BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

LÊ HOÀNG OANH

NHẬN BIẾT CHỦ ĐỀ CỦA TÀI LIỆU DỰA TRÊN WIKIPEDIA

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 4 năm 2015

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

LÊ HOÀNG OANH

NHẬN BIẾT CHỦ ĐỀ TÀI LIỆU DỰA TRÊN 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 4 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
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

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

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 11 tháng 4 năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: LÊ HOÀNG OANH Giới tính: Nữ

Ngày, tháng, năm sinh: 09/03/1985 Nơi sinh: Cần Thơ

Chuyên ngành: Công nghệ Thông tin MSHV: 1341860016

I- Tên đề tài:

Nhận biết chủ đề của tài liệu dựa trên Wikipedia

II- Nhiệm vụ và nội dung:

- Khảo sát, phân tích hệ thống chủ đề của tài liệu dạng văn bản lưu trữ trong Wikipedia .
- Khảo sát các nghiên cứu liên quan đến việc nhận biết chủ đề của văn bản trong Wikipedia.
- Phát triển (trên cơ sở kế thừa) hoặc cải tiến một phương pháp nhận biết chủ đề tài liệu (dạng văn bản), 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: .....

V- Cán bộ hướng dẫn: *(Ghi rõ học hàm, học vị, họ, tên)* .....

TS. Nguyễn Chánh Thành

TS. Lê Mạnh Hải

CÁN BỘ HƯỚNG DẪN **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) (Họ tên và chữ ký)

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. 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.

Học viên thực hiện Luận văn

(Ký và ghi rõ họ tên)

Lê Hoàng Oanh

LỜI CÁM ƠN

Lời cảm ơn chân thành em xin gởi Ban Giám Hiệu, toàn thể cán bộ nhân viên, giảng viên trường Đại Học HUTECH, Ban lãnh đạo Phòng Quản Lý Khoa Học và Đào Tạo Sau Đại Học, khoa Công Nghệ Thông Tin đã tạo điều kiện thuận lợi cho em học tập và nghiên cứu trong suốt học trình cao học

Với lòng tri ân sâu sắc, em muốn nói lời cảm ơn chân thành đến TS. Nguyễn Chánh Thành và TS. Lê Mạnh Hải đã rất tận tụy và nghiêm túc hướng dẫn em trong quá trình thực hiện nghiên cứu này.

Em xin chân thành cảm ơn Quý thầy cô ngoài trường đã tận tâm dạy bảo em trong suốt quá trình học tập và giúp đỡ em trong suốt quá trình nghiên cứu.

Em xin chân thành cảm ơn những người thân yêu trong gia đình cùng các anh chị em, bạn bè, đồng nghiệp đã giúp đỡ và động viên em trong quá trình thực hiện và hoàn thành luận văn này.

Học viên thực hiện Luận văn
(ký và ghi rõ họ tên)

LÊ HOÀNG OANH

TÓM TẮT

(Tóm tắt nội dung LV bằng Tiếng Việt)

Wikipedia là một bách khoa toàn thư tự do, là kết quả của sự cộng tác của chính những người đọc từ khắp nơi trên thế giới. Mục tiêu phát triển của Wikipedia là nâng cao chất lượng bài viết, thêm nhiều bài viết chọn lọc, bài viết chất lượng và ngày càng thu hút nhiều thành viên tham gia.

Với số lượng bài viết ngày càng gia tăng thì việc tìm kiếm một bài báo nào đó như mong muốn là rất khó khăn và tốn nhiều thời gian. Chẳng hạn, khi người dùng muốn tìm kiếm một thông tin nào đó thì kết quả thường trả về rất nhiều danh mục có chứa thông tin đó. Vậy làm thế nào để kết quả chỉ trả về danh mục phù hợp nhất mà không phải là tất cả danh mục có chứa thông tin đó. Việc này đã đặt ra thách thức cho luận văn là tìm kiếm một giải pháp giúp nhận diện được danh mục nào có trọng số cao nhất phù hợp với thông tin cần tìm kiếm.

Chính vì thế, trong nghiên cứu này chúng tôi sẽ trình bày một thuật toán được sử dụng chỉ để khai thác tiêu đề và phân nhóm các tiêu đề trong Wikipedia. Giúp cho việc tìm ra các danh mục phù hợp với các bài báo một cách tự động và đạt độ chính xác cao.

ABSTRACT

(Tóm tắt nội dung LV bằng tiếng Anh)

Wikipedia is a free encyclopedia, as a result of the collaboration of the readers from all over the world. The objective of development of Wikipedia is to improve the quality of articles; add more selected articles, quality articles and increasingly attract more participants.

As regards the increasing number of articles these days, it is very difficult and time-consuming to find a specific article. For instance, when a user wants to search some information, the results are often returned a lot of catalogues containing that information. Thus, how the results are returned the most relevant catalogues related to information instead of all catalogues. This is sue has rise to the challenge to the thesis for seeking a solution identifying the most significant catalogue being suitable for the required information.

Therefore, in this study, we will represent an algorithm used to exploit only the titles and divide titles into many groups in Wikipedia. This helps to find the suitable catalogues to the articles automatically and accurately.

MỤC LỤC

Danh mục các từ viết tắt

STT	Từ hoặc cụm từ	Viết tắt
1.	Wikipedia	Wiki
2.	World Wide Web	WWW
3.	Wikipedia Category Graph	WCG
4.	Support vector machine	SVM
5.	Latent Dirichlet Allocation	LDA

Danh mục các bảng

Danh mục hình ảnh

Danh mục các biểu đồ

CHƯƠNG 1. MỞ ĐẦU

1.1 Lý do chọn đề tài

Khả năng cung ứng dữ liệu lớn trong World Wide Web (WWW) đã phát triển theo cấp số nhân từ vài năm qua, việc tìm kiếm, trích xuất và duy trì các thông tin là một nhiệm vụ khó khăn và tốn thời gian. Để khắc phục vấn đề quá tải thông tin, một công cụ nhận biết chủ đề của tài liệu rất cần thiết cho người dùng theo dõi và xác định chính xác vị trí của chủ đề tài liệu mà mình cần tìm kiếm.

Wikipedia [28] chính thức bắt đầu vào ngày 15 tháng 01 năm 2001 nhờ hai người sáng lập Jimmy Wales và Larry Sanger cùng với vài người cộng tác nhiệt thành và 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. Tính đến ngày nay đã có 4.847.953 bài viết tiếng Anh,763.384.059 chỉnh sửa, 24.524.565 người dùng đăng ký và 1.358 nhà quản trị (Cập nhật 30-3-2015)

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. Do số lượng bài viết ngày càng tăng, người dùng khó có thể tìm kiếm tài liệu một cách nhanh chóng và phân loại tiêu đề theo mong muốn. Vì thế, một thách thức mới được đặt ra là làm thế nào để nhận biết chủ đề có trong một tài liệu văn bản một cách hiệu quả, mà cụ thể là làm thế nào để máy tính có thể trợ giúp xử lý tự động được chúng.

Nhận biết chủ đề của tài liệu dựa vào các động cơ tìm kiếm là một vấn đề hết sức quan trọng trong việc tra cứu tài liệu hằng ngày của người sử dụng giúp cho người dùng tiết kiệm được nhiều thời gian tìm kiếm tài liệu, giúp người dùng tìm kiếm tài liệu một cách nhanh chóng, chính xác thông tin mình cần.

Ngoài ra, việc nhận biết chủ đề tài liệu dựa trên Wikipedia còn giúp người dùng kiểm soát lượng thông tin mình tìm kiếm, tìm kiếm được những đặc trưng của tài liệu một cách nhanh chóng và chính xác.

Trong những năm gần đây, qui mô và tầm cỡ bách khoa toàn thư trực tuyến miễn phí Wikipedia đã đạt đến tầm mức như một ontology (bản thể luận) và có thể phân loại sử dụng để nhận diện chủ đề có trong một tài liệu văn bản nào đó. Vì vậy đề tài ***"Nhận biết chủ đề của tài liệu dựa trên Wikipedia"*** giúp hỗ trợ người dùng nhận biết được chủ đề tài liệu mình tìm kiếm một cách nhanh chóng và chính xác.

1.2. Mục tiêu, nội dung và phương pháp nghiên cứu

1.2.1. Mục tiêu nghiên cứu

Mục tiêu của việc nhận diện chủ đề của văn bản nhằm để tìm nhãn hoặc phân nhóm, có thể giúp mô tả tốt nhất về vấn đề cốt lõi của văn bản phục vụ cho việc phân loại và xếp nhóm. Tìm ra được những danh mục có trọng số phù hợp với bài báo một cách tự động và đạt độ chính xác cao.

Nhiệm vụ của luận văn:

- Khảo sát, phân tích hệ thống chủ đề của tài liệu dạng văn bản lưu trữ trong Wikipedia

- Khảo sát các nghiên cứu liên quan đến việc nhận biết chủ đề của văn bản trong Wikipedia

- Phát triển (trên cơ sở kế thừa) hoặc cải tiến một phương pháp nhận biết chủ đề tài liệu (dạng văn bản), dựa trên nguồn dữ liệu tên thể loại sẵn có của Wikipedia.

- Thực nghiệm, đánh giá và viết báo cáo.

1.2.2. Nội dung nghiên cứu

Wikipedia bao gồm các bài viết, trang hình ảnh tách biệt, những ý kiến tranh luận về nội dung bài viết, về tác giả, các trang thiết kế mẫu…Mỗi bài viết đều có tiêu đề, xếp theo phân loại và có liên hệ đến các bài viết khác. Một số bài viết có thể truy

xuất với nhiều tiêu đề, trong trường hợp này, các tiêu đề phụ được xem như bài viết đặc biệt hoặc chuyển hướng chỉ gồm một liên kết duy nhất đến bài viết chính. Việc phân loại được tổ chức theo phân tầng theo hạng mục hạ tầng (hạng mục con) và hạng mục thượng tầng (hạng mục cha). Nội dung nghiên cứu của luận văn tập trung vào hai vấn đề cơ bản:

Thứ nhất, luận văn sẽ trình bày một phương pháp chỉ để khai thác tiêu đề bài viết và phân loại các bài viết trong Wikipedia, để quyết định những yếu tố đặc trưng nhất của tài liệu. Trước tiên, chúng ta xem xét tất cả các bài viết liên quan đến tài liệu bằng việc kết hợp tìm tiêu đề và những từ khóa trong tài liệu. Sau đó những bài viết này được xem xét theo ba yếu tố như sau:

- Từ khóa chia sẻ chung giữa tài liệu và tiêu đề, chẳng hạn tần suất hoặc số lượng phân nhóm mà từ khóa đó xuất hiện.
- Cường độ kết hợp giữa tài liệu và bài viết, chẳng hạn số lượng từ khóa phổ biến, tỷ lệ phần trăm tên tiêu đề xuất hiện trong tài liệu.
- Bản thân bài viết, chẳng hạn số lượng bài viết trong Wikipedia có tiêu đề tương tự.

Thứ hai, thu thập việc phân loại gắn liền với bài viết để hình thành nên sự phân loại chủ yếu dựa trên tính chất của bài viết, đồng thời cũng xem xét có bao nhiêu từ có mối liên hệ hỗ tương trong tài liệu và xem xét mức độ mối liên hệ hỗ tương từ khóa chia sẻ thuộc tính với các phân nhóm khác. Luận văn không khai thác sức mạnh tiềm năng của Wikipedia vì luận văn không sử dụng thông tin chứa trong đoạn văn của bài viết, sự liên kết giữa các bài viết, hay sự phân tầng trong phân loại tài liệu. Luận văn chỉ giải quyết hai bài toán lớn sau:

Bài toán 1: Loại bỏ từ dừng và dư thừa, loại bỏ cả những từ không xuất hiện trong tiêu đề của các bài viết. Thu thập các từ trong tài liệu và thu thập tiêu đề trong Wikipedia (ngoại trừ những tiêu đề chỉ có một từ) có xuất hiện trong tài liệu. Tiếp theo là thu thập bài viết trong Wikipedia dẫn kết đến tiêu đề. Cuối cùng là thu thập phân

nhóm trong Wikipedia gắn với tiêu đề.

Bài toán 2: Tinh giảm mức độ các phân nhóm có từ chia sẻ với các phân nhóm khác có trị R_c cao nhất. Sau đó chọn phân nhóm chiếm tỷ lệ cao nhất.

1.2.3. Phương pháp nghiên cứu

- Phương pháp nghiên cứu lý thuyết: nghiên cứu các tài liệu mô tả cách thức làm việc Wikipedia, cách thức phân nhóm của Wikipedia.
- Phương pháp thu thập số liệu: thống kê tổng số phân nhóm trong Wikipedia tiếng anh và tiếng việt tính đến ngày 03/03/2015.
- Phương pháp thực nghiệm: tiến hành phân loại và xếp nhóm đối với số tài liệu để tìm ra những tài liệu thuộc phân loại của Wikipedia.

1.3 Cấu trúc của luận văn

Chương 1. Mở đầu

Trình bày lý do chọn đề tài, mục tiêu nội dung và phương pháp nghiên cứu, cấu trúc của luận văn

Chương 2. Nghiên cứu tổng quan

Phân tích, đánh giá các công trình nghiên cứu đã có của các tác giả trong và ngoài nước liên quan mật thiết đến đề tài; nêu những vấn đề còn tồn tại; chỉ ra những vấn đề mà đề tài cần tập trung nghiên cứu, giải quyết.

Chương 3. Phương pháp nhận biết và rút trích chủ đề

Trình bày các cơ sở lý thuyết, lý luận, giả thuyết khoa học và phương pháp nghiên cứu đã được sử dụng trong Luận văn.

Chương 4. Thực nghiệm và đánh giá

Mô tả công việc nghiên cứu khoa học đã tiến hành, các số thực nghiệm. Đánh giá độ chính xác của thuật toán

Chương 5. Kết luận và hướng phát triển

Kết luận những việc đã đạt được và còn hạn chế của luận văn. Đề xuất hướng phát triển

CHƯƠNG 2. NGHIÊN CỨU TỔNG QUAN

Chương 2 phân tích một số nghiên cứu trong và ngoài nước có liên quan đến luận văn. Mục đích nhằm xác định những ưu điểm hạn chế và các khó khăn của những nghiên cứu có liên quan đến luận văn để từ đó luận văn đưa ra các giải pháp nhằm phát triển hệ thống đạt kết quả. Trong chương này, 2.1 trình bày tóm lược các nghiên cứu trên thế giới, phần 2.2 trình bày tóm lược về các nghiên cứu trong nước có liên quan đến luận văn, trong đó tập trung các nghiên cứu có liên quan đến Wikipedia để nghiên cứu trong luận văn.

2.1 Tình hình nghiên cứu trên thế giới

Trên thế giới, có rất nhiều mô hình phân nhóm chủ đề tài liệu ra đời, một số thì đã được thương mại hóa, số còn lại là xây dựng riêng cho mình một hệ thống phân nhóm chủ đề tài liệu hay chỉ đóng góp một phần nhỏ cho khoa học.

M. Aery, N. Ramamurthy, and Y. A. Aslandogan [11] Nhận diện chủ đề trong văn bản động với mức độ phức tạp cao. Vấn đề phân tích tự động phát hiện dữ liệu văn bản đã phát triển trong vài năm qua. Một ví dụ về dữ liệu đó là các cuộc thảo luận xuất hiện trong dòng chat Internet. Trong nghiên cứu này đề cập đến một phương pháp tách nguồn được giới thiệu gần đây, được gọi là theo dõi mức độ phức tạp, được áp dụng cho các vấn đề tìm kiếm chủ đề trong văn bản động học và được so sánh ngược lại với một số thuật toán tách mù đối với nội dung xem xét. Theo dõi mức độ phức tạp là khái niệm tổng quát của một phép chiếu chuỗi thời gian và nó có thể sử dụng cả hai biện pháp thống kê bậc cao và thông tin phụ thuộc thời gian trong việc tách các chủ đề. Kết quả thực nghiệm trên dữ liệu dòng chat và nhóm tin đã chứng minh rằng chuỗi thời gian tối thiểu đáp ứng các chủ đề có ý nghĩa vốn có trong dữ liệu văn bản động, và cũng cho thấy khả năng ứng dụng của phương pháp để thu hồi từ một văn bản tạm thời thay đổi truy vấn dựa trên dòng.

C.-Y. Lin [15] nhận diện tự động chủ đề dựa trên tri thức. Như là bước đầu tiên trong một thuật toán tổng hợp văn bản động, tác giả đã giới thiệu một phương pháp mới nhằm tự động xác định các ý tưởng trung tâm trong một văn bản dựa trên một khái niệm đếm mô hình tri thức. Để tiện cho việc trình bày, khái quát các khái niệm, tác giả sử dụng khái niệm phân loại theo cấp bậc WordNet bằng cách thiết lập các giá trị cắt phù hợp với các thông số, như khái niệm tổng quát và tần số mối quan hệ cha-con để kiểm soát số lượng và mức độ tổng quát của khái niệm trích xuất từ các văn bản

M. Ruiz-Casado, E. Alfonseca, and P. Castells [20] Tự động khai thác các mối quan hệ ngữ nghĩa cho WordNet bằng bách khoa toàn thư Wikipedia. Tác giả giới thiệu cách tiếp cận nhằm tự động kết hợp các mục từ trong bách khoa toàn thư trực tuyến với các khái niệm trong hệ thống ngữ nghĩa từ vựng. Cách tiếp cận này đã được thử nghiệm với Wikipedia tiếng Anh đơn giản và WordNet, mặc dù nó có thể được sử dụng với các nguồn khác nhau. Độ chính xác trong việc nhận diện lưỡng nghĩa của các mục từ điển bách khoa đạt 91,11% (83,89% cho các từ đa nghĩa). Bước tiếp cận này sẽ được áp dụng để làm phong phú thêm bản thể học với kiến thức bách khoa. Trong bài báo này, tác giả trình bày một thủ tục tự động làm giàu một mạng ngữ nghĩa từ trong hiện tại với thông tin bách khoa toàn thư giúp định nghĩa các khái niệm. Mạng được chọn là WordNet, vì nó hiện đang được sử dụng, ứng dụng trong nhiều lĩnh vực khác nhau, mặc dù các thủ tục nói chung là đủ khái quát hóa để được sử dụng với bản thể học khác. Wikipedia cũng được chọn với phiên bản tiếng Anh đơn. Các cấu trúc cú pháp đơn giản trong tiếng Anh dễ dàng xử lý và phân tích thông tin dễ hơn so với văn bản hoàn toàn không bị giới hạn, từ đó xử lý các định nghĩa được dễ dàng hơn trong tương lai.

M. Ruiz-Casado, E. Alfonseca, and P. Castells [21] Tự động khai thác các mối quan hệ ngữ nghĩa đối với WordNet bằng phương tiện học tập mô hình mẫu từ Wikipedia. Bài viết mô tả một cách tiếp cận tự động xác định mẫu từ vựng mà đại diện cho mối quan hệ ngữ nghĩa giữa các khái niệm, từ một bách khoa toàn thư trực tuyến.

Tiếp theo, các mô hình có thể được áp dụng để mở rộng bản thể hiện có hoặc mạng ngữ nghĩa với mối quan hệ mới. Các thí nghiệm đã được thực hiện với Wikipedia tiếng Anh đơn giản và WordNet 1.7. Một thuật toán mới đã được đặt ra cho các mô hình tự động việc tổng quát từ vựng được tìm thấy trong các mục bách khoa toàn thư. Tác giả đã tìm thấy mô hình chung của các mối quan hệ thượng tầng vị, hạ tầng vị, bộ phận và tổng thể. Tác giả đã rút ra hơn 1200 mối quan hệ mới không xuất hiện trong WordNet ban đầu. Độ chính xác của những mối quan hệ trong khoảng giữa 0,61 và 0,69, tùy thuộc vào mối quan hệ.

B. Stein and S. M. zu Eien [23]. Xác định chủ đề là điều cần thiết để kết nối trong phân loại các ứng dụng tìm kiếm, trong đó bộ tài liệu được cung cấp và những mô tả ý nghĩa đối với mỗi loại được xây dựng. Những đóng góp của bài viết này gồm 3 nội dung. (1) Đưa ra một khung chuẩn chính thức xác định chủ đề cùng với đặc tính mong muốn của mình, (2) giới thiệu một hệ thống phân loại cho các thuật toán xác định chủ đề và đề xuất các thuật toán tương ứng của các công cụ tìm kiếm, (3) đề xuất một cách tiếp cận để xác định chủ đề, dựa vào kiến thức phân loại các bản thể hiện có.

S. Tiun, R. Abdullah, and T. E. Kong [24]. Bài viết này đề xuất một phương pháp sử dụng hệ thống phân cấp bản thể trong xác định chủ đề tự động. Ý tưởng cơ bản của cách tiếp cận này là khai thác một cấu trúc phân cấp bản thể để tìm một chủ đề của một văn bản. Các từ khóa được trích xuất từ một văn bản sẽ được ánh xạ vào các khái niệm tương ứng của phân cấp trong bản thể học. Bằng cách tối ưu các khái niệm tương ứng, chúng tôi sẽ chọn một điểm nút duy nhất trong số các nút khái niệm mà chúng tôi tin là chủ đề của nghiên cứu này. Tuy nhiên, từ vựng hạn chế là vấn đề gặp phải khi lập bản đồ các từ khóa vào các khái niệm tương ứng của phân cấp bản thể. Tình trạng này buộc chúng ta phải mở rộng bản thể học để làm phong phú mỗi khái niệm những khái niệm mới bằng cách sử dụng ngôn ngữ bên ngoài kiến thức cơ bản (WordNet). Sử dụng từ khóa ánh xạ lên các khái niệm bản thể là kỹ thuật xác định chủ đề mà chúng tôi tin rằng là phương cách thực hiện hiệu quả nhất

Tuoi T. Phan, Chau Q. Nguyen [27] đề xuất một giải pháp trích xuất cụm từ khóa trong văn bản tiếng Việt trong đó khai thác từ điển bách khóa Wikipedia tiếng Việt và khai thác những đặc tính riêng biệt của tiếng Việt trong giai đoạn chọn lựa từ khóa để trích xuất. Bài báo cũng tìm hiểu kỹ thuật xử lý ngôn ngữ tự nhiên tiếng Việt đề xuất để phân tích văn bản tiếng Việt, tập trung gắn thẻ vào các cụm từ, cũng như loại từ. Cuối cùng, xem xét kết quả thử nghiệm để kiểm tra sự tác động của chiến lược đã chọn trong việc trích xuất cụm từ khóa tiếng Việt.

2.2 Tình hình nghiên cứu trong nước

Các nghiên cứu liên quan:

Đinh Quang Định [2] đưa ra được cái nhìn khái quát việc triển khai mô hình Web3.0 trên thế giới đồng thời đánh giá hiện trạng việc sử dụng Web 2.0 trong nước từ đó đưa ra lộ trình thực hiện áp dụng công nghệ Web 3.0 tại Việt Nam.

Nguyễn Đình Bình [5] Nghiên cứu khai phá dữ liệu Web và ứng dụng tìm kiếm trích chọn thông tin theo chủ đề. Mục đích của đề tài là nghiên cứu áp dụng tìm kiếm và trích chọn mẫu mới, hữu ích, hiểu được, tiềm ẩn trong Web. Những thông tin theo chủ đề nhanh, chính xác và đầy đủ, thông tin tiềm ẩn bên trong nội dung trang Web đó và những thông tin quan trọng hay những luồng thông tin tốt nhất trên trang Web tìm kiếm trả về kết quả phù hợp với yêu cầu người dùng. Tác giả trích chọn thông tin dựa trên mô hình phân cụm, gán nhãn, CRFs, mô hình Latent Dirichlet Allocation (LDA) và thuật toán Viterbi. Tác giả khai phá dữ liệu Web (chủ yếu là kho dữ liệu Google), trích chọn thông tin theo chủ đề, cho ra kết quả rất khả quan về mặt khoa học và mặt thực tiễn, giúp cho người dùng nắm được những chủ đề thời sự nổi bật và có thêm giải pháp hỗ trợ về công tác quản lý.

Nguyễn Thị Hồng Nhung và Nguyễn Thị Tuyết Mai [6] đã xây dựng một hệ thống tìm kiếm thông tin ấn tượng với 3 ngôn ngữ Việt-Anh-Hoa dựa trên từ điển bởi

rất nhiều ưu điểm. Tuy vậy kết quả đạt không cao bởi số lượng các mục từ còn hạn chế (liên quan đến lĩnh vực tin học và bài báo tiếng Hoa) nên việc chuyển ngữ chưa có độ chính xác cao. Hướng phát triển bổ sung một số kho ngữ liệu ở nhiều lĩnh vực khác để khử nhập nhằng, cho hiệu suất cao.

Nguyễn Tiến Thanh [7] Luận văn nghiên cứu về trích chọn quan hệ thực thể trên Wikipedia Tiếng Việt dựa vào cây phân tích cú pháp. Trên cơ sở phân tích ưu và nhược điểm của các phương pháp trích chọn quan hệ, luận văn áp dụng phương pháp trích chọn quan hệ dựa trên đặc trưng để giải quyết bài toán này. Các đặc trưng biểu thị quan hệ được trích chọn dựa trên cây phân tích cú pháp tiếng Việt, sau đó được đưa vào bộ phân lớp SVM tìm được loại quan hệ tương ứng, từ đó trích chọn được các thể hiện của quan hệ. Hơn nữa, nhằm giảm công sức cho giai đoạn xây dựng tập dữ liệu học, luận văn khai thác tính giàu cấu trúc của dữ liệu trên Wikipedia tiếng Việt để xây dựng tập dữ liệu học bán tự động.

Trần Ngọc Phúc [8] Phân loại nội dung tài liệu Web. Luận văn đã trình bày một số thuật toán phân lớp tiêu biểu và đưa ra hướng thực nghiệm cho hệ thống phân lớp. Luận văn áp dụng phân tích chủ đề ẩn cụ thể là thuật toán Latent Dirichlet Allocation để xác định chủ đề phục vụ cho việc tiến hành phân lớp.

2.3 Tóm lược

Có nhiều phương pháp tiếp cận trong việc nhận diện chủ đề bằng việc sử dụng nguồn dữ liệu đã có sẵn theo phương cách nhân thể luận và phân loại để định danh một vài ví dụ mẫu: so sánh từ khóa quan trọng của một tài liệu với tiêu đề thư mục của Yahoo [24]; tìm kiếm khái niệm từ WordNet trong văn bản và ước lượng tầm quan trọng dựa trên tần suất hoặc khái niệm liên quan xuất hiện [15]; so sánh mẫu ngôn ngữ của tài liệu với mẫu ngôn ngữ của Yahoo, Google [11]; tìm khái nhiệm WordNet gần giống với tài liệu, đo lường tính tương đồng qua từ ngữ trung gian [20]; xác định điểm nút đối với việc xếp nhóm tài liệu [23]. Mặc dù một số phương pháp như đã nêu, và

phương pháp đặc thù [15] và [24] khá giống với phương pháp của luận văn, nhưng phương pháp tính toán việc phân bổ tài liệu và xử lý cấu trúc theo nhân thể luận có sự khác biệt rõ rệt.

CHƯƠNG 3. PHƯƠNG PHÁP NHẬN BIẾT VÀ RÚT TRÍCH CHỦ ĐỀ

Trong chương 3, ở phần 3.1 tác giả trình bày cái nhìn tổng quát về Wikipedia và cách thức hoạt động của Wikipedia. Phân biệt một số khái niệm về thực thể, mục phân loại và thể loại. Ở phần 3.2 tác giả đưa ra phương pháp nghiên cứu đề nghị của luận văn và cách giải quyết. Phần cuối cùng 3.3 sẽ trình bày về một số cải thiện của phương pháp đề xuất

3.1 Khái niệm về Wikipedia

Wikipedia gọi tắt là Wiki (phát âm như "Uy-ki"; từ tiếng Hawaii wikiwiki, có nghĩa "nhanh"; cũng được gọi là công trình mở), là một loại ứng dụng xây dựng và quản lý các trang thông tin do nhiều người cùng phát triển được đưa ra vào năm 2001 bởi Jimmy Wales và Larry Sanger [28]. Wiki được xây dựng theo nguyên tắc phân tán: Ai cũng có thể chỉnh sửa, thêm mới, bổ sung thông tin lên các trang tin và không ghi lại dấu ấn là ai đã cung cấp thông tin đó. Đây được xem là một "Bách khoa toàn thư" – bộ tra cứu lớn nhất và phổ biến nhất trên Internet hiện nay. Wikipedia tiếng Việt được thành lập vào tháng 10 năm 2003. Tính đến ngày 12.3.2015 đã có 1.113.602 bài, với 3.029.046 trang tất cả

Wikipedia - trí tuệ đám đông

Nguồn: *i.a.cnn.net/*

Nhờ đặc trưng biểu diễn thông tin rất giàu ngữ nghĩa được thể hiện ở các mẫu định dạng dữ liệu, các liên kết giữa các thực thể trang Wiki và cách phân mục các trang Wiki mà Wikipedia trở thành một đối tượng được quan tâm đặc biệt trong lĩnh vực khai phá dữ

liệu và xử lý ngôn ngữ tự nhiên. Các lĩnh vực trong Wiki là:

Khoa học tự nhiên

- Địa chất học
- Địa lý học
- Hóa học
- Khoa học máy tính
- Logic
- Sinh học
- Thiên văn học
- Toán học
- Vật lý học
- Y học

Khoa học xã hội

- Chính trị học
- Giáo dục
- Kinh tế học
- Lịch sử
- Luật pháp

- Ngôn ngữ học
- Nhân chủng học
- Tâm lý học
- Thần học
- Triết học
- Xã hội học

Kỹ thuật

- Công nghiệp
- Cơ học
- Điện tử học
- Giao thông
- Kiến trúc
- Năng lượng
- Người máy
- Nông nghiệp
- Quân sự
- Y tế

Văn hóa

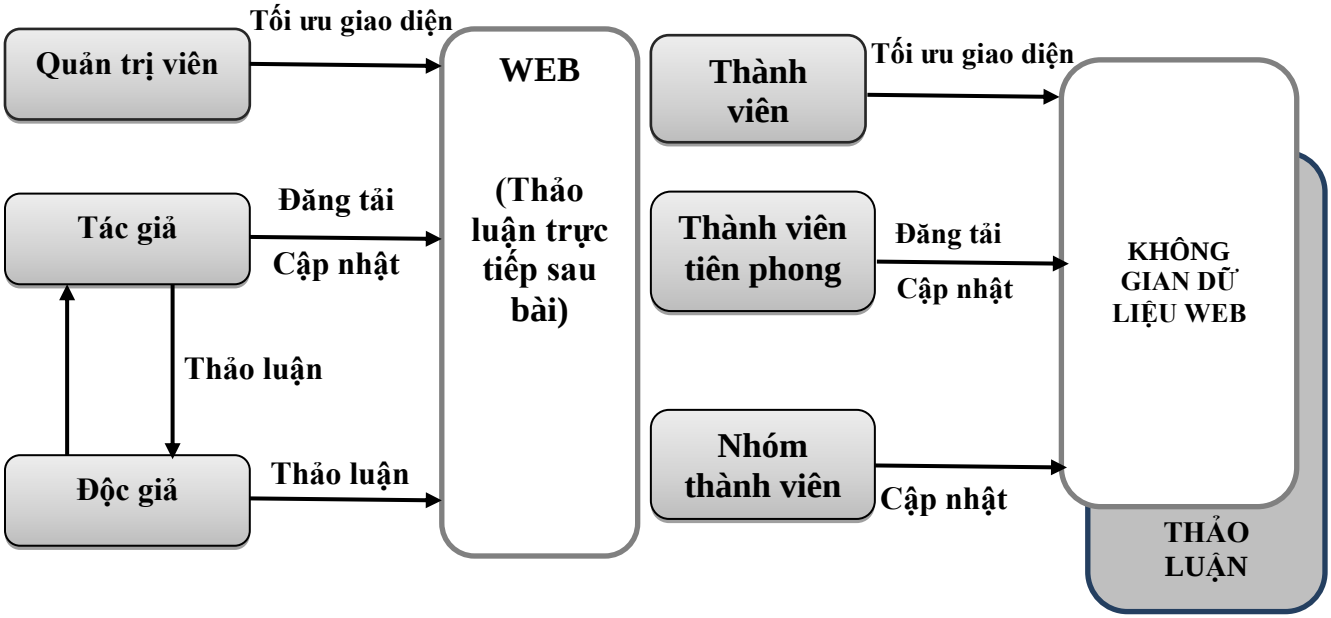
- Âm nhạc
- Chính trị
- Du lịch
- Điện ảnh
- Giải trí
- Khiêu vũ
- Nghệ thuật
- Phong tục tập quán
- Thần thoại
- Thể thao
- Thời trang
- Tôn giáo
- Văn học

3.1.1 Những ưu điểm của mô hình Web Wiki

Dễ dàng tìm hiểu và trình bày bài viết bằng mã wiki. Tất cả các thành viên tham gia đều có khả năng đóng góp vào các dự án bằng nhiều cách thức đa dạng, phù hợp với năng lực của từng người như sửa đổi, bổ sung, viết mới, tải lên, chữa lỗi chính,…

Nguyên tắc hoạt động của nó dựa vào mô hình mở cả về nội dung và mã nguồn đối với mọi thành viên. Wiki là mô hình bình đẳng về cộng đồng: mở về nội dung, đồng

cấp về quyền hạn sử dụng, không phân biệt giữa thành viên, khách và cả người quản lý



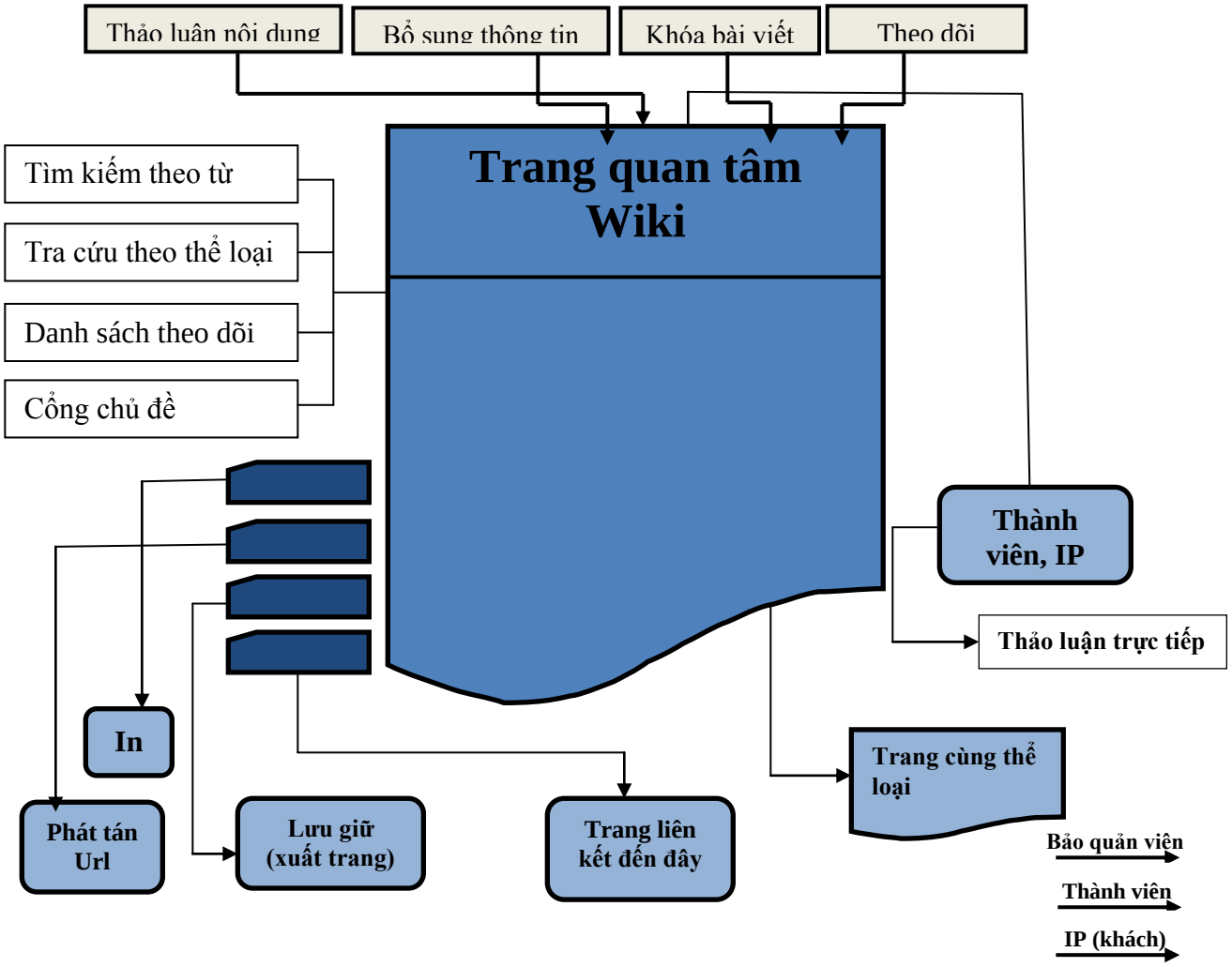
Hình 3.1 Mô hình Web Wiki

Nguồn: https://voer.edu.vn/m/nhung-uu-diem-cua-mo-hinh-web-wiki/40d9cfad

Các thành viên web Wiki đóng góp theo năng lực, đồng thuận, tôn trọng lẫn nhau và không công kích cá nhân là những nền tảng của web Wiki.

Giao diện của Wiki hướng nội dung hơn: nội dung chính được đặt vào trọng tâm của trang, phần thảo luận được tách biệt thành một trang đi kèm giúp người đọc tiếp cận thông tin trực tiếp, không bị nhiễu.

Mọi phiên bản theo thời gian của trang viết đều được lưu lại tách rời và có khả năng phục hồi.



Hình 3.2 Giao diện của Wiki

Nguồn: https://voer.edu.vn/m/nhung-uu-diem-cua-mo-hinh-web-wiki/40d9cfad

Với các dự án toàn cầu đa ngôn ngữ, web Wiki cho phép khả năng tham chiếu nội dung bài viết trong các phiên bản thuộc ngôn ngữ khác nhờ kết nối Interwiki. Độc giả biết nhiều ngoại ngữ có thể tham chiếu cùng một đề mục được nhìn nhận như thế nào về nội dung và hình thức ở mỗi cộng đồng ngôn ngữ.

Bên cạnh đó, Wiki cũng có những nhược điểm của nó. Wiki là mô hình hợp tác xã hội mở, bất kỳ ai cũng có thể sửa đổi và viết bài trên Wiki nên không khó tránh khỏi

những hành vi phá hoại, đưa thông tin quảng cáo, vu khống hoặc đùa cợt. Tinh thần tự nguyện là nền tảng thu hút mọi người tham gia web Wiki nhưng cũng là khó khăn trong hoạt động bảo quản (đảm bảo chất lượng, hình thức bài viết cũng như khắc phục hành vi phá hoại bài viết) và đề ra những nguyên tắc, quy định phát sinh trong quá trình phát triển ở từng dự án.

3.1.2 Wikipedia hoạt động như thế nào

Một yếu tố được người sáng lập Wales thấm nhuần là nguyên tắc tự quản trị và tôn trọng người khác. Wikipedia còn có tính minh bạch, ai cũng có thể xem và nhận xét lời biên tập của bất kỳ ai

Tuy nhiên, cơ sở thực tế của quản trị Wikipedia là tập hợp các chính sách và những hướng dẫn đã được xây dựng qua nhiều năm để xác định mọi thứ, từ các tiêu chuẩn đánh giá bài viết cho đến các quy ước xung quanh việc tranh luận. Điều này thật sự cho thấy các Wikipedia dựa nhiều vào các nền tảng này ra sao - đó thực sự là các nền tảng mà Wikipedia sử dụng.

3.1.3 Kiểu cách và định dạng

Wikipedia không có giới hạn thực sự nào cho số lượng chủ đề mà Wikipedia có thể bao phủ, cũng không giới hạn về lượng nội dung chứa đựng, ngoài việc chúng cần phải kiểm chứng được cùng những điểm được ghi tại trang này.

Bài viết có độ lớn vừa phải là một điều quan trọng giúp Wikipedia dễ truy cập, đặc biệt khi người đọc kết nối bằng quay số hoặc trình duyệt di động vì nó ảnh hưởng trực tiếp đến thời gian tải trang về. Sau khi kết thúc một vấn đề, tách bài viết thành các bài viết rời nhau và để lại một tóm tắt vừa phải là một cách phát triển chủ đề rất tự nhiên. Ngoài ra, Wikipedia có thể đưa vào nhiều thông tin hơn, cung cấp thêm các liên kết ngoài, cập nhật chúng nhanh chóng hơn, và nhiều điều khác nữa.

3.1.4 Thực thể trong Wikipedia

Trên Wiki, một thực thể thường được liên kết tới một trang Wiki mô tả thực thể đó (đôi khi được gọi là thực thể trang Wiki) theo cách: khi một thực thể được tạo ra trên wiki, tác giả tạo ra một liên kết giữa thực thể và trang web Wiki mô tả thực thể đó, đồng thời, với mỗi thực thể xuất hiện trong trang Wiki này, liên kết tới trang Wiki mô tả thực thể đó cũng tạo ra. Đây là một đặc trưng quan trọng của Wiki cho phép dễ dàng xác định các thực thể. Ví dụ sau được trích ra từ trang "Trường Đại học Công nghệ Thành phố Hồ Chí Minh - HUTECH" trên Wiki , bao gồm các liên kết tới thực thể "trường đại học", "Bộ Giáo dục và Đào tạo", "Thủ tướng Chính phủ", "Bình Thạnh", "đại học tự chủ tài chính",…

Trường Đại học Công nghệ Thành phố Hồ Chí Minh - HUTECH (tiền thân là Trường Đại học Kỹ thuật Công nghệ Thành phố Hồ Chí Minh) là một trường đại học trực thuộc Bộ Giáo dục và Đào tạo. Trường được thành lập ngày 26 tháng 4 năm 1995 theo quyết định 235/TTg của Thủ tướng Chính phủ. Trường có trụ sở tại 475A (số cũ144/24) đường Điện Biên Phủ, Phường 25, quận Bình Thạnh, trường hiện hoạt động theo quy chế đại học tự chủ tài chính.

Hình 3.3 Thực thể trong Wikipedia

3.1.5 Mục phân loại

Wikipedia cũng cung cấp các mục phân loại, cho phép các tác giả phân nhóm và tạo các liên kết từ các trang tới các mục phân loại tương ứng. Một trang có thể liên kết tới nhiều mục. Một mục trên Wikipedia có một tên duy nhất. Một mục mới có thể được tạo ra bởi một tác giả tuân theo những khuyến cáo của Wiki trong việc tạo một mục mới và liên kết các trang tới nó. Một vài thuộc tính quan trọng của mục trên Wikipedia gồm có:

- Một mục có thể có nhiều mục con và nhiều mục cha
- Một mục có thể có chứa rất nhiều trang nhưng cũng có những mục chỉ có một lượng nhỏ các trang.

- Một trang mà thuộc về mục mở rộng thường không thuộc về các mục cha của mục mở rộng đó. Ví dụ trang Spain không thuộc mục “Người châu Âu”
- Quan hệ “mục con của một mục” không phải luôn luôn là quan hệ cha con.
- Ví dụ “Bản đồ Châu Âu” là mục con của mục “Châu Âu” nhưng hai mục này không có quan hệ is-a
- Có chu trình trong đồ thị biểu diễn các mục.

3.1.6 Thể loại

3.1.6.1 Thể loại là gì?

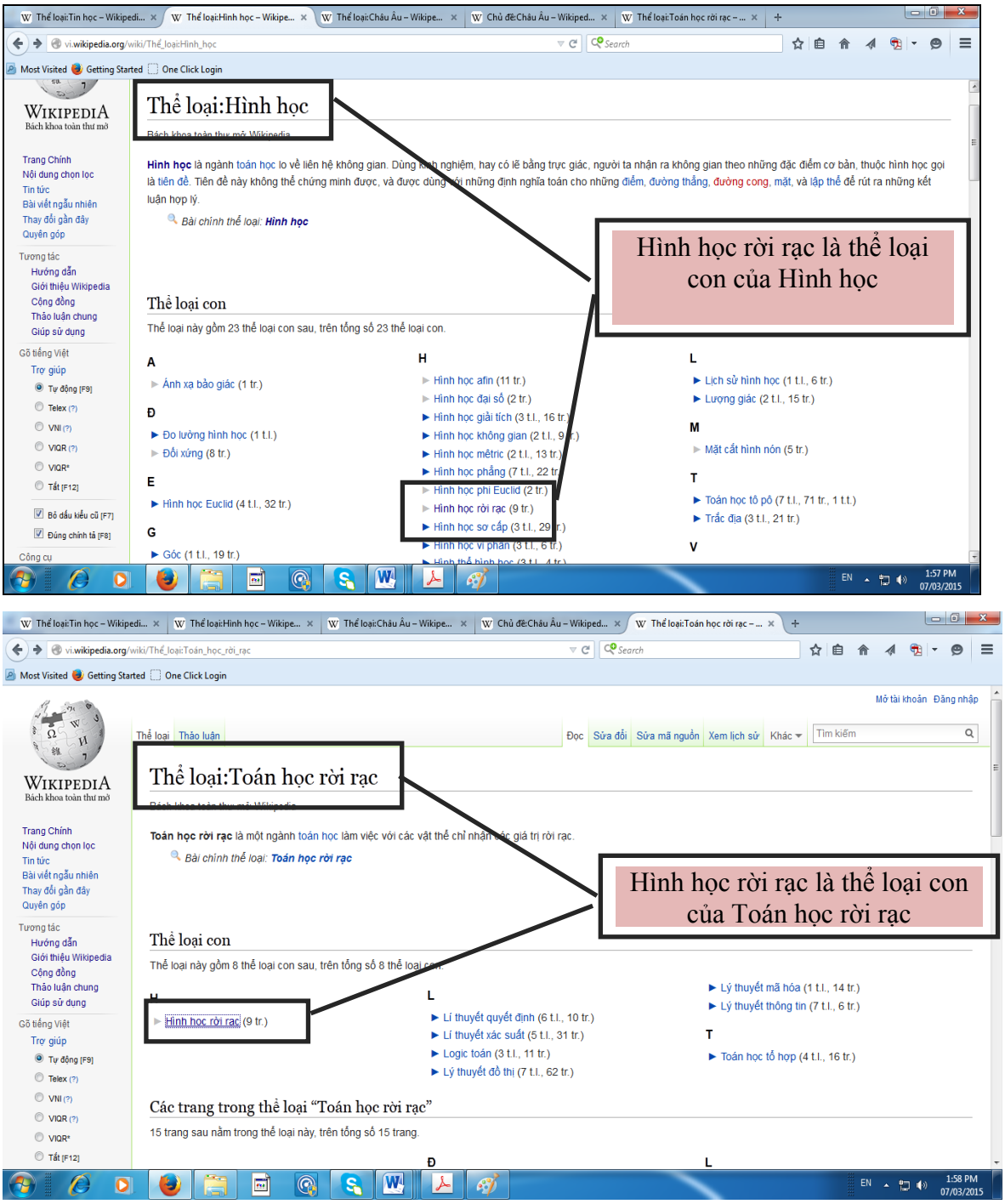
Thể loại là các trang có tên bắt đầu bằng chữ "Thể loại:" (còn gọi là nằm trong không gian tên Thể loại). Ví dụ Thể loại: Tin học.

Các thể loại chứa trong nó liên kết đến các bài viết hay hình ảnh đã được xếp vào thể loại đó. Nó cũng chứa các thể loại con của nó (còn gọi là tiểu thể loại), nếu có.

Một bài viết đã được xếp thể loại thì khi đọc sẽ thấy bên dưới liên kết đến thể loại chứa nó. Ví dụ trang này đã được xếp vào thể loại Thể loại: Tin học, bạn có thể thấy liên kết đến thể loại đó ở dưới cùng trang này. Khi ấn vào liên kết chúng ta sẽ được dẫn đến trang thể loại tương ứng.

Bản thân các thể loại cũng có thể được xếp loại vào thể loại lớn hơn. Tất cả những thể loại đều là thể loại con của một trong những thể loại được liệt kê tại Trang Chính. Cũng như bài viết, thể loại cũng có thể có các liên kết giữa ngôn ngữ...

Một loại trang của vùng tên miền không gian Category, nơi mà các bài có mục đề liên quan được liệt kê. Một bài có thể được xếp vào nhiều thể loại (thí dụ, Hình học rời rạc có thể thuộc cả Thể loại: Toán rời rạc và Thể loại: Hình học). Một thể loại cũng có thể thuộc một hay nhiều thể loại lớn hơn; thể loại lớn là "thể loại cha" và thể loại nhỏ là "thể loại con" (thí dụ, Thể loại: Hình học đại số và Thể loại: Hình học giải tích đều nằm trong Thể loại: Hình học).



Hình 3.4 Thể loại trong Wikipedia

Mọi bài viết hay mọi chủ đề, do cấu trúc Wiki có thể sắp xếp theo nhiều hướng phân loại. Mỗi một bài viết có thể thuộc vào nhiều thể loại tùy theo nội dung và có thể dễ dàng dịch chuyển hay điều cách phân loại theo mô hình cấu trúc "đa gốc, phân nhánh, liên kết đan nhau" bởi những người tham gia viết bài hay bởi sysop, qua đó người đọc có thể liên hệ được nội dung bài viết với bất kỳ khía cạnh liên quan nào với các bài viết khác hay cũng có thể truy nguyên đến các chủ đề xuất phát gốc của bài viết. Ngoài ra với cấu trúc sắp xếp hợp lý, người tham khảo còn có thể thấy được vị trí và vai trò của đề tài so với sơ đồ hình tổng quan tương đối của tổng thể.

Các chủ đề hay bài viết đều có thể dễ dàng tìm thấy nhờ vào máy truy tìm dữ liệu sẵn có trên hệ thống Wiki (search engine build-in), độc giả còn có thể tìm ra bài viết theo các hệ thống phân loại cổ điển. Nhiều bài viết tương cận và liên hệ đến cùng một chủ đề cũng có thể tìm ra cùng một lúc nếu biết sử dụng bộ từ khoá hợp lý bằng Việt ngữ qua đó có thể thấy được đề tài mình muốn trong tầm nhìn rộng hơn. Điều này giúp những người học tập hay nghiên cứu chưa đủ trình độ ngoại ngữ được tiếp cận kiến thức mà không bị trở ngại do ngoại ngữ.

3.1.6.2 Cách sắp xếp thể loại

Việc xếp các bài mới viết vào các thể loại rất có ích. Giúp người đọc tra cứu dễ dàng theo chuyên ngành và phân ngành. Giúp bài viết mới được quảng bá nhanh hơn khi được xếp vào thể loại chứa các bài liên quan. Do đó những người soạn bài nên chú ý xếp công trình của mình vào thể loại tương ứng.

Khi xếp bài vào thể loại, chúng ta cố gắng đưa chúng vào các thể loại chi tiết nhất có thể. Đừng để ở thể loại chung chung quá. Điều này có ích vì nó sẽ giúp các thể loại lớn không bị đầy tràn, gây khó khăn cho tra cứu.

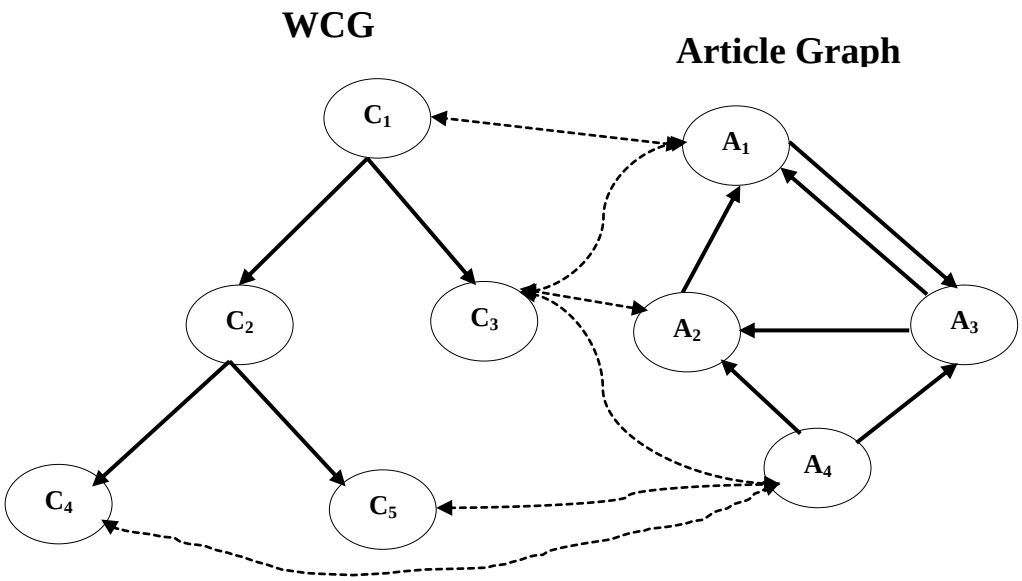
Ví dụ, nếu bạn mới viết bài Bộ nhớ RAM, đừng xếp nó vào Thể loại: Tin học, hãy thử xếp vào thể loại chi tiết hơn như Thể loại: Tin học đại cương; hay chi tiết hơn nữa như Thể loại: Phần cứng máy tính (một thể loại con của Thể loại: Tin học đại

cương). Khi chúng ta sắp xếp bài viết vào những thể loại con như thế sẽ giúp cho người dùng hay khách dễ tìm kiếm đến những bài báo mà mình cần tìm kiếm.

3.1.7 Kiến trúc Wikipedia

Các trang thông tin của Wikipedia được lưu trữ trong một cấu trục mạng. Chi tiết hơn, các bài viết của Wikipedia được tổ chức dạng một mạng các khái niệm liên quan với nhau về mặt ngữ nghĩa và các mục chủ đề (category) được tổ chức trong một cấu trúc phân cấp (taxonomy) được gọi là đồ thị chủ đề Wikipedia (Wikipedia Category Graph - WCG).

Đồ thị bài viết (Article graph): Giữa các bài viết của Wikipedia có các siêu liên kết với nhau, các siêu liên kết này được tạo ra do quá trình chỉnh sửa bài viết của người sử dụng. Nếu ta coi mỗi bài viết như là một nút và các liên kết từ một bài viết đến các bài viết khác là các cạnh có hướng chạy từ một nút đến các nút khác thì ta sẽ có một đồ thị có hướng các bài viết trên Wikipedia (phía bên phải của hình 3.5).



Hình 3.5 Mối quan hệ giữa đồ thị bài viết và đồ thị chủ đề Wikipedia

Đồ thị chủ đề (Category graph): Các chủ đề của Wikipedia được tổ chức giống như cấu trúc của một taxonomy (phía bên trái của hình 3.5). Mỗi một chủ đề có thể có một số lượng tùy ý các chủ đề con.

3.2 Phương pháp nghiên cứu đề nghị

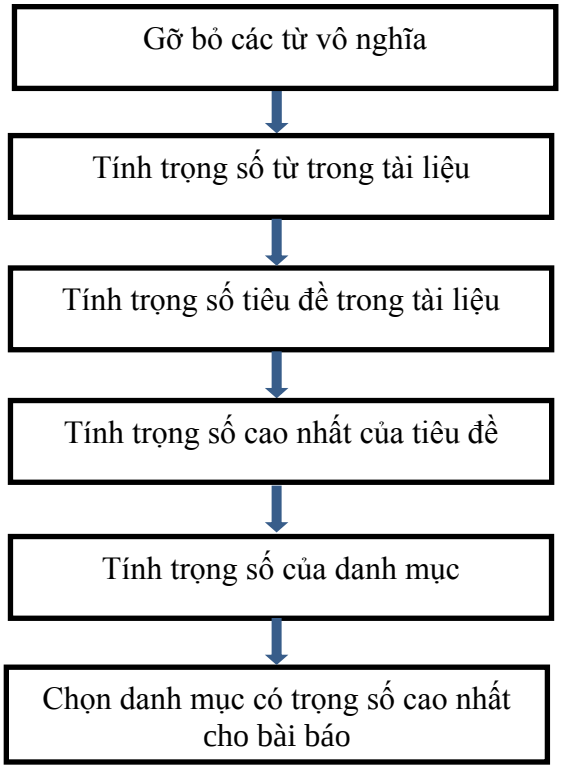
Mục tiêu của phương pháp này là tìm ra những thuộc tính đặc trưng nhất trong phân loại của Wikipedia đối với tài liệu tìm kiếm. Để đạt được điều này, chúng ta phải tiến hành thu thập tất cả phân loại nhóm của Wikipedia qua từ khóa hiển thị trong tài liệu, sau đó xác định phân nhóm nào của Wikipedia có thuộc tính đặc trưng nhất giữa các bài viết đó.

3.2.1 Hướng nghiên cứu chính của luận văn

Luận văn cần giải quyết hai bài toán sau

Bài toán 1: Loại bỏ từ dừng và dư thừa, loại bỏ cả những từ không xuất hiện trong tiêu đề của các bài viết. Thu thập các từ trong tài liệu và thu thập tiêu đề trong Wikipedia (ngoài trừ những tiêu đề chỉ có một từ) có xuất hiện trong tài liệu. Tiếp theo là thu thập bài viết trong Wikipedia dẫn kết đến tiêu đề. Cuối cùng là thu thập phân nhóm trong Wikipedia gắn với tiêu đề

Bài toán 2: Tinh giảm mức độ các phân nhóm có từ chia sẻ với các phân nhóm khác có trị R_c cao nhất. Sau đó chọn phân nhóm chiếm tỷ lệ cao nhất.



Hình 3.6 Sơ đồ thực nghiệm

3.2.2 Việc chuẩn bị thu thập

Wikipedia ở dạng nguyên mẫu bao gồm những tập hợp các trang siêu văn bản dạng HTML được cung cấp bởi máy chủ Wikipedia, hoặc những tập tin XML có thể tải được, hoặc những tập tin chứa các trang được Wiki đánh dấu trang. Khi tải những tập tin này về, chúng ta cần có những bước chuẩn bị như sau:

Để dễ dàng nhận ra phân nhóm Wikipedia trong tài liệu, chúng ta cần phải loại bỏ những từ dừng và tiêu đề gắn kết với bài viết. Như thế, có khả năng xảy hai hay nhiều hơn tiêu đề sẽ dẫn đến những bài viết khác nhau được dò tìm trên cơ sở chuỗi từ tương tự. Trong trường hợp này, tiêu đề được thống nhất và đối tượng mới sẽ dẫn đến tất cả bài viết. Cuối cùng, chỉ mục từ khóa được thực hiện dựa trên tiêu đề.

Lưu ý rằng một từ có thể liên kết đến nhiều bài viết khác nhau, tương tự, một tiêu đề có thể dẫn kết đến nhiều bài viết khác nhau, và cũng có thể nhiều từ khác hoặc nhiều tiêu đề dẫn kết đến cùng một bài viết.

3.2.3 Nhận diện chủ đề của tài liệu

Để chuẩn bị cho quá trình nhận diện chủ đề của tài liệu, các công thức được tham khảo từ [9] Peter Sch¨onhofen. Identifying document topics using the Wikipedia category network. Computer and Automation Research Institute Hungarian Academy of Sciences Kende u. 13–17, H-1111 Budapest.

Sau khi đã chuẩn bị cơ chế lọc Wikipedia, mọi thứ đã sẵn sàng cho tiến trình lọc chúng ta tiến hành tiếp các bước như sau

Bài toán 1: Loại bỏ từ dừng và dư thừa, loại bỏ cả những từ không xuất hiện trong tiêu đề của các bài viết. Thu thập các từ trong tài liệu và thu thập tiêu đề trong Wikipedia (ngoài trừ những tiêu đề chỉ có một từ) có xuất hiện trong tài liệu. Tiếp theo là thu thập bài viết trong Wikipedia dẫn kết đến tiêu đề. Cuối cùng là thu thập phân nhóm trong Wikipedia gắn với tiêu đề.

Bài toán 2: Tinh giảm mức độ các phân nhóm có từ chia sẻ với các phân nhóm khác có trị R_c cao nhất. Sau đó chọn phân nhóm chiếm tỷ lệ cao nhất.

- Loại bỏ từ dừng và dư thừa, loại bỏ cả những từ không xuất hiện trong tiêu đề của các bài viết.
- Thu thập các từ trong tài liệu và xem xét chúng theo công thức

$$R_\omega = tf_\omega \times \log \frac{N}{cf_\omega}$$

- Thu thập tiêu đề trong Wikipedia (ngoài trừ những tiêu đề chỉ có một từ) có xuất hiện trong tài liệu theo công thức

$$R_t = \sum_{\omega \in t} R_\omega \times \frac{1}{t_\omega} \times \frac{1}{a_t} \times \frac{S_t}{L_t}$$

- Thu thập bài viết trong Wikipedia dẫn kết đến tiêu đề và xem xét chúng theo công thức

$$R_a = \max_{t \in a} R_t$$

- Thu thập phân nhóm trong Wikipedia gắn với tiêu đề và xem xét chúng theo công thức

$$R_c = \frac{v_c}{d_c} \times \sum_{a \in c} R_a$$

- Tinh giảm mức độ các phân nhóm có từ chia sẻ với các phân nhóm khác có trị R_c cao nhất.
- Chọn phân nhóm chiếm tỷ lệ cao nhất.

Trước khi xử lý dữ liệu, chúng ta cần phải định nghĩa một vài khái niệm. Phân nhóm **c** được chỉ định cho bài viết **a**, hoặc **c** là một trong phân nhóm chính thức của **a,** và theo cấu trúc Wikipedia, **a** thuộc **c**. Từ **w** chỉ dẫn đến tiêu đề **t,** nếu xảy ra, tiêu đề **t** sẽ dẫn kết đến bài viết **a** nếu nó là một trong những tiêu đề của **a**. Cuối cùng, tập hợp từ xảy ra trong tiêu đề của những bài viết trong phân nhóm **c** sẽ gọi là trường từ vựng của **c**

Bài toán 1: chúng ta tiến hành loại bỏ dừng từ và dư từ trên tài liệu gốc, tương tự như cách chúng ta tiến hành chuẩn bị lọc trên Wikipedia để sắp xếp các trường từ vựng cả hai mặt. Những từ trong tài liệu không xuất hiện trong Wikipedia sẽ được bỏ qua.

Tiếp theo, Theo [9] chúng ta đặt biến số R_ω đối với mỗi từ ω

$$R_\omega = tf_\omega \times \log \frac{N}{cf_\omega} \quad (3.1)$$

Trong đó

R_ω: Trọng số của một từ trong tài liệu.

tf_ω*:* Số lần từ đó xuất hiện trong tài liệu.

N: Số lượng danh mục

cf_ω *:* Trọng số của một từ trong danh mục.

cf_ω tần suất xuất hiện của phân nhóm, tìm ra bao nhiêu phân nhóm chứa từ ω trong trường từ vựng. Yếu tố thứ hai là tần suất phân nhóm nghịch đảo, icf_ω xác định phân nhóm qua trường từ vựng đối với tần suất xuất hiện tài liệu nghịch đảo. Lưu ý đã có vài nghiên cứu đã định nghĩa tần suất phân nhóm nghịch đảo theo những cách khác nhau, chúng đếm phân nhóm gốc, chứ không phải xem xét những từ khóa đã xuất hiện trong phân nhóm Wikipedia.

Trong công thức (3.1), yếu tố đầu tiên nhấn mạnh từ khóa xuất hiện nhiều lần trong tài liệu, được xem là từ trọng yếu trong tài liệu. Yếu tố thứ hai đưa ra sự lựa chọn đối với những từ trong số ít các phân nhóm, vì thế, không nên đưa ra những yếu tố không chắc chắn vào những phân tích sau đó. Chúng ta cũng không sử dụng trị đo lường idf bởi vì mục tiêu của nghiên cứu là xác định phân nhóm mà mô tả tài liệu một cách tốt nhất, chứ không phải những phân nhóm thuận tiện cho việc phân loại, sắp xếp hay những thuật toán truy xuất dữ liệu trên nguồn dữ liệu đã cho.

Tiếp theo, chúng ta thu thập tiêu đề Wikipedia hỗ trợ bằng những từ xuất hiện trong tài liệu. Từ ω có trong tiêu đề t nếu (1) ω xuất hiện trong t, và (2) không thuộc M từ của t, tối thiểu M-1 từ xuất hiện trong tài liệu. Tất nhiên, nếu tiêu đề chỉ gồm một từ, thì điều kiện thứ hai bỏ qua.

Lưu ý trong bước này, chúng ta cho phép từ đơn không gắn liền giữa tiêu đề và tài liệu để xử lý những tài liệu liên quan đến người, nơi chốn, và thuật ngữ kỹ thuật theo cách hợp lý.

Ví dụ, “Boris Yelsin” có thể xuất hiện như “Yelsin”, hay “Paris, France” như “Paris”. Ngoài ra, tiêu đề Wikipedia thường bao gồm những miêu tả phụ nằm trong dấu ngoặc hoặc sau dấu phảy. Những thông tin phụ không cần thiết xuất hiện trong tài liệu, bởi vì nó là bằng chứng từ ngữ cảnh hoặc tài liệu sử dụng từ khác để hình thành nên

một định nghĩa.

Tương tự như từ [9], tiêu đề cũng được xem xét trong công thức:

$$R_t = \sum_{\omega \in t} R_\omega \times \frac{1}{t_\omega} \times \frac{1}{a_t} \times \frac{S_t}{L_t} \quad (3.2)$$

Trong đó

t_ω : Số lượng tiêu đề chứa các từ cần tính

a_t : Số lượng bài báo trỏ đến tiêu đề cần tính

L_t : Kích thước của tiêu đề

S_t : Số lượng từ trong tài liệu được miêu tả trong bài báo

R_t : Trọng số của các tiêu đề trong tài liệu

Mặc dù, yếu tố thứ hai trong công thức (3.2) tiêu đề được ưu tiên hay loại bớt tùy theo mức độ quan trọng từ khóa hỗ trợ. Yếu tố cuối cùng trong công thức đơn giản để đo lường tỷ lệ phần trăm từ tiêu đề xuất hiện trong tài liệu. Lý do chính đáng để củng cố cho các bài viết với tiêu đề dài hơn là xác suất kiểm tra lỗi sẽ thấp hơn.

Mục đích của yếu tố thứ hai và thứ ba trong công thức (3.2) là nhằm tránh trường hợp các từ thông thường dẫn đến nhiều tiêu đề và tiêu đề dẫn đến những bài viết trong quá trình phân tích sau đó. Các chủ đề trong Wikipedia cung cấp phần chi tiết không tương đồng nhau, chẳng hạn chủ đề Album âm nhạc có số lượng bài viết nhiều hơn chủ đề nhiếp ảnh. Tương tự, do ảnh hưởng số lượng "dư từ", có nhiều tiêu đề gắn với số lượng lớn những bài viết khác, chẳng hạn, trong cụm từ "Architecture in X", trong đó X là năm, sẽ gộp thành "Architecture". Bởi vì những bài viết có cùng chủ đề, cũng sẽ ở cùng nhóm phân loại, và không có tác động cân bằng bởi yếu tố thứ ba, những bài viết này có thể bao phủ những khái niệm quan trọng tương đương khác.

Bài toán 2: chúng ta thu thập bài viết dẫn kết đến tiêu đề đã đề cập ở bước

trước. Nếu cùng một bài viết dẫn kết đến những tiêu đề khác nhau do có liên kết chuyển hướng, biến số tối đa. Tham khảo từ [9] Peter Sch¨onhofen. Identifying document topics using the Wikipedia category network. Computer and Automation Research Institute Hungarian Academy of Sciences Kende u. 13–17, H-1111 Budapest, để tính trọng số cao nhất của bài báo và tính trọng số của danh mục ta có:

$$R_a = \max_{t \in a} R_t \qquad (3.3)$$

Trong đó

R_a *:* Là trọng số cao nhất của bài báo trong tài liệu.

Lưu ý chúng ta không bổ sung biến số số tiêu đề đối với một bài viết, phản ánh cấu trúc Wikipedia chứ không không phải tầm quan trọng của bài viết.

Bước tiếp theo, chúng ta sẽ tạo một danh sách các phân nhóm chỉ định cho những bài viết đã thu thập được, và chúng ta xem xét từng phân nhóm với tổng số bài viết liên quan, theo công thức:

$$R_c = \sum_{a \in c} R_a \qquad (3.4)$$

Trong đó:

R_c *:* Trọng số của danh mục

Cuối cùng, đơn giản chúng ta chọn H phân nhóm với biến số cao nhất; và chủ đề phân nhóm này cần được xem xét tính tiêu biểu đặc trưng nhất trong nội dung của tài liệu.

3.3 Một số cải thiện của phương pháp đề xuất

Bằng việc giới thiệu hai phương pháp bổ sung cho phương pháp nghiên cứu này đã được mô tả trong phần trước, chúng ta có thể đạt được độ chính xác cao, phương pháp bổ sung chỉ ảnh hưởng ở bước tính toán biến số phân nhóm R_c. Để dễ giải thích cho những phần sau, chúng ta cần định nghĩa từ hỗ trợ thuộc phân nhóm **c** như tập hợp

từ hỗ trợ bài viết mà dẫn kết đến **c**.

Đối với phương pháp bổ sung thứ nhất, chúng ta cố gắng loại bỏ những phân nhóm có trị R_c cao do trường từ vựng cực kỳ lớn như từ "actors' và "films". Điều này được xem là nỗ lực để tìm ra các yếu tố thứ hai và thứ ba trong công thức (3.2). Phần bổ sung được xem là phần bổ sung cho công thức (3.4). Tham khảo từ [9] Peter Sch¨onhofen. Identifying document topics using the Wikipedia category network. Computer and Automation Research Institute Hungarian Academy of Sciences Kende u. 13–17, H-1111 Budapest, ta có:

$$R_c = \frac{v_c}{d_c} \times \sum_{a \in c} R_a \qquad (3.5)$$

v_c : là số từ hỗ trợ của phân nhóm c

d_c : là số từ trong từ vựng của phân nhóm c.

Với phần bổ sung thứ hai giúp chúng ta loại bỏ được những trường hợp những từ như "consumed" hoặc "accounted for" thuộc nhóm nổi trội lại gắn với những phân nhóm yếu hơn. Chẳng hạn, từ "ban" đã hỗ trợ khái niệm "comprehensive test ban treaty", rõ ràng sẽ mắc sai lầm khi xem xét "ban" trong khái niệm huyền bí học với cùng mức độ.

Phần bổ sung thứ hai giới thiệu một bước phụ sau bước tính R_c, giai đoạn thu thập phân nhóm và tính toán biến số của mỗi phân nhóm. Trước tiên, chúng ta đặt d_w là giá trị suy giảm, khởi đầu bằng 1 cho mỗi từ của tài liệu. Kế tiếp, chúng ta phân loại phân nhóm theo biến số, và xem xét những phân nhóm có biến số cao nhất. Đối với mỗi phân nhóm, chúng ta sẽ tính toán tỷ trọng lần nữa, đồng thời xem xét giá trị suy giảm cho tập hợp từ hỗ trợ B_c theo công thức

$$R'_c = R_c \times \frac{\sum_{\omega \in B_c} d_\omega}{|B_c|} \qquad (3.6)$$

$$d'_\omega = \frac{d_\omega}{2},\ \omega \in B_c \qquad (3.7)$$

Trong đó, R_c được nhân với giá trị suy giảm trung bình của nhóm từ trong phân nhóm c, với giá trị suy giảm chia hai. Nếu không có từ hỗ trợ nào chia sẻ với phân nhóm được thử nghiệm trước đó, giá trị R_c còn nguyên, không biến thiên.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Trong chương 4, tác giả tiến hành thực nghiệm và đánh giá các kết quả đạt được

4.1 Tiến hành thực nghiệm

4.1.1 Môi trường thực nghiệm

Luận văn tiến hành thực nghiệm trên máy cấu hình như sau:

Bảng 4.1 Cấu hình yêu cầu

Thành phần	Chỉ số
CPU	Core I7 2.5 GHz
HDD	500 Ghz
RAM	8Ghz
OS	Window 8.1
Công cụ lập trình	Visual studio 2013
Ngôn ngữ lập trình	C#
Cơ sở dữ liệu	Microsoft SQL sever 2012

4.1.2 Nguồn dữ liệu

Nguồn dữ liệu của luận văn lấy từ Wikipedia bao gồm 2588 bài báo và 150435 danh mục, tất cả dữ liệu được lưu vào tập tin XML sau khi tải về từ [33] tác giả tiến hành giải nén dữ liệu thu được những tập tin về các bài báo trên Wikiapia có dạng như sau:

```
<root>
  <page>
    <title>Anarchism</title>
    <ns>0</ns>
    <id>12</id>
    <revision>
      <id>612787421</id>
      <parentid>612717160</parentid>
      <timestamp>2014-06-13T16:52:37Z</timestamp>
      <contributor>
        <username>Malik Shabazz</username>
        <id>3020778</id>
      </contributor>
      <comment>Reverted [[WP:AGF|good faith]] edits by [[Special:Contributions/66.66.225.114|66.66.225.114]] ([[User talk:66.66.225.114|talk]]): See [[WP:ENGVAR]]. ([[
      <text xml:space="preserve">{{Redirect2|Anarchist|Anarchists|the fictional character|Anarchist (comics)|other uses|Anarchists (disambiguation)}}
{{Use British English|date=January 2014}}
{{pp-move-indef}}
{{Anarchism sidebar}}

'''Anarchism''' is a [[political philosophy]] that advocates [[stateless society|stateless societies]] often defined as [[self-governance|self-governed]] voluntary ins
view that a society without the state, or government, is both possible and desirable.&quot;&lt;/ref&gt;&lt;ref&gt;Sheehan, Sean. Anarchism, London: Reaktion Books Ltd.
d enlarge the precious kernel of social customs without which no human or animal society can exist.&quot; [[Peter Kropotkin]]. [http://www.theanarchistlibrary.org/HTML
{{cite journal |last=Malatesta|first=Errico|title=Towards Anarchism|journal=MAN!|publisher=International Group of San Francisco|location=Los Angeles|oclc=39304443|url=h
{{cite journal |url=http://www.theglobeandmail.com/servlet/story/RTGAM.20070514.wxlanarchist14/BNStory/lifeWork/home/
|archiveurl=http://web.archive.org/web/20070516094548/http://www.theglobeandmail.com/servlet/story/RTGAM.20070514.wxlanarchist14/BNStory/lifeWork/home |archivedate=16
{{cite web |url=http://www.britannica.com/eb/article-9117285|title=Anarchism|year=2006|work=Encyclopædia Britannica|publisher=Encyclopædia Britannica Premium Service|a
{{cite journal |year=2005|title=Anarchism|journal=The Shorter [[Routledge Encyclopedia of Philosophy]]|page=14|quote=Anarchism is the view that a society without the s
```

Cấu trúc của tập tin trên gồm các thẻ như sau:

- <page>: chứa những bài báo riêng biệt.
- <title>: chứa nội dung tiêu đề bài báo.
- <id>: diễn tả mã của mỗi bài báo.
- <Username>: tác giả của bài báo
- <text>: nội dung của bài báo

.....

Đặc biệt trong nội dung của mỗi bài báo phần cuối có các thẻ category để đánh dấu bài báo đó thuộc những danh mục nào như sau:

[[Category:Anarchism]]
[[Category:Political culture]]
[[Category:Political ideologies]]
[[Category:Social theories]]
[[Category:Anti-fascism]]
[[Category:Anti-capitalism]]
[[Category:Far-left politics]]
……..

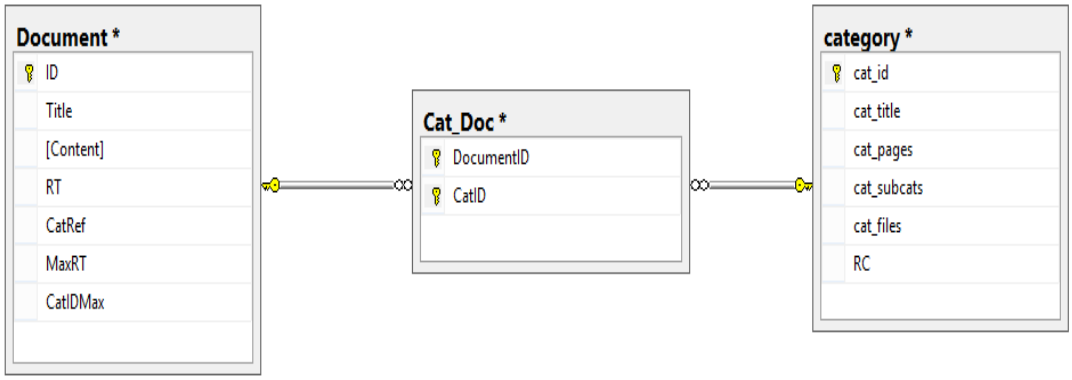
Sau khi đã có được tập tin XML mô tả các bài báo, tác giả tiến hành tải tập tin về các danh mục của Wikipedia. Cấu trúc của tập tin về các danh mục của wikipedia như sau:

```
DROP TABLE IF EXISTS 'category';
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE 'category' (
  'cat_id' int(10) unsigned NOT NULL AUTO_INCREMENT,
  'cat_title' varbinary(255) NOT NULL DEFAULT '',
  'cat_pages' int(11) NOT NULL DEFAULT '0',
  'cat_subcats' int(11) NOT NULL DEFAULT '0',
  'cat_files' int(11) NOT NULL DEFAULT '0',
  PRIMARY KEY ('cat_id'),
  UNIQUE KEY 'cat_title' ('cat_title'),
  KEY 'cat_pages' ('cat_pages')
) ENGINE=InnoDB AUTO_INCREMENT=210429427 DEFAULT CHARSET=binary;
/*!40101 SET character_set_client = @saved_cs_client */;

--
-- Dumping data for table 'category'
--

/*!40000 ALTER TABLE 'category' DISABLE KEYS */;
INSERT INTO 'category' VALUES (127497,'FLCL_images',0,0,0),(127498,'FMA_aircraft',29,0,0),(127499,'FM_Bats_albums',0,0,0),(127500,'FM_Radio',
INSERT INTO 'category' VALUES (151145,'Guildford_Grammar_School',11,1,3),(151146,'Guildhall,_Vermont',2,0,0),(151147,'Guilds',68,9,0),(151148
INSERT INTO 'category' VALUES (175336,'Johannesburg_Region_10',0,0,0),(175337,'Johannesburg_Region_11',0,0,0),(175338,'Johannesburg_Region_2'
INSERT INTO 'category' VALUES (199399,'Members_of_the_Confederate_House_of_Representatives',15,13,0),(199400,'Members_of_the_Congress_of_Colo
INSERT INTO 'category' VALUES (221837,'North_American_election_results_templates',0,0,0),(221838,'North_American_election_years_navigational_
INSERT INTO 'category' VALUES (244435,'People_from_the_Province_of_Ascoli_Piceno',21,2,0),(244436,'People_from_the_Province_of_Asti',22,2,0),
INSERT INTO 'category' VALUES (267749,'Robin_Hood_ballads',38,0,0),(267750,'Robin_Hood_by_medium',8,6,0),(267751,'Robin_Hood_characters',21,2
```

Sau khi đã có được các tập tin trên tác giả tiến hành xây dựng cơ sở dữ liệu và dùng ngôn ngữ lập trình C# để xây dựng chương trình lưu các nội dung trên vào cơ sở dữ liệu.

4.1.3. Cấu trúc cơ sở dữ liệu

Cấu trúc cơ sở dữ liệu của luận văn được tổ chức như sau:



Bảng 4.2 Cấu trúc cơ sở dữ liệu

Tên Bảng	Tên cột	Kiểu dữ liệu	Ý nghĩa
Document	ID	Int	Mã của bài báo
	Title	nvarchar(1000)	Tên bài báo
	Content	text	Nội dung bài báo
	RT	real	Trọng số của tiêu đề bài báo
	CatRef	Nvarchar(max)	Các danh mục mà bài báo đó thuộc vào
	MaxRT	real	Trọng số cao nhất của tiêu đề bài báo
Cat_Doc	DocumentID	Int	Khóa ngoại của mã bài báo
	CatID	Int	Khóa ngoại của danh mục
Category	Cat_id	Int	Mã danh mục
	Cat_title	nvarchar(350)	Tên danh mục
	Cat_pages	Bigint	
	Cat_subcats	bigint	

	Cat_files	bigint	
	RC	real	

4.2 Thực hiện chương trình

4.2.1 Gỡ bỏ các từ vô nghĩa

Sau khi đưa các bài báo và danh mục vào cơ sở dữ liệu tác giả dùng store của SQL để tiến hành gỡ bỏ các từ vô nghĩa từ tài liệu để tăng hiệu năng cho chương trình danh sách các từ vô nghĩa bao gồm các từ sau:

Bảng 4.3 Một số từ vô nghĩa

a	About	Above
After	Again	Against
All	Am	An
And	Any	Are
aren't	As	At
Be	because	Been
Before	Being	Below
Between	Both	But
By	can't	Cannot
Here	Have	Having
Her	he's	he'll
he'd	He	haven't

hasn't	Has	hadn't
Had	further	From
For	Few	Each
During	Down	don't
....	...	...

4.2.2 Tính trọng số của các từ trong tài liệu

Sau khi gỡ bỏ hết các từ vô nghĩa chương trình tiến hành tính trọng số các từ của tiêu đề tài liệu theo công thức sau:

$$R_\omega = tf_\omega \times \log \frac{N}{cf_\omega}$$

Trong đó:

R_ω : Trọng số của một từ trong tài liệu.

tf_ω : Số lần từ đó xuất hiện trong tài liệu.

N: Số lượng danh mục

cf_ω : Trọng số của một từ trong danh mục.

Sau khi áp dụng công thức trên tác giả thu được trọng số của các từ trong tài liệu cho các bước sau như:

R_ω: 6.13325465792414 a_t: 53 S_t : 353 L_t: 2 strTitle: Absolute majority

R_ω: 10.8634889748633 a_t: 103 S_t : 449 L_t: 2 strTitle: Absolute value

R_ω: 1.93183399934809 a_t: 4 S_t : 55 L_t: 5 strTitle: List of Atlas Shrugged characters

……

4.2.3 Tính trọng số tiêu đề của tài liệu

Sau khi có trọng số của tất cả các từ trong tài liệu luận văn tiến hành tính trọng

số của các tiêu đề của tài liệu theo công thức sau:

$$R_t = \sum_{\omega \in t} R_\omega \times \frac{1}{t_\omega} \times \frac{1}{a_t} \times \frac{S_t}{L_t}$$

Trong đó:

t_ω : Số lượng tiêu đề chứa các từ cần tính

a_t : Số lượng bài báo trỏ đến tiêu đề cần tính

L_t : Kích thước của tiêu đề

S_t : Số lượng từ trong tài liệu được miêu tả trong bài báo.

R_t : Trọng số của các tiêu đề trong tài liệu

Sau khi tác giả tiến hành thực hiện bước này tác giả tiến hành lưu kết quả thu được của R_t của công thức trên vào cơ sở dữ liệu cột RT trong bảng Document như sau:

	ID	Title	Content	Redirect	RT
1	46	AbacuS	#REDIRECT [[Abacus]] {{RCamelCase}}	Abacus	3.617074
2	49	AbbesS	#REDIRECT [[Abbess]] {{RCamelCase}}	Abbess	3.035079
3	51	AbbeY	#REDIRECT [[Abbey]] {{RCamelCase}}	Abbey	0.2027313
4	52	AbboT	#REDIRECT [[Abbot]] {{RCamelCase}}	Abbot	0.1128703
5	248	AbeL	#REDIRECT [[CainAbel]] {{RCamelCase}}	Cain and Abel	0.006521846
6	278	AmericA	#REDIRECT [[America]] {{Rambiguous page}} {{RCa...	America	0.0002765126
7	280	AndorrA	#REDIRECT [[Andorra]] {{RCamelCase}}	Andorra	1.636997
8	290	A	{{Twouses|the letterthe alphabet|the English indefinite a...		2.161996E-06
9	303	Alabama	{{about|the U.S. state|Alabama (disambiguation)}} {{pp...		0.1758476
10	304	AfricA	#REDIRECT [[Africa]] {{RCamelCase}}	Africa	0.006205478
11	307	Abraham Lincoln	{{About|the American president}} {{Use mdy dates|date...		0.5559488
12	309	An AmericanP...	{{About|the 1928 George Gershwin music|the 1951 mus...		0.003784668
13	316	Academy Awar...	{{Use mdy dates|date=June 2013}} {{Infobox award | ...		1.836936
14	324	Academy Awar...	{{Redirect2|Oscars|The Oscar|the film|The Oscar (film)|o...		0.2310035
15	325	Action Film	#REDIRECT [[Action film]]{{Rother capitalisation}}	Action film	0.02892464
16	332	Animalia (book)	{{Use dmy dates|date=June 2013}} {{Infobox book|<!-- ...		141.8899
17	334	International At...	{{Refimprove|date=July 2012}} International Atomic Ti...		0.1716782
18	339	Ayn Rand	{{Use mdy dates|date=June 2013}} {{Good article}} {{I...		0.2751858
19	340	Alain Connes	{{Use dmy dates|date=June 2013}} {{More footnotes|d...		88.14375
20	344	Allan Dwan	{{Use dmy dates|date=October 2013}} {{Infobox perso...		5.554792

4.2.4 Tính trọng số cao nhất của tài liệu

Một tài liệu có thể chứa nhiều tiêu đề bao gồm tiêu đề của tài liệu đó và của tài liệu khác cho nên trong bước này tác giả sẽ tiến hành tìm những tiêu đề mà nội dung bài báo đó chứa có trọng số được tính ở bước trên là cao nhất, nếu bài báo đó chỉ chứa một tiêu đề duy nhất thì trọng số của bài báo đó chính là trọng số của tiêu đề được tính ở bước trên.

Tác giả tiến hành tính trọng số cao nhất của bài báo của tài liệu theo công thức sau:

$$R_a = \max_{t \in a} R_t$$

Trong đó:

R_a : Là trọng số cao nhất của bài báo trong tài liệu.

Trong công thức này tác giả tiến hành tìm R_t (tính ở bước trước đó) cao nhất cho tiêu đề của tài liệu và được lưu vào cơ sở dữ liệu cho cột MaxRT trong bảng Document.

Results | Messages

	ID	title	Rt	MaxRT
1	3338	The Bronx	580.936	6713.301

Trong thực nghiệm trên, MaxRT chính là RA

	ID	Title	Content	Redirect	RT	CatRef	MaxRT
1	46	AbacuS	#REDIRECT [[Abacus]] {{RCamelCase}}	Abacus	3.617074	NULL	99.5819
2	49	AbbesS	#REDIRECT [[Abbess]] {{RCamelCase}}	Abbess	3.035079	NULL	1095.99
3	51	AbbeY	#REDIRECT [[Abbey]] {{RCamelCase}}	Abbey	0.2027313	NULL	3347.16
4	52	AbboT	#REDIRECT [[Abbot]] {{RCamelCase}}	Abbot	0.1128703	NULL	3154.84
5	248	AbeL	#REDIRECT [[CainAbel]] {{RCamelCase}}	Cain and Abel	0.006521846	NULL	8282.32
6	278	AmericA	#REDIRECT [[America]] {{Rambiguous page}} {{RCa...	America	0.0002765126	NULL	17881.2
7	280	AndorrA	#REDIRECT [[Andorra]] {{RCamelCase}}	Andorra	1.636997	NULL	54.7886
8	290	A	{{Twouses|the letter|the alphabet|the English indefinite a...		2.161996E-06	,ISO basic Latin letters,Vowel letters	19516.8
9	303	Alabama	{{about|the U.S. state|Alabama (disambiguation)}} {{pp...		0.1758476	,Alabama,Former French colonies,Former Spanish colo...	1336.60
10	304	AfricA	#REDIRECT [[Africa]] {{RCamelCase}}	Africa	0.006205478	NULL	6720.89
11	307	Abraham Lincoln	{{About|the American president}} {{Use mdy dates|date...		0.5559488	,Abraham Lincoln,1809 births,1865 deaths,American p...	1193.83
12	309	An AmericanParis	{{About|the 1928 George Gershwin music|the 1951 mus...		0.003784668	,Compositions by George Gershwin,Symphonic poems,...	64.6563
13	316	Academy AwardBest Production Design	{{Use mdy dates|date=June 2013}} {{Infobox award | ...		1.836936	,Academy AwardsArt Direction,Best Art Direction Acad...	6720.89
14	324	Academy Awards	{{Redirect2|Oscars|The Oscar|the film|The Oscar (film)|o...		0.2310035	,Academy Awards,American film awards,Awards establ...	17881.2

4.2.5 Tính trọng số của danh mục

Trong bước này tác giả tiến hành tính tổng các trọng số của các bài báo trong một danh mục cho trọng số của mỗi danh mục

Tác giả tiến hành tính trọng số của danh mục theo công thức sau:

$$R_c = \sum_{a \in c} R_a$$

Trong đó: R_c : Trọng số của danh mục

4.2.6 Chọn danh mục phù hợp cho bài báo với trọng số của chúng

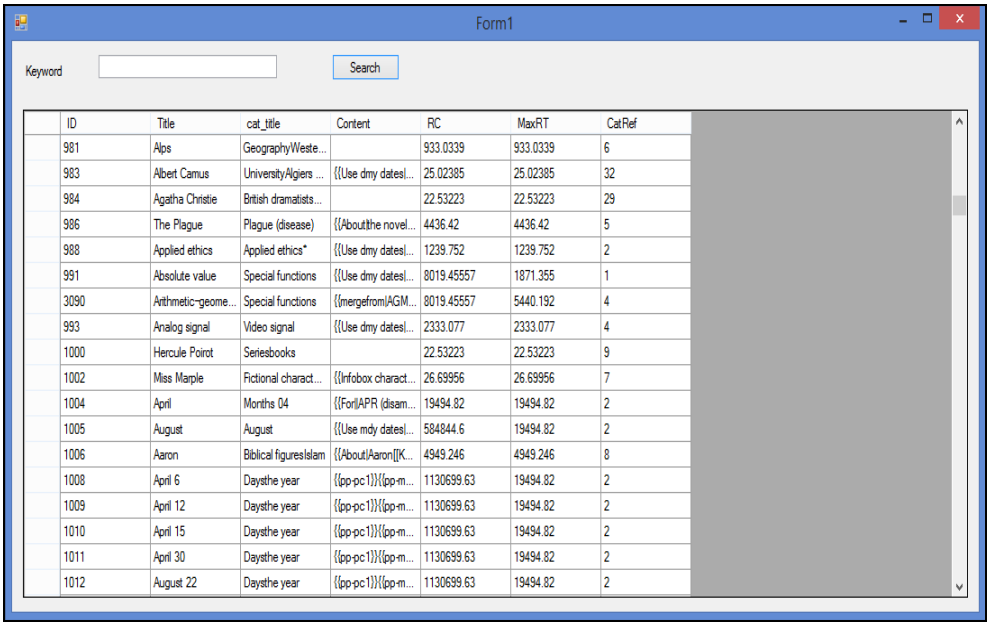
Một bài báo có thể thuộc nhiều danh mục, trong bước này tác giả tiến hành chọn danh mục có trọng số cao nhất cho bài báo đó là danh mục chính cho bài báo đó.

4.3 Chương trình thực nghiệm

Tác giả tiến hành xây dựng chương trình hỗ trợ tìm kiếm các bài báo của Wiki và danh mục tốt nhất của bài báo.

Sau khi người dùng nhập vào nội dung cần tìm chương trình sẽ trả về các bài báo với nội dung tương ứng cùng với danh mục có trọng số cao nhất của bài báo. Chương trình trả về bảy cột theo thứ tự như sau: Mã bài báo, tiêu đề bài báo, danh mục, trọng

số danh mục phù hợp. Trọng số danh mục phù hợp nhất và số danh mục của bài báo đó.



Trong thực nghiệm trên tác giả tìm theo phương pháp tìm những nội dung mà người dùng nhập vào có xuất hiện trong nội dung hoặc trong tiêu đề bài báo không. Nếu từ nào nhập vào mà không xuất hiện thì chứng tỏ là từ đó không có trong nội dung hoặc tiêu đề bài báo.

4.4 Trường hợp thành công và thất bại

Luận văn áp dụng thuật toán của bài báo giúp tìm ra những danh mục phù hợp nhất trong bài báo.

Trong trường hợp thành công thì khi áp dụng công thức luôn có ít nhất một danh mục được tìm ra cho bài báo, vấn đề là độ chính xác cao hay thấp mà thôi.

Trong trường hợp thất bại là do chúng ta phải áp dụng đến bảy bước mới hoàn thành thuật toán này nên nếu trong bảy bước trên mà có một bước có giá trị là không thì sẽ dẫn đến kết quả của thuật toán là không. Do đó, nếu áp dụng cả bảy bước trên vào thuật toán thì sẽ có rất nhiều bài báo sẽ có giá trị là không trong bảy bước đó. Từ đó sẽ giảm độ chính xác cho các danh mục ở các bài báo đó. Nếu có quá nhiều bài báo

không đáp ứng đủ bảy bước trên thì thí nghiệm sẽ thất bại. Và đó chính là điểm yếu của thuật toán do xử lý phức tạp, rườm rà.

$R_t: 0\ t_\omega: 0\ R_\omega: 3.70774243398595\ a_t: 15\ S_t: 1\ L_t: 1$ *strTitle: Anarchism*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanHistory*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanGeography*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanPeople*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanCommunications*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanTransportations*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanMilitary*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AfghanistanTransnationalIssues*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AssistiveTechnology*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 1\ L_t: 1$ *strTitle: AmoeboidTaxa*

$R_t: 0\ t_\omega: 0\ R_\omega: 5.73940755071869\ a_t: 13\ S_t: 2\ L_t: 1$ *strTitle: Autism*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 2\ L_t: 1$ *strTitle: AlbaniaHistory*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 2\ L_t: 1$ *strTitle: AlbaniaPeople*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 2\ L_t: 1$ *strTitle: AsWeMayThink*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 2\ L_t: 1$ *strTitle: AlbaniaGovernment*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 2\ L_t: 1$ *strTitle: AlbaniaEconomy*

$R_t: 0\ t_\omega: 0\ R_\omega: NaN\ a_t: 6\ S_t: 3\ L_t: 1$ *strTitle: Albedo*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 3\ L_t: 1$ *strTitle: AfroAsiaticLanguages*

$R_t: 0\ t_\omega: 0\ R_\omega: 0\ a_t: 0\ S_t: 3\ L_t: 1$ *strTitle: ArtificalLanguages*

4.5 Đánh giá

4.5.1 Dữ liệu đánh giá

Luận văn đánh giá thuật toán với dữ liệu bao gồm 2588 bài báo và 150435 danh

mục, sau khi thực hiện thuật toán kết quả được lưu vào cơ sở dữ liệu phục vụ cho việc tìm kiếm bài báo cùng với danh mục tốt nhất của nó. Luận văn tiến hành đánh giá thuật toán với dữ liệu trên.

4.5.2 Độ chính xác của chương trình

Để đo lường độ chính xác của thuật toán tác giả tính trong top n danh mục có bao nhiêu phần trăm các danh mục chính thức của bài báo đó. Trong top n các danh mục có rất nhiều bài báo mà các danh mục khác lại có trọng số là cao hơn các danh mục đó.

Tác giả dùng câu truy vấn SQL sau để lấy ra top 10 danh mục có chứa nhiều bài báo nhất:

```
select top 10 category.cat_id,category.cat_title, count(CatIDMax) as TotalDocument
from Document,Cat_Doc,category
where Document.ID=Cat_Doc.DocumentID and Cat_Doc.CatID=category.cat_id and
CatIDMax =category.cat_id
group by category.cat_id,category.cat_title
order by TotalDocument desc
```

Kết quả truy vấn như sau: Mã danh mục, tên danh mục, số lượng bài báo của danh mục đó

	cat_id	cat_title	TotalDocument
1	65358076	Daysthe year	58
2	65357824	Place name disambiguation pages	6
3	65357288	Functional groups	4
4	65363217	Batting statistics	4
5	65357571	Angiosperm orders	4
6	65360558	Binary trees	3
7	65360260	Amiga	3
8	65358268	GeographyAmerican Samoa	2
9	65360229	Spacecraft launched1971	2
10	65359962	Inorganic chemistry	2

Để đánh giá độ chính xác của thuật toán tác giả sử dụng công thức sau:

$$U = \frac{A}{D} \times 100\%$$

Trong đó :

U: Độ chính xác của thuật toán.

A: Số danh mục chính thức

D: Tổng số danh mục bao gồm danh mục chính thức và danh mục khác trong cùng bài báo.

Tiến hành thực nghiệm với danh mục thứ 1 có ID là ‘65358076’ tác giả tiến hành kiểm tra xem trong danh mục thuộc top n trên có bao nhiêu bài báo có danh mục trên nhưng lại có danh mục ngoài top n trên lại có trọng số cao hơn.

Với câu truy vấnSQL tác giả thu được kết quả như sau:

```
select id,Document.Title,CatIDMax
from Document,Cat_Doc,category
where Document.ID=Cat_Doc.DocumentID and Cat_Doc.CatID=category.cat_id and
category.cat_id =65358076
order by CatIDMax
```

Với câu truy vấn trên kết quả thu được 58 danhh mục chính thức trên 58 danh mục của bài báo :

	id	Title	CatIDMax
43	1974	April 17	65358076
44	1990	August 5	65358076
45	2192	August 11	65358076
46	2194	April 5	65358076
47	2195	April 20	65358076
48	2196	April 19	65358076
49	2224	April 8	65358076
50	2279	April 3	65358076
51	2315	August 10	65358076
52	2326	April 27	65358076
53	2418	August 4	65358076
54	2734	April 24	65358076
55	2735	April 7	65358076
56	2483	April 21	65358076
57	2564	April 10	65358076
58	2733	April 25	65358076

Với kết quả trên ta có được 58 danh mục chính thức so với tổng 58 danh mục áp dụng công thức $U = \frac{A}{D} \times 100\%$ ta có được kết quả của độ chính xác thuật toán như sau:

$$U = \frac{58}{58} \times 100\% = 100\%$$

Tiến hành thực nghiệm với danh mục thứ 2 có ID là ‘65357824’ tác giả tiến hành kiểm tra xem trong danh mục thuộc top n trên có bao nhiêu bài báo có danh mục trên nhưng lại có danh mục ngoài top n trên lại có trọng số cao hơn.

Với câu truy vấn SQL tác giả thu được kết quả như sau:

```
select id,Document.Title,CatIDMax
from Document,Cat_Doc,category
```

where Document.ID=Cat_Doc.DocumentID and Cat_Doc.CatID=category.cat_id and category.cat_id =65357824

order by CatIDMax

Với câu truy vấn trên thu được kết quả như sau :

	id	Title	CatIDMax
1	885	Altenberg	65357824
2	1919	Antwerp (disambiguation)	65357824
3	2827	Abingdon	65357824
4	2833	Abitibi	65357824
5	3121	Andersonville	65357824
6	3371	Bach (disambiguation)	65362681
7	3748	Bourbon	65357824

Với kết quả trên ta có được 6 danh mục chính thức so với tổng 7 danh mục đó áp dụng công thức $U = \frac{A}{D} \times 100\%$ ta có được kết quả của độ chính xác thuật toán như sau:

$$U = \frac{6}{7} \times 100\% = 85\%$$

Tương tự như trên tác giả tiến hành thực nghiệm với mã danh mục ‘65357288’ thu được kết quả như sau:

	id	Title	CatIDMax
1	639	Alkane	65357288
2	1422	Amide	65357288
3	2761	Alkene	65357288
4	2763	Alkyne	65357288
5	1014	Alcohol	65358077
6	2762	Allene	65361713

Với kết quả trên ta có được 4 danh mục chính thức so với tổng 6 danh mục áp dụng công thức $U = \frac{A}{D} \times 100\%$ ta có được kết quả của độ chính xác thuật toán như sau:

$$U = \frac{4}{6} \times 100\% = 66\%$$

Tương tự như trên tác giả tiến hành thực nghiệm với mã danh mục ‘65363217’ thu được kết quả như sau:

	id	Title	CatIDMax
1	3802	Baseballs	65363217
2	3808	On-base percentage	65363217
3	3810	On-base plus slugging	65363217
4	3812	Plate appearance	65363217
5	3806	Hitpitch	65363220
6	3807	Hit (baseball)	65363220

Với kết quả trên ta có được 4 danh mục chính thức so với tổng 6 danh mục áp dụng công thức $U = \frac{A}{D} \times 100\%$ ta có được kết quả của độ chính xác thuật toán như sau:

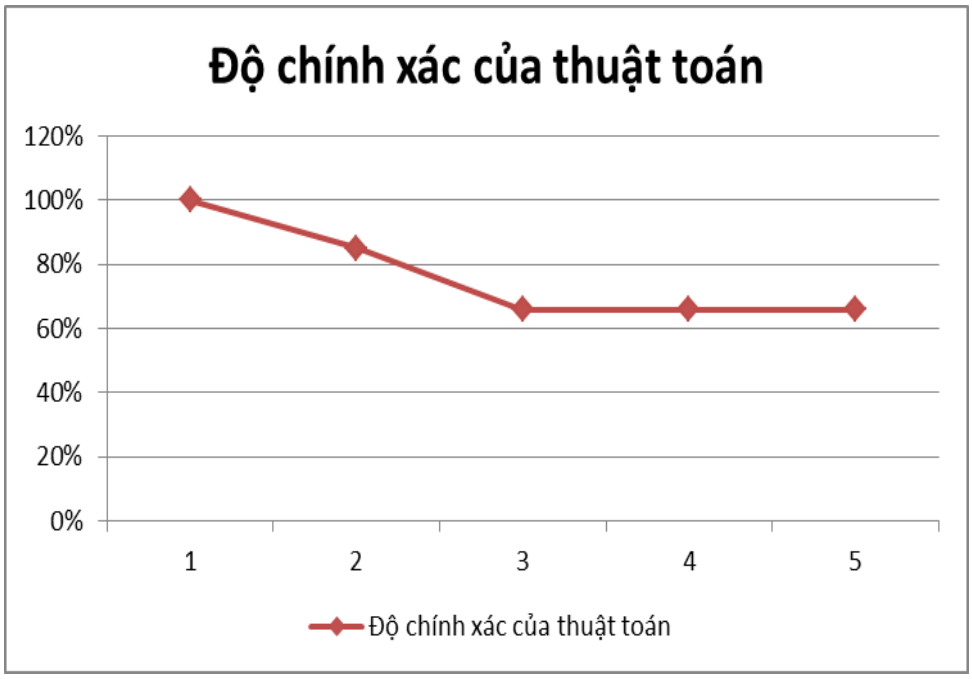
$$U = \frac{4}{6} \times 100\% = 66\%$$

Tương tự với hai danh mục còn lại tác giả thu được độ chính xác U đều là 66%.

Theo thực nghiệm trên ta có độ chính xác của thuật toán như sau:

Bảng 4.4 Độ chính xác của thuật toán

Top 10 danh mục	Độ chính xác
Daysthe year	100%
Place name disambiguation pages	85%
Functional groups	66%
Batting statistics	66%
Angiosperm orders	66%
Độ chính xác trung bình của thuật toán là : 76.6%	



Biểu đồ 4.1 Đánh giá độ chính xác của thuật toán

Biểu đồ trên thể hiện độ chính xác của thuật toán khi tác giả tiến hành thí nghiệm theo tứ tự của 5 danh mục được lấy trong bảng 4.4. Trong biểu đồ trên các dòng được thể hiện theo trục X và độ chính xác của thuật toán được thể hiện theo trục Y. Sau khi nhìn biểu đồ trên chúng ta có thể thấy được độ chính xác giảm dần theo số lượng danh mục

Chúng ta có thể dễ dàng thấy được đường màu đỏ có chiều hướng đi xuống theo số lượng danh mục.

Từ đánh giá trên ta thấy được độ chính xác của chương trình khá cao giúp tăng cường tính tự động trong phân loại tài liệu.

4.6 Độ phản hồi của chương trình

Độ phản hồi của chương trình được xác định bằng tỉ lệ danh mục chính thức trong top n danh mục. Độ phản hồi được tính theo công thức sau:

$$H = \frac{E}{F} \times 100\%$$

Trong đó:

H: Độ phản hồi chương trình

E: Số danh mục chính thức

F: Tổng số danh mục

Theo kết quả thực nghiệm tác giả thu được 84 danh mục có các bài báo. Với kết quả thu được từ các thí nghiệm trên áp dụng công thức ta có được độ phản hồi như sau:

Trường hợp lấy top 1 danh mục theo thực nghiệm trên ta thu được 58 danh mục chính thức

$$H = \frac{58}{84} \times 100\% = 69\%$$

Trường hợp lấy top 2 danh mục theo thực nghiệm ở phần trên ta có thêm 6 danh mục chính thức nữa vậy áp dụng công thức ta có kết quả như sau:

$$H = \frac{64}{84} \times 100\% = 76\%$$

Trường hợp lấy top 3 tác giả thu được thêm 4 danh mục chính thức nữa vậy áp dụng công thức ta sẽ thu được kết quả như sau:

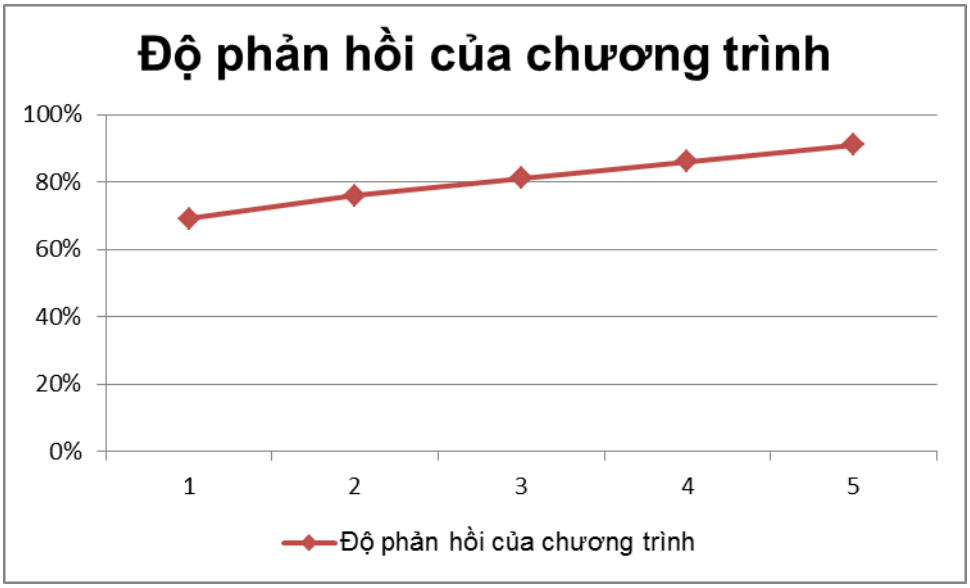
$$H = \frac{68}{84} \times 100\% = 81\%$$

Áp dụng cho các trường hợp còn lại tác giả thu được kết quả như bảng sau:

Trong top năm danh mục trên tác giả tính được độ phản hồi như sau:

Bảng 4.5 Độ phản hồi của chương trình

Số danh mục	Độ phản hồi
1	69%
2	76%
3	81%
4	86%
5	91%
Trung bình	80.6%



Biểu đồ 4.2 Độ phản hồi của chương trình

Biểu đồ trên thể hiện độ phản hồi của chương trình với các dữ liệu thí nghiệm được lấy từ bảng 4.5 với trục X thể hiện số chủ đề và trục Y thể hiện độ phản hồi của thuật toán, chúng ta thấy được độ phản hồi của thuật toán tăng dần theo số lượng các chủ đề qua chiều tăng dần của các cột hoặc hướng đi lên của đường màu đỏ, theo thực nghiệm trên ta có độ phản hồi trung bình của chương trình là 80,6%

4.7 Kết luận:

Phương pháp này được sử dụng thuần túy cho việc phân loại và xếp nhóm các tiêu đề và phân nhóm các bài viết Wikipedia, thuật toán giúp bỏ qua giai đoạn khai thác thông tin phong phú được cung cấp danh mục bài viết, bỏ qua đường kết nối giữa các tài liệu, hoặc ngay cả cấu trúc phân tầng các phân nhóm.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Luận văn đã giải quyết được những nhiệm vụ mà luận văn đã đưa ra

- Khảo sát, phân tích hệ thống chủ đề của tài liệu dạng văn bản lưu trữ trong Wikipedia

- Khảo sát các nghiên cứu liên quan đến việc nhận biết chủ đề của văn bản trong Wikipedia

- Phát triển (trên cơ sở kế thừa) hoặc cải tiến một phương pháp nhận biết chủ đề tài liệu (dạng văn bản), dựa trên nguồn dữ liệu tên thể loại sẵn có của Wikipedia.

- Thực nghiệm, đánh giá và viết báo cáo.

Qua những kết quả thực nghiệm đạt được cho thấy đề tài nhận biết chủ đề của tài liệu dựa trên Wikipedia là khả thi và có thể áp dụng được. Giúp tìm ra các danh mục phù hợp cho các bài báo một các tự động và đạt độ chính xác cao.

Bên cạnh đó, do hạn chế về mặt thời gian và kiến thức đề tài vẫn còn những hạn chế sau:

- Trong một số trường hợp, kết quả thực nghiệm chưa cao.

- Đối với dữ liệu lớn thì thời gian thực hiện tìm danh mục phù hợp cho bài báo sẽ rất lâu do chương trình khá phức tạp.

5.2. Hướng phát triển

Tìm giải pháp giảm thời gian thực hiện, tăng độ chính xác và tìm kiếm thuật toán đơn giản

TÀI LIỆU THAM KHẢO

Trong nước

[1] Nguyễn Chánh Thành (2010). *Xây dựng mô hình mở rộng truy vấn trong truy xuất thông tin văn bản*, Luận án tiến sĩ kỹ thuật. Đại học Bách khoa TP.HCM.

[2] Đinh Quang Định (2013). *Nghiên cứu công nghệ Web 3.0 (Semantic Web) và khả năng triển khai áp dụng*. Học viện công nghệ bưu chính viễn thông

[3] Phạm Đình Hồng (2013). *Nghiên cứu phương pháp phân nhóm dữ liệu động áp dụng vào truy vấn thông tin*. Đại học Đà Nẵng

[4] Nguyễn Thị Bích Phương (2012). *Nghiên cứu phương pháp mở rộng truy vấn trong truy xuất thông tin (Information Retrieval).* Học viện công nghệ bưu chính viễn thông

[5] Nguyễn Đình Bình (2012). *Nghiên cứu khai phá dữ liệu web và ứng dụng tìm kiếm trích chọn thôn tin theo chủ đề.* Đại học Đà Nẵng

[6] Nguyễn Thị Hồng Nhung, Nguyễn Thị Tuyết Mai. *Hệ thống tìm kiếm thông tin xuyên ngôn ngữ Việt – Anh – Hoa*.

[7] Nguyễn Tiến Thanh (2010)- *Trích chọn quan hệ thực thể trên Wikipedia Tiếng Việt dựa vào cây phân tích cú pháp.* Trường Đại học Công nghệ

[8] Trần Ngọc Phúc (2012) – *Phân loại nội dung tài liệu Web*. Trường Đại học Lạc Hồng

Ngoài nước

[9] Peter Sch¨onhofen. Identifying document topics using the Wikipedia category network. Computer and Automation Research Institute Hungarian Academy of Sciences Kende u. 13–17, H-1111 Budapest

[10] S. F. Adafre and M. de Rijke. Discovering missing links in Wikipedia. In Proc. of the 3rd int'l workshop on Link discovery, pages 90–97, 2005.

[11] M. Aery, N. Ramamurthy, and Y. A. Aslandogan. Topic identification of textual data. Technical Report CSE-2003-25, University of Texas at Arlington, Department of Computer Science and Engineering, 2003.

[12] D. Ahn, V. Jijkoun, G. Mishne, K. M¨uller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA track. In Proc. of the 13rd Text Retrieval Conf. (TREC), 2004.

[13] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[14] F. Bellomi and R. Bonato. Network analysis for Wikipedia. In Proc. of Wikimania 2005, the 1st Int'l Wikimedia Conf.,2005.

[15] C.-Y. Lin. Knowledge-based automatic topic identification. In Meeting of the Association for Computational Linguistics, pages 308–310, 1995.

[16] C.-Y. Lin. Robust automated topic identification. PhD thesis, University of Southern California, 1997.

[17] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/˜mccallum/bow, 1996.

[18] G. Mishne, M. de Rijke, and V. Jijkoun. Using a reference corpus as a user model for focused information retrieval. J. of Digital Information Management, 3(1):47–52, 2005.

[19] R. Navigli. Automatically extending, pruning and trimming general purpose ontologies. In Proc. of the 2nd IEEE Int'l Conf. on Systems, Man and Cybernetics, 2002.

[20] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatic assignment of Wikipedia encyclopedic entries to wordnet synsets. In Proc. of the 3rd Int'l Atlantic Web Intelligence Conf. (AWIC), pages 380–386, 2005.

[21] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatic extraction of semantic relationships for wordnet by means of pattern learning from Wikipedia. In Proc. of the 10th Int'l Conf. on Applications of Natural Language to Information Systems (NLDB), pages 67–79, 2005.

[22] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In Proc. of the Int'l Conf. on New Methods in Language Processing, Manchester, UK, 1994.

[23] B. Stein and S. M. zu Eien. Topic identification: Framework and application. In Proc. of the 4th Int'l Conf. on Knowledge Management (I-KNOW 04), pages 353–360, 2004.

[24] S. Tiun, R. Abdullah, and T. E. Kong. Automatic topic identification using ontology hierarchy. In Proc. of the 2nd Int'l Conf. on Computational Linguistics and Intelligent Text Processing, pages 444–453, London, UK, 2001.

[25] M. V¨olkel, M. Kr¨otzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic Wikipedia. In Proc. of the 15th int'l conf. on World Wide Web. WWW2006, 2006.

[26] J. Voss. Measuring Wikipedia. In Proc. of the Int'l Conf. Of the Int'l Society for Scientometrics and Informetrics, Stockholm, Sweden, 2005.

[27] Chau Q. Nguyen, Tuoi T. Phan. An Ontology–Based Approach for Key Phrase Extraction

Trang web

[28] http://vi.wikipedia.org/wiki/Wikipedia:Gi%E1%BB%9Bi_thi%E1%BB%87u

[29] http://www.google.com.vn//giaidap/thread?tid=4a6585a2692334fa

[30] http://dantri.com.vn/blog/tu-wiki-co-nghia-la-gi-443030.htm

[31] https://voer.edu.vn/m/nhung-uu-diem-cua-mo-hinh-web-wiki/40d9cfad

[32] http://tuanvietnam.vietnamnet.vn/wikipedia-hoat-dong-nhu-the-nao-phan-i

[33] http://dumps.wikimedia.org/enwiki/latest/