BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN THỊ HỒNG MỸ

RÚT TRÍCH TRI THỨC NGỮ NGHĨA TỪ TÊN THỂ LOẠI WIKIPEDIA

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ Thông tin

Mã số ngành : 60480201

TP. HỒ CHÍ MINH, tháng 04 năm 2015

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN THỊ HỒNG MỸ

RÚT TRÍCH TRI THỨC NGỮ NGHĨA TỪ TÊN THỂ LOẠI WIKIPEDIA

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ Thông tin
Mã số ngành : 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: **TS. NGUYỄN CHÁNH THÀNH**
TS. LÊ MẠNH HẢI

TP. HỒ CHÍ MINH, tháng 04 năm 2015

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : TS. NGUYỄN CHÁNH THÀNH
TS. LÊ MẠNH HẢI

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày … tháng 4 năm 2015

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:
(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	Họ và tên	Chức danh Hội đồng
1	PGS. TSKH. Nguyễn Xuân Huy	Chủ tịch
2	PGS. TS. Lê Hoài Bắc	Phản biện 1
3	PGS. TS. Quản Thành Thơ	Phản biện 2
4	TS. Vũ Thanh Hiền	Ủy viên
5	TS. Cao Tùng Anh	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM

PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày…… tháng….. năm 20……

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên:	Nguyễn Thị Hồng Mỹ	Giới tính:	Nữ
Ngày, tháng, năm sinh:	03/9/1984	Nơi sinh:	Khánh Hòa
Chuyên ngành:	Công nghệ Thông tin	MSHV:	1341860013

I- Tên đề tài:

RÚT TRÍCH TRI THỨC NGỮ NGHĨA TỪ TÊN THỂ LOẠI WIKIPEDIA

II- Nhiệm vụ và nội dung:

- Khảo sát, phân tích cấu trúc thể loại và tài liệu lưu trữ trong Wikipedia
- Khảo sát các nghiên cứu liên quan đến việc rút trích ngữ nghĩa từ tên thể loại
- Phát triển trên cơ sở kế thừa hoặc cải tiến một phương pháp rút trích ngữ nghĩa từ tên thể loại, dựa trên nguồn dữ liệu tên thể loại sẵn có của Wikipedia
- Thực nghiệm, đánh giá và viết báo cáo

III- Ngày giao nhiệm vụ: 18/8/2014

IV- Ngày hoàn thành nhiệm vụ: 10/3/2015

V- Cán bộ hướng dẫn: TS. Nguyễn Chánh Thành - TS. Lê Mạnh Hải

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

TS. Lê Mạnh Hải

KHOA QUẢN LÝ CHUYÊN NGÀNH
(Họ tên và chữ ký)

LỜI CAM ĐOAN

Tôi xin cam đoan nội dung của luận văn là công trình nghiên cứu của bản thân. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo.

Học viên thực hiện Luận văn

Nguyễn Thị Hồng Mỹ

LỜI CẢM ƠN

Trước tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới TS. Nguyễn Chánh Thành và TS. Lê Mạnh Hải, hai thầy đã trực tiếp hướng dẫn tận tình cho tôi trong suốt quá trình thực hiện luận văn tốt nghiệp này.

Tôi cũng xin chân thành cảm ơn các thầy, cô Khoa Công nghệ Thông tin, Phòng Quản lý Sau Đại học và các phòng ban của trường Đại học Công nghệ TP.HCM đã hỗ trợ và tạo điều kiện thuận lợi cho tôi trong suốt quá trình học tập và làm luận văn.

Và tôi xin được gửi lời cảm ơn tha thiết tới ba mẹ, anh chị, tất cả bạn bè và người thân yêu của tôi, là những người đã giúp đỡ, khuyến khích và động viên tôi trong suốt quá trình thực hiện Luận văn.

Tôi xin chân thành cảm ơn!

Tác giả Luận văn

Nguyễn Thị Hồng Mỹ

TÓM TẮT

Hệ thống Wikipedia miễn phí, được mở rộng và cập nhật thường xuyên. Hiện nay, trên thế giới đã có một số công trình nghiên cứu khai thác dữ liệu từ hệ thống bách khoa toàn thư này. Tuy nhiên, các công trình nghiên cứu về Wikipedia chủ yếu là phân tích nội dung các trang bài viết. Một số nhóm nghiên cứu rút trích thông tin từ infobox – là bảng được định dạng cố định ở góc trên bên phải của bài viết, bảng này trình bày tóm tắt nội dung chính của bài viết; một số công trình khác thì nghiên cứu về hệ thống phân loại thể loại của Wikipedia (Wikipedia Category Network - viết tắt là WCN).

Việc khảo sát Wikipedia cho thấy hệ thống phân loại trong Wikipedia có nhiều mối liên hệ, là nguồn dữ liệu ngữ nghĩa tiềm năng cho nghiên cứu của luận văn. Do vậy, luận văn tập trung vào việc nghiên cứu đề xuất một phương pháp để trích xuất thông tin hữu ích từ Wikipedia thông qua các đặc trưng ngữ nghĩa từ hệ thống tên thể loại của Wikipedia. Luận văn thực hiện với cách tiếp cận xử lý dữ liệu ít hơn: chỉ xử lý tên thể loại và tiêu đề bài viết mà không cần phải xử lý các trang bài viết. Luận văn còn đề xuất đề xuất mô hình mở rộng truy vấn dựa vào phương pháp trích rút đặc trưng ngữ nghĩa để mở rộng và cải thiện các kết quả truy vấn.

ABSTRACT

Wikipedia is a free encyclopedia which is frequently expanded and updated. Up to now, there are a number of researches on extracting data from Wikipedia. However, some of them focus on article content analysis; some study how to extract information from infobox which is a fixed-format table designed on the top right-hand corner of articles, presenting a summary of articles; the others work on categories taxonomy which is called Wikipedia Category Network (WCN).

Our investigation into Wikipedia indicates that Wikipedia's categories taxonomy has a large amount of correlations which is a potential resource to extract semantic knowledge. Therefore, this thesis concentrates on studying to propose a method to extract useful information from Wikipedia using semantic features derived from Wikipedia categories. Our approach only processes categories' names and articles' titles instead of full-text articles. The thesis also presents a query expanding model using derived semantic features to expand and improve query results.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
1	IE	Information Extraction	Rút trích thông tin
2	r	Pearson correlation coefficient	Hệ số tương quan Pearson
3	SVM	Support Vector Machine	Máy học vectơ hỗ trợ
4	WCN	Wikipedia Category Network	Hệ thống thể loại Wikipedia

DANH MỤC CÁC BẢNG

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH

CHƯƠNG 1. MỞ ĐẦU

1.1 Lý do chọn đề tài

Wikipedia được biết đến như một bách khoa toàn thư trực tuyến với nội dung mở, được viết bằng nhiều ngôn ngữ. Hệ thống này được xây dựng và phát triển bởi người dùng trên khắp thế giới cùng cộng tác. Nội dung bài viết được cập nhật thường xuyên và trên phạm vi rộng. Dữ liệu Wikipedia ngày càng lớn mạnh và trở thành cơ sở tri thức đầy tìm năng để khai thác.

Wikipedia ngày một lớn dần, miễn phí, cập nhật thường xuyên và là tiêu điểm của nhiều nghiên cứu gần đây. Các công trình nghiên cứu chủ yếu tập trung phân tích bài viết trong hệ thống Wikipedia. Nội dung các bài viết trong Wikipedia chứa nhiều thông tin để khai thác. Tuy nhiên dung lượng thông tin bài viết khá lớn, ngược lại hệ thống phân loại Wikipedia với dung lượng lưu trữ nhỏ hơn nhiều nhưng đầy tính ngữ nghĩa. Do vậy việc khai thác dữ liệu ở mảng này sẽ có nhiều ưu điểm về thời gian và hiệu quả hơn so với sử dụng toàn bộ bài viết của hệ thống Wikipedia.

Thêm vào đó, với sự phát triển mạnh mẽ và không ngừng của công nghệ thông tin, dữ liệu trên internet trở thành nguồn thông tin đồ sộ của nhân loại. Nhu cầu tìm kiếm, truy xuất thông tin từ đó cũng gia tăng, mà chủ yếu là người dùng tìm kiếm nội dung của các trang trên internet.

Để đáp ứng nhu cầu tìm kiếm thông tin của người sử dụng, nhiều hệ thống truy xuất thông tin đã được nghiên cứu và phát triển; Trong đó phải kể đến một số máy tìm kiếm phổ biến như Google [22], Yahoo [28], Bing [18], Ask [17] và một số công cụ tìm kiếm khác. Tuy nhiên các hệ thống này vẫn chưa đáp ứng tốt cho nhu cầu tìm kiếm thông tin của người sử dụng. Thực trạng này do nhiều nguyên nhân khác nhau, trong đó có nguyên nhân do người sử dụng gặp khó khăn trong việc diễn đạt nội dung của vấn đề cần tìm kiếm, dẫn đến yêu cầu truy vấn chỉ bao gồm một vài từ chính, không thể hiện đủ ngữ nghĩa cần thiết. Do đó, kết quả tìm kiếm có thể

không thỏa mãn mong muốn của người dùng về vấn đề tìm kiếm. Để giải quyết vấn đề này, việc mở rộng truy vấn ban đầu của người dùng là yêu cầu cần thiết.

Từ các phân tích trên, luận văn nghiên cứu "**rút trích tri thức ngữ nghĩa từ tên thể loại wikipedia**" và xây dựng mô phỏng ứng dụng mở rộng truy vấn sử dụng các đặc trưng ngữ nghĩa được rút trích từ hệ thống Wikipedia.

1.2 Mục đích

Mục tiêu của luận văn là khai thác kho dữ liệu đồ sộ của Wikipedia với chủ đích xử lý nhanh, ít tốn kém. Luận văn kế thừa và cải tiến phương pháp sử dụng hệ thống tên loại Wikipedia (Wikipedia Category Network - WCN) để tính độ tương quan giữa hai từ. Độ đo này có thể được sử dụng cho nhiều lĩnh vực: học máy có giám sát, tóm tắt văn bản, rút trích thông tin, truy xuất thông tin, mở rộng truy vấn.

Luận văn tập trung nghiên cứu tìm hiểu hệ thống phân loại Wikipedia để rút trích tri thức ngữ nghĩa. Từ đó, xây dựng ứng dụng thực nghiệm mở rộng truy vấn tìm kiếm để cải tiến kết quả tìm kiếm tiến gần mong muốn người dùng.

Để thực hiện mục tiêu trên, luận văn cần giải quyết các vấn đề sau:

+ Phân tích mối liên quan của các thể loại trong hệ thống phân cấp thể loại của Wikipedia – WCN để rút ra được các đặc trưng hữu ích. Luận văn cần tập trung khai thác dữ liệu về số lượng các bài viết của một thể loại, số lượng các thể loại của bài viết và các mối liên kết giữa chúng được chuyển thành các đặc trưng ngữ nghĩa.

+ Phân tích cú pháp tên thể loại: Tên thể loại là các cụm danh từ, luận văn dùng thư viện Opennlp để gán nhãn và tách từ, phân tích tên thể loại để chia nhỏ cụm danh từ để tạo thành cặp từ.

+ Đề xuất phương pháp để tự động tính độ tương quan ngữ nghĩa cặp từ vựng từ hệ thống thể loại Wikipedia dựa vào các đặc trưng rút trích được.

+ Mô phỏng ứng dụng áp dụng độ đo đã đề xuất cho bài toán mở rộng truy vấn tìm kiếm

Từ những mục tiêu nêu trên, luận văn xác định nhiệm vụ của đề tài là:

+ Khảo sát, phân tích cấu trúc thể loại và tài liệu lưu trữ trong Wikipedia

+ Khảo sát các nghiên cứu liên quan đến việc rút trích ngữ nghĩa từ tên thể loại

+ Phát triển trên cơ sở kế thừa hoặc cải tiến phương pháp rút trích ngữ nghĩa từ tên thể loại, dựa trên nguồn dữ liệu tên thể loại sẵn có của Wikipedia.

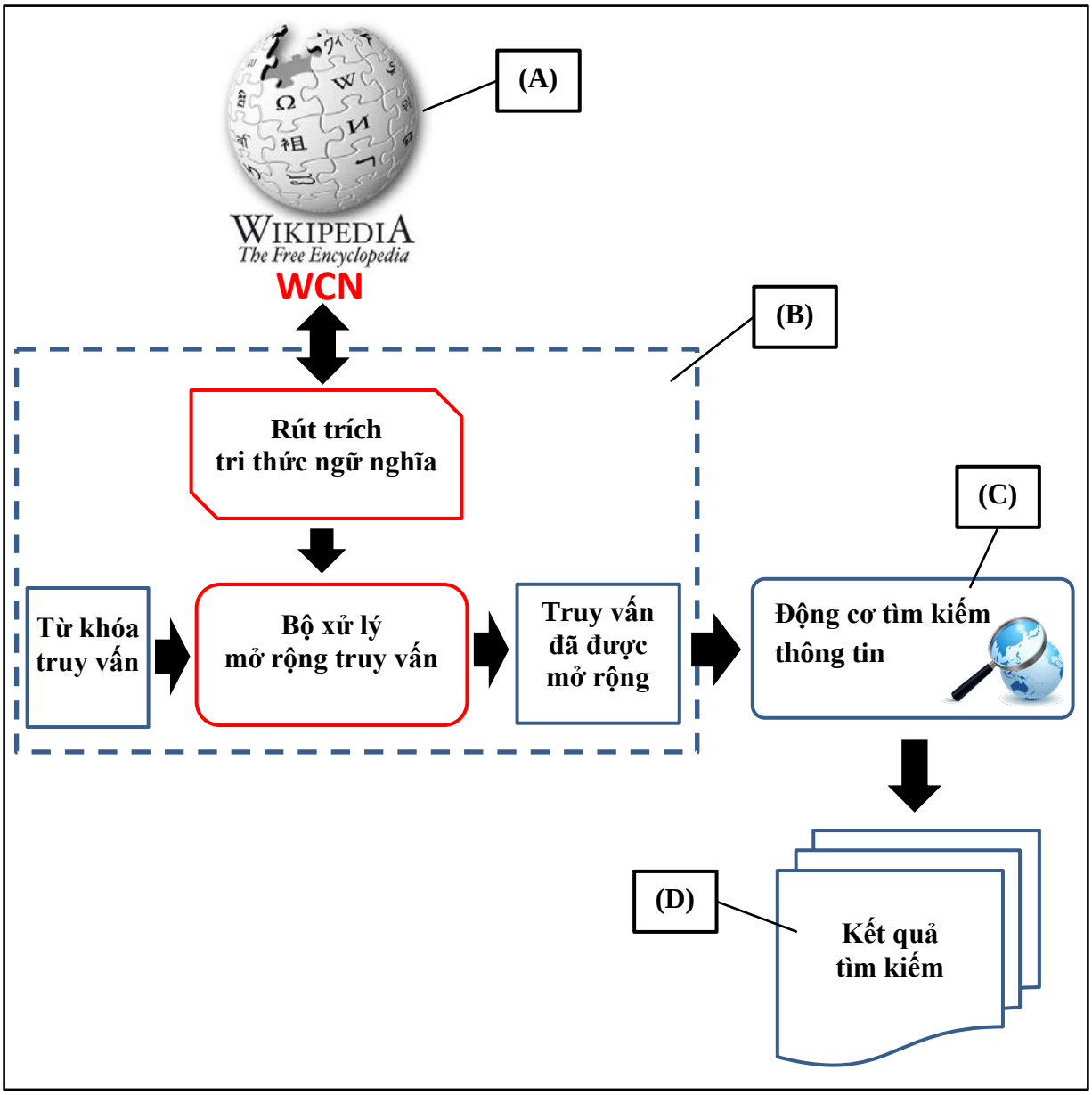
+ Thực nghiệm, đánh giá

+ Xây dựng mô phỏng ứng dụng mở rộng truy vấn sử dụng các độ đo vừa rút trích được.

1.3 Đối tượng, phạm vi nghiên cứu

Từ mục đích nghiên cứu, luận văn xác định khai thác hệ thống tên thể loại của Wikipedia (Wikipedia Category Network - WCN) để rút trích tri thức ngữ nghĩa và tính độ tương đồng từ vựng và áp dụng trong mô phỏng mở rộng truy vấn tìm kiếm.

Theo định hướng nêu trên, phạm vi nghiên cứu của luận văn được thể hiện trong hình 1.1, trong khung đường nét đứt.



Hình 1.1 Mô hình phạm vi luận văn và các hệ thống liên quan

(A) Hệ thống bách khoa toàn thư mở Wikipedia

(B) Mô hình phạm vi nghiên cứu của luận văn

(C) Động cơ tìm kiếm thông tin của hệ thống truy xuất thông tin

(D) Kết quả tìm kiếm thông tin từ yêu cầu của các câu truy vấn đã mở rộng

Luận văn tập trung nghiên cứu dữ liệu bách khoa toàn thư mở Wikipedia. Trong phạm vi hệ thống phân cấp thể loại với các tiêu đề bài viết thuộc các loại đó. Luận văn tìm hiểu và sử dụng dữ liệu Wikipedia cập nhật tháng 11 năm 2014.

Ngoài hệ thống phân cấp thể loại của Wikipedia, luận văn còn nghiên cứu về Wordnet, máy học hỗ trợ vectơ (Support Vector Machine - SVM), gán nhãn từ loại (Part of Speech – POS tagging) và các đối tượng liên quan khác.

Để hoàn thành mục tiêu của đề tài, cần giải quyết các bài toán:

1. Chọn lọc dữ liệu từ Wikipedia
2. Phân tích các tên thể loại, tiêu đề bài viết thành các cặp từ vựng
3. Rút trích các đặc trưng từ hệ thống phân cấp thể loại Wikipedia
4. Tính độ tương đồng từ vựng dựa trên các đặc trưng rút trích từ Wikipedia
5. Mô phỏng ứng dụng mở rộng truy vấn sử dụng các đặc trưng đã rút trích được.

1.4 Ý nghĩa khoa học của đề tài

Các đóng góp chính của đề tài:

Khai thác đặc trưng ngữ nghĩa từ hệ thống Wikipedia; luận văn chỉ sử dụng tiêu đề bài viết và tên thể loại, không cần phân tích toàn bộ nội dung bài báo.

Kết hợp các đặc trưng rút trích từ Wikipedia tính độ tương đồng của từ vựng.

Xây dựng ứng dụng thực nghiệm mở rộng truy vấn tìm kiếm sử dụng các đặc trưng rút trích được để cải tiến kết quả tìm kiếm.

1.5 Cấu trúc của luận văn

Luận văn được bố cục thành 6 chương và được trình bày như sau:

Chương 1: Mở đầu

Trình bày lý do chọn đề tài, mục đích, đối tượng và phạm vi nghiên cứu, ý nghĩa khoa học và thực tiễn của đề tài nghiên cứu rút trích tri thức ngữ nghĩa từ tên thể loại Wikipedia.

Chương 2: Tổng quan

Nội dung chương này trình bày việc phân tích, đánh giá các công trình nghiên cứu về rút trích thông tin Wikipedia của các tác giả trong và ngoài nước; nêu những vấn đề còn tồn tại và đưa ra định hướng mà đề tài tập trung nghiên cứu, giải quyết đó là tập trung khai thác tính ngữ nghĩa từ hệ thống phân loại thể loại của Wikipedia.

Chương 3: Rút trích đặc trưng từ Wikipedia

Chương này tập trung chủ yếu trình bày các cơ sở lý thuyết, lý luận, và các phương pháp đề xuất đã được sử dụng trong Luận văn. Các phương pháp nghiên cứu được trình bày theo từng khái niệm thông qua các tính chất và ví dụ minh hoạ.

Chương 4: Thực nghiệm

Nội dung chương 4 trình bày quá trình thực nghiệm tính các độ đo WordNet, tính hệ số tương quan Pearson (ký hiệu là r). Trình bày phương pháp huấn luyện dữ liệu sử dụng mô hình hàm nhân phi tuyến (**Radial Basis Function - RBF**), kiểm tra chéo 5 phần (5 folds cross-validation). Chương 4 đồng thời trình bày mô phỏng ứng dụng mở rộng truy vấn sử dụng các đặc trưng đã rút trích được từ Wikipedia.

Chương 5: Đánh giá

Chương này, luận văn trình bày mô tả ngắn gọn công việc thực nghiệm của đề tài và trình bày các số liệu các kết quả của quá trình thực nghiệm và nhận xét đánh giá kết quả thực nghiệm. Cụ thể là so sánh kết quả tính độ tương quan của các độ đo chuẩn WordNet và độ tương quan khi có thêm các đặc trưng ngữ nghĩa Wikipedia.

Chương 6: Kết luận và hướng phát triển

Nội dung của chương 6 là phần tổng kết, trong đó trình bày tóm lược các kết quả của luận văn, một số vấn đề còn tồn tại và hướng phát triển trong tương lai, liên quan đến đề tài.

Phần cuối của luận văn là các phụ lục. Trong đó, phụ lục A trình bày tóm lược về hệ thống bách khoa toàn thư mở Wikipedia. Phụ lục B trình bày danh mục các từ loại tiếng Anh.

CHƯƠNG 2. TỔNG QUAN

2.1 Trong nước

Hệ thống bách khoa toàn thư mở Wikipedia được xem như một cơ sở tri thức, việc khai thác dữ liệu từ hệ thống Wikipedia đã trở thành tiêu điểm của nhiều nghiên cứu gần đây trong lĩnh lực rút trích thông tin (Information Extraction - IE) và việc xây dựng cơ sở tri thức. Tuy nhiên, việc rút trích thông tin ngữ nghĩa nói chung và rút trích thông tin ngữ nghĩa từ hệ thống dữ liệu Wikipedia nói riêng vẫn là công việc đầy khó khăn thử thách.

Trong nghiên cứu "Mô hình rút trích cụm từ đặc trưng ngữ nghĩa trong tiếng Việt" [3] nhóm tác giả Nguyễn Quang Châu, PGS.TS. Phan Thị Tươi đã đề xuất mô hình xác định cụm từ đặc trưng ngữ nghĩa ViKEa dùng phương pháp so trùng mẫu dựa trên việc khai thác Vi.Wikipedia như một Ontology tiếng Việt. Đề xuất phương pháp khai thác Vi.Wikipedia như một ontology tiếng Việt không chỉ để phục vụ cho việc xác định cụm danh từ đặc trưng ngữ nghĩa cho câu tiếng Việt mà còn mở ra một hướng giải quyết cho vấn đề thiếu hụt về kho ngữ liệu của các công trình nghiên cứu về xử lý ngôn ngữ tiếng Việt bằng máy tính hiện nay.

Ở nghiên cứu "Tóm tắt đa văn bản dựa vào trích xuất câu" [3] của nhóm tác giả Trần Mai Vũ, PGS. TS. Hà Quang Thụy đã đề xuất Phương pháp tính độ tương đồng câu dựa vào Wikipedia, nghiên cứu này sử dụng các trang bài viết trong Wikipedia.

Ở nghiên cứu [7] của nhóm tác giả Hien T Nguyen, Tru H Cao. đã khai thác dữ liệu từ Wikipedia phục vụ bài toán khử nhập nhằng tự động cho thực thể có tên. Nhóm nghiên cứu định hướng xây dựng và phát triển chuyên sâu về bài toán thực thể có tên và ontology.

2.2 Nước ngoài

Trên thế giới hiện nay có khá nhiều đề tài, công trình nghiên cứu sử dụng tài nguyên Wikipedia trong các lĩnh vực rút trích thông tin, truy xuất thông tin. Tuy

nhiên nhiều tính năng của tài nguyên Wikipedia này vẫn chưa được khai thác hết tiềm năng. Đặc biệt là hệ thống tên loại Wikipedia với dung lượng nhỏ nhưng hàm chứa nhiều tính năng ngữ nghĩa. Một số nghiên cứu rút trích thông tin từ Wikipedia sử dụng phương pháp học máy có giám sát (Supervised Machine Learning) lấy thông tin từ hệ thống phân cấp tên loại Wikipedia. Trong nghiên cứu [9], chỉ ra tầm quan trọng của thứ tự trong danh sách các thể loại (Category) mà bài viết thuộc về. Vị trí của category trong danh sách đó cho biết độ liên quan tới bài viết và mức độ quan trọng của nó. Nghiên cứu [5] công bố một ontology mới – YAGO, vừa có phạm vi rộng vừa có chất lượng cao. Đề xuất này lợi dụng các trang category, đưa ra kỹ thuật rút trích thông tin sự kiện từ Wikipedia kết hợp Wordnet.

Giải pháp [12] của nhóm Maria Ruiz - Casado cho phép khai thác các quan hệ ngữ nghĩa của Wikipedia để bổ sung cho WordNet thông qua sử dụng các mẫu từ vựng xác định để thể hiện quan hệ ngữ nghĩa giữa các khái niệm. Kết quả đạt được bao gồm 270 câu cho quan hệ hạ danh, 158 câu cho quan hệ thượng danh, 247 câu cho quan hệ bộ phận và 222 câu cho quan hệ toàn thể. Việc phân tích tiếp cho 1.204 quan hệ dạng hạ danh với 852 quan hệ chưa tồn tại trong WordNet với độ chính xác bình quân là 0,69; 418 quan hệ dạng bộ phận với 303 chưa có trong WordNet dẫn đến độ chính xác 0,61, và 184 quan hệ mới dạng toàn thể với độ chính xác 0,61.

Giải pháp [14] đề xuất sử dụng mạng ngữ nghĩa Wikipedia để thay thế Wordnet. Vì rằng các phương pháp tính độ tương đồng câu sử dụng kho ngữ liệu Wordnet được đánh giá cho ra kết quả cao. Tuy nhiên, kho ngữ liệu Wordnet chỉ hỗ trợ ngôn ngữ tiếng Anh, việc xây dựng kho ngữ liệu này cho các ngôn ngữ khác đòi hỏi sự tốn kém về mặt chi phí, nhân lực và thời gian. Trong giải pháp này Simone Paolo Ponzetto và cộng sự tập trung vào việc áp dụng và cải tiến một số độ đo phổ biến về tính độ tương đồng từ trên tập ngữ liệu Wordnet cho việc tính độ tương đồng giữa các khái trên mạng ngữ nghĩa Wikipedia.

Luận văn này tập trung tìm hiểu các đặc trưng ngữ nghĩa của WCN dùng cho việc trút trích thông tin từ Wikipedia. Sau đó trình bày mô phỏng ứng dụng mở rộng truy vấn sử dụng các đặc trưng này.

CHƯƠNG 3. RÚT TRÍCH ĐẶC TRƯNG NGỮ NGHĨA TỪ TÊN LOẠI WIKIPEDIA

3.1 Cơ sở lý luận

Wikipedia được xem như một ontology mở, được xây dựng bởi những người tình nguyện theo hướng tiếp cận từ dưới lên, với các khái niệm được hình thành từ một tập từ vựng tự do và các thoả thuận mang tính cộng đồng.

Trong quá trình phân loại bài viết Wikipedia, người ta luôn xếp một bài viết vào loại có liên quan. Điều đó cũng có nghĩa là người ta luôn cố gắng dùng tên phân loại sao cho bao gồm được các tên bài viết thuộc tên phân loại đó.

Mỗi bài viết Wikipedia chứa một danh sách các thể loại mà nó thuộc về. Tiêu đề bài viết và danh sách các tên thể loại mà bài viết đó thuộc về có quan hệ ngữ nghĩa với nhau.

Ví dụ, mối quan hệ một bài viết với các thể loại chứa nó, cụ thể với bài viết ‘Eraser’ trong Wikipedia như hình 3.1. Bài viết ‘Eraser’ thuộc các thể loại ‘Stationery’, ‘Writing implements’, ‘Art materials’.

Bài viết ‘Eraser’ và thể loại ‘Stationery’ có quan hệ ngữ nghĩa với nhau.

Bài viết ‘Eraser’ và thể loại ‘Writing implements’ có quan hệ ngữ nghĩa với nhau.

Bài viết ‘Eraser’ và thể loại ’Art materials’ có quan hệ ngữ nghĩa với nhau.

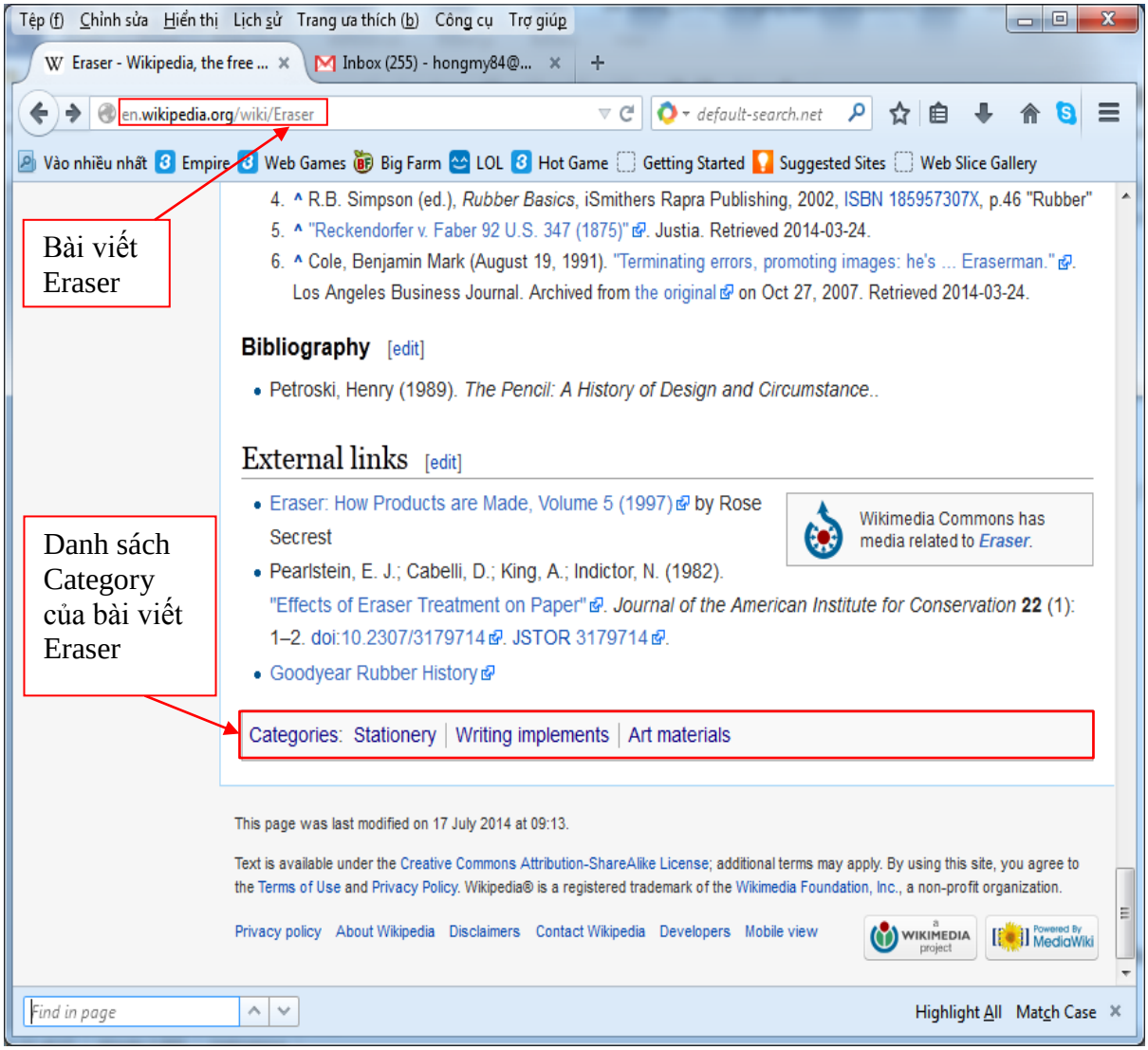
Thể loại ‘Stationery’ và thể loại ‘Writing implements’ có quan hệ ngữ nghĩa.

Thể loại ‘Stationery’ và thể loại ‘Art materials’có quan hệ ngữ nghĩa.

Thể loại ‘Writing implements’ và thể loại ‘Art materials’có quan hệ ngữ nghĩa.

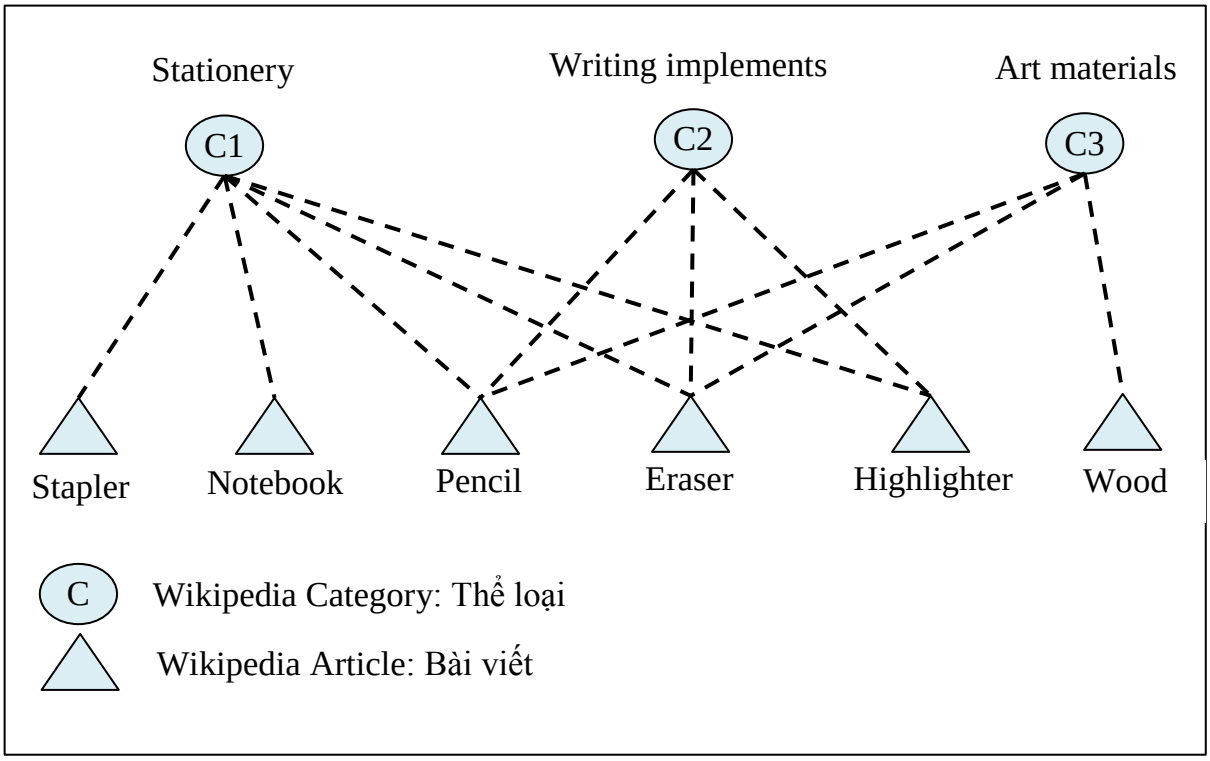
Ngoài ra, thứ tự của các thể loại trong danh sách thể loại này mang ý nghĩa nhất định đối với bài báo. Các thể loại bên trái trong danh sách có quan hệ ngữ nghĩa cao hơn, điều này được chứng minh trong nghiên cứu [9]. Do đó, luận văn xem xét khai

thác đặc trưng thứ tự vị trí của thể loại trong danh sách thể loại của bài viết. Đây là cơ sở để tính đặc trưng ngữ nghĩa 'Leftness' ở các phần sau.



Hình 3.1 Các thể loại của một bài viết trong hệ thống Wikipedia

Các thể loại (category) và bài viết (article) của Wkipedia được tổ chức lưu trữ dưới dạng mạng lưới các khái niệm liên quan ngữ nghĩa với nhau gọi là Wikipedia Category Network (WCN) [16].



Hình 3.2 Mô hình Wikipedia Category Network (WCN)

Từ những phân tích trên cho thấy WCN là nguồn dữ liệu mang tính ngữ nghĩa cao. Vì vậy, luận văn tập trung khai thác thông tin từ danh sách tên thể loại và tên tiêu đề bài viết của mạng lưới thể loại Wikipedia. Luận văn này trình bày một phương pháp với các đặc trưng ngữ nghĩa của WCN dùng cho việc trút trích thông tin từ Wikipedia.

Luận văn đồng thời trình bày đánh giá các đặc trưng ngữ nghĩa trên các tập dữ liệu chuẩn.

3.2 Phân tích hệ thống cấp bậc

Phần này tập trung phân tích vị trí của một thể loại trong WCN, đầu tiên là phân tích thể loại đơn và sau đó là xem xét cặp thể loại.

3.2.1 Category đơn

3.2.1.1 NormalizedRepresentation (NR_1)

Tên thể loại biểu diễn các bài viết bằng một khái niệm hoặc chủ đề. Vì vậy, số lượng bài viết dưới một thể loại được xem như độ biểu diễn của nó. Để tránh chênh lệch với thể loại lớn luận văn chuẩn hóa phép đo này bằng cách chia nó cho tổng của các biểu diễn của các thể loại mà các bài viết thuộc về. Với tham chiếu đến WCN trong hình 3.2, có thể được viết như sau:

$$NR_1(C_1) = \frac{\#article\ (C_1)}{\#article\ (C_1) + \#article\ (C_2) + \#article\ (C_3)}$$

Công thức 3-1. Tính đặc trưng NR1

Trong đó:

#article(C_1) là số bài viết của thể loại C_1

#article(C_2) là số bài viết của thể loại C_2

#article(C_3) là số bài viết của thể loại C_3

3.2.1.2 $Leftness_1$

Độ đo leftness của một thể loại (ký hiệu là Leftness1) là độ liên quan của thể loại và bài viết. Giả sử có N bài viết trong thể loại C, độ leftness của thể loại C được tính bởi công thức sau:

$$Leftness_1(C) = \sum_{i=1}^{N} \frac{\#categories(Ai) - pos(C)}{\#categories(Ai)}$$

Công thức 3-2. Tính đặc trưng Leftness1

Trong đó:

$\#categories(Ai)$ là số các thể loại của bài viết Ai

pos(*C*) là vị trí của thể loại C trong danh sách thể loại của bài viết Ai.

3.2.2 Cặp category

3.2.2.1 NormalizedRepresentation (NR_2)

Biểu diễn chuẩn của hai thể loại được tính bằng số bài viết chung của hai thể loại chia cho tổng số lượng các bài viết của từng loại

$$NR_2(C_1, C_2) = \frac{\#article\ (C_1, C_2)}{\#article\ (C_1) + \#article\ (C_2)}$$

Công thức 3-3. Tính đặc trưng NR_2

Trong đó:

$\#article\ (C1, C2)$ là số các bài viết giữa C_1 và C_2

#article(C_1*)* là số bài viết trong thể loại C_1

#article(C_2*)* là số bài viết trong thể loại C_2

3.2.2.2 $Leftness_2$

Độ đo leftness của cặp thể loại được tính là giá trị leftness1 nhỏ nhất của các thể loại thành phần.

$$Leftness_2(C_1, C_2) = Min(Leftness_1(C_1), Leftness_1(C_2))$$

Công thức 3-4. Tính đặc trưng Leftness2

Trong đó:

Độ đo $Leftness_1$ của một thể loại được tính bằng công thức 3-2.

3.3 Phân tích cú pháp

Các tên thể loại là các cụm danh từ. Phân tích các thẻ từ loại (Part-of-Speech) POS của cụm từ để xác định từ loại của mỗi từ trong tên. Tên thể loại có dạng một trong hai mẫu sau: *NounPhrase* hoặc *NounPhrase*(IN *NounPhrase*)+ trong đó:

- Mẫu *NounPhrase* bao gồm bất kỳ thẻ POS trong tập {NN, NNP, NNPS, NNS, VBN, VBZ, VBD, VB, CD, DT, JJ , CC}.
- IN là đại diện cho tập {*by*, *in*, *from*, *for*, *out*, *of*, *with*, *at*, *about*}.
- Dấu + có nghĩa là có thể xuất hiện lặp lại nhiều hơn một lần.

Ví dụ phân tích mẫu **NounPhrase**: cụm danh từ 'Guinea pig' được chia nhỏ ra thành các từ đơn 'guinea' và 'pig'. Rồi các từ đơn gộp lại thành cặp 'guinea_pig'

Ví dụ phân tích mẫu **NounPhrase (IN NounPhrase)+** : 'Cats in art' và 'Films about cats' sẽ được phân tích thành các cặp: 'cats_films', 'cats_art' và 'art_films'.

3.4 Cơ sở lý thuyết kiến thức liên quan

3.4.1 Thư viện libsvm

Libsvm là thư viện hỗ trợ hiệu quả trong việc phân lớp SVM và hồi quy. Thư viện này có thể sử dụng để phân lớp C-SVM, nu-SVM, hồi quy epsilon-SVM và hồi quy nu-SVM. Thư viện này có thể tải về ở địa chỉ http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Định dạng của tập tin dữ liệu huấn luyện (train) và tập tin kiểm tra (test) là:

<label><index1>:<value1> <index2>:<value2>...

Trong đó:

<label> là giá trị đích của tập huấn luyện. Đối với hồi quy, nó là một số thực bất kỳ

<index> là một số nguyên bắt đầu từ 1

<value> là một số thực

Có một tập huấn luyện đơn giản đối với việc phân lớp trong gói này: heart_scale.

Gõ ‘svm-train heart_scale’ thì chương trình sẽ đọc dữ liệu huấn luyện và xuất ra file mô hình heart_scale.model. sau đó ta có thể gõ ‘svm-predict heart_scale heart_sacle.model output’ để xem tỉ lệ phân lớp trên tập huấn luyện. File output chứa giá trị dự đoán của mô hình.

Một số công cụ khác:

svm-scale: là một công cụ đối với việc xác định file dữ liệu vào.

Ví dụ:

> svm-scale -l -1 -u 1 -s range train

> train.scale

> svm-scale -r range test

> test.scale

Scale mỗi đặc trưng của dữ liệu huấn luyện là [-1,1].

svm-toy: là một giao diện đồ họa đơn giản thể hiện dữ liệu phân tách SVM trong mặt phẳng. Trong giao diện này có thể vẽ các điểm dữ liệu, đọc dữ liệu từ file, tạo mô hình SVM, nhưng chỉ áp dụng cho trường hợp phân lớp, không áp dụng cho trường hợp hồi quy.

Sử dụng thư viện svm-train:

svm-train [các tùy chọn] tập_tin_huấn_luyện [kiểu_tập_tin]

các tùy chọn:

-s : kiểu của SVM (default 0)

0 -- C-SVC

1 -- nu-SVC

2 -- one-class SVM

3 -- epsilon-SVR

4 -- nu-SVR

-t : kiểu của hàm kernel(default 2)

0 -- linear: u'*v

1 -- polynomial: (gamma*u'*v + coef0)^degree

2 -- radial basis function: exp(-gamma*|u-v|^2)

3 -- sigmoid: tanh(gamma*u'*v + coef0)

-d degree : bậc của hàm kernel (default 3)

-g gamma : giá trị gamma trong kernel function (default 1/k)

-r coef0 : giá trị coef0 trong kernel function (default 0)

-c cost : tham số C của C-SVC, epsilon-SVR, and nu-SVR (default 1)

-n nu : tham số nu của nu-SVC, one-class SVM, and nu-SVR (default 0.5)

-p epsilon : giá trị epsilon trong hàm loss của epsilon-SVR (default 0.1)

-m cachesize : kích thước cache bộ nhớ tính theo MB (default 40)

-e epsilon : dung sai (tolerance) của tiêu chuẩn thoát (default 0.001)

-h shrinking: có sử dụng shrinking(co lại) heuristics hay không, 0 or 1 (default 1)

-wi weight: tham số C của lớp i tới trọng số weight*C in C-SVC (default 1)

-v n: n-fold cross validation mode (chia phần kiểm tra chéo)

Giá trị k trong tùy chọn -g nghĩa là số thuộc tính trong dữ liệu đầu vào.

Tùy chọn -v phân chia ngẫu nhiên dữ liệu thành n phần

Ví dụ:

> svm-train -s 0 -c 1000 -t 2 -g 0.5 -e 0.00001 data_file

Huấn luyện phan lớp RBF kernel exp(-0.5|u-v|^2) và dung sai là 0.00001

> svm-train -s 3 -p 0.1 -t 0 -c 10 data_file

Thực hiện hồi quy SVM vói hàm nhân u'v và C=10, epsilon = 0.1

> svm-train -s 0 -c 500 -g 0.1 -v 5 data_file

Thực hiện kiểm tra chéo 5 phần trong phân lớp sử dụng tham số C = 500 và gamma = 0.1

3.4.2 Thư viện ws4j

Thư viện ws4j có tên tiếng Anh đầy đủ là WordNet Similarity for Java. Thư viện này cung cấp Java API để tính các độ đo WordNet, như là HSO, LCH, LESK, WUP, RES, JCN, LIN. Để sử dụng thư viện này, yêu cầu máy tính có cài đặt JDK 6 trở lên.

+ Độ đo Path Length (PATH):

Độ đo Path Length được Rada và cộng sự đề xuất năm 1989, độ đo này sử dụng độ dài khoảng cách ngắn nhất giữa hai khái niệm trên đồ thị để thể hiện sự gần nhau về mặt ngữ nghĩa. Độ dài khoảng cách ngắn nhất giữa hai khái niệm là số cạnh giữa hai khái niệm.

$$PATH(n1, n2) = 1 / path\_length(n1, n2)$$

Công thức 3-5. Độ đo Path Length

Trong đó:

- n1, n2 là hai khái niệm cần tính toán
- path_length(n1, n2): khoảng cách ngắn nhất giữa hai khái niệm

+ Độ đo Leacock & Chodorow (LCH)

Độ đo LeacockChodorow được Leacock và Chodorow đề xuất năm 1998 chuẩn hóa độ dài khoảng cách giữa hai node bằng độ sâu của đồ thị

LCH(n1, n2) = -log(path_length(n1, n2))/ (2 * depth)

Công thức 3-6. Độ đo Leacock & Chodorow

Trong đó:

- n1, n2: là hai khái niệm cần tính toán

- depth: là độ dài lớn nhất trên đồ thị

- path_length(n1, n2): khoảng cách ngắn nhất giữa hai khái niệm

+ Độ đo WuPalmer (WUP)

Độ đo WUP được Wu và Palmer đề xuất năm 1994

WUP(n1, n2) = 2*depth(LCS) / path_length(n1, LCS) + path_length (n2, LCS) + 2*depth(LCS)

Công thức 3-7. Độ đo WuPalmer

Trong đó:

- n1, n2: là hai khái niệm cần tính toán

- LCS: Khái niệm thấp nhất trong hệ thống cấp bậc quan hệ is-a hay nó là cha của hai khái niệm n1 và n2

- depth(LCS): là độ sâu của khái niệm cha

+ Độ đo Resnik (RES)

Độ đo Resnik được Resnik đề xuất 1995. Độ tương đồng ngữ nghĩa Resnik giữa hai khai niệm được xem như nội dung thông tin trong nút cha gần nhất của hai khái niệm.

$$RES(n1, n2) = IC(LCS(n1, n2))$$

Công thức 3-8. Độ đo Resnik

Trong đó:

- n1, n2: là hai khái niệm cần tính toán

- IC được tính như công thức: IC(n) = 1-(log(hypo(n)+1)/log(C)

- hypo(n) là số các khái niệm có quan hệ thượng hạ vi (hyponym) với khái niệm n và C là tổng số các khái niệm có trên cây thể loại

+ Độ đo JiangConrath (JCN)

Độ đo JCN được Jiang và Conrath đề xuất năm 1997:

$$JCN(n1, n2) = IC(n1) + IC(n2) + 2*IC(LCS(n1, n2))$$

Công thức 3-9. Độ đo JiangConrath

Trong đó:

- n1, n2: là hai khái niệm cần tính toán

- IC được tính như công thức: IC(n) = 1-(log(hypo(n)+1)/log(C)

+ Độ đo Lin

Độ đo Lin được Lin đề xuất năm 1998:

$$LIN(n1, n2) = 2*IC(LCS(n1, n2) / (IC(n1) + IC(n2))$$

Công thức 3-10. Độ đo Lin

Trong đó:

- n1, n2: là hai khái niệm cần tính toán

- IC được tính như công thức: IC(n) = 1-(log(hypo(n)+1)/log(C)

3.4.3 Độ tương quan (correlation)

Độ tương quan (correlation) đo mối liên hệ tương đối giữa hai biến. Hệ số tương quan (correlation coefficient) cho biết độ mạnh yếu của mối quan hệ tuyến tính giữa hai biến số ngẫu nhiên. Hệ số tương quan Pearson (kí hiệu r) là một chỉ số thống kê dùng để đo mức độ tương quan giữa hai biến số. Hệ số tương quan giữa 2 biến có thể dương hoặc âm. Hệ số tương quan dương cho biết rằng giá trị 2 biến tăng cùng nhau còn hệ số tương quan âm thì nếu một biến tăng thì biến kia giảm. Hệ số tương quan r có giá trị từ -1 đến 1. Khi $r < 0$ có nghĩa là nếu giá trị của biến này tăng thì giá trị của biến còn lại giảm. Ngược lại, khi $r > 0$ có nghĩa là nếu giá trị của biến này tăng thì giá trị của biến kia cũng tăng. Khi hệ số tương quan bằng 0 hay gần 0 có nghĩa là hai biến số ít có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có mối liên hệ tuyệt đối.

Tính hệ số này theo công thức Pearson (Pearson's Correlation):

Gọi x và y là hai biến

Bước 1: Tính trung bình của x và y

Bước 2: Tính độ lệch của mỗi giá trị của x với trung bình của x (lấy các giá trị của x trừ đi trung bình của x) và gọi là "a", làm tương tự như vậy với y và gọi là "b"

Bước 3: Tính: $a \times b$, a^2 và b^2 cho mỗi giá trị

Bước 4: Tính tổng $a \times b$, tổng a^2 vả tổng b^2

Bước 5: Chia tổng của a × b cho căn bậc 2 của [(sum a^2) × (sum b^2)]

Hệ số tương quan Pearson r của hai biến số x và y từ n mẫu, được tính bằng công thức sau:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) - (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Công thức 3-11. Hệ số tương quan Pearson (r)

Trong đó:

$\bar{x}$ là giá trị trung bình của biến số x

$\bar{y}$ là giá trị trung bình của biến số y

Độ tương quan được sử dụng được sử dụng để tính và so sánh độ tương quan các độ đo WordNet và kết hợp các độ đo WordNet với các đặc trưng rút trích từ tên thể loại Wikipedia, dùng dữ liệu là các tập dữ liệu chuẩn WS353 và TSA287. Khi thực hiện tính toán so sánh, luận văn đã sử dụng hàm Pearson trong Excel.

CHƯƠNG 4. THỰC NGHIỆM

4.1 Môi trường thực nghiệm

Các quá trình thực nghiệm của luận văn được thực hiện trên máy tính cá nhân hệ điều hành Window, với các chi tiết về phần cứng và phần mềm như sau:

Bảng 4.1 Cấu hình máy tính

STT	Thành phần	Thông số kỹ thuật
1	CPU	Intel Core i5-3230M 2.60 GHz
2	RAM	4.00 GB
3	HDD	500 GB
4	OS	Windows 7 Ultimate 64-bit

Bảng 4.2 Danh sách phần mềm

STT	Tên phần mềm	Tác giả; nguồn
1	Eclipse	https://eclipse.org
2	JDK 6	https://jdk6.java.net
3	Mysql	http://www.mysql.com
4	MS Excel	Microsoft Office
5	Phần mềm R	http://cran.r-project.org
6	Thư viện Libsvm 3.20	http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.2 Dữ liệu

Các đặc trưng ngữ nghĩa (Sematic Features) được đánh giá bằng cách sử dụng độ đo tương quan ngữ nghĩa. Luận văn thực nghiệm 2 tập dữ liệu chuẩn: WordSimilarity-353 Test Collection [11] và Temporal Semantic Analysis [10]. Tập dữ liệu WordSimilarity có mức đánh giá liên quan của con người về 353 cặp từ vựng. Tập dữ liệu Temporal Semantic Analysis có 287 cặp từ với mức đánh giá của

con người. Để tiện cho việc sử dụng trong luận văn, tên các tập dữ liệu viết tắt tương ứng là WS353 và TSA287.

Tập dữ liệu WS353, chứa tập hợp các mức đánh giá liên quan của con người về 353 cặp từ vựng. Tập dữ liệu này có thể được dùng để huấn luyện (train) hoặc kiểm tra (test) các thuật toán tính độ tương đồng ngữ nghĩa, ví dụ như là các thuật toán đánh giá độ tương đồng của các từ vựng trong ngôn ngữ tự nhiên. Tập dữ liệu WS353 được ghép từ 2 tập dữ liệu con. Tập đầu tiên chứa 153 cặp từ cùng với độ đánh giá tương tự trong 13 đối tượng. Tập thứ hai chứa 200 cặp từ với các độ đánh giá tương tự trong 16 đối tượng. Tập WS353 kết hợp hai tập con trên và mỗi cặp từ vựng có độ đánh giá tương tự là trung bình của các độ đo từ các tập con.

4.3 Thực nghiệm

Luận văn sử dụng các phép đo từ WikiRelate [14] làm cơ sở cho nghiên cứu và thực hiện thực nghiệm đề tài, các độ đo tương quan ngữ nghĩa trên WordNet được điều chỉnh áp dụng cho phù hợp với Wikipedia.

Các đặc trưng ngữ nghĩa rút trích từ Wikipedia được thực nghiệm và đánh giá bằng cách sử dụng các độ đo WordNet Similarity thông qua thư viện WS4J. Sau đó so sánh đánh giá độ tương đồng bằng máy học vectơ hỗ trợ (Support Vector Machine - SVM).

Mạng thể loại Wikipedia (WCN) giống như là một mạng ngữ nghĩa giữa các từ tương tự như Wordnet. Mặc dù mạng thể loại không hoàn toàn được xem như là một cấu trúc phân cấp do vẫn có các thể loại không có liên kết đến các thể loại khác tuy nhiên số lượng này là khá ít.

Phương pháp tính độ tương đồng giữa các khái niệm trong mạng ngữ nghĩa Wikipedia được khá nhiều các nghiên cứu đưa ra như Simone Paolo Ponzetto và cộng sự năm 2006 [14], Torsten Zesch và cộng sự năm 2007 [16]. Các nghiên cứu này tập trung vào việc áp dụng và cải tiến một số độ đo phổ biến về tính độ tương

đồng từ trên tập ngữ liệu Wordnet cho việc tính độ tương đồng giữa các khái trên mạng ngữ nghĩa Wikipedia.

Luận văn sử dụng thư viện WS4J (WordNet Similarity for Java) tính các độ đo liên quan giữa hai từ. Các độ đo được chia thành hai loại độ đo, nhóm độ đo dựa vào khoảng cách giữa các khái niệm như độ đo Path Length (PL, năm 1989), HirstStOnge, LeacockChodorow (LC, năm 1998), Lesk, WuPalmer (WP, năm 1994); và nhóm các độ đo dựa vào nội dung thông tin như độ đo Resnik (Res, năm 1995), JiangConrath (JC, năm 1997), Lin (Lin, năm 1998). Các độ đo có giá trị tính toán giữa hai khái niệm càng lớn thì độ tương đồng càng cao.

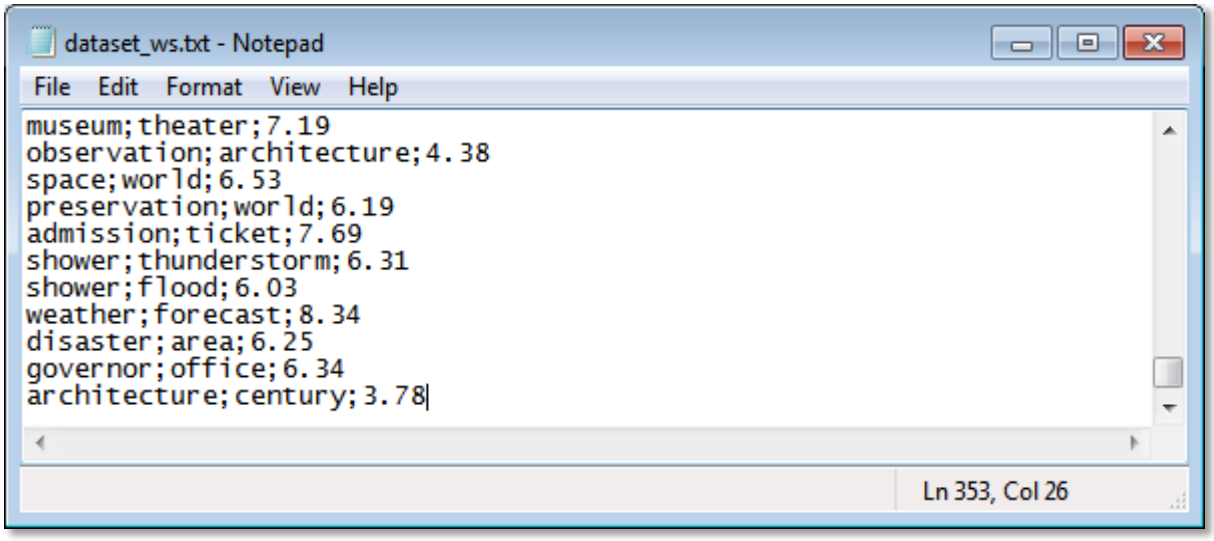
Đề tài dùng thư viện WS4J để tính độ đo similarity của các cặp từ trong các tập dữ liệu WS353 và TSA287. Trong các tập dữ liệu, với mỗi cặp từ luận văn tính các độ đo tương đồng WordNet Similarity bằng cách sử dụng thư viện WS4J. Gồm các độ đo: HirstStOnge, LeacockChodorow, Lesk, WuPalmer, Resnik, JiangConrath, Lin, Path Length.

Minh họa đoạn code tính độ đo HirstStOnge (HSO) sử dụng thư viện WS4J

```
double HirstStOnge_relatedness(String word1, String word2){
    ILexicalDatabase db = new NictWordNet();
    HirstStOnge hso = new HirstStOnge(db);
    double hso_r = hso.calcRelatednessOfWords(word1, word2);
    return hso_r;
}
```

Theo giới chuyên gia về lĩnh vực tương quan ngữ nghĩa, chúng ta đánh giá bằng cách tính hệ số tương quan Pearson giữa các giá trị độ đo với đánh giá của con người. Luận văn áp dụng tính độ tương quan với các đánh giá của các dữ liệu độ đo WordNet được trình bày ở phần trên. Luận văn sử dụng hệ số tương quan Pearson để so sánh trên các tập dữ liệu chuẩn: WS353, TSA287.

Sau khi được các giá trị độ đo của các cặp từ trong tập dữ liệu, tính hệ số tương quan r cho các độ đo cho từng tập dữ liệu.

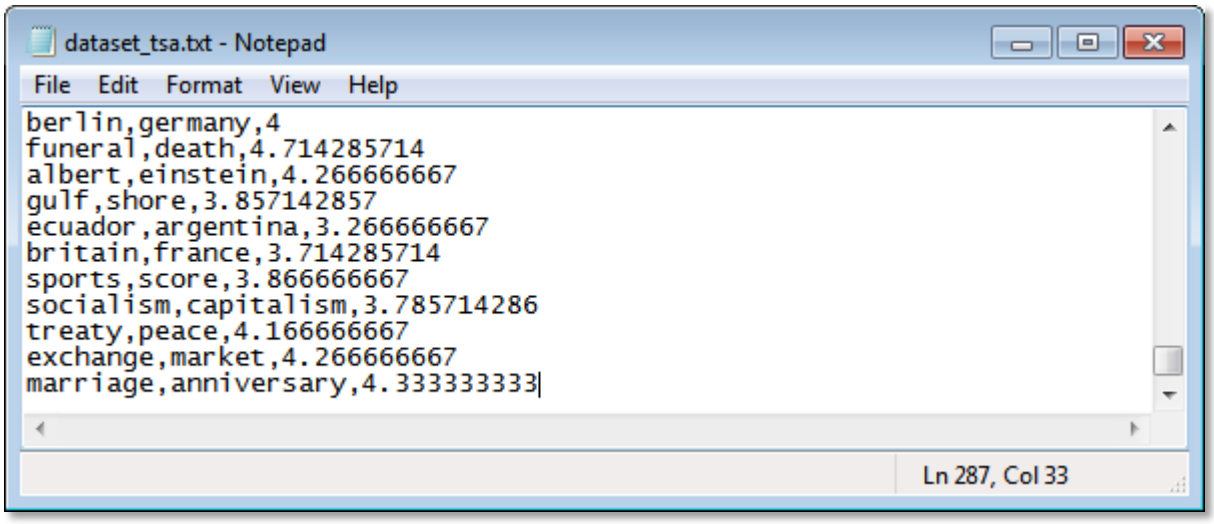


Hình 4.1 Dữ liệu WS353

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	pairword	WS353	Hirst	Leacock	Lesk	WuPalmer	Resnik	Jiang	Lin	Path		Độ đo	r(WS353)
2	love-sex	6.77	0	1.54	0	0.29	0	0.06	0	0.2		Hirst	0.278
3	tiger-cat	7.35	0	1.29	0	0.58	1.82	0.06	0.2	0.1		Leacock	0.219
4	book-paper	7.46	5	1.39	0	0.35	0.61	0.08	0.1	0.1		Lesk	0.272
5	computer-keyboard	7.62	4	2.3	7	0.84	4.37	0.08	0.4	0.3		WuPalmer	0.221
6	computer-internet	7.58	5	1.61	0	0.67	3.45	0.08	0.3	0.1		Resnik	0.325
7	plane-car	5.77	5	1.49	3	0.69	5.53	0.26	0.7	0.1		Jiang	0.217
8	train-car	6.31	3	1.61	0	0.7	5.46	0.23	0.7	0.1		Lin	0.255
9	telephone-communication	7.5	0	1.05	0	0.24	0	0.07	0	0.1		Path	0.248
10	television-radio	6.77	5	2.59	2	0.91	8.27	0.57	0.9	0.3			
11	media-radio	7.42	0	0	0	0	0	0	0	0		Tính độ tương đồng Pearson (ô M2)	
12	drug-abuse	6.85	0	1.25	0	0.29	0	0.06	0	0.1		=PEARSON($B$2:$B$353,C$2:C$353)	
13	bread-butter	6.19	5	2.3	2	1.5	6.97	0.19	0.7	0.3			
14	cucumber-potato	5.92	3	0.86	0	0.11	0.61	0	0	0.1			
15	doctor-nurse	7	4	2.3	0	0.88	6.89	0.26	0.8	0.3			
16	professor-doctor	6.62	0	1.74	0	0.77	6.1	0.29	0.8	0.1			
17	student-professor	6.81	0	1.61	0	0.7	1.9	0.08	0.2	0.1			
18	smart-student	4.62	0	0.86	0	0.2	0	0	0	0.1			
19	smart-stupid	5.81	0	0.86	0	0.2	0	0	0	0.1			
20	company-stock	7.08	0	1.39	0	0.38	0.78	0.07	0.1	0.1			
21	stock-market	8.08	0	1.72	0	0.38	0.78	0.06	0.1	0.2			

ws_corelation2 Sheet1

Hình 4.2 Các độ đo WordNet trên dữ liệu WS353



Hình 4.3 Dữ liệu TSA287

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	pairword	TSA287	Hirst	Leacock	Lesk	WuPalme	Resnik	Jiang	Lin	Path		Độ đo	r(TSA287)	
2	episcopal-russia	2.75	0	0	0	0	0	0	0	0		Hirst	0.298	
3	water-shortage	2.714286	0	1.29	0	0.17	0	0.06	0	0.09		Leacock	0.337	
4	horse-wedding	2.266667	0	0.6	0	0.16	0	0.06	0	0.05		Lesk	0.245	
5	plays-losses	3.2	0	0	0	0	0	0	0	0		WuPalmer	0.407	
6	classics-advertiser	2.25	0	0.74	0	0.18	0	0.04	0	0.05		Resnik	0.355	
7	latin-credit	2.0625	0	1.49	0	0.5	3.07	0.08	0.3	0.11		Jiang	0.266	
8	ship-ballots	2.3125	0	0	0	0	0	0	0	0		Lin	0.305	
9	mistake-error	4.352941	16	3.69	13	1	7.72	1.28	1	1		Path	0.36	
10	disease-plague	4.117647	4	2.08	0	0.85	6.85	0	0	0.2				
11	sake-shade	2.529412	0	1.39	0	0.47	2.4	0.09	0.3	0.1		Tính độ tương đồng Pearson (ô M2)		
12	saints-observatory	1.9375	0	0	0	0	0	0	0	0		=PEARSON($B$2:$B$288,C$2:C$288)		
13	treaty-wheat	1.8125	0	0.74	0	0.1	0	0.05	0	0.05				
14	texas-death	1.533333	0	1.49	0	0.2	0	0.06	0	0.11				
15	republicans-challenge	2.3125	0	0	0	0	0	0	0	0				
16	body-peaceful	2.058824	0	0	0	0	0	0	0	0				
17	admiralty-intensity	2.647059	0	1.05	0	0.32	0.78	0	0	0.07				
18	body-improving	2.117647	0	0	0	0	0	0	0	0				
19	heroin-marijuana	3.375	2	0.98	0	0.12	0.61	0	0	0.07				
20	scottish-commuters	2.6875	0	0	0	0	0	0	0	0				
21	apollo-myth	2.6	0	1.29	0	0.17	0	0	0	0.09				
22	film-cautious	2.125	0	1.49	0	0.46	0.78	0.06	0.1	0.11				

tsa_corelation2

Hình 4.4 Các độ đo WordNet trên dữ liệu TSA287

Để thuận tiện cho việc so sánh, dữ liệu tổng hợp độ tương đồng Pearson của các độ đo Wordnet ở hình 4.2 và 4.4 được trình bày ở bảng 4.1 bên dưới.

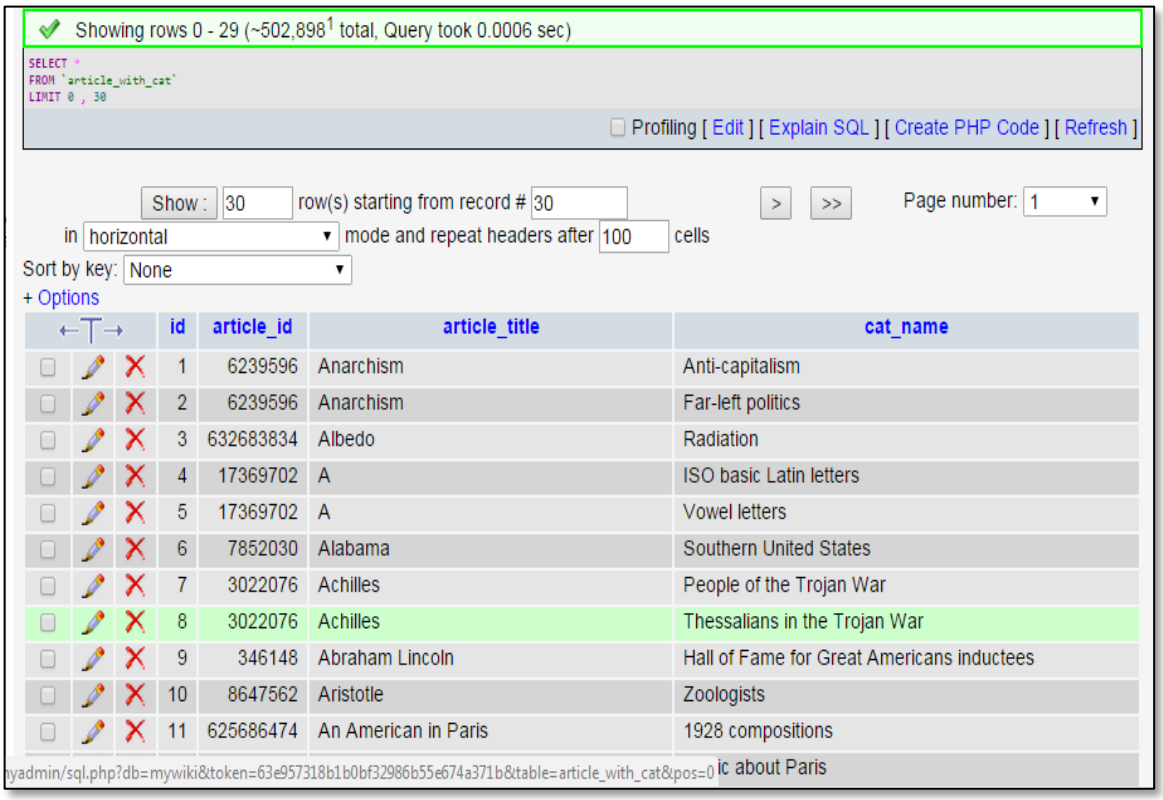
Bảng 4.3 Sự tương quan các độ đo Wordnet similarity

Độ đo	r(WS353)	r(TSA287)
HirstStOnge	0.278	0.298
LeacockChodorow	0.219	0.337
Lesk	0.272	0.245
WuPalmer	0.221	**0.407**
Resnik	**0.325**	0.355
JiangConrath	0.217	0.266
Lin	0.255	0.305
Path	0.248	0.36

Trong bảng 4.3 trình bày các độ đo tương tự từ WordNet trên hai tập dữ liệu là WS353 và TSA287. Độ tương đồng cao nhất trong mỗi tập dữ liệu được in đậm.

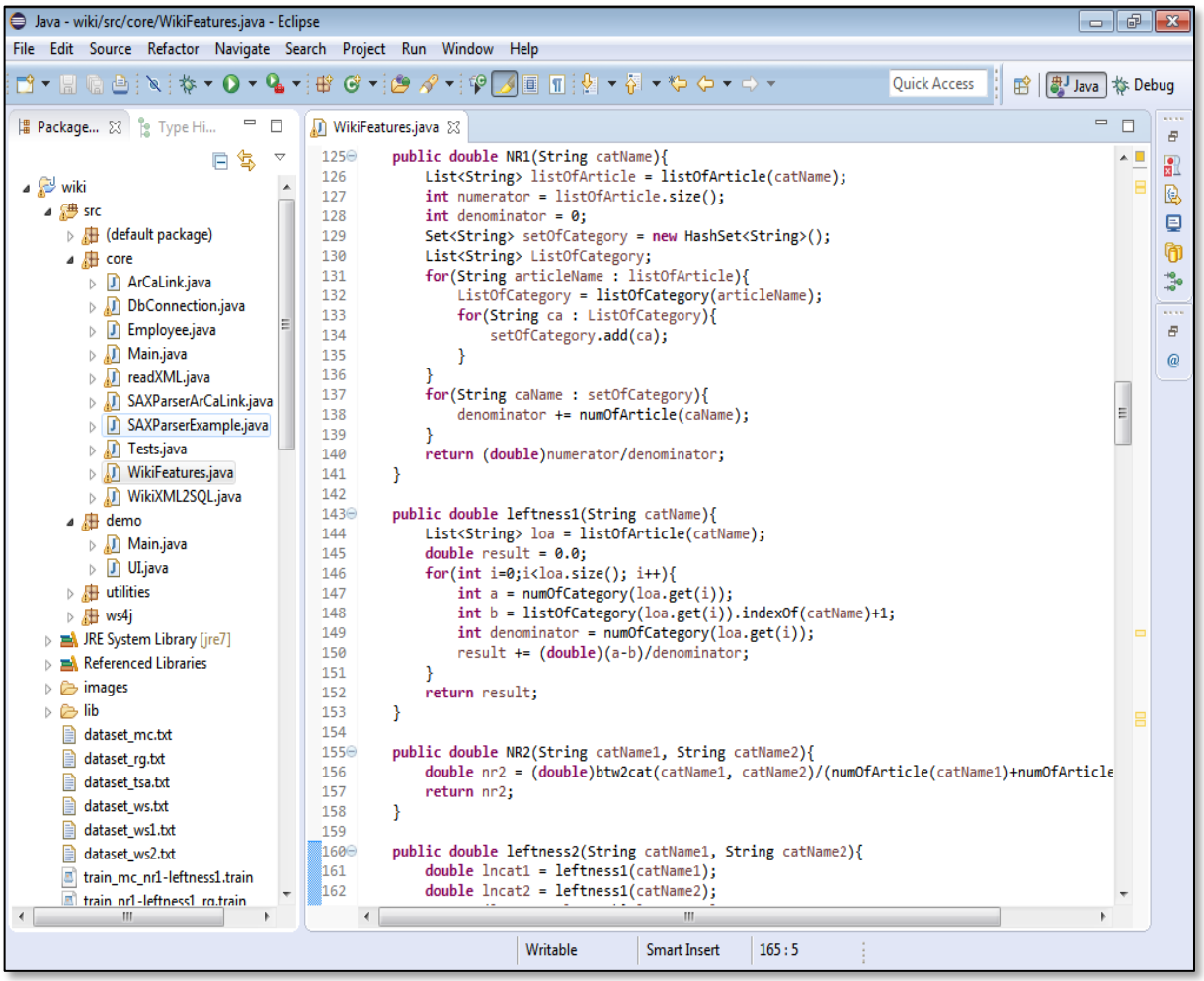
Luận văn sử dụng ngôn ngữ Java, JDK6. Phân tích các tập tin dữ liệu Wikipedia dạng XML, chỉ lấy tên bài viết và danh sách các thể loại của bài viết đó. Dữ liệu tên bài viết và danh sách tên thể loại được lưu vào cơ sở dữ liệu MySql để thực hiện tính các đặc trưng Wikipedia.

Đề tài sử dựng dữ liệu Wiki dump bản ngôn ngữ tiếng Anh thời gian 20141106. Hiện tại đề tài chỉ sử dụng 10 tập tin trên tổng số 27 tập tin ‘enwiki-20141106-pages-articles.xml’. Sau khi phân tích các trang XML để lấy tiêu đề bài viết và tên thể loại, dữ liệu này được lưu vào cơ sở dữ liệu MySql với hơn 500.000 mẩu tin. Trong đó, dữ liệu có hơn 360.000 tiêu đề bài viết và hơn 190.000 tên thể loại.



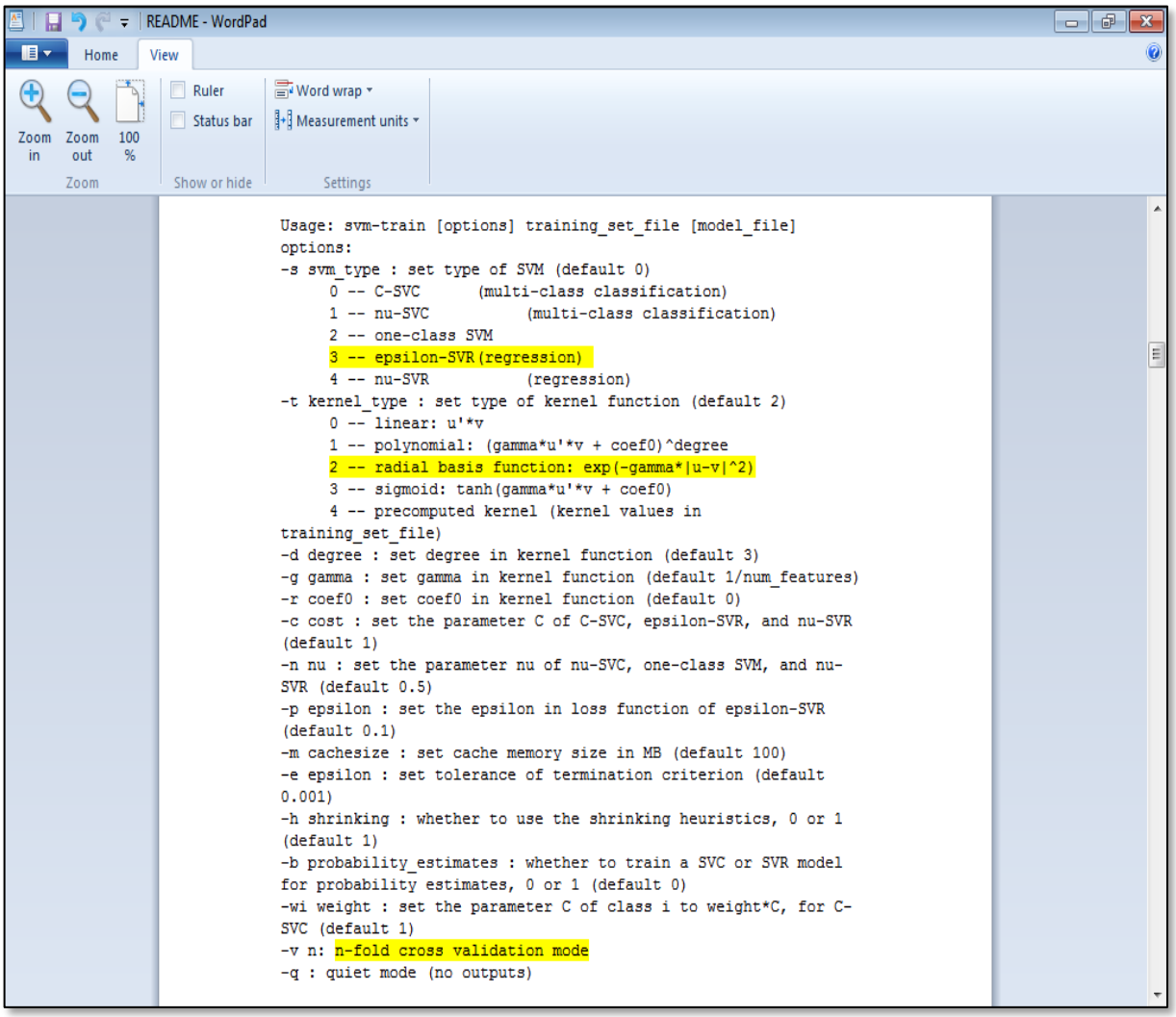
Hình 4.5 Dữ liệu tiêu đề bài viết cùng tên thể loại

Dưới đây là các đoạn mã thực hiện rút trích đặc trưng từ dữ liệu tiêu đề bài viết và tên thể loại Wikipedia.

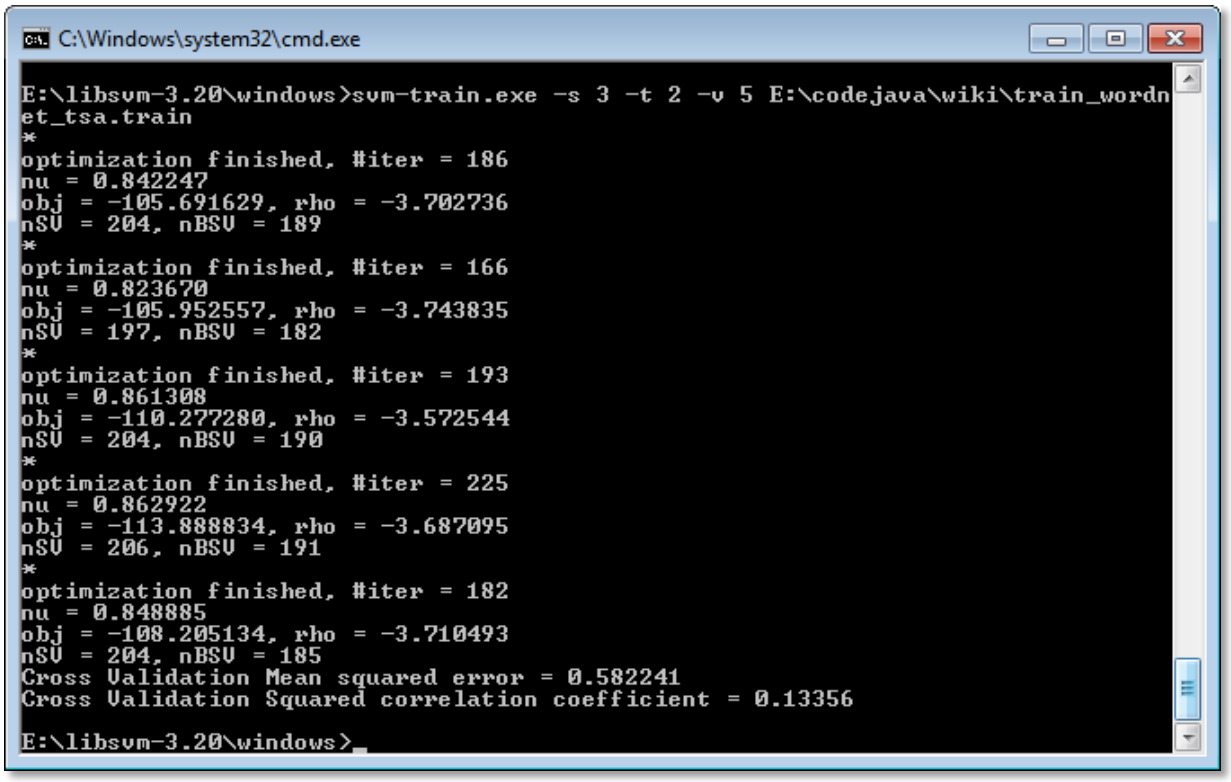


Hình 4.6 Phương thức tính các đặc trưng từ Wikipedia

Các độ đo WordNet được hợp nhất lại bằng cách thực hiện hồi quy sử dụng giải thuật máy học vector hỗ trợ (Support Vector Machine - SVM) để ước tính mức độ phụ thuộc theo đánh giá con người trên nhiều độ đo mối liên quan. Dữ liệu huấn luyện và kiểm tra là tất cả các độ đo WordNet trong bảng 4-1. Sau đó thêm các đặc trưng ngữ nghĩa vào các độ đo WordNet và lặp lại quá trình huấn luyện. Luận văn sử dụng mô hình hàm nhân phi tuyến (**Radial Basis Function - RBF**), kiểm tra chéo 5 phần trên tập dữ liệu huấn luyện. Luận văn sử dụng thư viện Libsvm, cú pháp huấn luyện là svm-train –s 3 –t 2 –v 5.



Hình 4.7 Sử dụng thư viện Libsvm



Hình 4.8 Huấn luyện dữ liệu sử dụng hàm nhân RBF kiểm tra chéo 5 phần

Trong hình 4.5 trình bày quá trình huấn luyện tập dữ liệu các độ đo WordNet trên dữ liệu TSA287 sử dụng hàm nhân RBF kiểm tra chéo 5 phần. Kết quả được bình phương độ tương quan là 0.13356, nên độ tương quan $r = \sqrt{0.13356} \approx 0.3655$. Các số liệu độ tương quan của các độ đo được trình bày ở bảng 5.1.

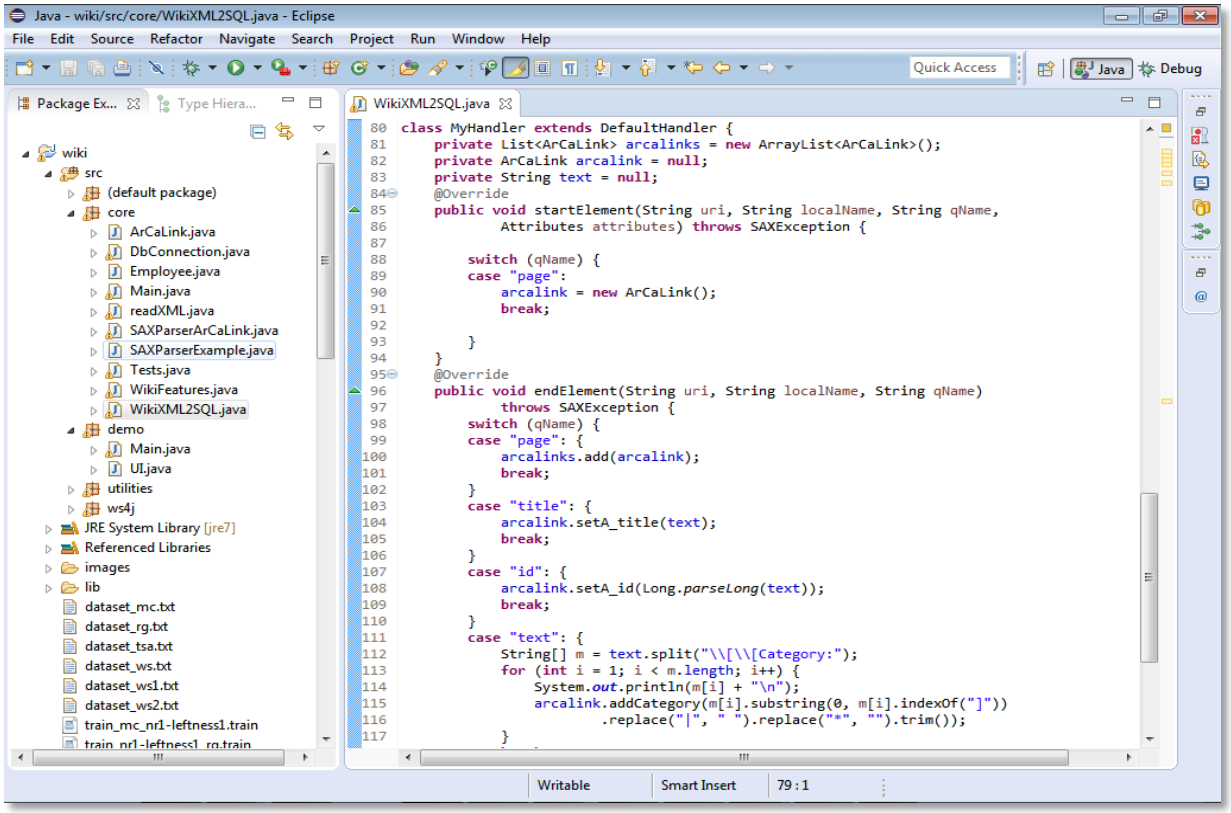
4.4 Mô hình mở rộng truy vấn

Tìm kiếm thông tin trên web là nhu cầu không thể thiếu trên thế giới cũng như ở Việt Nam. Tuy nhiên, với tốc độ internet phát triển nhanh chóng và mạnh mẽ tạo ra một kho tàng dữ liệu đồ sộ, đi đôi với sự phát triển này thì nó cũng gây ra nhiều khó khăn trong việc tìm kiếm và chọn lọc thông tin cần thiết. Người dùng cần một công cụ để tìm kiếm những thông tin một cách hiệu quả nhất. Trong các hệ thống tìm kiếm, đa phần là tìm theo từ khóa, cụm từ khóa. Điều này dẫn đến kết quả tìm kiếm liệt kê rất nhiều tài liệu nhưng chưa gần sát với yêu cầu của người dùng, chưa đáp ứng mong muốn của người dùng. Vì vậy, rất cần có ứng dụng xử lý tìm kiếm

cho kết quả cải tiến hơn. Luận văn xây dựng mô phỏng hệ thống mở rộng truy vấn sử dụng những đặc trưng ngữ nghĩa được khai thác hệ thống tên thể loại của Wikipedia.

Để chuẩn bị dữ liệu cho mô hình mở rộng truy vấn, luận văn sử dụng các dữ liệu bài báo wikipedia được lưu với định dạng XML. Tiến hành phân tích dữ liệu từ các tập tin XML để lấy tiêu đề bài viết và danh sách các thể loại mà bài viết thuộc về. Lưu các tiêu đề bài viết và danh sách category tương ứng vào cơ sở dữ liệu để thuận tiện cho việc truy xuất tính toán.

Ví dụ minh họa phân tích tập tin bài viết XML, chỉ lấy thông tin tiêu đề bài viết và danh sách các tên thể loại mà bài báo thuộc về:



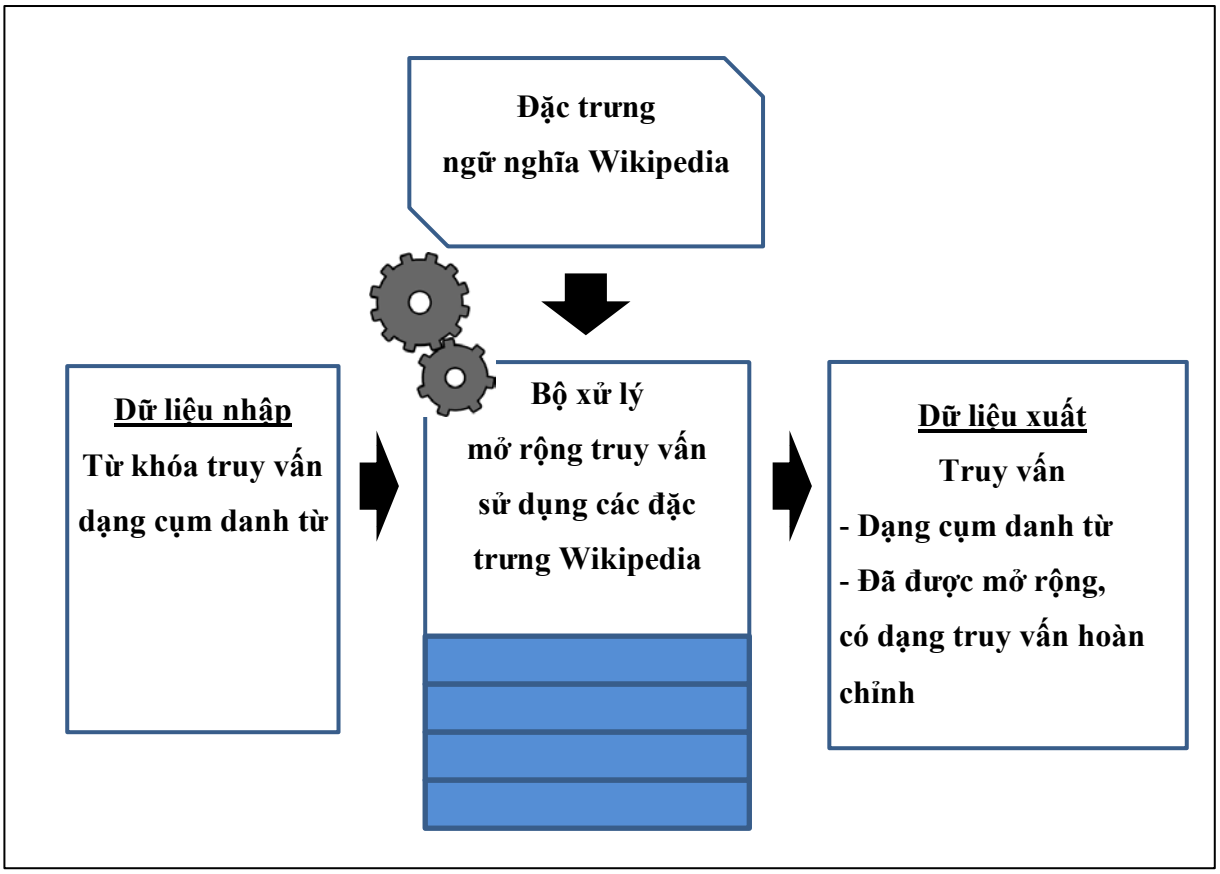

```
class MyHandler extends DefaultHandler {
    private List<ArCaLink> arcalinks = new ArrayList<ArCaLink>();
    private ArCaLink arcalink = null;
    private String text = null;
    @Override
    public void startElement(String uri, String localName, String qName,
            Attributes attributes) throws SAXException {

        switch (qName) {
        case "page":
            arcalink = new ArCaLink();
            break;

        }
    }
    @Override
    public void endElement(String uri, String localName, String qName)
            throws SAXException {
        switch (qName) {
        case "page": {
            arcalinks.add(arcalink);
            break;
        }
        case "title": {
            arcalink.setA_title(text);
            break;
        }
        case "id": {
            arcalink.setA_id(Long.parseLong(text));
            break;
        }
        case "text": {
            String[] m = text.split("\\[\\[Category:");
            for (int i = 1; i < m.length; i++) {
                System.out.println(m[i] + "\n");
                arcalink.addCategory(m[i].substring(0, m[i].indexOf("]"))
                        .replace("|", " ").replace("*", "").trim());
            }
```

Hình 4.9 Xử lý phân tích lấy tiêu đề bài viết và tên thể loại từ tập tin XML

Nội dung một tập tin XML mẫu:

```
<xml>
 …
<page>
    <title>Anarchism</title>
    <ns>0</ns>
    <id>12</id>
    <revision>
      <id>605992325</id>
      <parentid>605973323</parentid>
      <timestamp>2014-04-27T06:00:47Z</timestamp>
      <contributor>
        <username>BG19bot</username>
        <id>14508071</id>
      </contributor>
      <minor />
      <comment>[[WP:CHECKWIKI]] error fix for #61.  Punctuation
goes before References. Do [[Wikipedia:GENFIXES|general fixes]] if
a problem exists. - using [[Project:AWB|AWB]] (10084)</comment>
      <text xml:space="preserve">{{Redirect|Anarchist|the
fictional character|Anarchist (comics)}}
{{Redirect|Anarchists}}
{{Use British English|date=January 2014}}
{{pp-move-indef}}
{{Anarchism sidebar}}
'''Anarchism''' is a …….
{{Anarchism}}
{{Philosophy topics}}
{{Political culture}}
{{Political ideologies}}
{{Social and political philosophy}}
{{Aspects of Capitalism}}
{{Good article}}
[[Category:Anarchism| ]]
[[Category:Political culture]]
[[Category:Political ideologies]]
[[Category:Social theories]]
[[Category:Anti-fascism]]
[[Category:Anti-capitalism]]
[[Category:Far-left politics]]
{{Link FA|eo}}
{{Link FA|id}}</text>
      <sha1>2dde9i4r47ubvko1imjyrec1ffbchr2</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
    </revision>
  </page>…
</xml>
```

Quá trình mở rộng truy vấn:



Hình 4.10 Mô hình hệ thống mở rộng truy vấn tìm kiếm với động cơ tìm kiếm

Quá trình mở rộng truy vấn được thực hiện theo các bước sau:

Bước 1: Phân tích câu truy vấn thành các từ chính, loại bỏ các từ dừng (stop words) bằng cách sử dụng thư viện ‘opennlp’ để gán nhãn từ loại - POS tagging.

Bước 2: Tìm các từ này thuộc các tên thể loại hay tiêu đề bài viết nào, kết quả bước này cho ra danh sách các tên thể loại.

Bước 3: Tính các đặc trưng cho các tên thể loại đó, sau đó chọn các tên thể loại có giá trị tốt nhất.

Bước 4: Sử dụng POS taging các tên thể loại có độ tương đồng cao, lấy được các từ có độ tương đồng cao với các từ của truy vấn. Kết quả cho ra danh sách các từ mới tìm được vào danh sách các từ của truy vấn ban đầu.

Giao diện mô phỏng ứng dụng mở rộng truy vấn tìm kiếm cho phép nhập và truy vấn vào ô 'Input query', sau khi nhấn nút 'Expand' thì truy vấn đã được mở rộng sẽ hiển thị ở ô 'Expanded query'.

Hình 4.11 Giao diện mô phỏng ứng dụng mở rộng truy vấn sử dụng các đặc trưng rút trích từ Wikipedia

4.5 Xử lý dữ liệu lớn của Wikipedia

Trong phạm vi giới hạn về thời gian, phần cứng máy tính chưa mạnh và việc xử lý dữ liệu lớn, luận văn chưa thực nghiệm được trên dữ liệu đầy đủ của Wikipedia. Tuy nhiên, tác giả đã tìm hiểu và xử lý bước đầu cho việc xử lý dữ liệu lớn - dữ liệu đầy đủ của Wikipedia. Cụ thể là đã tải về các tập tin dữ liệu của Wikipedia dưới dạng sql và thực hiện dump sql và cơ sở dữ liệu MySQL.

Tải về dữ liệu Wikipedia cập nhật ngày 06 tháng 11 năm 2014:

http://dumps.wikimedia.org/enwiki/20141106/

- enwiki-20141106-page.sql.gz 1.1 GB
- enwiki-20141106-category.sql.gz 27.2 MB
- enwiki-20141106-categorylinks.sql.gz 1.4 GB

Sau khi tải về, giải nén và thực hiện dump sql (vì dữ liệu lớn không thể import trực tiếp vào cơ sở dữ liệu) có được kết quả dữ liệu như sau:

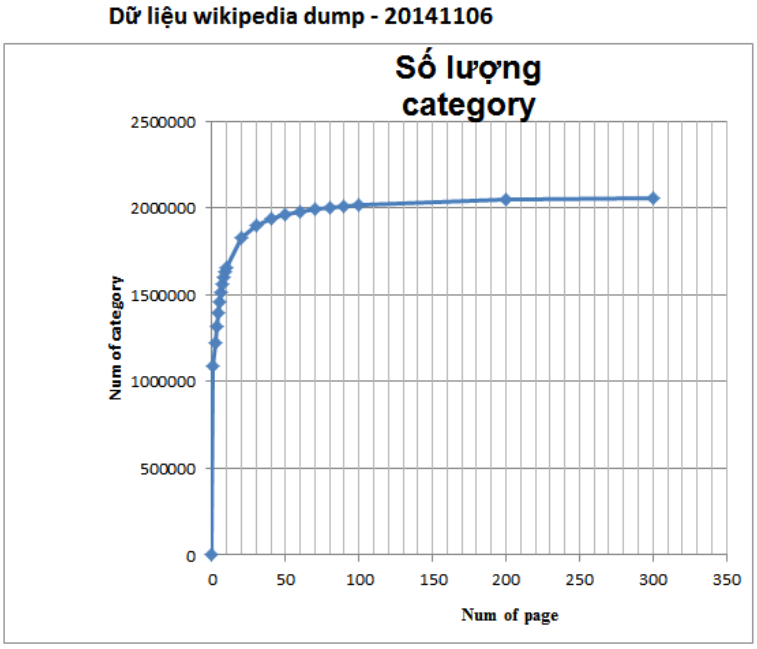
Table	Action	Records [1]	Type	Collation	Size
page		~26,919,905	InnoDB	binary	6.2 GiB
categorylinks		~43,806,297	InnoDB	binary	15.5 GiB
category		~2,099,271	InnoDB	binary	310.2 MiB
3 table(s)	Sum	~72,825,473	InnoDB	latin1_swedish_ci	22.0 GiB

Hình 4.12 Dữ liệu Wikipedia 20141106

Dung lượng của ba bảng ‘page’, ‘categorylinks’ và ‘category’ trong cơ sở dữ liệu MySQL chiếm khoảng 22 GB. Với dữ liệu này, chương trình phân tích và tính toán viết bằng ngôn ngữ Java không thực thi được, lỗi timeout vì dữ liệu quá lớn.

Giải pháp đầu tiên là giảm bớt số lượng category không cần thiết – là các thể loại quản trị của Wikipedia (administrative category), bằng cách thống kê số page của category và chọn ngưỡng giới hạn.

Ngưỡng (Số page)	Số lượng category
0	0
1	1089405
2	1220170
3	1319456
4	1398355
5	1461296
6	1515091
7	1558399
8	1595701
9	1628488
10	1658447
20	1823874
30	1894596
40	1933762
50	1958628
60	1976110
70	1989344
80	1999426
90	2007404
100	2013947
200	2044053
300	2054457



Hình 4.13 Thống kê để giới hạn dữ liệu

Sau khi chọn ngưỡng số trang, việc tiếp theo là xử lý chọn những thể loại có chứa bài viết và loại bỏ những thể loại khác. Cách này có thể giảm nhiều số lượng cần xử lý. Hướng xử lý tiếp theo là dùng các công nghệ xử lý dữ liệu lớn để xử lý và tính toán truy xuất dữ liệu. Định hướng có thể dụng công nghệ Hadoop và Map-Reduce để giảm thời gian xử lý.

CHƯƠNG 5. ĐÁNH GIÁ

5.1 Đánh giá kết quả thực nghiệm

Để so sánh đánh giá độ tương quan của các đặc trưng với đánh giá của con người, trên các tập dữ liệu WS353 và TSA287, luận văn tính các độ tương quan. Đầu tiên là chỉ tính với độ liên quan chuẩn là các độ đo WordNet, sau đó thêm các đặc trưng ngữ nghĩa vào và tính độ tương quan.

Bảng 5.1 Độ tương quan của các đặc trưng với đánh giá của con người

Độ đo	WS353	TSA287
WN	0.2806	0.3655
WN + NR1	0.2776	0.3596
WN + Leftness1	0.2767	0.3672
WN + NR2	0.2791	0.3667
WN + Leftness2	0.2779	0.3673

Dữ liệu trong bảng 5.1 trình bày các hệ số tương quan của các đặc trưng ngữ nghĩa với các giá trị đánh giá của con người. Kết quả thực hiện cao nhất là 0.3673 khi kết hợp các độ đo WordNet với đặc trưng leftness2, kết quả này lớn hơn so với khi chỉ sử dụng các độ đo WordNet mặc dù chưa sự chênh lệnh chưa nhiều ; dữ liệu này có ý nghĩa thống kê, được kiểm định t hai đuôi (2 tailed t-test) với độ tin cậy 95% (mức ý nghĩa $\alpha = 0.05$).

5.2 Đánh giá chung

Với cách tiếp cận của luận văn là phân tích thông tin phân cấp và ngữ nghĩa của hệ thống thể loại Wikipedia gồm các tên thể loại và tiêu đề bài viết, nên việc xử lý toàn bộ nội dung bài viết là không cần thiết. Điều này làm cho quá trình xử lý nhanh và hiệu quả, ít tốn chi phí hơn so với các công trình nghiên cứu phải xử lý

toàn bộ nội dung thông tin bài viết hoặc nội dung các infobox trong trang Wikipedia.

Các đặc trưng ngữ nghĩa rút trích hệ thống tên thể loại Wikipedia có thể được sử dụng trong các thuật toán học máy có giám sát để rút trích thông tin từ cơ sở tri thức bán cấu trúc như Wikipedia.

CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Từ chương 1 đến chương 5, luận văn đã trình bày toàn bộ nghiên cứu của để tài về rút trích tri thức ngữ nghĩa từ tên loại Wikipedi và áp dụng vào ứng dụng mở rộng truy vấn. Các chương đầu lần lượt cung cấp cơ sở lý thuyết làm cơ sở cho toàn bộ các phương pháp phân tích rút trích tri thức ngữ nghĩa từ hệ thống phân cấp thể loại Wikipedia được trình bày ở các chương tiếp theo. Các chương kế tiếp đề xuất mô hình, phương pháp cũng như những giải thuật xử lý phù hợp dựa trên cơ sở lý thuyết đã được trình bày ở các chương trước. Những nội dung được trình bày ở các chương đã bám sát mục tiêu đề ra. Điều này cũng thể hiện thông qua những kết quả đạt được về mặt lý thuyết và thực tiễn của luận văn. Các kết quả nghiên cứu của luận văn đã góp phần xác định những vấn đề cần nghiên cứu, phát triển trong thời gian tới.

6.1 Kết luận

Luận văn này đã kế thừa và cải tiến phương pháp để trích xuất thông tin hữu ích từ Wikipedia, sử dụng tính năng ngữ nghĩa được lấy từ hệ thống tên thể loại của Wikipedia. Phương pháp này cho kết quả khả quan. Các đặc trưng ngữ nghĩa lấy được từ phương pháp này có mối tương quan tốt với đánh giá của con người.

Từ việc khảo sát, phân tích cấu trúc thể loại và tài liệu lưu trữ trong Wikipedia, luận văn đã thực hiện phương pháp khai thác rút trích các đặc trưng ngữ nghĩa từ tên thể loại. Phát triển trên cơ sở kế thừa một phương pháp rút trích ngữ nghĩa từ tên thể loại, dựa trên nguồn dữ liệu tên thể loại sẵn có của Wikipedia. Kết quả thực nghiệm, đánh giá cho thấy phương pháp đề xuất là khả quan, có ý nghĩa thực tiễn.

Đóng góp của luận văn là xác định độ tương quan giữa các bài viết trên Wikipedia bằng cách áp dụng các độ đo khác nhau dựa trên WordNet. Luận văn

cũng thực hiện ứng dụng mô phỏng việc sử dụng các đặc trưng rút trích từ Wikipedia bằng ứng dụng mở rộng truy vấn.

Kết quả khoa học của luận văn là luận văn đã áp dụng một hướng tiếp cận kết hợp giữa WordNet và đặc trưng wikipedia để đánh giá độ tương đồng giữa các khái niệm.

6.2 Hướng phát triển

Trong khoảng thời gian giới hạn, luận văn đã thực hiện được các bài toán đặt ra. Tuy nhiên, một số vấn đề cần được nghiên cứu trong giai đoạn tiếp theo:

- Cải tiến công thức tính đặc trưng Wikipedia để truy xuất các từ tương đồng có độ chính xác cao hơn.
- Tối ưu các thuật giải, để xử lý dữ liệu lớn trong quá trình thực nghiệm và đánh giá.
- Hiệu chỉnh một số bước tiền xử lý để có thể áp dụng cho tiếng Việt hoặc ngôn ngữ khác. Cụ thể như là việc phân đoạn từ và gán nhãn từ loại cho truy vấn tiếng Việt.

Nhìn chung, các vấn đề trên là tập hợp những bài toán không quá phức tạp nhưng cần được xem xét và nghiên cứu trong tương lai, để có thể hỗ trợ cho việc rút trích thông tin ngữ nghĩa tiếng Việt từ Wikipedia và xây dựng một hệ thống truy xuất thông tin hướng ngữ nghĩa cho tiếng Việt.

TÀI LIỆU THAM KHẢO

- Tiếng Việt:

[1]. Nguyễn Chánh Thành. (2010). *Xây dựng mô hình mở rộng truy vấn trong truy xuất thông tin văn bản*. Luận văn Tiến sĩ Kỹ thuật. Chuyên ngành Khoa học máy tính, Đại học Bách khoa tp HCM.

[2]. Nguyễn Quang Châu, Phan Thị Tươi. (2008). *Nhận diện cụm từ đặc trưng ngữ nghĩa trong tiếng Việt*. Tạp chí Bưu chính Viễn thông và Công nghệ thông tin, số19, 2/2008.

[3]. Trần Mai Vũ. (2009). *Tóm tắt đa văn bản dựa vào trích xuất câu*. Luận văn Thạc sĩ. Đại học Công nghệ, Đại học quốc gia Hà Nội.

- Tiếng Anh:

[4]. D. Milne and I. H. Witten. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In In Proceedings of AAAI 2008.

[5]. F. M. Suchanek, G. Kasneci, and G. Weikum. (2007). 'Yago: a core of semantic knowledge'. In Proceedings of the 16th international conference on World Wide Web, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[6]. F. Wu and D. S. Weld. (2007). 'Autonomously semantifying wikipedia'. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, pages 41–50, New York, NY, USA, 2007. ACM.

[7]. Hien Thanh Nguyen, Tru Hoang Cao. (2010). 'Enriching Ontologies for Named Entity Disambiguation'. SEMAPRO 2010 : The Fourth International Conference on Advances in Semantic Processing, Vietnam.

[8]. Jun Cui. (2009). 'Query Expansion Research and Application in Search Engine Based on Concepts Lattice'. Master Thesis in Computer Science, Thesis no: MCS-2009: 28. School of Computing, Blekinge Institute of Technology, Soft Center, SE-37225 RONNEBY, SWEDEN.

[9]. K. Gyllstrom and M.-F. Moens. (2011). 'Examining the "leftness" property of wikipedia categories'. In Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, pages 2309–2312, New York, USA, 2011. ACM.

[10]. K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. (2011). 'A word at a time: computing word relatedness using temporal semantic analysis'. In Proceedings of the 20th international conference on World wide web, WWW'11, pages 337–346, New York, NY, USA, 2011. ACM

[11]. L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. (2001). 'Placing search in context: the concept revisited'. In WWW, pages 406–414, 2001

[12]. Maria Ruiz-Casado, Enrique Alfonseca and Pablo Castells. (2007). 'Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia'. 186 Data & Knowledge Engineering archive, Volume 61, Issue 3 (June 2007), pp.484-499. 2007. ISSN: 0169-023X

[13]. Priya Radhakrishnan, Vasudeva Varma. (2013). 'Extracting Semantic Knowledge from Wikipedia Category Names'. The 3rd Wordshop on Knowledge Extraction at CIKM 2013, San Francisco.

[14]. Strube, M. & S. P. Ponzetto (2006). 'WikiRelate! Computing semantic relatedness using Wikipedia', In Proc. of AAAI-06, 2006.

[15]. S. Banerjee and T. Pedersen. (2003). 'Extended gloss overlaps as a measure of semantic relatedness'. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pages 805–810, 2003.

[16]. T. Zesch and I. Gurevych. (2007). ‘Analysis of the wikipedia category graph for nlp applications’. In Proceedings of the TextGraphs-2 Workshop, NAACL-HLT, pages 1–8, Rochester, Apr. 2007. Association for Computational Linguistics

- Trang web:

[17]. Ask, http://www.ask.com/

[18]. Bing, http://www.bing.com/

[19]. Dữ liệu TSA287: http://www.technion.ac.il/~kirar/Datasets.html

[20]. Dữ liệu WS353: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[21]. Gate UK, http://gate.ac.uk

[22]. Google, http://www.google.com

[23]. Microsoft Corporation, http://www.microsoft.com

[24]. Support Vector Machines, http://www.support-vector.net

[25]. Wikipedia dump: http://dumps.wikimedia.org/enwiki/latest/

[26]. Wordnet, http://wordnet.princeton.edu/

[27]. XML, http://www.w3.org/XML

[28]. Yahoo, http://www.yahoo.com

PHỤ LỤC

Phụ lục A. Tóm lược về Wikipedia

Wikipedia là một bách khoa toàn thư mở, là thành quả cộng tác của chính những người đọc từ khắp nơi trên thế giới. Trang mạng này tất cả mọi người đều có thể sửa đổi ở bất cứ trang nào bằng cách bấm vào các liên kết "sửa đổi" có ở hầu hết các trang, ngoại trừ những trang bị khóa.

Wikipedia chính thức bắt đầu vào ngày 15 tháng 1 năm 2001 nhờ hai người sáng lập Jimmy Wales và Larry Sanger, chỉ có phiên bản tiếng Anh. Chỉ hơn ba năm sau, vào tháng 3 năm 2004, đã có 6.000 người đóng góp tích cực cho 600.000 bài viết với 50 thứ tiếng. Cho đến hôm nay đã có hơn 4.300.000 bài viết ở riêng phiên bản tiếng Anh, hơn 30.000.000 bài viết ở tất cả phiên bản ngôn ngữ. Mỗi ngày hàng trăm nghìn người ghé thăm từ khắp nơi để thực hiện hàng chục nghìn sửa đổi cũng như bắt đầu nhiều bài viết mới. Hiện tại, hệ thống Wikipedia đã có 427.009 thành viên đăng ký, trong đó 26 bảo quản viên, 3 hành chính viên, 214 robot.

Biểu trưng của Wikipedia là "quả bóng ghép chữ", hiện nay thuộc bản quyền của Quỹ hỗ trợ Wikimedia.

Nguồn tham khảo: http://vi.wikipedia.org/

Phụ lục B. Danh mục từ loại tiếng Anh

STT	Nhãn từ loại	Tên đầyđủ (tiếng Anh)	Ý nghĩa
1	CC	Coordinating conjunction	Liên từ kết hợp
2	CD	Cardinal number	Số đếm
3	DT	Determiner	Định từ
4	EX	Existential there	"Có"
5	FW	Foreign word	Từ tiếng nước ngoài
6	IN	Preposition or subordinating conjunction	Giới từ hoặc liên từ
7	JJ	Adjective	Tính từ
8	JJR	Adjective, comparative	Tính từ so sánh hơn
9	JJS	Adjective, superlative	Tính từ so sánh nhất
10	LS	List item marker	Dấu liệt kê
11	MD	Modal	Động từ tình thái
12	NN	Noun, singular or mass	Danh từ số ít hoặc không đếm được
13	NNS	Noun, plural	Danh từ số nhiều
14	NNP	Proper noun, singular	Danh từ riêng số ít
15	NNPS	Proper noun, plural	Danh từ riêng số nhiều
16	PDT	Predeterminer	Tiền chỉ định từ
17	POS	Possessive ending	Dấu sở hữu cách
18	PRP	Personal pronoun	Đại từ nhân xưng
19	PPS	Possessive pronoun (prolog version PRP-S)	Đại từ sở hữu
20	RB	Adverb	Trạng từ
21	RBR	Adverb, comparative	Trạng từ so sánh hơn
22	RBS	Adverb, superlative	Trạng từ so sánh nhất

23	RP	Particle	Tiểu từ
24	SYM	Symbol	Ký hiệu
25	TO	to	“to”
26	UH	Interjection	Thán từ
27	VB	Verb, base form	Động từ nguyên mẫu không to
28	VBD	Verb, past tense	Động từ thì quá khứ
29	VBG	Verb, gerund or present participle	Hiện tại phân từ
30	VBN	Verb, past participle	Quá khứ phân từ
31	VBP	Verb, non-3rd person singular present	Động từ không phải ngôi thứ 3 số ít
32	VBZ	Verb, 3rd person singular present	Động từ ngôi thứ 3 số ít
33	WDT	Wh-determiner	Định từ bắt đầu bằng Wh-
34	WP	Wh-pronoun	Đại từ bắt đầu bằng Wh-
35	WPZ	Possessive wh-pronoun (prolog version WP-S)	Đại từ sở hữu bắt đầu bằng Wh-
36	WRB	Wh-adverb	Trạng từ bắt đầu bằng Wh-
37	ADJP	Adjective Phrase.	Cụm tính từ
38	NP	Noun Phrase	Cụm danh từ
39	VP	Verb Phrase	Cụm động từ
40	ADVP	Adverb Phrase	Cụm trạng từ
41	CONJP	Conjunction Phrase	Cụm liên từ
42	RRC	Reduced Relative Clause	Mệnh đề tương đối thu giảm
43	UCP	Unlike Coordinated Phrase	Cụm phối hợp khác
44	WHADJP	Wh-adjective Phrase	Cụm tính từ bắt đầu với Wh-
45	WHAVP	Wh-adverb Phrase	Cụm trạng từ bắt đầu với Wh-
46	WHNP	Wh-noun Phrase	Cụm danh từ bắt đầu với Wh

47	WHPP	Wh-prepositional Phrase	Cụm giới từ bắt đầu với Wh-
48	PP	Prepositional Phrase	Cụm giới từ

Nguồn tham khảo: [1]