kiến trúc Web ngữ nghĩa

1.1.3.3. Lớp RDF - RDF Schema: định dạng biểu diễn dữ liệu nồng cốt của web ngữ nghĩa là RDF. RDF là một khung biểu diễn thông tin tài nguyên dưới dạng một hình ảnh.

RDFS (RDF Schema) là một ngôn ngữ ontology đơn giản, là một ngôn ngữ cơ sở của web ngữ nghĩa. RDFS là ngôn ngữ mô tả bộ từ vựng trên các bộ ba RDF.

1.1.3.4. OWL: các ontology chi tiết hơn có thể được tạo ra với OWL. OWL là một ngôn ngữ bắt nguồn từ hình thức biểu diễn logic và cấu trúc hơn RDFS. Nó được nhúng vào RDF nhằm cung cấp thêm các từ vựng được chuẩn hóa, do đó nó giống như RDFS.

1.1.3.5. RIF: Để cung cấp các luật cho các ngôn ngữ RDF và OWL. Các luật được chuẩn hóa cho web ngữ nghĩa.

1.1.3.6. SPARQL : để truy vấn dữ liệu RDF, RDFS và các ontology OWL cùng với các cơ sở tri thức. SPARQL là một ngôn ngữ giống như SQL nhưng sử dụng các bộ ba RDF, tài nguyên để so khớp các thành phần truy vấn và trả kết quả cho câu truy vấn đó.

1.1.3.7. Logic: Việc biểu diễn các tài nguyên dưới dạng các bộ từ vựng ontology giúp máy có thể lập luận được. Cơ sở của việc lập luận chủ yếu dựa vào logic. Chính vì vậy, các ontology được ánh xạ sang logic.

***1.1.3.8: 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. Để có được suy luận này thì cơ sở là FOL. Tầng này hiện nay các nhà nghiên cứu đang xây dựng các ngôn ngữ luật cho nó như SWRL, RuleML.

***1.1.3.9: Trust*:** Đảm bảo sự tin cậy của các ứng dụng.

1.1.4. Ngôn ngữ cho Web ngữ nghĩa

Ngôn ngữ biểu diễn dữ liệu và tri thức là một khía cạnh quan trọng của Web ngữ nghĩa. Có nhiều ngôn ngữ cho Semantic Web, hầu hết dựa trên XML hay sử dụng XML làm cú pháp. Một số ngôn ngữ sử dụng RDF và RDFschema.

1.1.4.1. XML và XML Schema

XML là một siêu ngôn ngữ sử dụng để biểu diễn các ngôn ngữ web ngữ nghĩa khác. XML cho phép đặc tả và đánh dấu các tài liệu mà máy tính có thể đọc được. Nó giống với HTML ở điểm chứa các chuỗi ký tự, các thẻ dùng để đánh dấu nội dung tài liệu, và dữ liệu XML được lưu trữ dưới dạng văn bản thuần túy. Không giống như HTML, XML có thể được sử dụng để biểu diễn các tài liệu có cấu trúc tùy ý, và không có các thẻ cố định.

Mỗi XML Schema cung cấp một khung làm việc cần thiết cho việc tạo ra một danh mục tài liệu XML. Schema mô tả các thẻ, các element và các thuộc tính của một tài liệu XML của danh mục chỉ định, cấu trúc tài liệu đúng, các ràng buộc, và các loại dữ liệu cơ

sở. Ngôn ngữ XML schema cũng cung cấp một số hỗ trợ bị hạn chế về việc chỉ định số lượng xuất hiện các element con, các giá trị mặc định, ... Cú pháp mã hóa ngôn ngữ XML schema là XML.

1.1.4.2. RDF và RDF Schema

Khung biểu diễn tài nguyên RDF là ngôn ngữ cung cấp mô hình biểu diễn dữ liệu về "những gì tồn tại trên web" có nghĩa là tài nguyên dưới dạng bộ ba: ***"chủ đề – thuộc tính – đối tượng"*** và mạng ngữ nghĩa. Biểu diễn tài nguyên trong RDF là một danh sách các mệnh đề gồm các bộ ba, bao gồm chủ đề là tài nguyên web, các thuộc tính của chủ đề và đối tượng. Đối tượng có thể là văn bản hoặc tài nguyên khác. Mỗi một đặc tả RDF cũng có thể được biểu diễn dưới dạng các hình ảnh được gắn nhãn trực tiếp (mạng ngữ nghĩa).

RDF Schema cung cấp từ vựng dựa trên cơ sở XML để chỉ rõ các lớp và các mối quan hệ giữa chúng, định nghĩa các thuộc tính và kết hợp các thuộc tính với các lớp, cho phép tạo các nguyên tắc phân loại.

RDF và RDF schema cung cấp một mô hình chuẩn để mô tả về tài nguyên web, nhưng những mô hình này thường cần chỉ rõ ngữ nghĩa của tài nguyên web. RDFS được so sánh khá đơn giản với các ngôn ngữ biểu diễn tri thức đầy đủ.

1.1.4.3. OWL

OWL kế thừa trực tiếp của DAML, là một ngôn ngữ web ngữ nghĩa được ghép hai ngôn ngữ ontology khác là DAML và OIL.

Các từ vựng OWL bao gồm các element và thuộc tính của XML được định nghĩa đúng. Chúng được sử dụng để định nghĩa miền các bộ ba và các mối quan hệ giữa chúng trong một ontology. Thực tế, từ vựng của OWL được xây dựng dựa trên từ vựng của RDF. OWL được chia thành hai thành phần là *datatype domain* và

object domain . Tương tự, có hai loại thuộc tính của OWL: những đối tượng này quan hệ với những đối tượng khác được chỉ định bằng owl:ObjectProperty và những đối tượng quan hệ với những giá trị của kiểu dữ liệu được chỉ định bởi owl:DatatypeProperty. Cú pháp dành cho các lớp và các thuộc tính tương tự như DAML và OIL.

Ngày nay, OWL là ngôn ngữ được sử dụng để biểu diễn các ontology và là ngôn ngữ web ngữ nghĩa mà máy tính có thể đọc và hiểu dữ liệu và đưa ra các suy luận từ nó. Thêm vào đó nó đưa ra các luật và các định nghĩa tương tự như RDF, OWL cũng cho phép chỉ rõ các ràng buộc và các mối quan hệ giữa các tài nguyên, bao gồm lượng số, các ràng buộc về miền và phạm vi, các luật hợp nhất, luật phân tách, luật nghịch đảo và luật ngoại động từ.

Một đặc điểm quan trọng của từ vựng OWL là sự phong phú để mô tả các mối quan hệ giữa các lớp, thuộc tính và đối tượng.

1.1.4.4. SPARQL

SPARQL sử dụng để truy vấn dữ liệu web. Chính xác hơn nó là một ngôn ngữ truy vấn RDF. Để hiểu rõ về SPARQL, chúng ta hãy xem các tài nguyên RDF dưới dạng các mạng ngữ nghĩa. SPARQL được sử dụng để: trích lọc thông tin từ các lược đồ RDF, trích lọc các lược đồ con của RDF, xây dựng các lược đồ RDF mới dựa trên các thông tin có được khi truy vấn các lược đồ RDF.

SPARQL truy vấn so khớp các khuôn mẫu lược đồ với lược đồ đích của truy vấn. Khuôn mẫu giống như các lược đồ RDF, nhưng có thể chứa các biến được đặt tên trong không gian của các node hoặc các liên kết / vị ngữ. Khuôn mẫu lược đồ đơn giản nhất tương tự như một bộ ba RDF đơn. Các khuôn mẫu lược đồ đơn giản có thể được kết hợp sử dụng các toán tử khác nhau tạo thành các khuôn mẫu lược đồ phức tạp hơn.

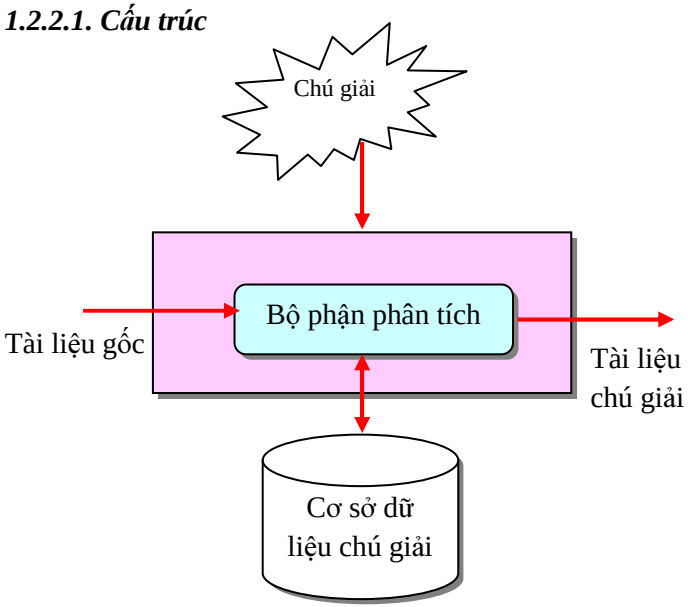
1.2. HỆ THỐNG CHÚ GIẢI CHO WEB NGỮ NGHĨA

1.2.1. Chú giải ngữ nghĩa

Chú giải là những bình luận, ghi chú, giải thích, những nhận xét ngoài mà có thể được gán cho một tài liệu hay một phần được chọn của tài liệu

1.1.2. Mô hình tổng quát cho hệ thống chú giải ngữ nghĩa tự động

1.2.2.1. Cấu trúc



1.2.2.2. Các giai đoạn làm việc của quá trình chú giải

Quá trình chú giải ngữ nghĩa tổng quát bao gồm 3 giai đoạn.

a. Giai đoạn 1 : Ontology mô tả miền ứng dụng cần quan tâm. Thông thường để thực hiện điều này người ta sử dụng các công cụ soạn thảo Ontology. Ontology này được chuyển thành các mô tả dựa vào RDF và chứa trong kho ngữ nghĩa.

b. Giai đoạn 2 : Nhận dạng sự thể hiện dữ liệu khám phá trong tài liệu Web đích. Giai đoạn này gồm 3 pha: Phân tích văn bản, lập chỉ mục và khôi phục tài liệu, trích lọc thông tin trả về.

1.2.2.3. Một số phương pháp phân tích câu

Hiện nay tồn tại 2 hướng tiếp cận chính cho việc tách từ:

- Hướng tiếp cận dựa trên từ (Word - based approaches): Mục tiêu của hướng tiếp cận này là tách thành các từ hoàn chỉnh trong câu. Nó có các hướng chính: *dựa vào thống kê* (statistics-base), *dựa vào tự điển* (dictionarry - base), *hybrid* (kết hợp nhiều phương pháp, hy vọng đạt được những ưu điểm của các phương pháp này).

- Hướng tiếp cận dựa trên ký tự (Character- based approaches): Chia các văn bản ra các một ký tự đơn (unigram) hoặc nhiều ký tự (n-gram) để thực hiện tách từ. Hiện nay phương pháp tách văn bản theo từng ký tự đơn không còn sử dụng nữa. Đối với cách n-gram, văn bản được chia thành các chuỗi, mỗi chuỗi từ 2 đến 3 ký tự trở lên. Cách tiếp cận này cho kết quả ổn định hơn, dễ thực hiện trong ứng dụng và nhất là ít tốn chi phí trong lập chỉ mục và thực hiện truy vấn. Những kết quả nghiên cứu gần đây cho thấy hướng tiếp cận này được xem là sự lựa chọn thích hợp, tuy nhiên độ chính xác không cao bằng phương pháp dựa trên từ. Chúng ta có một số các phương pháp tách từ thông dụng như sau: Phương pháp so khớp tối đa (*Maximum Matching),* phương pháp biến đổi dựa vào việc học *(Transformation-based Learning, TBL),* mô hình tách từ bằng WFST và mạng Neural, phương pháp thống thê dựa trên Internet. Một số phương pháp lập chỉ mục và khôi phục: phương pháp lập chỉ mục theo từ khóa, phương pháp lập chỉ mục ngữ nghĩa tiềm tàng *(LSI-Latent Semantic Indexing).*

CHƯƠNG 2 - HỆ THỐNG QUẢN LÝ THÔNG TIN VÀ TRI THỨC KIM

2.1. GIỚI THIỆU KIM

Phần này giới thiệu sơ lược về KIM.

2.2. HỆ THỐNG KIM

2.2.1. Kiến trúc KIM

Nền tảng KIM bao gồm các nguồn tài nguyên tri thức chính thức, KIM Server cùng với các front end. KIM Server bao gồm các thành phần chính sau: kho ngữ nghĩa, chú giải ngữ nghĩa, persistence tài liệu, lập chỉ mục và truy vấn.

KIM được xây dựng dựa trên cơ sở các nền tảng mã nguồn mở mạnh mẽ: **GATE**, **Sesame** và **Lucene** tương ứng với ba lĩnh vực khác nhau: kho RDF(S), HLT (đặc biệt là IE) và IR. Tài nguyên tri thức được lưu trữ trong kho RDF của Sesame, cung cấp cơ sở hạ tầng lưu trữ và khả năng truy vấn. Kho Sesame được nạp với hàng triệu câu lệnh RDF(S).

GATE làm cơ sở cho quá trình trích lọc thông tin và cũng được sử dụng cho việc quản lý nội dung và chú giải. Nó cung cấp các công nghệ phân tích văn bản thiết yếu, trên những công nghệ này KIM đã được xây dựng với các thành phần mở rộng nhận thức về ngữ nghĩa, đặc biệt cho quá trình trích lọc thông tin của KIM.

Máy phục hồi thông tin Lucene đã được thêm vào để lập chỉ mục, phục hồi thông tin và đánh giá nội dung liên quan theo các thực thể có tên, điều này cho phép các phương thức truy cập ngữ nghĩa.

2.2.2. KIM Ontology (KIMO)

KIM Ontology cung cấp một ontology tối thiểu nhưng đầy đủ, thích hợp cho miền mở và mục đích chung là chú giải ngữ nghĩa. KIMO là một ontology ở mức cao đơn giản, bắt đầu với một số cơ sở

khác biệt về triết học giữa các loại thực thể. Ngoài ra, ontology còn đi vào chi tiết hơn như một phần mở rộng của các loại thực thể có tầm quan trọng trong thế giới thực. Có ontology này làm cơ sở, chúng ta có thể dễ dàng mở rộng các miền, để cấu hình các chú giải ngữ nghĩa cho các ứng dụng cụ thể.

Sự phân bố của các thực thể thường được gọi thay đổi rất nhiều qua các lĩnh vực khác nhau. Mặc dù có sự khác nhau về sự phân bố của các loại nhưng có nhiều loại thực thể chung xuất hiện trong tất cả các kho ngữ liệu như Người, tổ chức, địa điểm, tiền bạc, ngày tháng, ...Định vị và biểu diễn các loại cơ sở này thích hợp là một trong các mục tiêu đằng sau việc thiết kế KIMO. Hơn nữa, KIM Ontology định nghĩa các loại thực thể cụ thể hơn nữa .

Sự mở rộng về chuyên môn hóa ontology được xác định dựa trên cơ sở nghiên cứu các loại thực thể trong kho ngữ liệu tin tức tổng hợp bao gồm cả chính trị, thể thao và tài chính. Hiện nay, KIMO bao gồm khoảng 250 lớp và khoảng 100 thuộc tính và quan hệ. Các lớp ở đỉnh là Entity, EntitySource, và LexicalResource

2.2.3. Cơ sở tri thức KIM

2.2.3.1. Cơ sở tri thức định nghĩa sẵn của KIM

KIM bao gồm hơn 200.000 thực thể, được thu thập từ một số lượng lớn nguồn dữ liệu, và khoảng 36000 địa điểm bao gồm các lục địa, các vùng miền trên toàn cầu, các quốc gia cùng với các thủ đô, 4400 thành phố, núi, sông lớn, đại dương, biển ...

Các tổ chức có tầm quan trọng to lớn đã được xây dựng sẵn trong cơ sở tri thức của KIM. Bao gồm các tổ chức lớn trên thế giới như liên hợp quốc, NATO, OPEC, hơn 140000 công ty quốc tế, 140 sàn giao dịch thị trường chứng khoán, với tổng số 147000 tổ chức.

Cuối cùng, để cho phép quá trình trích lọc thông tin mà các

thực thể và các mối quan hệ mới, không phải là một phần của cơ sở tri thức KIM được nhận diện, một tập hợp các tài nguyên từ vựng (GATE) cũng được biểu diễn trong cơ sở tri thức của KIM. Nó bao gồm các hậu tố tổ chức, tên người, thời gian, tiền tố tiền tệ,...

2.2.3.2. Điều khiển chất lượng và độ bao phủ cơ sở tri thức của KIM

Cơ sở tri thức của KIM được xác thực lặp đi lặp lại nhiều lần bằng cách sử dụng một quá trình xây dựng cơ sở tri thức bao gồm các thực thể và các quan hệ một cách độc lập.

a. *Xác minh chất lượng, cơ sở tri thức định nghĩa sẵn của KIM*

Độ bao phủ tri thức KIM được đảm bảo với quá trình xử lý và phân tích thường xuyên các tiêu đề tin tức, sử dụng các bộ thu thập tin tức – một dịch vụ thu thập khoảng từ 500 đến 2000 đầu câu chuyện một ngày từ khoảng 20 nguồn tin tức phổ biến toàn cầu.

b. *Tầm hiểu biết và nhận thức – các tài nguyên tin tức và cách thức giao tiếp của con người thông qua các phương tiện thông tin đại chúng*

Việc sử dụng các nguồn tin cho việc làm giàu cơ sở tri thức của KIM có thể là một sự lựa chọn gây tranh cãi do các nguồn tin trên thế giới không bao giờ trung lập, mà là một cách khác xoay quanh việc hầu hết các tin tức khá thành kiến và khăng khăng đến một mức độ nhất định mà thay đổi phụ thuộc vào đất nước, chính trị, xã hội và chuyên môn của nguồn tin tương ứng, ...

2.2.4. Trích lọc thông tin trong KIM

2.2.4.1 Đánh giá quá trình trích lọc thông tin trong KIM

Mặc định, trích lọc thông tin trong KIM dựa trên từ điển ngữ nghĩa, phân tích văn bản và các ngữ pháp so khớp mẫu. Lý do để đánh giá lại corpora của các thực thể được đặt tên là không có các số

liệu tốt để chú giải ngữ nghĩa. Ngoài ra, không có bất kỳ corpora được chú thích bởi con người nào có các chú giải tuân theo một hệ thống các thực thể được đặt tên mà có thể được ánh xạ tới KIMO và do đó cung cấp một tiêu chuẩn vàng cho các đánh giá chú giải ngữ nghĩa.

2.2.4.2 Tiếp cận trích lọc thông tin truyền thống và tùy biến trích lọc thông tin trong KIM

Khác biệt giữa quá trình trích lọc thông tin ngữ nghĩa và trích lọc thông tin truyền thống là không phát hiện ra loại của thực thể được trích xuất nhưng nhận diện thực thể. Điều này cho phép các thực thể được truy tìm thông qua các tài liệu và các đặc tả của chúng được làm giàu thông qua quá trình trích lọc thông tin.

Những gì mà quá trình trích lọc thông tin truyền thống tiếp cận là cung cấp chú thích cho các văn bản tương. Tuy nhiên, kiểu chú giải này không liên quan đến ngữ nghĩa. Mặc dù những loại này biểu diễn là quan trọng đối với các kiểu thực thể được đặt tên trong miền độc lập, nhưng một người được đào tạo trung bình có thể phân loại các thực thể thành các loại cụ thể. KIM đã tạo ra những khác biệt to lớn bằng cách thêm ngữ nghĩa vào quá trình trích lọc thông tin. KIM liên kết các chú giải mà nó đưa ra, không chỉ là các điểm của quá trình phân loại mà là một mô hình chính thức về toàn bộ các miền tương ứng: các ontology, các logic nội bộ, các luật và các quan hệ. Hơn thế nữa, hướng tiếp cận này cho phép nhận diện các thực thể cụ thể diễn ra cùng với chú giải.

Quá trình trích lọc thông tin trong KIM dựa trên nền tảng GATE. Một số các thành phần xử lý ngôn ngữ tự nhiên được sử dụng để xác định từ, xác định từ loại cho từ, ... và những thành phần khác được sử dụng trực tiếp trong KIM. Từ điển ngữ nghĩa KIM sẽ tra cứu

các thành phần tìm kiếm thông qua các bí danh thực thể và các nguồn từ vựng khác. Ngữ pháp so khớp khuôn mẫu trong GATE đã được sửa đổi để xử lý thông tin lớp thực thể và cho phép tổng quát hóa các luật. Các nguyên tắc nền tảng là đơn giản – một tham chiếu đến một thực thể của một lớp cụ thể, có thể so khớp một khuôn mẫu được chỉ ra với một lớp tổng quát hơn.

2.2.5. Lập chỉ mục và khôi phục thông tin

KIM cung cấp việc đánh chỉ mục đối với các chú giải ngữ nghĩa, được phát sinh cho một tài liệu tức là lập chỉ mục đối với siêu dữ liệu. Phương pháp lập chỉ mục này cho phép các phương thức truy cập tin tức (đã được bổ sung ngữ nghĩa). Do đó người dùng có thể chỉ định truy vấn, bao gồm các ràng buộc liên quan đến loại thực thể, mối quan hệ giữa các thực thể, các thuộc tính của thực thể.

Bước đầu tiên trong quá trình lập chỉ mục là tiền xử lý về mặt ngữ nghĩa cho mỗi tài liệu sẽ được đưa vào kho ngữ liệu của các tài liệu cho việc phục hồi thông tin. Quá trình tiền xử lý tìm ra các từ ngữ phụ thuộc hoặc các liên kết của một định danh chuỗi bên trong duy nhất (một chú giải ngữ nghĩa) tới các thành phần văn bản mà chúng ta biết nghĩa của nó tùy theo các ontology và cơ sở tri thức mà chúng ta sử dụng.

Siêu dữ liệu này phục vụ dưới dạng một con trỏ đến thực thể tương ứng trong quá trình phục hồi thông tin. Sau đó đến bước tiếp theo: tài liệu để lập chỉ mục được gởi tới máy lập khôi phục thông tin Lucene cùng với các chuỗi ID và một thủ tục lập chỉ mục được thực hiện. Sau đó chúng ta có thể thực hiện việc tìm kiếm sử dụng các chuỗi ID này dưới dạng một chỉ mục. Việc lập chỉ mục của KIM có một sự khác biệt nhỏ so với lập chỉ mục văn bản chuẩn bởi vì KIM sử dụng nhận diện duy nhất các loại cụ thể. Tuy nhiên, lập lập chỉ

mục không tự nó sử dụng trực tiếp cơ sở tri thức đặc tả thực thể mà chỉ được sử dụng trong quá trình phục hồi thông tin đối với các truy vấn có cấu trúc.

Lợi ích của việc tiền xử lý này là: Có thể tìm thấy tham chiếu đến một thực thể trong văn bản mà không quan tâm đến bí danh có được sử dụng hay không, mức độ liên quan với các thực thể tương ứng là cao hơn.

Độ chính xác phục hồi thông tin của KIM vẫn chưa được đánh giá so với các cỗ máy phục hồi thông tin truyền thống, đây là một chủ đề sẽ được nghiên cứu trong tương lai. Tuy nhiên, KIM có tiềm năng để thực hiện tốt hơn, không chỉ hướng tới việc giảm các tài liệu không liên quan trong kết quả trong khi vẫn phục hồi thông tiên liên quan (nâng cao độ chính xác như với một hệ thống lập chỉ mục các thực thể được đặt tên) mà còn hướng tới việc tăng số lượng tài liệu liên quan của các thực thể mà không chứa các bí danh, được sử dụng cho các thực thể giới hạn về tên.

2.2.6. Đầu cuối của KIM

KIM Server API cho phép xây dựng giao diện người sử dụng đầu cuối khác nhau. Các đầu cuối này có thể cho phép truy cập đầy đủ đến các chức năng của KIM Server bao gồm: tính năng khôi phục thông tin, kho ngữ nghĩa, các dịch vụ chú giải ngữ nghĩa, và cơ sở hạ tầng quản lý tài liệu và siêu dữ liệu. Một số đầu cuối đã được xây dựng sẵn trong KIM: plug in cho trình duyệt (KIM plug in), KIM Web UI, KIM Explorer và Graph View.

2.2.7. Hiệu suất

Tốc độ chú giải phụ thuộc vào kích thước của tài liệu và có xu hướng trở nên chậm hơn với các tài liệu lớn với độ phụ thuộc logarit.

CHƯƠNG 3 – XÂY DỰNG ỨNG DỤNG CHÚ GIẢI NGỮ NGHĨA TỰ ĐỘNG

3.1. KIẾN TRÚC TỔNG THỂ CỦA HỆ THỐNG CHÚ GIẢI

3.1.1. Kiến trúc hệ thống

Trong ứng dụng thử nghiệm này, chúng ta xây dựng cơ sở tri thức, định nghĩa các Ontology cho KIM sử dụng nó để chú giải ngữ nghĩa trên Web.

Các nguồn dữ liệu về các thực thể, các lớp được thu thập từ Internet được tổng hợp. Những thông tin này được GATE quản lý nội dung và những chú giải, sau đó được sắp xếp chỉ mục và lưu trữ trong hệ thống OWLIM.

OWLIM cũng cho phép chúng ta cập nhật dữ liệu từ ứng dụng tạo Ontology thứ ba. Vậy nhiệm vụ của chúng ta là tổng hợp dữ liệu tạo các Ontology và đưa vào nên tảng KIM để thực hiện chú giải.

3.1.2. Các thành phần của hệ thống

3.1.2.1. Server KIM

Server KIM được xây dựng trên nền tảng Java. Sau khi khởi động, KIM server chạy dịch vụ trên máy chủ localhost và cổng 1099.

3.1.2.2. Popular Import

Công cụ này cho phép Import các thực thể được nhận dạng từ các văn bản Text chúng ta thu thập được qua hệ thống thông tin. Các dạng định dạng cho phép là .DOC, .HTML, .XML, .TXT …

3.1.2.3 RDF import

Công cụ RDF Import cho phép cập nhật các nguồn tài nguyên thu nhập được lên các máy chủ chứa định nghĩa các URI.

3.2 THIẾT LẬP KIM ONTOLOGY VÀ CƠ SỞ TRI THỨC

KIM 3 dựa trên PROTON Ontology phát triển trong phạm vi ngữ nghĩa của dự án SEKT. KIM phụ thuộc hoàn toàn vào mô-đun

hệ thống của proton đó là tiếp tục mở rộng bằng KIMSO. Các bản thể học liên quan khác là một phần của hệ thống phân phối. Chúng ta có thể thay thế, thay đổi và bổ sung thêm cơ sở tri thức.

3.2.1. PROTON

Proton là một cấp trên của Ontology định nghĩa về 300 lớp và 100 thuộc tính, bao gồm hầu hết các khái niệm cần thiết cho việc chú thích ngữ nghĩa, lập chỉ mục, và phản hồi. Proton được chia thành ba phân hệ: System module chứa một meta cấp vài nguyên bản, Top module là mô-đun cao nhất chung nhất, khái niệm cấp, bao gồm khoảng 20 lớp đảm bảo một sự cân bằng tốt của tiện ích độc lập, và cách sử dụng dễ hiểu, Upper module - hơn 200 lớp của các thực thể, thường xuất hiện trong nhiều tên .

KIMSO và KIMLO là mô-đun tùy chọn mở rộng ontology proton, một phần của KIM.

3.2.2 Mở rộng Ontology

Để tích hợp một phần mở rộng ontology, các lớp mới phải kế thừa http://proton.semanticweb.org/2006/05/protons#Entity một cách trực tiếp hoặc gián tiếp.

Thiết kết lớp kế thừa từ :

- http://proton.semanticweb.org/2006/05/protont#Person
- http://proton.semanticweb.org/2006/05/protont#Organization
- http://proton.semanticweb.org/2006/05/protont#Location

3.2.3. Giới thiệu Protégé

Protégé là một công cụ mã nguồn mở Java được phát triển tại khoa tin học y học Stanford. Protégé - OWL là một trong các công cụ chính trong Protégé, là một thư viện cho ngôn ngữ Web Ontology (OWL) và RDF(S). Nó cung cấp các lớp và các phương thức để nạp và ghi các tệp OWL, cung cấp khả năng xây dựng các mô hình dữ

liệu OWL và thực hiện lập luận trên DL. Bên cạnh đó nó còn cung cấp một giao diện đồ hoạ trực quan, dễ sử dụng.

Cụ thể Protégé- OWL cung cấp các khả năng chính sau:

- Soạn thảo các Ontology cho OWL
- Duy trì, phát triển và kiểm tra Ontology

3.3 THIẾT KẾ HỆ THỐNG

3.3.1 Giới thiệu khái quát

Ứng dụng phân tích các tài liệu hoặc văn bản qua việc sử dụng các mẫu từ ngữ quy chuẩn và nhận dạng các thành tố ngữ nghĩa tương đương, chú thích lớp tự động cho các thực thể có tên trên các trang web theo miền Ontology đã được định nghĩa. Các thành phần chính của ứng dụng sử dụng các thư viện:

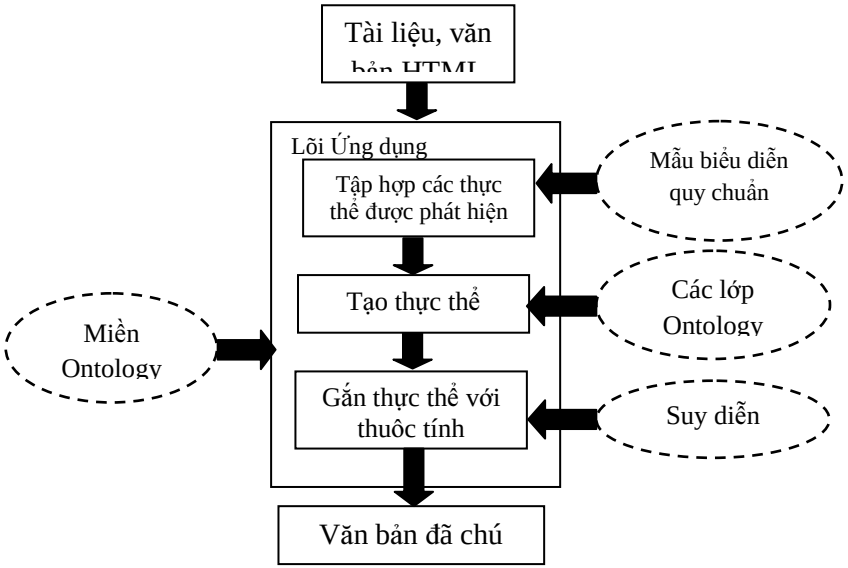
- Thư viện Web ngữ nghĩa trên Seasame.
- Thư viện khôi phục thông tin Lucence.
- Chú giải ngữ nghĩa: Nhận dạng các đối tượng chuẩn hóa trong văn bản.
- Ontology: Chuẩn hóa các mô hình để máy tính hiểu được.
- Biểu diễn mẫu quy chuẩn: là một chuỗi để mô tả và so khớp theo một số quy tắc cú pháp.

3.3.2 Phương pháp

Ứng dụng làm việc sử dụng các văn bản sau khi đã chuyển về định dạng chung, ở các miền đặc biệt được mô tả bởi miền Ontology sử dụng cho việc chuẩn hóa mẫu cho chú giải ngữ nghĩa. Ứng dụng sẽ phát hiện các thành tố ontology trong ứng dụng hoặc trong miền hiện hành của mô hình Ontology.

3.3.3 Cấu trúc tổng quát và nguyên lý hoạt động

3.3.3.1 Cấu trúc tổng quát



Cấu trúc của công cụ bao gồm 4 phần:

Phần 1: Là các nguồn văn bản đầu vào như HTML, email, văn bản gốc cần phải được chú giải.

Phần 2: Là đầu ra của hệ thống, chứng là những thực thể Ontology mới tương ứng với những chú giải văn bản. Thuộc tính của các thực thể này được làm đầy bằng cách phát hiện các thực thể Ontology thông qua các mẫu được định nghĩa.

Phần 3: Các miền thực thực thể được định nghĩa, các mẫu biểu diễn quy chuẩn, thực thể kết quả, các tham chiếu từ bên ngoài.

Phần 4: Lõi công cụ gồm các giải thuật chính của công cụ như : phát hiện, tạo chú giải, gắn các thực thể với các chú giải tương ứng từ miền Ontology đang xét.

3.3.3.2 Nguyên lý hoạt động

Hoạt động của ứng dụng thực hiện tuần tự theo các bước sau:

1. Nạp văn bản của một tài liệu.
2. Xác định biểu thức quy chuẩn nếu chúng được tìm thấy tương ứng với các thể ontology theo các thuộc tính mẫu, chúng được bổ sung vào tập hợp các cá thể ontology được tìm thấy.
3. Nếu không có cá thể được tìm thấy bằng phép so khớp mẫu thì thuộc tính createInstance được thiết lập, một cá thể của một kiểu lớp bao gồm thuộc tính *hasClass* thì chỉ được tạo ra với thuộc tính *rfs:label* chứa trong văn bản so khớp.
4. Quá trình trên lặp lại cho tất cả các biểu thức quy chuẩn, kết quả là một tập các cá thể được tìm thấy.
5. Một cá thể của lớp rỗng biểu diễn cho văn bản gốc được tạo ra và có thể tất cả các thuộc tính của lớp ontology được phát hiện từ lớp định nghĩa.
6. Cá thể được phát hiện được so sánh với các kiểu thuộc tính và nếu kiểu thuộc tính là tương tự như kiểu cá thể, thì thực thể được quy cho thuộc tính này.
7. Việc so sánh được thực hiện cho tất cả các thuộc tính của một cá thể mới tương ứng với các văn bản/tài liệu.

3.3.4 Giới thiệu một số lớp quan trọng trong ứng dụng

3.3.4.1 Lớp SemanticQuery

3.3.4.2 Lớp SemanticQueryResult

3.3.4.3 Lớp DocumentQuery

3.3.4.4 Lớp DocumentQueryResult

3.3.5 Xây dựng ontology danh nhân lịch sử Việt Nam

3.4. CÀI ĐẶT THỬ NGHIỆM

3.4.1. Môi trường

3.4.2. Cài đặt các công cụ

3.5. KẾT QUẢ VÀ ĐÁNH GIÁ

3.5.1. Kết quả chạy thử nghiệm

3.5.2. Đánh giá các kết quả đạt được

Việc xây dựng hệ thống chú giải ngữ nghĩa trong Web ngữ nghĩa làm giảm thiểu đáng kể thời gian, sai sót so với chú giải bằng tay, đặc biệt khi miền ngữ liệu lớn và thay đổi.

Hệ thống cài đặt thử nghiệm thành công Server KIM trên một server bất kỳ, cập nhật thành công các dữ liệu có sẵn trên miền KIM và PROTON đồng thời cho phép định nghĩa miền dữ liệu và cơ sở tri thức riêng.

Ứng dụng chú giải chạy trên hệ thống Server Apache Tomcat với các hàm KIM API có sẵn cho phép thực hiện nhiều ứng dụng trên nền khác nhau.

Hướng mở rộng của hệ thống là cài đặt nhiều server KIM khác nhau, kết nối thông qua môi trường Java RMI, cho phép nhiều ứng dụng khác nhau kết nối trên môi trường Internet.

KẾT LUẬN

Luận văn đã giới thiệu về thế hệ sắp tới của Web là Web ngữ nghĩa, trình bày các lý thuyết liên quan đến Web ngữ nghĩa cũng như hệ thống chú giải ngữ nghĩa. Bên cạnh đó, hệ thống quản lý thông tin và tri thức KIM cũng được tìm hiểu và trình bày khá chi tiết giúp chúng ta có thể hình thành khung chung cho việc triển khai các ứng dụng Web ngữ nghĩa. Đặc biệt đối với Web ngữ nghĩa dành cho tiếng việt, việc xử lý tính toán đòi hỏi nhiều quy trình phức tạp như lưu trữ và truy xuất trên hàng trăm ngàn thực thể ở nhiều lĩnh vực khác nhau, với các miền giá trị khác nhau.

Việc kết hợp nhiều kỹ thuật, công cụ hỗ trợ là cần thiết. Nó giúp chúng ta giảm thiểu đáng kể thời gian và giúp vận hành dễ dàng hơn với nhiều hệ thống công cụ khác nhau.Luận văn cũng đã xây dựng thành công hệ thống chú giải ngữ nghĩa tự động giúp người sử dụng tiết kiệm được nhiều thời gian, công sức và tiền bạc.

Luận văn cũng mở ra một hướng mới trong việc khám phá tri thức từ kho tri thức khổng lồ của nhân loại trên Internet, tiếp cận tri thức theo lĩnh vực mà mình yêu thích.

Tuy nhiên, vì thời gian nghiên cứu tìm hiểu trong thời gian ngắn nên luận văn vẫn còn tồn tại những điểm yếu như lượng tri thức trong cơ sở dữ liệu còn khiêm tốn.Từ những nhìn nhận trên, tác giả cũng mạnh dạn đề xuất các hướng nghiên cứu và phát triển tiếp luận văn trong tương lai như sau:

Thứ nhất, thử nghiệm trên nhiều bộ trích lọc khác nhau.

Thứ hai, nâng cấp giao diện tương tác với người dùng để thuận tiện hơn cho người sử dụng.

Thứ ba, tăng lượng tri thức trong dữ liệu và mở rộng ra các lĩnh vực nghiên cứu khác.