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

NGUYỄN ĐÀO MINH THƯƠNG

XÂY DỰNG MÔ HÌNH CÁC CHỦ ĐỀ VÀ CÔNG CỤ TÌM KIẾM NGỮ NGHĨA

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 ĐÀO MINH THƯƠNG

XÂY DỰNG MÔ HÌNH CÁC CHỦ ĐỀ VÀ CÔNG CỤ TÌM KIẾM NGỮ NGHĨA

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công Nghệ Thông Tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN THỊ THANH SANG

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 THỊ THANH SANG

LuậnvănThạcsĩđượcbảovệtạiTrườngĐạihọcCôngnghệTP.HCM

ngày …tháng… năm …

Thành phần hội đồng đánh giá luận văn thạc sĩ gồm:

TT	Họ và tên	Chức danh hội đồng
1	**GS.TSKH Hoàng Văn Kiếm**	Chủ tịch
2	**TS.Lê Tuấn Anh**	Phản biện 1
3	**TS.Nguyễn Văn Mùi**	Phản biện 2
4	**PGS.TS Lê Trọng Vĩnh**	Uỷ viên
5	**TS. Võ Đình Bảy**	Uỷ viên, Thư ký

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày..... tháng..... năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: Nguyễn Đào Minh Thương — Giới tính:Nam
Ngày, tháng, năm sinh: 19/01/1984 — Nơi sinh:Long An
Chuyên ngành: Công nghệ thông tin — MSHV:1341860027

I- Tên đề tài:Xây Dựng Mô Hình Các Chủ Đề Và Công Cụ Tìm Kiếm Theo Ngữ Nghĩa

II- Nhiệm vụ và nội dung:
- Xây dựng mô hình các chủ đề
- Áp dụng mô hình các chủ đề xây dựng công cụ tìm kiếm theo ngữ nghĩa

III- Ngày giao nhiệm vụ: 15/09/2014
IV- Ngày hoàn thành nhiệm vụ: 08/03/2015
V- Cán bộ hướng dẫn: TS. Nguyễn Thị Thanh Sang

Cán Bộ Hướng Dẫn — **Khoa Quản Lý Chuyên Ngành**

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. TôixincamđoanrằngmọisựgiúpđỡchoviệcthựchiệnLuậnvănnày 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

Nguyễn Đào Minh Thương

LỜI CẢM ƠN

Tôi xin bày tỏlòng biết ơn sâu sắc đến TS Nguyễn Thị Thanh Sang đã hướng dẫnnhiệt tình, tận tâm trong suốt quá trình tôi thực hiện luận văn này.Tôi xin chân thành cảmơn Quý thầy cô trong Khoa Công nghệ thôngtin trường Đại Công Nghệ đã tạo điều kiện thuận lợi cho tôi trong suốt thờigian học tập và nghiên cứu tại trường.Tôi cũng xin chân thành cảmơn Quý thầy cô ngoài trườngđã tận tâmdạy bảo tôi trong suốt quá trình học tập và giúp đỡ tôi trong quá trình nghiêncứu.Xin chân thành cảm ơn những người thân trong giađình, cùng các anhchịem, bạn bè,đồng nghiệpđã giúpđỡ,động viên tôi trong quá trình thựchiện và hoàn thành luận văn này.

HCM, ngày 14 tháng 3 năm 2015

Học viên

Nguyễn Đào Minh Thương

TÓM TẮT

Ngày nay với lượng thông tin lớn từ internet đã đặt ra vấn đề về tìm kiếm và xử lý dữ liệu, phải có một công cụ đảm bảo về độ chính xác trong việc tìm kiếm và đồng thời cũng phải trả về một lượng kết quả phong phú cho người dùng. Ngoài việc trảvề những tài liệu chứa những từ mà người dùng cần tìm kết quả trả về có thể bao gồm những tài liệu có nội dung gần với nội dung mà người dùng tìm giúp tạo nên sự phong phú về kết quả của việc tìm kiếm.Với vấn đề trên luận văn tiến hành xây dựng mô hình các chủ đề nhằm phục vụ cho việc tìm kiếm theo ngữ nghĩa và đồng thời cũng xây dựng chương trình áp dụng mô hình trên bằng ngôn ngữ ontology cho việc tìm kiếm theo ngữ nghĩa.

ABSTRACT

Today the large amount of information from the Internet rises special problems of search and data processing, it is crucial to have to a tool allowing to efficiently search and return a large amount of correct and sound results for users.Beside returning the documents containing the words that user is searchingreturned results should include documents whose content is related to the user's topics,that helps to increase the richness of the search results. It is expected thatresultingcontentreturn are not only interestingbut also semanticallyrich. Therefore, this thesis has proposed solutions ofconstructing topic models served for semantically searching in some specificwebsites and building a program which can automatically generate the ontology-based topic model for facilitating the Web search.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
CRFs	Conditional Random Fields
FIFO	First In First Out
HDP	Hierarchical Dirichet Process
LDA	Latent Drichlet Allocation
LSI	latent semantic indexing
NLP	neuro-linguistic programming.
RDF	Resource Description Framework
SPARQL	Simple Protocol and RDF Query Language
SQL	Structured Query Language
SVMs	Support Vector Machines
URL	Uniform Resource Locator
WebCrawler	Web crawler
XML	Extensible Markup Language

DANH SÁCH CÁC TỪ TIẾNG ANH

TIẾNG ANH	Ý NGHĨA
NameSpace	Không gian tên
Schame	Lược đồ
Proof	Thực hiện các luật
Trust	Kiểm ta ứng dụng tin tưởng hay không
Stopword	Từ vô nghĩa
Thread	Luồng
RDF Data Access Working Group	Nhóm phát triển ngôn ngữ truy vấn dữ liệu RDF
Cache	Bộ nhớ
Repository	Nơi lưu trữ dữ liệu

DANH MỤC CÁC BẢNG

DANH MỤC CÁC ĐỒ THỊ, HÌNH ẢNH

CHƯƠNG 1: MỞ ĐẦU

❖ TÍNH CẤP THIẾT CỦA ĐỀ TÀI:

Với sự phát triển nhanh của công nghệ thông tin dẫn đến lượng thông tin ngày càng dày đặt với lượng thông tin dày đặt như vậy để tìm kiếm thông tin một cách chính xác và nhanh chóng đang được nghiên cứu và phát triển khá phổ biến hiện nay. Tuy nhiên việc tìm kiếm nội dung theo ngữ nghĩa bằng ngôn ngữ tiếng Việt không được phát triển nhiều ở Việt Nam. Do việc xử lý ngôn ngữ tiếng Việt chưa được phổ biến và còn nhiều phức tạp tạo nên tạo sự khó khăn trong việc xây dựng công cụ tìm kiếm theo ngữ nghĩa.

Hiện tại trong nước các chương trình tìm kiếm theo ngữ nghĩa chưa được nghiên cứu nhiều, trong quá trình nghiên cứu và phát triển luận văn tác giả chưa tìm được chương trình tìm kiếm theo ngữ nghĩa hỗ trợ tiếng Việt.

Để cho việc tìm kiếm được chính xác và kết quả trả về phong phú cho người dùng với lượng thông tin lớn như trên tác giả tiến hành nghiên cứu và xây dựng mô hình các chủ đề cùng với chương trình tìm kiếm áp dụng mô hình trên phục vụ cho việc tìm kiếm được chính xác hơn và kết quả phong phú hơn.

Luận văn góp phần xây dựng và phát triển công cụ hỗ trợ cho việc tìm kiếm theo ngữ nghĩa bằng ngôn ngữ tiếng Việt. Tuy nhiên để xây dựng công cụ tìm kiếm theo ngữ nghĩa cần giải quyết một số vấn đề như:

- Thu thập dữ liệu trên mạng để hỗ trợ cho việc tìm kiếm.
- Loại bỏ những từ không có ý nghĩa, xử lý tiếng Việt thành những cụm từ có ý nghĩa hỗtrợ cho việc tìm kiếm và gom nhóm từ v.v.
- Thực hiện việc gom nhóm các từ có cùng ý nghĩa vào cùngchủ đề , và dựa vào tỉ lệ xuất hiên của các từ trong các tài liệu Web v.v.
- Xây dựng mô hình chủ đề các tài liệu, mối liên hệ, các từ và các trọng số của nó v.v.
- Xây dựng công cụ tìm kiếm các tài liệu theo ngữ nghĩa dựa trên mô hình xây dựng được

Với các vấn đề trên em quyết định chọn đề tài xây dựng mô hình các chủ đề và công cụ tìm kiếm theo ngữ nghĩa

1. Mục tiêu của luận văn:

Xây dựng mô hình các chủ đề thể hiện mối liên hệ giữa các từ và cụm từ,các tài liệu,và các chủ đề, v.v. Mối liên hệ giữa các thành phần trên được thể hiện bằng các trọng số của các thành phần đó.

Xây dựng công cụ tìm kiếm theo ngữ nghĩa dựa trên mô hình các chủ đề đã xây dựng.

2. Đối tượng nghiên cứu:

Các tài liệu nghiên cứu phục vụ cho việc xây dựng mô hình các chủ đề và tìm kiếm các tài liệu có thể là văn bản hoặc thu thập các tài liệu này từ các trang web tin tức v.v. Các tài liệu trên phải chuẩn tiếng Việt các trang web tài liệu hoặc tin tức phải không bao gồm những trang chỉ hình ảnh hoặc âm thanh vì chương trình chỉ hỗ trợ tìm kiếm các tài liệu văn bản tiếng Việt.

3. Phạm vi nghiên cứu:

Các tài liệu văn bản trên các trang web cũng như các bài báo điện tử hiện nay bao gồm tất cả các thể loại (không bao gồm các bài báo chỉ hình ảnh, video hoặc âm thanh),hiện luận văn tiến hành thực nghiệm trên các bài báo của trang web www.docbao.vn. Do trang web bao gồm các bài báo chuẩn tiếng Việt nội dung phong phú và số lượng các bài báo lớn phục vụ tốt cho việc xây dựng mô hình các chủ đề và tìm kiếm.

4. Bố cục trình bày của luận văn:

Chương 1: Mở đầu

Chương 2: Giới thiệu tổng quan về mô hình các chủ đề và xây dựng công cụ tìm kiếm các tài liệu theo ngữ nghĩa.

Chương 3: Một sốkỹ thuật tạo mô hình các chủ đề và xây dựng công cụ tìm kiếm tài liệu theo ngữ nghĩa đồng thời đề cập đến các vấn đề liên quan.

Chương 4: Xây dựng mô hình các chủ đề và công cụ tìm kiếm theo ngữ nghĩa.

Chương 5: Đánh giá kết quả thực nghiệm đồng thờichỉ ra những điểm cần khắc phục đồng thời đặt ra hướng cần phát triển trong tương lai.

CHƯƠNG 2: GIỚI THIỆU TỔNG QUAN VỀ MÔ HÌNH CÁC CHỦ ĐỀ VÀ XÂY DỰNG CÔNG CỤ TÌM KIẾM CÁC TÀI LIỆU THEO NGỮ NGHĨA

2.1. Giới thiệu về mô hình các chủ đề:

Với số lượng thông tin ngày một lớn thì việc tìm kiếm dữ liệu trở nên rất quan trọng và cấp thiết, và việc tìm kiếm dữ liệu cũng đang được phát triển rất mạnh và đa dạng.

Giữa một lượng thông tin khổng lồ thì việc tìm kiếm dữ liệu chính xác và nhanh nhất luôn là vấn đề cần thiết và rất quan trọng trong tình hình hiện nay.

Hiện trên thế giới cũng có những chương trình tìm kiếm rất mạnh và chính xác phục vụ cho công việc tìm kiếm trên Intenet của hàng triệu người trên thế giới mỗi ngày như: Google, Bing,…. Các công cụ này phục vụ cho quá trình tìm kiếm online trên Internet rất hữu dụng và được dùng rộng rãi.

Nếu chúng ta có một nguồn dữ liệu lớn cho riêng mình và chúng ta cần tìm kiếm trên nguồn dữ liệu đó thì chúng ta có thể lưu vào các cơ sở dữ liệu phổ biến hiện nay như Oracle, SQL, MySQL,…các công cụ đó điều hỗ trợ tìm kiếm dữ liệu rất tốt và đa dạng tuy nhiên nếu chúng ta cần một sự tìm kiếm thông minh như tìm kiếm theo ngữ nghĩa thì chúng ta phải xây dựng một mô hình cho riêng mình để tiến hành việc tìm kiếm trên . Hiện trên thế giới cũng có nhiều công cụ và mã nguồn mở hỗ trợ việc tìm kiếm như: Lucene,… Tuy nhiên ở Việt Nam thì việc tìm kiếm theo ngữ nghĩa còn nhiều hạn chế.

Vì thế việc tìm kiếm theo ngữ nghĩa hỗ trợ tiếng Việt đang là vấn đề cần nghiên cứu và phát triển hiện nay đặc biệt là ở nước ta để giải quyết các vấn đề tìm kiếm dữ liệu theo ngữ nghĩa đang ngày một cấp thiết.

Mô hình các chủ đề được xây dựng và nghiên cứu phục vụ cho nhiều mục đích khác nhau, được xây dựng và phát triển khá phổ biến trong những năm gần đây. Tuy nhiên các mô hình hỗ trợ tiếng Việt khá hạn chế và chưa được phát triển nhiều.

Mô hình các chủ đề là xây dựng một mô hình quan hệ các chủ đề với nhau, các chủ đề đó liên quan với nhau dựa trên những mối quan hệ nào đó. Tùy mục đích khác nhau mà các mô hình các chủ đề được xây dựng khác nhau. Trong luận văn này mô hình các chủ đề được xây dựng dựa trên mối liên hệ giữa các từ, giữa các tài liệu với các tài liệu,giữa các từ với các chủ đề ,….Mô hình này xây dựng nhằm phục vụ cho quá trình tìm kiếm được tốt hơn và đặc biệt hỗ trợ tốt cho quá trình tìm kiếm theo ngữ nghĩa.

Mô hình các chủ đề được xây dựng cho ngôn ngữ tiếng Việt, Mô hình xây dựng trên các thuật toán tách từ CRF và SVM. Saukhi các tài liệu được thu thập trên mạng bằng WebCrawler các tài liệu đó sẽ được loại bỏ đi những từ dư thừa ít ảnh hưởng đến tài liệu và sau đó tiến hành tách từ thành từng cụm từ tiếng Việt có nghĩa.

Sau khi các tài liệu được tách thành những từ có nghĩa các tài liệu đó sẽ áp dụng thuật toán LDA để phân loại ra các chủ đề chứa các từ xuất hiện phổ biến trong chủ đề đó cùng với những trọng số của nó. Đồng thời thuật toán cũng hỗ trợ tìm ra các từ và trọng số của nó trong một tài liệu, số lần xuất hiện của tài liệu trong các topic,…

Quá trình thực hiện các bước trên hoàn toàn tự động giúp tiết kiệm được thời gian và tăng cường độ chính xác.

2.2. Tổng quan:

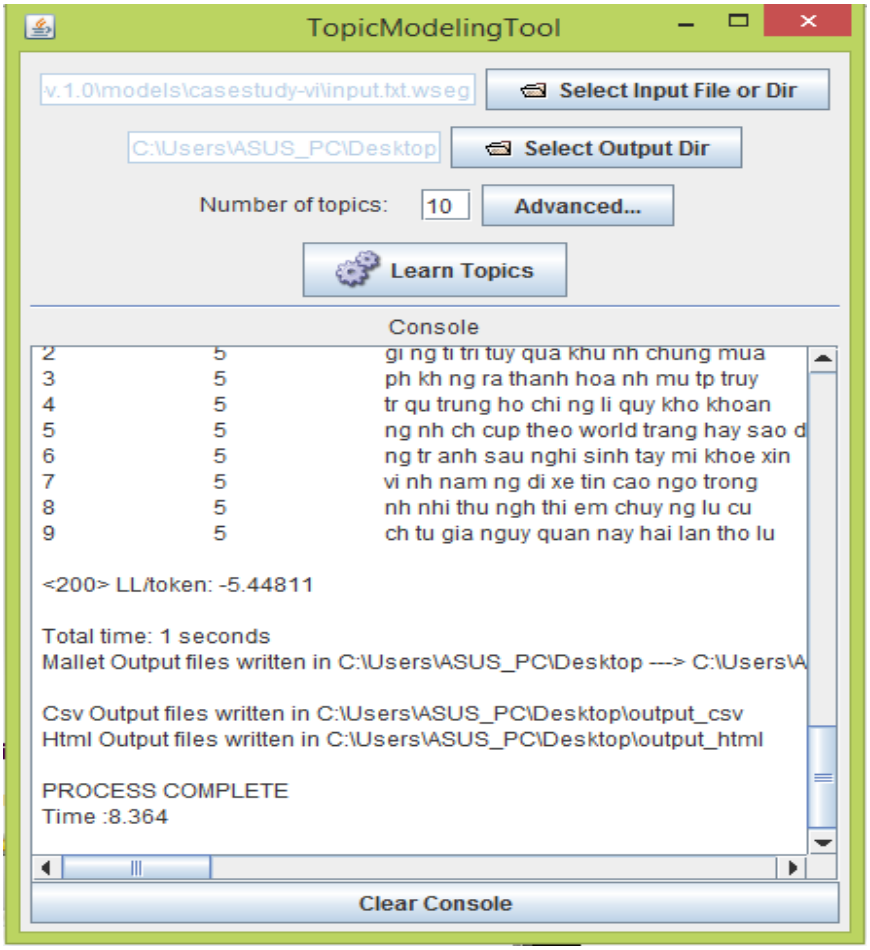
Với số lượng thông tin lớn như hiện nay và đòi hỏi độ chính xác cao của việc tìm kiếm, đòi hỏi phải có mô hình dữ liệu được xây dựng tốt để đáp ứng các yêu cầu trên, với yêu cầu cấp thiết trên mô hình dữ liệu đươc xây dựng để giúp việc tìm kiếm dữ liệu được tốt hơn. Trên thực tế các trang web hỗ trợ việc tìm kiếm n ội dung cho trang web mang tính chất tìm các đoạn văn chứa các từ cần tìm , điều này có độ chính xác không cao và không liệt kê được các câu có tuần suất từ đó xuất hiện nhiều hiển thị trên cùng cho người dùng ho ặc các nội dung liên quan với nội dung mà người dùng muốn tìm kiếm nhưng nội dung lại không chứa các từ mà người dùng nhập vào…Mô hình dữ liệu xây dựng mới sẽ đáp ứng các yêu cầu trên đồng

thời tăng tốc độ tìm kiếm , tăng độ chính xác, liệt kê các câu tìm được theo mức độ xuất hiện của các từ…

Mô hình các chủ đề đang được nghiên cứu và phát triển rộng rãi ở nước ngoài, ở Việt Nam cũng đang được nghiên cứu và phát triển.

Ngoài nước:

Trên trang web google code của google có hẳn một nhóm 50 người phát triển một phần mền về mô hình các chủ đề viết bằng ngôn ngữ java sử dụng thuật toán LDA và cho tải về miễn phí [9]giao diện người dùng như hình 2.1 phục vụ cho việt sử dụng và nghiên cứu. Công cụ hỗ trợ tạo ra những chủ đề với những từ thường xuyên xuất hiện cùng nhau, mô hình các chủ đề có thể kết nối các từ có nghĩa giống nhau và phân biệt giữa những từ nhiều nghĩa.



Hình 2.1. Công cụ mô hình các chủ đề của nhóm 50 người phát triển trên google code

Tuy nhiên công cụ không hỗ trợ tốt tiếng Việt, kết quả trả về là tập tinh HTML hơi chung chung nếu như muốn sử dụng phải chỉnh sửa lại theo đúng nhu cầu sử dụng.

Đề tài :"The Author-Topic Model for Authors and Documents" tạm dịch là "Mô hình tác giả - chủ đề cho tác giả và các tài liệu " của nhóm tác giả Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, Padhraic Smyth các tác giả này đều quốc tịch Mỹ địa chỉ trang Web tham khảo [1]

Đề tài sử dụng thuật toan LDA phiên bản năm 2003 , đề tài thể hiện mối quan hệ giữa tác giả với các tài liệu và mỗi tài liệu thì lại có nhiều tác giả,….

Đề tài : "Distributed Algorithms for Topic Models" tạm dịch là "Thuật toán phân tán cho mô hình các chủ đề" của nhóm tác giả David Newman, Arthur Asuncion, Padhraic Smyth, Max Welling của trường đại học khoa học máy tính California USA. Địa chỉ trang web tham khảo [2].

Đề tài mô tả thuật toán phân tán hai mô hình các chủ đề mô hình LDA vàmô hình HDP, đề tài mô tả thuật thuật toán phân tán và phân chia dữ liệu xử lý riêng biệt và song song.

Đề tài "Interactive Topic Modeling "tạm dịch là " Tích hợp mô hình các chủ đề " của nhóm tác giả Yuening Hu, Jordan Boyd-Graber, Brianna Satinoffthuộc trường đại học Maryland [3].

Đề tài mô tả một framework cho phép người dùng định nghĩa lại chủ đề bởi những mô hình như LDA bằng cách đưa thêm những ràng buộc tập hợp các từ phải xuất hiện cùng nhau trong cùngchủ đề.

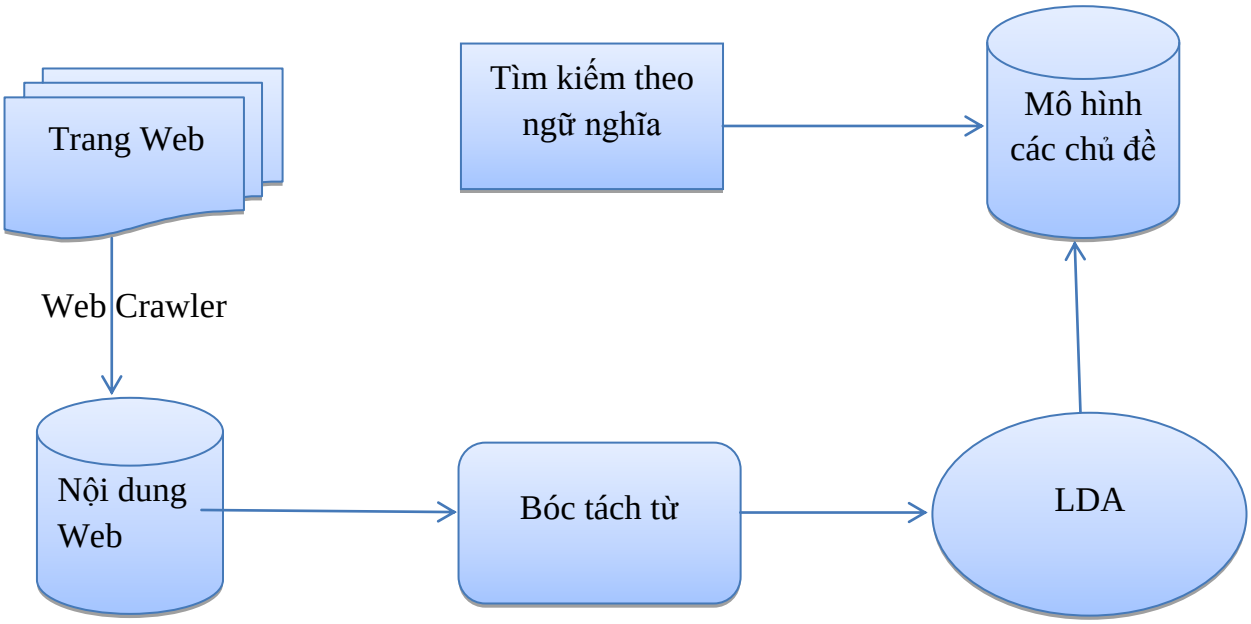
Hầu hết các mô hình trên áp dụng cho tiếng anh không hỗ trợ cho tiếng Việt nếu muốn sử dụng cho tiếng Việt phải dùng những thuật toán khác tạo ra những cụm từ tiếng Việt có nghĩa trước khi áp dụng những mô hình trên.Các bài báo trên cũng mô tả nhiều giải pháp tốt cho việc áp dụng mô hình các chủ đề cho các mục đích khác nhau tuy nhiên muốn có một mô hình các chủ đề phục vụ cho việc tìm kiếm theo ngữ nghĩa thì phải chuyển hóa lại những cái có sẵn theo mục đích tìm kiếm.

- Trong nước:

Hiện trong nước việc xây dựng mô hình các chủ đề chưa được phát triển nhiều, tác giả chỉ tìm hiểu được duy nhất mô hình các chủ đề JGibbLDA [10] của Nguyễn Cẩm Tú và Phan Xuân Hiếu, mô hình dùng để tìm các chủ đề cùng với các từ, cụm từ và trọng số của các từ, cụm từ trong mỗi chủ đề đó.Các công trình nghiên cứu về tìm kiếm theo ngữ nghĩa trong nước còn nhiều hạn chế.

2.3. Quy trình xây dựng mô hình các chủ đề và tìm kiếm theo ngữ nghĩa:

Qua tìm hiểu nghiên cứu, tác giả rút ra quy trình xây dựng mô hình các chủ đề phục vụ cho việc tìm kiếm tài liệu theo ngữ nghĩa.



Hình 2.2. Quy trình xây dựng mô hình các chủ đề và công cụ tìm kiếm theo ngữ nghĩa

Để tiến hành xây dựng mô hình các chủ đề hỗ trợ cho việc tìm kiếm theo ngữ nghĩa tác giả thực hiện các bước sau:

Bước 1: Dùng chương trình WebCrawler để tiến hành thu thập các nội dung web trên mạng để phục vụ cho việc xây dựng mô hình các chủ đềvà công cụ tìm kiếm theo ngữ nghĩa

Bước 2: Tiến hành bóc tách từ trong các bài báo gom nhóm các từ có nghĩa thành những từ hoặc cụm từ.

Bước 3: Dùng thuật toán LDA để tạo các chủ đề bao gồm các từ, số lần xuất hiện các tài liệu trong các chủ đề,v.v.

Bước 4: Dùng Ontologyxây dựng mô hình các chủ đề thể hiện mối liên hệ giữa các từ với các chủ đề, các chủ đề với các tài liệu,v.v.

Bước 5: Xây dựng công cụ tìm kiếm theo ngữ nghĩa dựa trên mô hình các chủ đề xây dựng.

2.4. Kết luận:

Từ những vấn đề trên cho ta thấy được việc tìm kiếm theo ngữ nghĩa trong tình hình dữ liệu lớn như hiện nay là rất cần thiết, tuy nhiên để phát triển nó cần có những quy trình phức tạp và đòi hỏi độ chính xác cao để cho được kết quả tìm kiếm cuối cùng có độ chính xác tốt nhất. Do độ phức tạp của việc xây dựng nên việc tìm kiếm theo ngữ nghĩa chưa được phát triển nhiều và rộng rãi ở trong nước, do đó việc xây dựng một công cụ tìm kiếm theo ngữ nghĩa ở thời điểm hiện tại sẽ góp phần thúc đẩy sự phát triển việc tìm kiếm theo ngữ nghĩa ở trong nước được phong phú hơn.

CHƯƠNG 3:MỘT SỐ KỸ THUẬT TRONG XÂY DỰNG MÔ HÌNH CÁC CHỦ ĐỀ VÀ TÌM KIẾM THEO NGỮ NGHĨA

3.1. Các kỹ thuật trong xây dựng mô hình các chủ đề và tìm kiếm theo ngữ nghĩa:

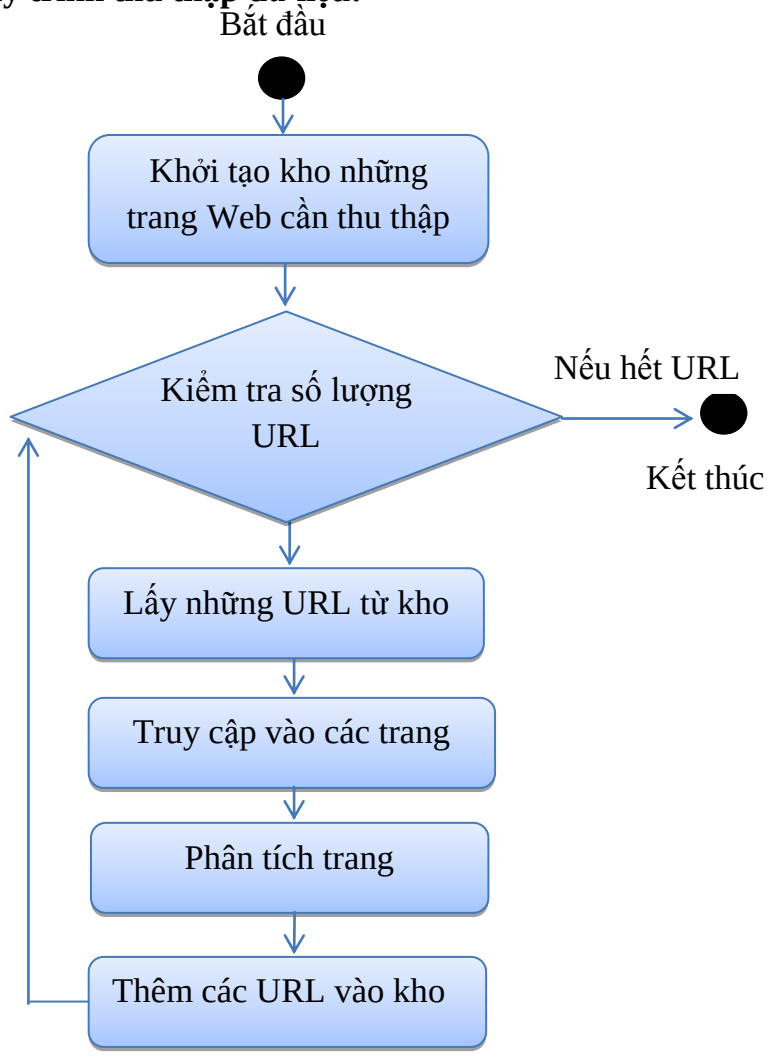
3.1.1. WebCrawler thu thập dữ liệu [4]:

Trình thu thập weblà một chương trình khai thác cấu trúc đồ thị của web di chuyển từ trang này qua trang khác. Thời kỳ đầu nó có những tên như bọ web, rô-bốt, nhện và sâu, nhưng ngày nay tên gọi phổ biến nhất là vẫn là trình thu thập web.

Động lực quan trọng thúc đẩy quá trình phát triển của việc thiết kế trình thu thập web là lấy được nội dung các trang web và thêm chúng hoặc đường dẫn của chúng vào một kho lưu trữ các trang – một kiểu kho lưu trữ có thể dùng để phục vụ cho các ứng dụng, cụ thể trong công cụ tìm kiếm web. Các trình thu thập thường bắt đầu bằng cách chọn một số các đường dẫn ứng với các trang web sẽ ghé thăm đầu tiên, các trang này được gọi là các trang hạt giống. Khi ghé thăm một trang hạt giống, trình thu thập sẽ đọc nội dung trang web, lọc ra tất cả các siêu liên kết có trong trang web đó và đưa các URL tương ứng với chúng vào một danh sách gọi là biên giới. Dựa vào danh sách này, trình thu thập tiếp tục quá trình duyệt đệ quy để ghé thăm tất cả các URL chưa được duyệt. Quá trình này chỉ dừng lại khi trình thu thập đã thu thập đủ số trang yêu cầu hoặc frontier là rỗng, tức là không còn URL để duyệt. Tuy mô tả này có vẻ đơn giản nhưng đằng sau chúng là khá nhiều vấn đề hóc búa liên quan đến kết nối mạng, bẫy nhện, tiêu chuẩn trích xuất URL, chuẩn hóa các trang HTML, bóc tách nội dung trang HTML... Sau khi đã có được một danh sách các URL dùng cho việc thu thập, ta sẽ thực hiện quá trình lấy trang. Tất cả các trang được lấy một lần và được lưu vào một kho lưu trữ giống như cơ sở dữ liệu của công cụ tìm kiếm, đến đây không cần thu thập thêm. Tuy nhiên web là một thực thể động với các không gian con liên tục phát triển và thay đổi nhanh một cách chóng mặt, vì thế thông tin phải liên tục được thu thập để giúp các ứng dụng luôn cập nhật, ví dụ như bổ sung các trang mới loại bỏ các trang đã bị xóa, di chuyển hoặc cập nhật các trang bị sửa đổi.

Hầu hết các trang web hiện nay chủ yếu được viết bằng các ngôn ngữ đánh dấu như HTML, XHTML và được nhắm đến đối tượng sử dụng là con người chứ không phải máy tính. Do đó, các trang web lại chứa đựng nhiều thông tin có ích mà con người có thể muốn thu thập và lưu trữ lại, vì vậy mà cần phải có các kỹ thuật bóc tách và trích xuất thông tin theo một cơ chế tự động. Các kỹ thuật bóc tách dữ liệu có thể ở mức đơn giản như việc bóc tách các siêu liên kết, hoặc ở mức phức tạp hơn một chút là bóc tách bất kỳ phần nội dung nào trong một trang web. Quá trình thu thập web chính là quá trình duyệt đệ quy một đồ thị. Các web được xem như là một đồ thị với các trang là các đỉnh và các siêu liên kết là các cạnh. Quá trình lấy trang và trích xuất các liên kết bên trong nó tương tự như việc mở rộng tìm kiếm một đỉnh trong đồ thị. Việc tìm kiếm này là khác nhau trong các trình thu thập sử dụng chiến lược tìm kiếm khác nhau.

3.1.2. Quy trình thu thập dữ liệu:



Hình 3.1. Vòng lặp thu thập dữ liệu từ Web

Trình thu thập chứa một danh sách các URL chưa được thăm gọi là biên giới . Danh sách được khởi tạo bởi một số các URL hạt giống – các URL này được cung cấp bởi một người dùng hoặc một chương trình khác. Mỗi vòng lặp là một quá trình gồm các bước :

- Lấy một URL tiếp theo từ frontier ra để thu thập.
- Lấy trang tương ứng với URL thông qua HTTP.
- Bóc tách trang vừa lấy để trích xuất ra các URL và các nội dung thông tin cụ thể.
- Cuối cùng là thêm các URL chưa thăm vào frontier.

Trước khi các URL được thêm vào frontier chúng có thể được đánh chỉ mục dựa trên số lượng truy cập vào trang web ứng với URL. Quá trình thu thập sẽ chấm dứt ngay khi trình thu thập đạt đủ số lượng trang nhất định hoặc frontier rỗng, đây được gọi là trạng thái kết thúc của trình thu thập.

3.1.3. Frontier:

Frontier là một danh sách chứa các URL của các trang chưa thăm. Trong thuật ngữ tìm kiếm đồ thị, frontier là một danh sách mở các đỉnh chưa được mở rộng. Đối với một trình thu thập lớn frontier có thể chứa hàng chục ngàn đến hàng trăm ngàn trang và phải lưu trữ trong ổ cứng. Tuy vậy frontier nào cũng có một miền giới hạn nhất định, miền giới hạn này lớn hay nhỏ phụ thuộc vào bộ nhớ của máy tính. Khi số lượng URL thu thập được vượt quá giới hạn này chúng ta sẽ cần một cơ chế để loại bỏ các URL ứng với các trang ít quan trọng và giữ lại các URL ứng với các trang quan trọng. Lưu ý rằng tốc độ thêm các URL vào frontier nhanh gần bằng tốc độ thu thập thông tin. Nó có thể thêm tới 60000 URL ngay khi trình thu thập thu thập dữ liệu của 10000 trang, giả định trung bình mỗi trang có khoảng 7 liên kết.

Frontier có thể coi như một hàng đợi làm việc theo cơ chế FIFO nghĩa là vào trước ra trước trong trường hợp chúng ta sử dụng thuật toán tìm kiếm theo chiều rộng để thu thập thông tin. Trình thu thập sử dụng chiến thuật tìm kiếm này gọi là trình thu thập theo chiều rộng . Các URL được lấy ra thu thập được chọn từ trên xuống dưới trong danh sách và các URL mới được thêm vào đuôi của danh sách.

Do miền giới hạn của frontier, ta phải đảm bảo các URL chỉ được lấy một lần. Để tìm kiếm xem một URL mới được trích xuất đã có trong danh sách chưa là khá phức tạp vì số lượng trang là rất lớn mỗi lần tìm kiếm là một lần chạy vòng lặp điều này là khá bất cập. Vì vậy có một giải pháp là sử dụng một phần bộ nhớ để duy trì một hàm băm với URL là khóa. Hàm băm này sẽ sinh ra các giá trị băm tương ứng với mỗi URL. Sở dĩ sử dụng hàm băm sẽ tìm kiếm nhanh hơn vì việc so sánh các giá trị băm nhanh hơn nhiều việc so sánh một giá trị với một khối dữ liệu lớn.

Hiện nay do bộ nhớ máy tính là rất lớn nên vấn đề về bộ nhớ là không mấy quan trọng so với vấn đề về tốc độ. Do vậy, cách sử dụng hàm băm được sử dụng rộng rãi vì tuy là tốn bộ nhớ hơn nhưng tốc độ tìm kiếm lại được cải thiện đáng kể.Khi frontier đạt đến miền giới hạn, thì các trình thu thập theo chiều rộng sẽ làm việc theo cơ chế sau : sau khi đưa một URL ra khỏi frontier để tiến hành quá trình thu thập trang tương ứng thay vì việc lấy tất cả URL trong trang này trình thu thập sẽ chỉ lấy URL chưa thăm đầu tiên và thêm vào frontier.Frontier có thể coi như một hàng đợi ưu tiên trong trường hợp chúng ta sử dụng thuật toán tìm kiếm theo lựa chọn tốt nhất . Trình thu thập sử dụng chiến thuật tìm kiếm này gọi là trình thu thập ưu tiên. Hàng đợi ưu tiên là một mảng với các phần tử là các URL được sắp xếp theo điểm đánh giá. Điểm đánh giá này được xác định dựa trên một số các phương pháp dựa trên kinh nghiệm. Trình thu thập ưu tiên sẽ làm việc theo cơ chế sau: URL được lấy ra khỏi frontier để tiến hành thu thập luôn là URL tốt nhất. Sau khi thu thập trang tương ứng, các URL được trích xuất ra được đưa vào frontier và các danh sách URL được sắp xếp lại theo điểm đánh giá. Để tránh việc trùng lặp URL chúng ta cũng duy trì một hàm băm với các khóa là URL để tra cứu. Khi frontier đạt đến miền giới hạn, cơ chế làm việc của trình thu thập tối ưu cũng giống với trình thu thập theo chiều rộng chỉ khác là các URL được lấy là các URL tốt nhất (là các URL có điểm đánh giá cao nhất).

Trong trường hợp trình thu thập nhận thấy frontier là danh sách rỗng (không thể lấy ra các URL tiếp theo để thu thập) thì quá trình thu thập sẽ kết thúc. Tuy vậy

trường hợp rất hiếm xảy ra vì với một số URL hạt giống và miền giới hạn khá lớn frontier hiếm khi đạt trạng thái rỗng.

Nhiều khi một trình thu thập có thể bắt gặp một bẫy nhệndẫn nó đến một lượng lớn các URL khác nhau nhưng trỏ đến cùng một trang web.Một cách để giảm bớt vấn đề này là hạn chế số lượng trang mà các trình thu thập truy cập từ một tên miền nhất định. Các mã liên kết với frontier có thể đảm bảo rằng trong một chuỗi liên tiếp các URL (khoảng 100 URL) trong frontier sẽ chỉ chứa một URL từ một tên miền máy chủ (ví dụ như www.cnn.com). Như vậy trình thu thập sẽ tốt hơn bởi không truy cập vào cùng một trang quá thường xuyên và các trang được thu thập cũng có xu hướng đa dạng hơn.

3.1.4. Cách lấy trang

Để lấy một trang web, chúng ta cần một máy khách HTTPgửi một yêu cầu HTTP cho trang đó và đọc các phản hồi . Client cần có thời gian trễ để đảm bảo rằng không bị mất thời gian không cần thiết vào các máy chủ chậm hoặc đọc các trang lớn. Trong thực tế chúng ta thường hạn chế vấn đề này bằng cách cho client tải về khoảng 10-20 KB đầu tiên của trang. Client cần bóc tách được tiêu đề phản hồi cho các mã trạng thái và chuyển hướng. Kiểm tra lỗi và xử lý ngoài luồng là rất quan trọng trong quá trình lấy trang vì chúng ta phải đối phó với hàng triệu máy chủ. Trong quá trình lấy trang, trình thu thập không thể tự quyết định tài liệu nào được lập chỉ mục và tài liệu nào không, do đó nó lấy tất cả những gì có thể. Thậm chí dù xác định được tài liệu vô ích thì nó cũng đã bỏ ra một chi phí đáng kể cho hoạt động thu thập. Tiêu chuẩn loại trừ robotra đời.

3.1.5. Bóc tách trang

Khi một trang đã được lấy, chúng ta cần phân tích nội dung của nó để trích xuất thông tin, lấy ra các URL để mở ra hướng đi tiếp theo của các trình thu thập. Phân tích nội dung có thể là quá trình khai thác hyperlink/URL đơn giản hoặc nó có thể bao gồm quá trình phức tạp hơn như lọc nội dung HTML để phân tích thành mô hình thẻ HTML dạng cây. Phân tích nội dung cũng có thể bao gồm các bước chuyển

đổi URL được trích xuất thành dạng tiêu chuẩn, loại bỏ những từ ở phần đầu nội dung của trang và lấy các từ còn lại ở phần thân.

3.1.6. Các chiến lược thu thập dữ liệu

Một số chiến lược thu thập dữ liệu bao gồm :

- Chiến lược thu thập dữ liệu theo chiều sâu : trong quá trình thu thập dữ liệu từ một trang web chương trình sẽ tiến hành ưu tiên thu thập dữ liệu trang con của URL trên trang web hiện tại rồi mới tiến hành thu thập trên các URL khác của trang đó.
- Chiến lược thu thập dữ liệu theo chiều rộng : trong quá trình thu thập dữ liệu từ một trang web chương trình sẽ tiến hành ưu tiên thu thập dữ liệu của các URL trên trang web hiện tại rồi mới tiến hành thu thập trên các con của URL đó.
- Chiến lược thu thập dữ liệu theo ngẫu nhiên : trong quá trình thu thập dữ liệu từ một trang web chương trình sẽ tiến hành thu thập dữ liệu của các URL trên trang web hiện tại hoặcthu thập trên các con của URL đó một cách ngẫu nhiên.
- Chiến lược thu thập dữ liệu theo lựa chọn tốt nhất ngây thơ: trong quá trình thu thập dữ liệu từ một trang web chương trình sẽ tiến hành thu thập dữ liệu của các URL trên trang web hiện tại hoặc thu thập trên các con của URL đó theo một thuật toán nào đó để quyết định URL nào sẽ duyệt tiếp theo là tốt nhất.

Quá trình thu thập web chính là quá trình duyệt đệ quy một đồ thị. Các web được xem như một đồ thị với các trang là các đỉnh và các siêu liên kết là các cạnh. Chính vì thế các chiến thuật thu thập dữ liệu cũng được xây dựng dựa trên các thuật toán tìm kiếm trên đồ thị. Các thuật toán tìm kiếm trên đồ thị bao gồm:

- Tìm kiếm theo chiều sâu: Là thuật toán tìm kiếm bằng cách mở rộng nút đồ thị theo chiều sâu.
- Tìm kiếm theo chiều rộng: Là thuật toán tìm kiếm bằng cách mở rộng nút đồ thị theo chiều rộng.
- Tìm kiếm theo lựa chọn tốt nhất: Là một thuật toán tìm kiếm tối ưu bằng cách mở rộng nút hứa hẹn nhất theo một quy tắc nào đó.

3.1.7. WebCrawler áp dụng cho luận văn:

Trên thế giới tồn tại nhiều công cụWebCrawler một sốcó phí như winwebcrawler [11],v.v.tuy nhiên cũng có một số open source miễn phínhư crawler4j [10],v.v.

Công cụ mã nguồn mở crawler4j [12] được viết bằng Java và hoàn toàn miễn phí với giao diện dễ sử dụng và dễ dàng can thiệp tùy chỉnh theo nhu cầu của mỗi người. Công cụ với một số đặc điểm nổi bật như :

Hỗ trợ chia thành nhiều luồn trong việc thu thập dữ liệu, xử lý dữ liệu,v.v. để tận dụng sức mạnh của CPU và sức mạnh của đường truyền mạng,v.v.

Công cụ hỗ trợ xử lý loại bỏ những hình ảnh, âm thanh,Video,v.v. Tuy nhiên nếu muốn giữ lại hình ảnh trên các trang web thu thập người dùng có thể tùy chỉnh để tải về một thư mục nào đó.

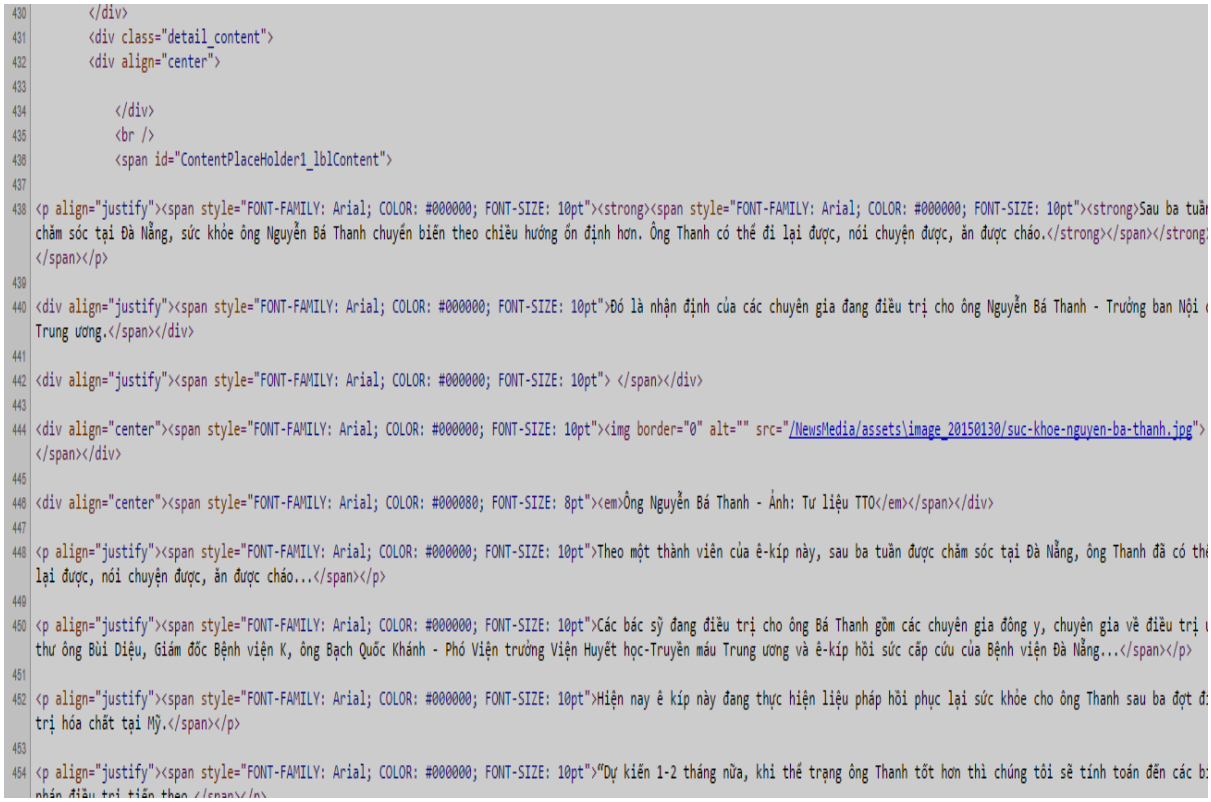
Để áp dụng công cụ WebCrawler cho luận văn cần phải chỉnh sửa lại công cụ theo mục đích riêng của luận văn như:

Định nghĩa cho công cụ biết trang web mình cần lấy thông tin về.

Tải về thư viện (jsoup-1.7.3.jar)hỗ trợ xử lý DOM để phân tích các trang web mà công cụ tải vềvà lấy ra các nội dung cần thiết.

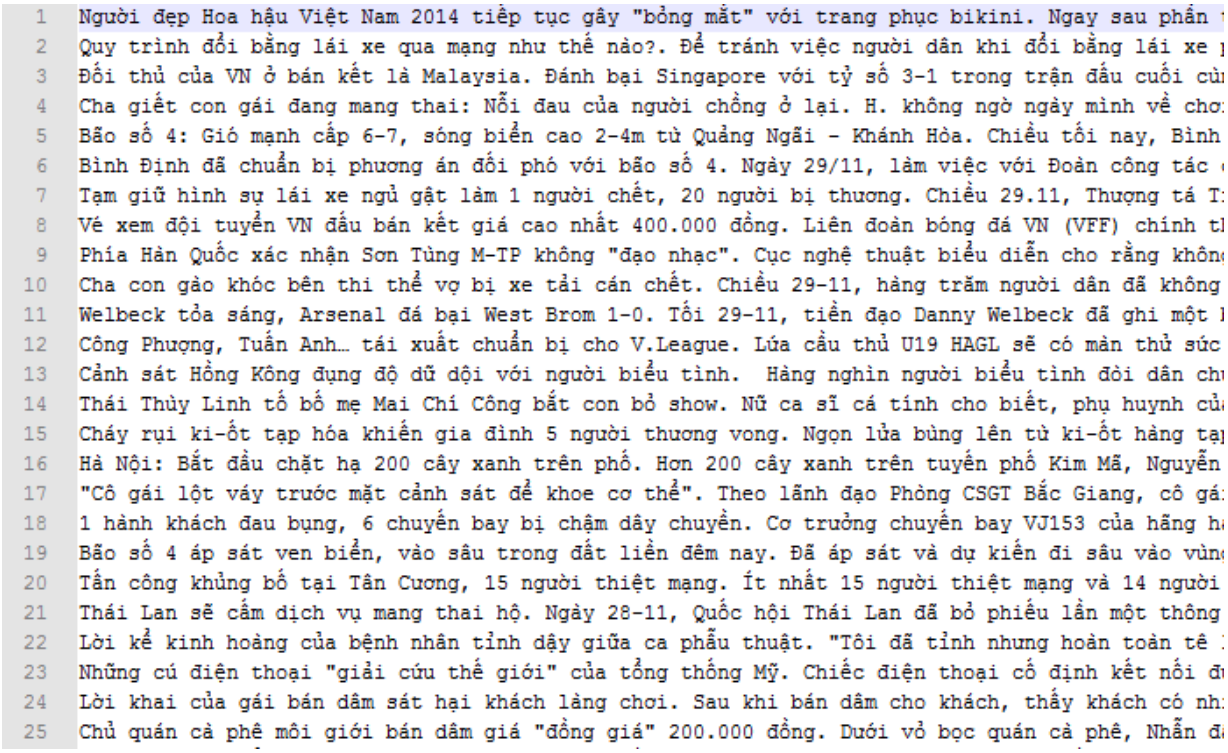
Chỉnh sửa Code lại giúp các kết quả lấy về xuất ra các tập tin XML phục vụ cho việc xử lý sau này cho luận văn.

Sau khi dùng công cụ WebCrawler thu thập dữ liệuchúng ta sẽ được các kết quả như sau:

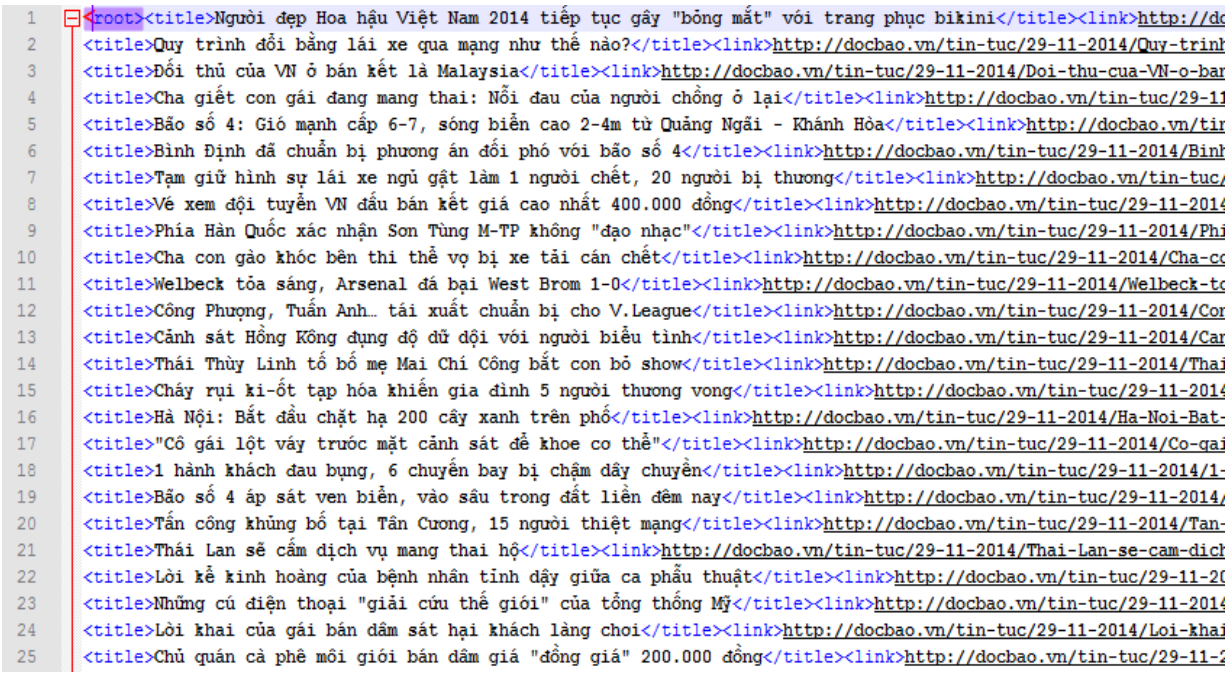


Hình 3.2. Dữ liệu lấy về bằng WebCrawler

Sau khi WebCrawler thu thập dữ liệu về dạng XML tác giả xử lý dữ liệu trên theo dạng XML để đọc ra các nội dung cần thiết cho luận văn bao gồm: tiêu đề, phần mô tả ngắn của bài báo, nội dung bài báo và ghi các nội dung đó ra một tập tin văn bản và tập tin XML phục vụ cho luận văn như sau:



Hình 3.3. Dữ liệu lấy về bằng WebCrawler sau khi đã xử lý



Hình 3.4. Tiêu đề và liên kết trang được lưu tập tin khác dưới dạng XML

3.2.Xử lý văn bản:

3.2.1. Đặc điểm của từ trong Việt:

Tiếng Việt là ngôn ngữđơn lập.Đặcđiểm này bao quát tiếng Việt cả về mặt ngữ âm, ngữnghĩa, ngữ pháp. Khác với các ngôn ngữ châu Âu, mỗi từlà một nhóm các ký tự có nghĩa được cách nhau bởi một khoảng trắng. Còn tiếng Việt, và các ngôn ngữđơn lập khác, thì khoảngtrắng không phải là căn cứđểnhận diện từ.

- Tiếng:

Trong tiếng Việt trước hết cần chú ýđếnđơn vị xưa nay vẫn quan gọi là tiếng. Về mặt ngữnghĩa, ngữâm, ngữpháp, đều có giá trịquan trọng.

Sửdụng tiếng đểtạo từcó hai trường hợp:

Trườnghợpmộttiếng:đâylàtrường hợpmộttiếngđược dùnglàm một từ, gọi là từ đơn. Tuy nhiên không phải tiếng nào cũng tạo thành một từ.

Trường hợp hai tiếng trởlên: đây là trường hợp hai hay nhiều tiếng kết hợp với nhau, cảkhối kết hợp với nhau gắn bó tươngđối chặt chẽ, mới có tưcách ngữpháp là một từ. Đây là trường hợp từghép hay từphức.

- Từ:

Có rất nhiều quan niệm về từ trong tiếng Việt, từ nhiều quan niệm về từ tiếng Việt khác nhauđó chúng ta có thể thấyđặc trưng cơ bản của "từ" là sự hoàn chỉnh về mặt nội dung, từ là đơn vị nhỏ nhất để đặt câu.

Người ta dùng "từ" kết hợp thành câu chứ không phải dùng "tiếng", do đó quá trình tách câu thành các "từ" cho kết quả tốt hơn là tách câu bằng “tiếng”.

3.2.2. Kỹ thuật tách từ trong tiếng Việt:

Luận văn sử dụng công cụ mã nguồn mở tách từ trong tiếng Việt được phát triển bởi Nguyễn Cẩm Tú và Phan Xuân Hiếu [13].

Công cụ sử dụng 2000 tên cá nhân và 707 tên của các vùng của Việt Nam đây được xem như là từ điển hỗ trợ mô hình CRF và SVM [5].

Bảng 3.1. Nội dung hỗ trợ mô hình CRF và SVM

Số thứ tự	Tên miền	Số tài liệu
1	Kinh tế	90
2	Công nghệ thông tin	59
3	Giáo dục	38
4	Xe cộ	35
5	Thể thao	28
6	Luật	31
7	Văn hóa xã hội	24
Tổng Cộng	**305 bài bào báo(7800 câu)**	

Vì công cụ là mã nguồn mở nên ta dễ dàng tùy chỉnh theo nhu cầu riêng của mình, Sau khi WebCrawer tiến hành thu thập dữ liệu, công cụ sẽ được lập trình thêm tính năngloại bỏ các stopword rồi tiến hành tách từ thành những cụm từ có nghĩa.

3.2.3. Công cụ áp dụng cho việc tách từ trong tiếng Việt:

Hiện ở trong nước lĩnh vực tách từ trong tiếng Việt khá phổ biến trong đó có hai công cụ tách từ khá mạnh là:

VNTokenizer của Lê Hồng Phương trường đại học khoa học tự nhiên Hà Nội [14]. Công cụ hướng dẫn đầy đủ cho việc chạy trên CMD của window hoặc dưới dạng thư viện hỗ trợ cho công việc lập trình có thể gọi các API đó và nó trả về kết quả,….Theo như tác giả công cụ có thể chạy tốt trên nền Window, Linux, Unix,…Theo tác giả thì kết quả thử nghiệm đạt được 98% trên kết quả tác giả đã thử nghiệm.

JvnSegmenter của Nguyễn Cẩm Tú trường đại học quốc tế Hà Nội và Phan Xuân Hiếu trường đại học khoa học thông tin Tohoku [13]công cụ mã nguồn mở viết bằng ngôn ngữ Java hỗ trợ lập trình tích hợp vào ứng dụng khác khá tốt và dễ dàng.

Sau khi tiến hành thử nghiệm và so sánh hai công cụ trên luận văn quyết định chọn công cụ JvnSegmenter do việc tách từ có độ chính xác nổi trội hơn VNTokenizer, công cụ mã nguồn mở Java hỗ trợ lập trình và tích hợp vào ứng dụng của luận văn tốt và dễ dàng,….

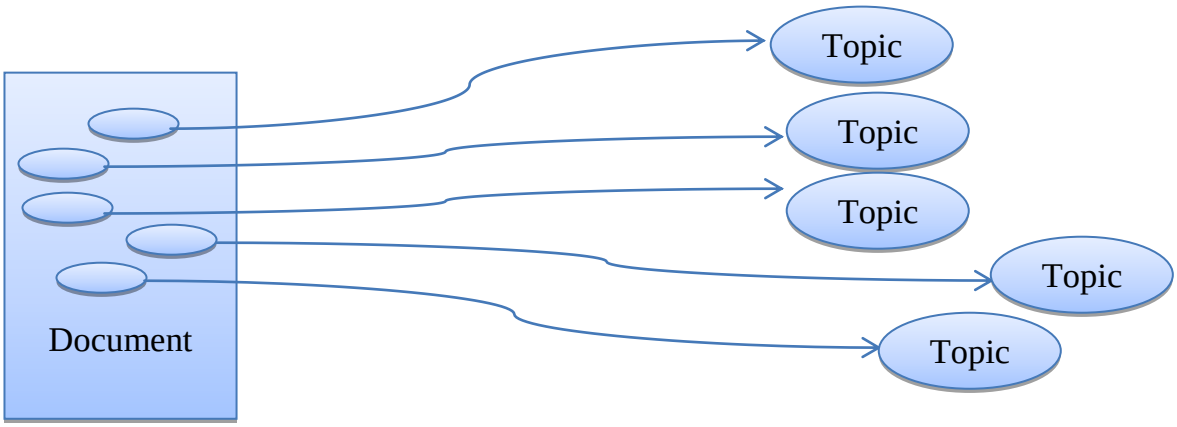
3.3. Phân chia các chủ đề và tính trọng số các từ trong chủ đề:

3.3.1. Thuật toán LatentDirichlet Allocation [6]:

LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc như text corpora. LDA dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủđề. Về bản chất, LDA là một mô hình Bayesian 3 cấp (corpus level, document level, word level) trong đó mỗi phần của mô hình được coi như một mô hình trộn hữu hạn trên cơ sở tập hợp các xác suất chủ đề.

Hofmann[6]người đã trình bày mô hình xác suất LSI còn được gọi là mô hình theaspect như là thay thế mô hình LSI. Do đó mỗi từ được tạo ra từ một chủ đề duy nhất, và các từ khác nhau trong một tài liệu có thể được tạo ra từ các chủ đề khác nhau. Mỗi tài liệu được biểu diễn như là một danh sách các tỷ lệ pha trộn cho các thành phần hỗn hợp và do đó giảm xuống còn một phân bố xác suất trên một tập cố định các chủ đề. Phân phối này là "giảm mô tả" liên quan đến tài liệu.

LDA là một mô hình sác xuấtcủa một corpus . Đó là tài liệu biểu diễn như một hỗn hợp ngẫu nhiên của các chủ đề tìm ẩn. Mỗi chủ đề có đặc điểm bởi sự phân phối của các từ.



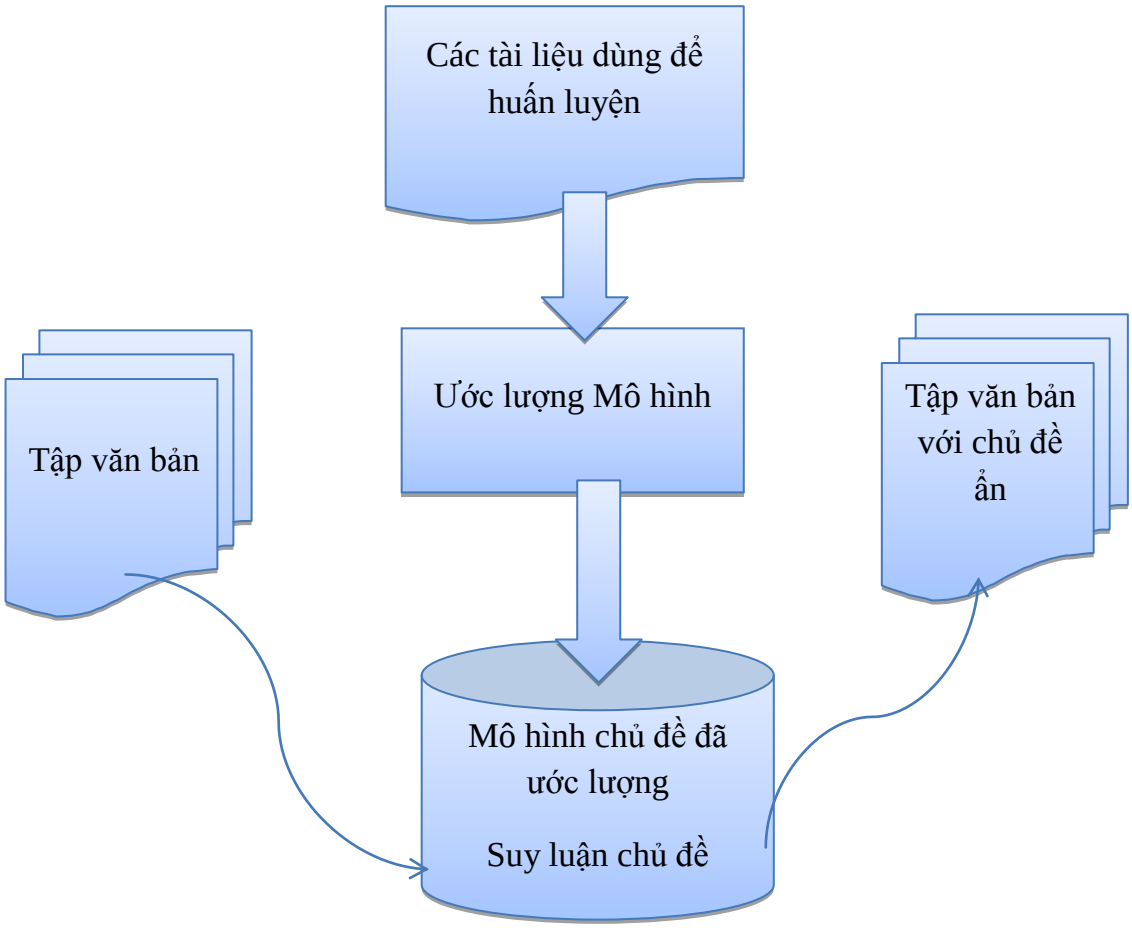
Hình 3.5. Phân chia chủ đề của một tài liệu

3.3.1.1.Suy luận chủ đề:

TheoNguyễnCẩmTú [7],vớimộtmôhìnhchủđềđãđượchuấn luyện tốt dựa trên tập dữ liệu toàn thể bao phủ miềnứng dụng, ta có thể thực hiện một tiến trình quá trình suy diễn chủ đề cho các tài liệu mới tương tự như quá trìnhước lượng tham số(là xácđịnhđược phân phối trên các chủđềcủa tài liệu qua tham sốtheta). Tác

giảcũng chỉ ra rằng sử dụng dữ liệu từ trangVnExpress.net huấnluyệnđượccácmôhìnhcóuthế hơn trong các phân tích chủđề trên dữ liệu tin tức, trong khi các mô hình được huấn luyện bởi dữ liệu từ Wiki tốt hơn trong phân tích chủđề các tài liệu mang tính học thuật.

Dựatrênnhữngnghiêncứuđó,tácgiả chọnmôhìnhchủđược huấnluyện bởi tậpdữ liệutoànthể thuthậptừ trangweb bất kỳcho phân tích chủđề. Một tiến trình phân tích chủđềtổng quátđược minh họa như sau:



Hình 3.6. Mô hình suy luận chủ đề

Với mô hình trên đầu tiên Nguyễn Cẩm Tú đã sưu tập các tài liệu thuộc nhiều lĩnh vực khác nhau để làm nguồn dữ liệu ước lượng cho các tập tin đầu vào, sau khi đã có dữ liệu để ước lượng thì các tập tin văn bản đưa vào sẽ được ước lượng ra mô hình các chủ đề tương ứng với nguồn dữliệu ước lượng. Như vậy với mô hình trên thì mô hình các chủ đề được tạo ra phụ thuộc vào nguồn dữ liệu dùng để ước

lượngnhư vậy dữ liệu đưa vào ước lượng càng phong phú thì độ chính xác của mô hình các chủ đề càng chính xác.

Công cụJGibbsLDA của Nguyễn Cẩm Tú đã hiện thực quá trìnhước lượng và suy luận chủđểẩn cho kết quả rất tốt, tác giả sử dụng công cụ nàyđể xây dựng tập đặc trưng cho từng thểloại và thu được kết quả khả quan.

3.3.1.2.Các kết quả thu được từ công cụ JGibbsLDA:

Sau khi thu thập được dữ liệu từ Internet thông qua công cụ WebCrawler dữ liệu được phân loại thành những cụm từ có nghĩa và được đưa vào công cụ JGibbsLDA để thu về các chủ đề và các từ cùng với trọng số của nó trong chủ đề đó. Tuy nhiên để dữ liệu trả về phù hợp với nhu cầu sử dụng của luận văn chúng ta cần chỉnh sửa lại những đoạn mã của công cụ đểkết quả trả về những định dạng phù hợp với mục đích sử dụng của luận văn. Để phục vụ dữ liệu cho luận văn cần một cấu trúc XML như sau:

<root>: Dùng để quản lý tất cả nội dung XML

<topic>: Dùng để chứa các chủ đề và các từ trong chủ đề đó.

<topicname>: Tên chủ đề

<worddetail>: Chứa các thông tin chi tiết của một từ trong chủ đề

<word>: Chứa một từ trong chủ đề

<rate>: Trọng số của từ đó trong chủ đề

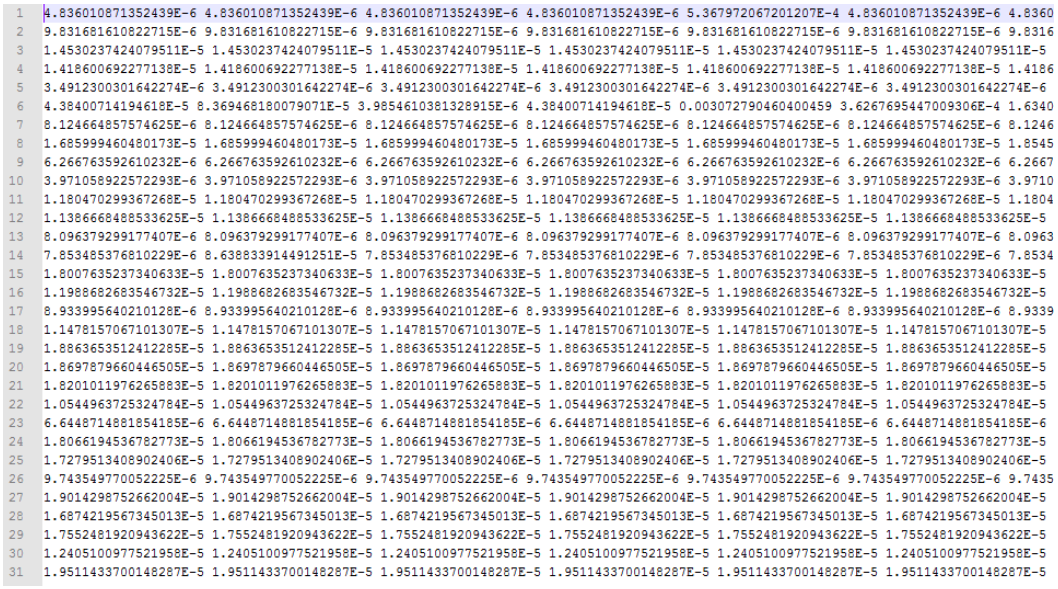
Kết quả trả về sau khi xử lý có dạng như sau:

```
<root><topic><topicname>Topic_0th</topicname>
    <worddetail><word></word><rate>0.2537503264307338</rate></worddetail>
    <worddetail><word>được</word><rate>0.017317754930313085</rate></worddetail>
    <worddetail><word> </word><rate>0.01606039210376145</rate></worddetail>
    <worddetail><word>người</word><rate>0.010498979601706144</rate></worddetail>
    <worddetail><word>Theo</word><rate>0.008612935361878692</rate></worddetail>
    <worddetail><word>từ</word><rate>0.0084678550357381 2</rate></worddetail>
    <worddetail><word>đến</word><rate>0.008419494927024596</rate></worddetail>
    <worddetail><word>khi</word><rate>0.007984253948602875</rate></worddetail>
    <worddetail><word>Việt_Nam</word><rate>0.007694093296321729</rate></worddetail>
    <worddetail><word>3</word><rate>0.007355572535327059</rate></worddetail>
    <worddetail><word>có</word><rate>0.006001489491348376</rate></worddetail>
    <worddetail><word>mạng</word><rate>0.005034287317077888</rate></worddetail>
    <worddetail><word>theo</word><rate>0.004792486773510267</rate></worddetail>
    <worddetail><word>số</word><rate>0.004744126664796742</rate></worddetail>
    <worddetail><word>ở</word><rate>0.004599046338656169</rate></worddetail>
    <worddetail><word>mang</word><rate>0.00450232612122912</rate></worddetail>
    <worddetail><word>đặt</word><rate>0.00372856438181273</rate></worddetail>
    <worddetail><word>trường</word><rate>0.003631844164385 6812</rate></worddetail>
    <worddetail><word>có_thể</word><rate>0.0035351239469586324</rate></worddetail>
    <worddetail><word>trang</word><rate>0.003438403729531 5836</rate></worddetail>
    <worddetail><word>phim</word><rate>0.00339004362081 8059</rate></worddetail>
    <worddetail><word>nhận</word><rate>0.003293323403391 0106</rate></worddetail>
    <worddetail><word>cần</word><rate>0.003196603185963962</rate></worddetail>
    <worddetail><word>đó,</word><rate>0.003051522859823389</rate></worddetail>
    <worddetail><word>1</word><rate>0.00295480264239634</rate></worddetail>
```

Hình 3.7. Kết quả thu được từ LDA

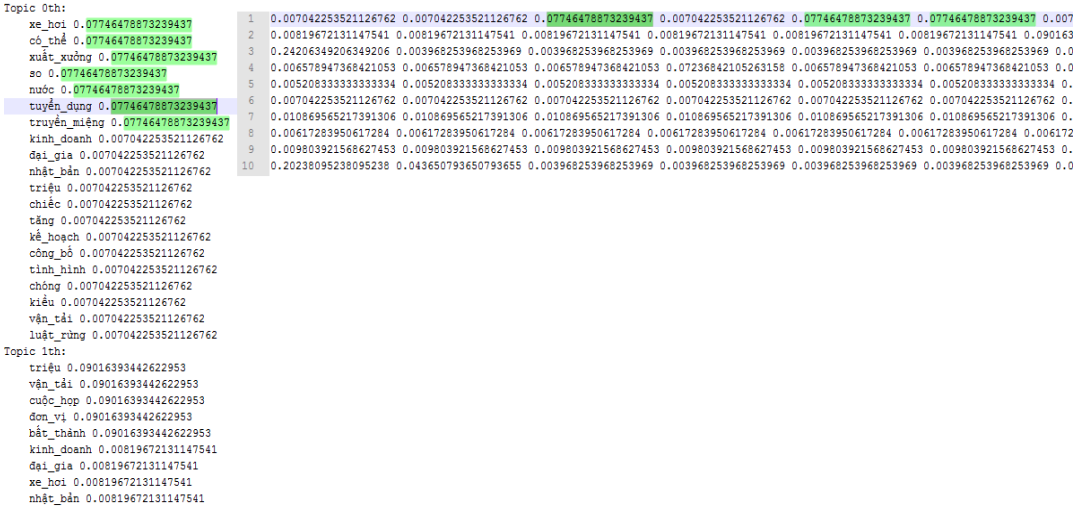
Ngoài việc trả về những chủ đề và các trọng số của các từ trong chủ đề đó công cụ còn trả về một số các thông tin khác như:

Trọng số của mỗi từ trong một chủ đề mỗi dòng là một chủ đề và mỗi cột là một từ:



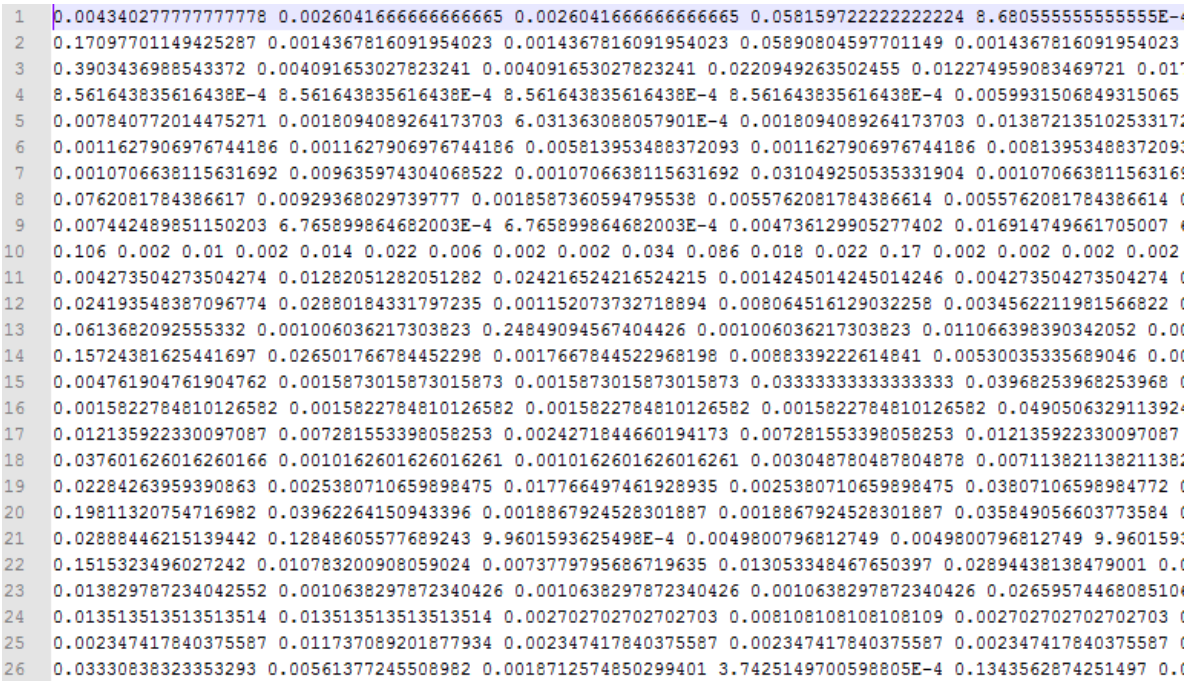
Hình 3.8. Trọng số của mỗi từ trong một chủ đề

Ví dụ trong hình sau tác giả chia tài liệu ra làm 10 chủ đề thì sẽ thu được 2 tập tin, tập tin thứ nhất gồm chủ đề và các từ cùng trọng số của nó trong chủ đề, tập tin thứ 2 sẽ bao gồm mỗi dòng là một chủ đề và mỗi cột là trọng số của một từ trên một chủ đề như sau:



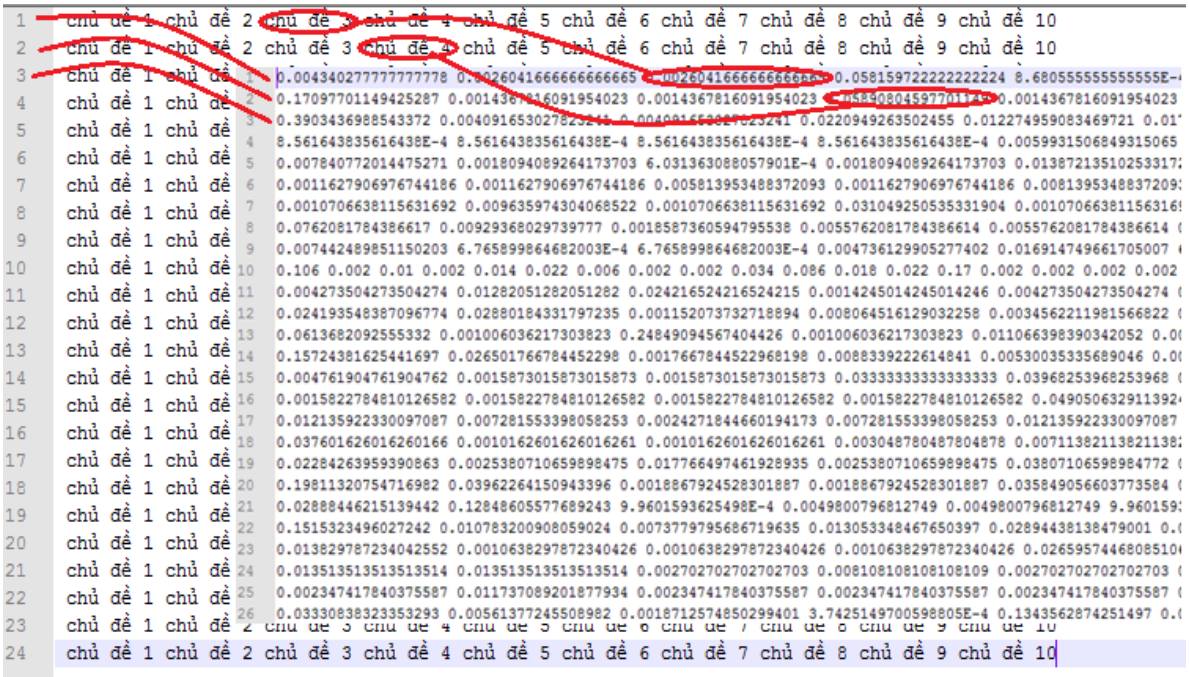
Hình 3.9. Mô tả dữ liệu thu được và trọng số của mỗi từ trong một chủ đề của 2 tập tin

Mối liên hệ giữa tài liệu và chủ đề mỗi dòng là một tài liệu và mỗi cột là trọng số một chủ đề:



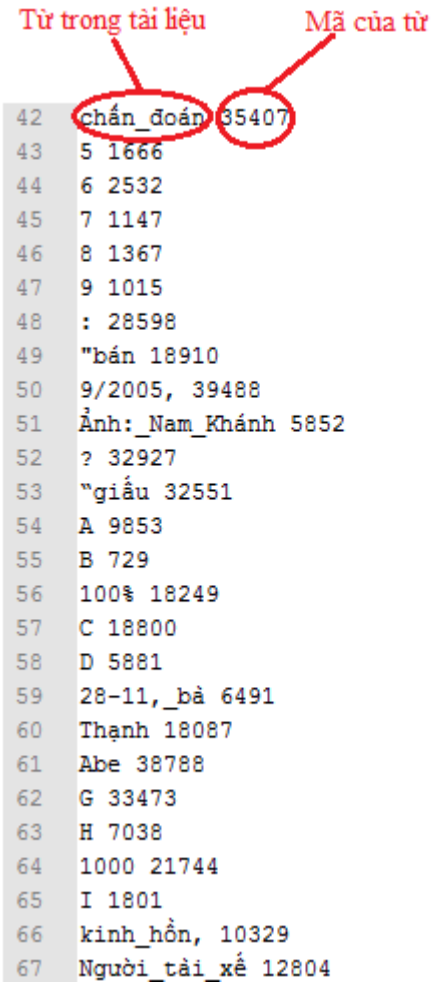
Hình 3.10. Trọng số của chủ đề trong tài liệu

Ví dụ sau đây sẽ cho ta thấy được mối liên hệ giữa tập tin trả về và tài liệu như mỏi dòng là một tài liệu và mỏi cột là một trọng số của chủ đề



Hình 3.11. Mối quan hệ giữa chủ đề và tài liệu

Ngoài ra công cụ còn trả về các mã của những từ trong một tập tin riêng, tập tin này gồm 2 cột, cột đầu tiên mô tả từ và cột thứ 2 là mã của từ đó, do công cụ quả lý các từ theo dạng mã cho từng từ để tiện việc sử lý và hiển thị:



Hình 3.12. Mã của các từ trong tài liệu

Ngoài những thông tin này công cụ còn trả về những thông tin khác hỗ trợ cho luận văn,v.v.

3.4.Web ngữ nghĩa [15]:

3.4.1. Tìm hiểu web ngữ nghĩa:

Tháng 12/1991, tạihội nghịHypertext91 ở San Antonio, lần đầu tiênTimBerners-Lee đưa ra khái niệm World Wide Web. Phát minh này có thể được xem là cột mốc làm thay đổi cách giao tiếp của con người vớidữliệu trên Internet, kéo theo sựra đời củacác trình duyệt Webnhư Mosaic(1993) hay Netcape

(1995)..Thaycho cácthao tácphức tạpbằng dòng lệnh, con người đã có thể truy cập các hìnhảnh đồhọa,di chuyển giữa các trang Web chỉbằng một cú click chuột. World Wide Web đã bùng nổngay sau đó. Từsốlượng khiêm tốn ban đầu, hàng triệu trang Web ra đời đã làm cho Internet trởthành một kho dữliệu khổnglồ và hỗn độn.Hệlụy kéo theo là việc tìm kiếm thông tin trên Web cũng trởnên khó khăn hơn.Con người thường xuyên phải đối đầu với mộtlượnglớnnhữngthôngtinkhônghợplýhoặckhôngliênquanđượctrảvềtừkếtquảtìm kiếm. Nguyên nhân lý giải cho thực tế trên xuất phát từ chính sự đơn giản của Web hiện tại, đã cản trở sự phát triển thông tin của nó. Trong mô hình này, các máy tính chỉlàm nhiệm vụ gởi nhận dữ liệu và thểhiện thông tin dưới dạng thô mà chỉcon người mớiđọchiểu được.Kết quảtất yếu là chính con người phải làm nhiệm vụ suy luận, tổng hợp và rút trích mọi thông tinmình cần.

Điều đó đã đặc ra thách thức làm sao để khai thác thông tin trên Web một cách hiệu quả, mà cụ thể là làm thếnào đểmáy tính có thểtrợgiúp xửlý tựđộng được chúng.Muốn vậy, Web phải có khả năng mô tả các sự vật theo cách mà máy tính có thể "hiểu" được. Động từ "hiểu" ở đây có ý nghĩahạn chế.Trong điều kiện hiện tại, nó dùng đểchỉkhảnăng của máy tính có thểphân tích cấu trúc của dữ liệu, xác định xem dữ liệu đó thuộc loại nào và từ đó có các hành động thích hợp. Lấy ví dụ,khi có yêu cầu tìm kiếm với từkhóa "Bill Clinton", máy tính mà cụthểlà ứng dụng chạy trên máy tính cần thể hiện các kết quả cho biết Clinton là cựu tổng tổng thống Mỹ, chứ không phải tổng thống Mỹ hiện tại hay một người có tên Bill Clinton nào khác.

Thách thức trên đã thúc đẩysựra đời của ý tưởng "Web có ngữnghĩa",thếhệtiếp theo của Web mà lộ trình của nó đã được Tim Berners-Lee phát thảo từ năm 1998. Theo Lee, "Web có ngữ nghĩa là sựmởrộng của Web hiện tại mà trong đó thông tin được định nghĩa rõ ràng sao cho con người và máy tính có thểcùng làm việc với nhau một cách hiệu quả hơn". Theo đó, mục tiêu của Web ngữnghĩa là phát triển các chuẩn chung về công nghệ, cải tiến Web hiện tại bằng cách thêm vào một lớp ngữ nghĩa đểmáy tính có thểhiểu được thông tin trên Webnhiều hơn, tăng cường khảnăng rút trích thông tin một cách tự động, tích hợp dữ liệu.

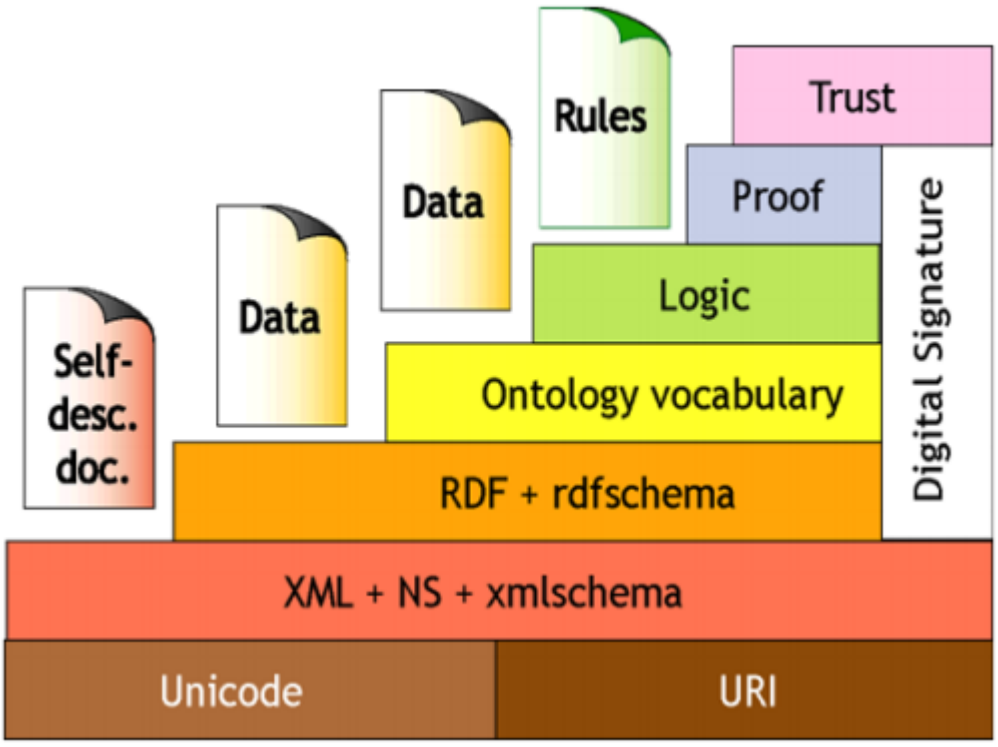
Có thể hình dung một số lợi ích của Web ngữ nghĩa so với Web hiện tại qua một số điểm sau:

- Máy tính có thểhiểu được thông tin trên Web:Web ngữnghĩa định nghĩa các khái niệm và bổ sung quan hệ dưới dạng máy tính có thểhiểu được. Do đó, việc tìm kiếm, đánh giá, xử lý, tích hợp thông tin có thể được tiến hành một cách tự động.
- Thông tin được tìm kiếm nhanh chóng và chính xác hơn:với Web ngữnghĩa, máy tính có thểxác định một thực thểthuộc lớp hay thuộc tính cụthểnào dựa trên ngữcảnh chứa nó. Do đó thu hẹp không gian tìm kiếm và cho kết quả nhanh, chính xác hơn.
- Khảnăng suy luận thông minh: dựa vào các luật suy diễn trên cơ sởtri thức vềcác thực thể, máy tính có khảnăng sinh ra những kết luận mới.Ứng dụng Web tương lai có thể sẽ trả lời được những câu hỏi kiểu như:"Tổng thống thứ42 của Hoa Kỳđã bình thường hóa quan hệ với nước nào vào năm 1995?".
- Dữ liệu liên kết động: thay thế cách liên kết sử dụng hyperlink tĩnh trong Web cũ, Web ngữ nghĩa liên kết dữliệu từnhiều nguồn khác nhaumột cách hiệu quảhơn dựa trên định danh của tài nguyênvà quan hệ giữa chúng. Cách liên kết này đôi khi còn được gọi là liên kết bằng siêu dữliệu .

3.4.2. Kiến trúc Web ngữ nghĩa:

Đểcóđượcnhữngkhảnăngtrên,Webngữnghĩa cần một hạ tầng chặt chẽvới nhiều lớp hỗtrợbêndưới. Mỗilớpcóvaitrònhấtđịnh:ởdướicùnglàlớpUnicodevàURInhằmđảm bảoviệcsửdụngtậpkýhiệuquốctếvàxác định cáctàinguyêntrênmạng.Kếđếnlàlớp XML,cung cấp cú pháp chungnhưng không ràng buộc vềngữnghĩa cho các tài liệu có cấu trúc, làm cơ sở cho sựtraođổidữliệutrênWeb.Bêntrên,lớp XML SCHEMA định nghĩa cấu trúc các tài liệu XML,chophépmởrộngXMLbằngcáckiểu dữ liệumới.Tiếptheo XML Schemalà lớp RDF,cungcấp cấu trúc mô tả các đối tượng hay tài nguyên trên m ạng và quan hệgiữa chúng. RDF cho phép gán kiểu cho các tài nguyên và làm nền tảng cho Ontology sẽđược nói trong phần tiếp theo. RDF và Ontologychínhlà hai thành phần quan trọng nhất trong kiến trúc Web ngữnghĩa. Kế

đến,l ớp RDF SCHEMAcung cấp một phương tiện để đặc tảcáctừvựngmôtảtínhchấtvàquan hệgiữacáctài nguyên RDF.Lớp tiếp theo trong mô hình phân cấp này là ONTOLOGY định nghĩacác từ vựng dùng đểmô tảcácthuộc tính, lớp trong mộtmiền ngữ vựng nhất định. Cuối cùng, Lớp LOGIC cung cấp các luật suy diễn, trong khi PROOF sử dụng các luật của lớp Logic đểkiểm tra tính đúng đắn của một suy diễn nào đó. Hailớp trên trong kiến trúc đãthểhiệnrõ hơngóc độngữnghĩa vàcung cấp chomô hìnhnàykhảnăng suy luận thông minh.Lớp TRUST hiện vẫn đang trong giai đoạn phát triển,nhằm mục đíchđánh giá mức độtin cậy và quyết định có nên tin tưởng cácbằng cớtừmộtkết quảsuy luậnnào đóhay không. Thông thường Trust chính là một hàm lượng giá áp dụng trên một tập các thông tin, thông tin nào có giá trị lượng giá cao hơn sẽ được chọn cho một mục đích nào đó, ví dụ để thể hiện trong kết quả tìm kiếm chẳng hạn.



Hình 3.13. Kiến trúc web ngữ nghĩa

- Lớp **Unicode & URI:** Bảo đảm việc sử dụng tập kí tự quốc tế và cung cấp phương tiện nhằm định danh các đối tượng trong Semantic Web. URI đơn giản chỉ là

một định danh Web giống như các chuỗi bắt đầu bằng “http” hay “ftp” mà chúng ta thường xuyên thấy trên mạng (ví dụ:http://www.cadkas.com). Bất kỳ ai cũng có thể tạo một URI, và có quyền sở hữu chúng. Vì vậy chúng đã hình thành nên một công nghệ nền tảng lý tưởng để xây dựng một hệ thống mạng toàn cầu thông qua đó.

- Lớp **XML** cùng với các định nghĩa về *namespace*và *schema*bảo đảm rằng chúng ta có thể tích hợp các định nghĩa Web ngữ nghĩa với các chuẩn dựa trên XML khác.

- Lớp **RDF [RDF] và RDFSchema [RDFS]:** ta có thể tạo các câu lệnhđể mô tả các đối tượng với những từ vựng và định nghĩa của URI, và các đối tượng này có thể được tham chiếu đến bởi những từ vựng và định nghĩa của URI ở trên. Đây cũng là lớp mà chúng ta có thể gán các kiểu cho các tài nguyên và liên kết. Và cũng là lớp quan trọng nhất trong kiến trúc Semantic Web .

- Lớp **Ontology**: hỗ trợ sự tiến hóa của từ vựng vì nó có thể định nghĩa mối liên hệ giữa các khái niệm khác nhau. Một Ontology định nghĩa một bộ từ vựng mang tính phổ biến & thông thường, nó cho phép các nhà nghiên cứu chia sẻ thông tin trong một hay nhiều lĩnh vực.

- Lớp **Digital Signature**: được dùng để xác định chủ thể của tài liệu (ví dụ: tác giảhay nhan đề của một loại tài liệu).

- Các lớp **Logic, Proof, Trust:** Lớp logic cho phép viết ra các luật trong khi lớp proofthi hành các luật và cùng với lớp trust đánh giá nhằm quyết định nên hay không nên chấp nhận những vấn đề đã thử nghiệm.

3.4.2.1.Giới thiệu RDF:

RDF hay khung mô tảtài nguyên, là nền tảng cho việc biểu diễn dữliệu tronglĩnh vực Web có ngữnghĩa. Thông tin biểudiễn theo mô hình RDF là một phátbiểu ởdạngcấutrúcbộbavìnógồmbathànhphầncơbảnlà: (subject,predicate,object). Trong đó:

- Subject chỉđối tượng đang được mô tả đóng vai trò là chủ thể.
- Predicate (còn được gọi là property) là kiểu thuộc tính hay quan hệ.
- Objectlàgiá trị thuộc tínhhay đối tượngcủa chủthểđã nêu.Objectcó thểlàmộtgiá trịnguyên thủy như số nguyên, chuỗi ,..hoặc cũng có thểlàmột tài nguyên.

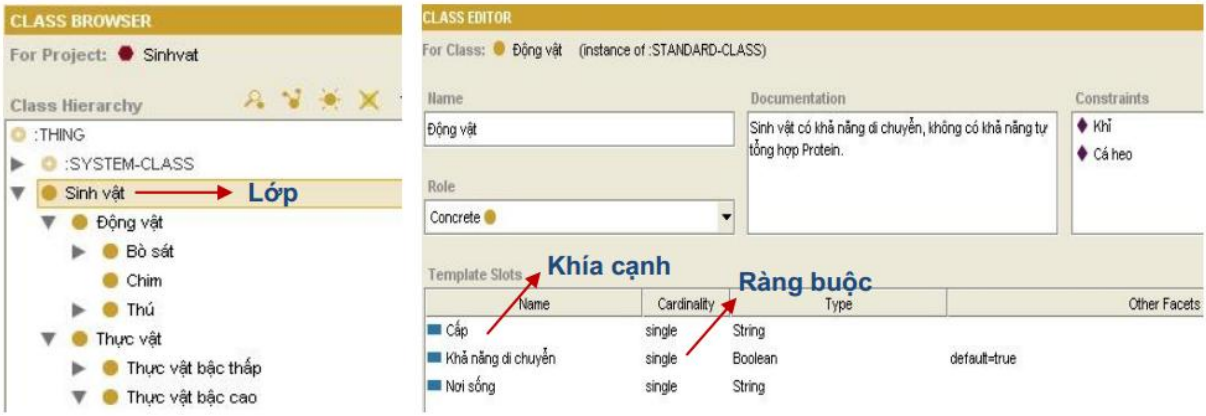
3.4.2.2.Ontology:

Ontology là một ngôn ngữ hay một tập các quy tắc được dùng đểxây dựng một hệthống Ontology.MộthệthốngOntologyđịnh nghĩa một tập các từvựng mang tính phổbiến trong lĩnh vực chuyên môn nào đó, và quan hệgiữa chúng. Sựđịnh nghĩa này có thểđược hiểubởi cả con ngườilẫn máy tính. Một Ontology bao gồm các thành phần sau:

Lớp:làthànhphầnquantrọngcủamộtOntology,còn đượcgọilàkháiniệm.Hầu hếtOntology đều tập trung xây dựng các lớpđượctổchức theo một cấu trúc phân cấp đểmô tả các loạivật trong một miền cần quan tâm.

Ví dụ “sinh vật” là một lớp trong ngữ cảnh sinh vật học. Bên dưới lớp này có thể có các lớp convídụ“động vật” và “thực vật”..

- Khía cạnh:mô tảcác thuộc tính của lớp và thực thể. Khía cạnhlà một mặt nào đó của sựvật,phân biệt với thuộc tínhchỉlà giá trịbiểu hiện của nó.Ví dụkhái niệmsinh vậtcó thểđược mô tảqua khía cạnh tình trạng chuyển độngvới các thuộc tính là chuyển độnghoặcđứng yên.Một cách hình thức ta gọi: khía cạnh là kiểu quan hệgiữathựcthể và thuộc tính, giữa thực thểvà lớp hoặc giữa các lớp với nhau. Mặc dù vậy, để thuận tiện trong một sốtrường hợpvẫn có thểdùng thuật ngữthuộctínhhoặcvai tròthay cho khía cạnh.
- Ràng buộc :mô tảmột sốràng buộc vềý nghĩa của các khái niệm và quan hệvới các khái niệm khác. Chẳng hạntình trạng chuyển độngtrong ví dụ trên chỉcó hai giá trị , không thể có sinh vật vừa chuyển động vừa phải đứng yên được.



Hình 3.14. Các thuộc tính của

3.4.2.3. Vai trò của Ontology:

Với ý nghĩa và cấu trúc như trên, Ontology đã trởthành một công cụquan trọng trong lĩnh vực Web ngữ nghĩa. Có thể kể ra một số l ợi ích của Ontology như:

Đểchiasẻnhữngnhiểuhiểubiếtchungvềcáckháiniệm,cấutrúcthôngtingiữa con ngườihoặcgiữacáchệthốngphầnmềm:đâylàvaitròquantrọngnhấtcủamột Ontology, không những trong lĩnh vực Web ngữnghĩa mà còn trong nhiều ngành và lĩnh vực khác. Vềphươngdiện này, có thểhình dung Ontology giống như một cuốn từđiển chuyênngành,cungcấpvàgiảithíchcácthuậtngữchongườikhôngcócùngchuyên môn khi được yêu cầu. Không chỉ được sử dụng bởi con người, Ontology còn hữu ích khi cần sựhợp tác giữa các hệthống phần mềm. Lấy ví dụ,Open Biological là bộOntology nổi tiếngđược phát triển bởi trường đại học Stanfordnhằm cung cấp các thuật ngữmột cách đầy đủtrong ngành sinh vật học. Ontology này hiện đã được tích hợp vào một số ứng dụngWeb trên Internet. Sau đó, mộtphần mềm tra cứu hoặcdạy sinh học trên máy tínhcóthểkết nối với cácứng dụng Web trên đểlấy thông tin cho mục tiêu chú giải.

- Cho phép tái sửdụng tri thức:đâylàmột vấn đềkhó vàlàmục tiêu nghiên cứu quan trọng trong những nămgần đây. Nó liên quan đến bài toán trộn hai hay nhiều Ontology thành một Ontologylớn và đầy đủhơn. Nhưng vấn đềởđây là tên các khái niệm được định nghĩa trong các Ontology này có thểgiống nhau trong khi chúng được dùng để mô tảcác loại vật hoàn toànkhác nhau. Tuy nhiên cũng có thểcó

trường hợp ngược lại, khi tên các khái niệm khác nhau nhưng cùng mô tảmột sựvật. Ngoài ra, làm thếnào đểbổ sung các quan hệ, thuộc tính có sẵn vào một hệthống mới càng làm cho vấn đềtrởnên phức tạp.

- Cho phép trithứcđộc lập với ngôn ngữ:đây cũng là vấn đềliên quan đến lĩnh vực tái sử dụng tri thức đãnóiởtrên, tuynhiênbài toán củanólàlàm thếnào đểmột hệthống Ontology có thểđược dùng bởi các ngôn ngữcủa các quốc gia khác nhau mà không phải xây dựnglại. Giải pháp mà Ontology mang lại là cho phép tên các khái niệm và quan hệ trongOntologymớithamkhảocáckháiniệm,địnhnghĩacủamộthệthốngOntology chuẩn thường được xây dựng bằng tiếng Anh.Điều này có thểsẽphá vỡphần nào rào cản vềmặt ngôn ngữkhi mà kết quảtìm kiếm sẽkhông bó gọn trong từkhóa và ngôn ngữmà nó sửdụng. Ngoài ra, Ontology có thểsẽtrởthànhhướng đi mới cho mộtlĩnh vực đã quen thuộc là dịch tài liệu tự động. Có thể nói như vậy, bởi ngữnghĩa các từ vựng trong văn bản sẽ được dịch chính xác hơn khi được ánh xạ vào đúng ngữ cảnh của nó.

- Chophéptri thức trởnên nhất quán và tường minh:các khái niệm khác nhau trong một hay nhiều lĩnh vực cụthểcó thểcùng tênvà gâynhập nhằngvềngữnghĩa, tuy nhiên khi được đưa vào một hệthống Ontology thìtênmỗi khái niệm là duy nhất.Một gi ải pháp cho vấn đềnàylà Ontology sẽsửdụng các tham khảo URI làm định danh thật sựcho khái niệm trong khi vẫn sử dụng các nhãn gợi nhớ bên trên để thuận tiện cho người dùng.

- Cungcấpmộtphươngtiệnchocôngviệcmôhìnhhóa:Ontologylàmộttậpcáckhái niệm phân cấp được liên kết với nhau bởi các quan hệ. Cơ bản mỗi khái niệm có thểxem như là một lớp, mà đối tượng của lớp đócùng các quan hệđãgóp phần tạo nên cấu trúc của bài toán hay vấn đề cần giải quyết.

- Cung cấp một phương tiện cho việc suy luận:hiệnnay,một sốngôn ngữOntology đã tích hợplớp Ontology suy luậnbên trongcho mục đích suy luận logic trên tập quan hệ giữa các đối tượng trong hệ thống.

3.4.2.4. Tìm hiểu ngôn ngữ truy vấn dữ liệu SPARQL :

Khung ứng dụng RDF được xem là công cụ để mô tả thông tin về các tài nguyên cho Web ngữ nghĩamột cách linh động. RDF có thể được sử dụng để biểu diễn thông tin cá nhân, mạng xã hội, siêu dữ liệu về tài nguyên số cũng như để cung cấp một phương tiện tích hợp các nguồn thông tin hỗn tạp. Với một nguồn tài nguyên phong phú và lớn như thế, làm thế nào để chúng ta có thể truy vấn chính xác và hiệu quả. Điều đó đã đặt ra một thách thức cho các nhà nghiên cứu, làm sao xây dựng một ngôn ngữ có thể đáp ứng được yêu cầu nói trên.

Tổ chức W3C đã phát triển và giới thiệu một ngôn ngữ chuẩn để truy vấn dữ liệu RDF. Ngôn ngữ truy vấn SPARQL. Đây là một ngôn ngữ được phát triển bởi nhóm RDF Data Access Working Group – một phần trong hoạt động của Semantic Web.

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

- Trích thông tin trong các dạng của URI, các blank node và các plain hay typed literals.
- Trích thông tin từ các đồ thị con
- Xây dựng một đồ thị RDF mới dựa trên thông tin trong đồ thị truy vấn.

Định dạng thông thường của một truy vấn SPARQL là:

PREFIX: Chỉ định tên cho một URI

SELECT: Trả về tất cả hoặc vài giá trị biến theo mệnh đề WHERE

CONSTRUCT: Trả về một đồ thị RDF với các biến liên quan

DESCRIBE: Trả về một "mô tả" của tài nguyên tìm được

ASK: Trả về kết quả tìm một mẫu đồ thị có hay không

WHERE: danh sách, tức là kết nối các mẫu (đồ thị) truy vấn

OPTIONAL: danh sách, tức là kết nối các mẫu (đồ thị) truy vấn tuỳ chọn

AND: biểu thức logic (để lọc các giá trị)

Một câu truy vấn chọn dữ liệu SPARQL-SELECT bao gồm 2 mệnh đề chính, mệnh đề SELECT và mệnh đề WHERE cùng các thành phần khác. Mệnh đề SELECT định danh các biến mà ứng dụng quan tâm và mệnh đề WHERE bao gồm các mẫu bộ ba, các thành phần khác sẽ được đề cập đến trong các phần tiếp theo. Cú pháp tổng quát của SPARQL-SELECT được liệt kê như sau:

PREFIX ns: <namespaceURI>

PREFIX : <.>

SELECT variables

[FROM <dataURI>]

[FROM NAMED <dataURI>]

WHERE { constraints [FILTER] [OPTIONAL] }

[ORDER BY variables] [OFFSET/LIMIT n] [DISTINCT]

Dữ liệu trong RDF được mô tả theo dạng các bộ ba. Tập hợp các bộ ba RDF tạo ra một đồ thị, gọi là đồ thị RDF. Ngôn ngữ truy vấn SPARQL lấy thông tin từ các đồ thị RDF, nó cung cấp các tính năng sau:

- Chiết xuất thông tin dưới dạng các URI, các node trắng, các plain literal và typed literal.
- Chiết xuất các đồ thị con RDF.
- Xây dựng các đồ thị RDF mới dựa trên thông tin của các đồ thị truy vấn.

3.5. Kết luận:

Như vậyđể xây dựng mô hình các chủ đề hoặc công cụ tìm kiếm theo ngữ nghĩa cần có sự kết hợp nhiều kỷ thuật và công nghệ với nhau để có được một sản phẩm hoàn chỉnh. Trong luận văn này các công cụ và kỹ thuật dùng để xây dựng điều hướng tới các công cụ mã nguồn mở để giúp cho chương trình ít tốn kém và sau này mọi người có thể phát triển, sử dụng hoặc xây dựng thêm được dễ dàng hơn.

CHƯƠNG 4: XÂY DỰNG MÔ HÌNH CÁC CHỦ ĐỀ VÀ CÔNG CỤ TÌM KIẾM THEO NGỮ NGHĨA

4.1Quy trình xây dựng mô hình các chủ đề và công cụ tìm kiếm theo ngữ nghĩa:

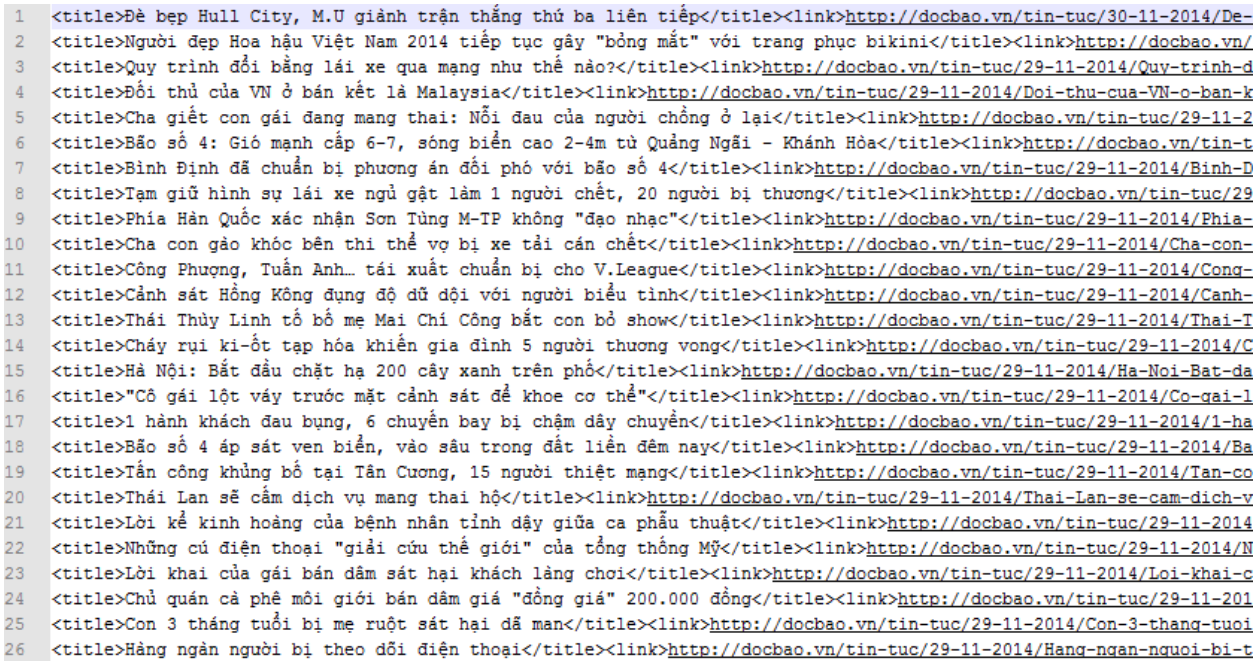
Trong quy trình này tác giả tiến hành các bước sau:

- Thu thập dữ liệu: Tiến hành thu thập các tài liệu trên mạng bao gồm các bài báo tiếng Việt nhằm phục vụ cho quá trình tìm kiếm. Trong luận văn này dữ liệu sẽ là các bài báo trên trang web docbao.vn và dùng công cụ Webcrawler để tiến hành thu thập các bài báo trên trang web này.
- Bóc tách dữ liệu: Sau khi tác giả đã thu thập các bài báo vềsẽ tiến hành bóc tách dữ liệu thu được bằng cách gỡ bỏ những từ vô nghĩa và tiến hành gom nhóm các từ vào cụm từ có nghĩa.
- Sử dụng mô hình LDA: Sau khi dữ liệu đã được bóc tách tác giả sử dụng mô hình LDA để tạo các chủ đề và các từ trong chủ đề đó cùng với trọng số của các từ,…Sau đó dùng công cụ lập trình để xây dựng mô hình ontology mô hình các chủ đề.
- Xây dựng chương trình tìm kiếm theo ngữ nghĩa: Sau khi tạo đươc tập tin ontology mô hình các chủ đề tác giả xây dựng chương trình dùng SPARQL để truy vấn dữ liệu và framework Jena để xử lý tập tin ontology phục vụ cho việc tìm kiếm.

4.1.1. Thu thập dữ liệu:

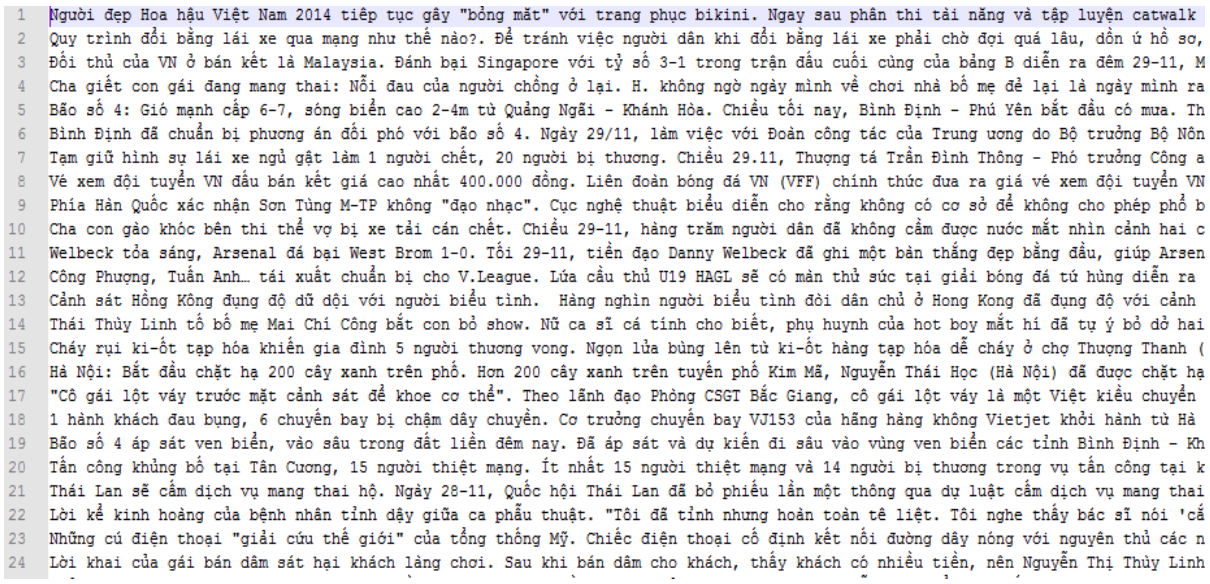
Dữ liệu thu thập được từ trang web www.docbao.vn thông qua công cụ WebCrawler được phát triển bởi google. Sau khi tải công cụ về để lấy dữ liệu ra được theo ý mình cần phải lập trình công cụ để chỉnh sửa kiểu dữ liệu trả về phù hợp với nhu cầu của luận văn.

Theo luận văn thì sau khi lập trình chỉnh sửa kết quả trả về của công cụ ta thu được hai tập tin với định dang XML như sau:*<root><title></title><link></link></root>*



Hình 4.1. Dữ liệu sau khi lấy về bằng công cụ Webcrawler bao gồm tiêu đề và địa chỉ

Dữ liệu sau khi lấy về được từ công cụ là một trang web tác giả dùng flugin hỗ trợ DOM đọc tựa đề và địa chỉ của trang web trả về dưới dạng XML theo cấu trúc như trên để hỗ trợ các bước tiếp theo cho luận văn.



Hình 4.2. Dữ liệu sau khi lấy về bằng công cụ Webcrawler

Dữ liệu lấy về của công cụ được phân tích để chỉ lấy về nội dung và phần miêu tả ngắn của trang web ta được tập tin trả về như hình 4.2 trên.

4.1.2. Bóc tách dữ liệu:

Với công cụ mã nguồn mở JvnSegmenter của Nguyễn Cẩm Túsẽ giúp việc bóc tách dữ liệu dễ dàng và chính xác hơn, trong phần mã nguồn bóc tách dữ liệu luận văn đã lập trình công cụ tích hợp thêm tính năng gỡ bỏ những Stopword để tăng cường hiệu năng cho các bước sau đó. Trong quá trình bóc tách dữ liệu nếu gặp các Stopword thì sẽ không xử lý bóc tách từ đó nữa và đồng thời cũng gỡ bỏ các từ đó khỏi tài liệu và tiến hành các bước tiếp theo.

Sau khi tiến hành bóc tách dữ liệu và gỡ bỏ Stopword ta được tập tin kết quả như sau:

```
Đè_bẹp Hull City, M.U giành trận_thắng thứ_ba liên_tiếp. Khuya 29-11,_Manchester_United chơi trận tưng
 Người đẹp Hoa_hậu Việt_Nam 2014 tiếp_tục gây bỏng mắt trang_phục bikini. Ngay sau phần thi tài_năng t
 Quy_trình đổi lái xe mạng nào?. Để tránh_việc người_dân khi đổi lái xe chờ_đợi lâu, dồn ứ hồ sơ,_đi_l
Đối thủ_của VN ở bán_kết là_Malaysia. Đánh_bại Singapore với_tỷ số_3-1 trậnđấu cuối_cùng của_bảng B di
 Cha giết gái mang thai: Nỗi_đau người_chồng ở lại. H. ngờ ngày mình chơi nhà bố_mẹ đẻ ngày mình mãi m
 Bão số 4: Gió_mạnh cấp 6-7, sóng biển cao 2-4m từ Quảng_Ngãi- Khánh Hòa. Chiều_tối nay, Bình_Định- Ph
 Bình_Định chuẩn_bị phương_án đối_phó bão số 4. Ngày 29/11, làm_việc Đoàn công_tác Trung_ương Bộ_trưởn
 Tạm_giữ hình_sự lái_xe ngủ_gật làm 1 người chết, 20 người thương. Chiều 29.11, Thượng_tá Trần_Đình_Th
 Phía Hàn_Quốc xác_nhận Sơn Tùng M-TP đạo_nhạc. Cục nghệ_thuật biểu_diễn có sở để phép phổ_biến hát Ch
 Cha_con gào_khóc bên thi_thể vợ xe_tải cán chết. Chiều 29-11, hàng trăm người_dân cầm được nước mắt n
CôngPhượng, Tuấn_Anh… tái_xuất chuẩn_bị V.League. Lứa cầu_thủ U19 HAGL có màn thử_sức giải bóng_đá tứ
 Cảnh_sát Hồng_Kông dụng_độ dội người biểu tình.  Hàng nghìn người_biểu_tình đòi dân_chủ ở Hong_Kong đ
 Thái_Thùy Linh tố bố_mẹ Mai_Chí_Công bắt bỏ show. Nữ_ca_sĩ cá_tính biết, phụ_huynh hot boy mắt hí tụ_
 Cháy_rụi ki-ốt tạp_hóa khiến gia_đình 5 người thương_vong. Ngọn_lửa bùng lên từ ki-ốt hàng tạp_hóa dễ
 Hà_Nội: Bắt_đầu chặt hạ 200 xanh phố. Hơn 200 xanh tuyến phố_Kim_Mã, Nguyễn_Thái_Học (Hà_Nội) được ch
 Cô gái lột váy mặt cảnh_sát để khoe thể. Theo lãnh_đạo Phòng CSGT Bắc_Giang, gái lột váy Việt_kiều ch
 1 hành_khách đau bụng, 6 chuyến_bay chậm dây chuyền. Cơ trưởng chuyến_bay VJ153 hãng hàng_không Vietj
 Bão số 4 áp_sát ven biển, sâu đất_liền đêm nay. Đã áp_sát dự_kiến sâu vùng ven_biển tỉnh Bình_Định- K
 Tấn_công khủng_bố Tân_Cương, 15 người thiệt mạng. Ít_nhất 15 người thiệt_mạng 14 người thương vụ công
 Thái_Lan cấm dịch_vụ mang thai hộ. Ngày 28-11,_Quốc_hội Thái_Lan bỏ_phiếu lần thông_qua dự_luật cấm d
 Lời kể kinh_hoàng bệnh_nhân tỉnh_dậy ca_phẫu thuật._Tôi tỉnh hoàn_toàn tê liệt. Tôi nghe_thấy bác_sĩ
 Những cú điện_thoại giải cứu_thế giới tổng_thống Mỹ. Chiếc điện_thoại cố_định kết_nối đường_dây nóng
 Lời_khai gái bán_dâm sát_hại khách làng chơi. Sau khi bán_dâm khách, thấy khách có_nhiều tiền, nên Ng
 Chủ_quán cà_phê môi_giới bán_dâm giá đồng_giá 200.000 đồng. Dưới vỏ bọc quán cà phê, Nhẫn tuyển số nh
 Con 3 tháng tuổi mẹ ruột sát_hại dã man. Uống rượu say thêm_chuyện giận chồng, người phụ_nữ nhẫn_tâm
```

Hình 4.3. Kết quả sau khi bóc tách dữ liệu

4.1.3. Sử dụng mô hình Latent Dirichlet Allocation:

Sau khi bóc tách dữ liệu tác giả sử dụng công cụ mã nguồn mở JGibbLDA [13] của Nguyễn Cẩm Tú dùng thuật toán LDA để trả về các chủ đề và các từ trong chủ đề đó cùng với các trọng số như đã mô tả ở phần trên.

Nguyễn Cẩm Tú đã sử dụng thuật toán LDA được cải tiến của nhóm tác giả David M. Blei, Andrew Y. Ng và Michael I. Jordan vào năm 2003.

Cú pháp để chạy được công cụ như sau:

$ java [-mx512M]-cp bin:lib/args4j-2.0.6.jar jgibblda.LDA -estc -dir <string> -model <string> [-niters <int>] [-savestep <int>] [-twords <int>]

Trong đó:

- -estc: Tiếp tục ước lượng mô hình từ mô hình ước lượng trước đó.
- -model <string>: Tên của mô hình ước lượng trước đó
- -niters <int>: Số lượng Gibbslặp lấy mẫu để tiếp tục ước lượng . Giá trị mặc định là 2000.
- -savestep <int>: Số bước (Tính theo số lượng Gibbs lặp lại lấy mẫu)mà tại đó các mô hình LDA được lưu vào ổ cứng
- -twords <int> : Số lượng từ quan trọng và có trọng số cao nhất từ trên xuống mà người dùng muốn lấy(Ví dụ ta gán nó là 20 thì trong mỗi chủ đề ta sẽ lấy được 20 từ cùng với trọng số của nó).
- -dfile <string> :Tên tập tin chứa đựng dữ liệu sau khi tính toán xong nội dung sẽ được lưu vào đó.

Dữ liệu đầu vào của chương trình sẽ có dạng như sau:

[M]

[Tài liệu 1]

[Tài liệu 2]

[Tài liệu 3]

……..

Trong đó:

[M] là số lượng tài liệu sẽ đưa vào cho việc tính toán.

[Tài liệu] là [từ 1] [từ 2] [từ 3]…..

Sau khi đầu vào phù hợp chạy chương trình thì ta sẽ thu được các tập tin sau:

<model_name>.others

<model_name>.phi

<model_name>.theta

<model_name>.tassign

<model_name>.twords

Trong đó:

<model_name>.others: Tập tin chứa đựng các tham số mà chúng ta cài đặt cho chương trình để chạy.

<model_name>.phi: Tập tin chứa đựng mối quan hệ giữa chủ đề và từ mỗi dòng là một chủ đề và mỗi cột là một từ hoặc cụm từ.

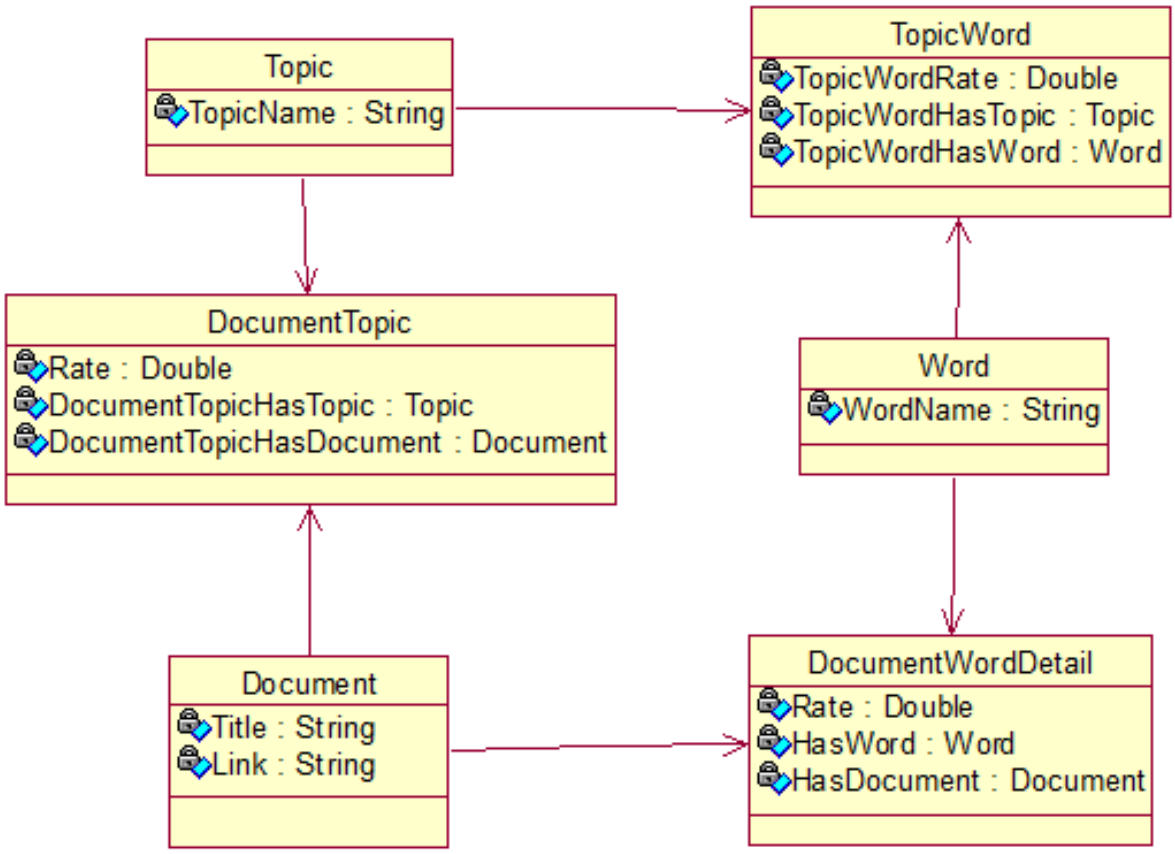
<model_name>.theta: Tập tin chứa đựng mối quan hệ giữa chủ đề và tài liệu .Mỗi dòng là một tài liệu, mỗi cột là một chủ đề.

<model_name>.tassign: Tập tin này chứa đựng các chủ đề cho các từ trong huấn luyện dữ liệu.Mỗi dòng là một tài liệu và mỗi cột là một từ và chủ đề của từ đó theo dạng <$word_{ij}$>:<topic of $word_{ij}$>

<model_name>.twords: Tập tin này chứa đựng các chủ đề và các từ trong chủ đề đó cùng với trọng số của nó.

4.2.Xây dựng mô hình các chủ đề:

Sau khi chạy chương trình ta thu được các tập tin trên, từ các thông tin thu được tác giảsử dụng công cụ protégé [16]tiến hành xây dựng mô hình các chủ đề theo cấu trúc sau:



Hình 4.4. Cấu trúc ontology cho mô hình các chủ đề

Mô hình trên bao gồm các lớp và các thuộc tính như sau:

Bảng 4.1. Các lớp và thuộc tính trong chủ đề

Lớp	Thuộc tính của lớp	Kiểu dữ liệu	Ý nghĩa thuộc tính
Topic	TopicName	String	Tên các chủ đề
DocumentTopic	Rate	Double	Trọng số của chủ đề trong tài liệu
	DocumentTopicHasTopic	Object	Các chủ đề trong tài liệu
	DocumentTopicHasDocument	Object	Các tài liệu chứa chủ đề
Document	Title	String	Tiêu đề của tài liệu
	Link	String	Liên kết trang của tài liệu
TopicWord	TopicWord Rate	Double	Trọng số của các từ trong chủ đề
	TopicWordHasTopic	Object	Chủ đề chứa các từ
	TopicWordHasWord	Object	Từ trong chủ đề
Word	WordName	String	Từ trong chủ đề
DocumentWord Detail	Rate	Double	Trọng số của các từ trong tài liệu
	HasDocument	Object	Tài liệu chứa các từ
	HasWord	Object	Từ trong tài liệu

Luận văn sử dụng công cụ Protégé [16] để xây dựng mô hình các chủ đề bằng ngôn ngữ Ontology, tuy nhiên do số lượng dữ liệu xây dựng rất lớn. Luận văn tiến hành thử nghiệm trên : 660 tài liệu và chia ra nhiều mô hình với số lượng các chủ đề và các từ khác nhau để kiểm tra độ chính xác của luận văn. Trong luận văn tác giả tiến hành thử nghiệm trong 3 trường hợp:

Trường hợp 1:

Mô hình bao gồm 20 chủ đề và 700 từ.

Trường hợp thứ 2:

Mô hình bao gồm 10 chủ đề và 700 từ.

Trường hợp thứ 3:

Mô hình bao gồm 10 chủ đề 400 từ.

Với số lượng dữ liệu lớn như trên nên việc xây dựng tập tin Ontology bằng cách nhập tay là không khả quan nên tác giả đã lập trình công cụ bằng ngôn ngữ C# hỗ trợ việc chạy dữ liệu tự động cho quá trình tạo tập tin Ontology. Tuy nhiên trong quá trình tạo tập tin Ontology tác giả đã phát hiện ra vấn đề cần khắc phục đó là khi tập tin đạt dung lượng trên 50MB tốc độ ghi tập tin sẽ rất chậm tác giả đã thí nghiệm trên máy cấu hình CPU Core I5 tốc độ 1,6Ghiz, RAM 12 Ghiz và hệ điều hành Window 8.1 thì mất đến 40 ngày để xây dựng tập tin với dung lượng 80 MB. Một vấn đề nữa là nếu hệ điều hành sử dụng không phải là Window sever thì chương trình chỉ ghi được tối đa 85MB và không thể ghi thêm được nữa nếu dùng Window Sever thì dung lượng ghi tập tin sẽ không bị hạn chế.

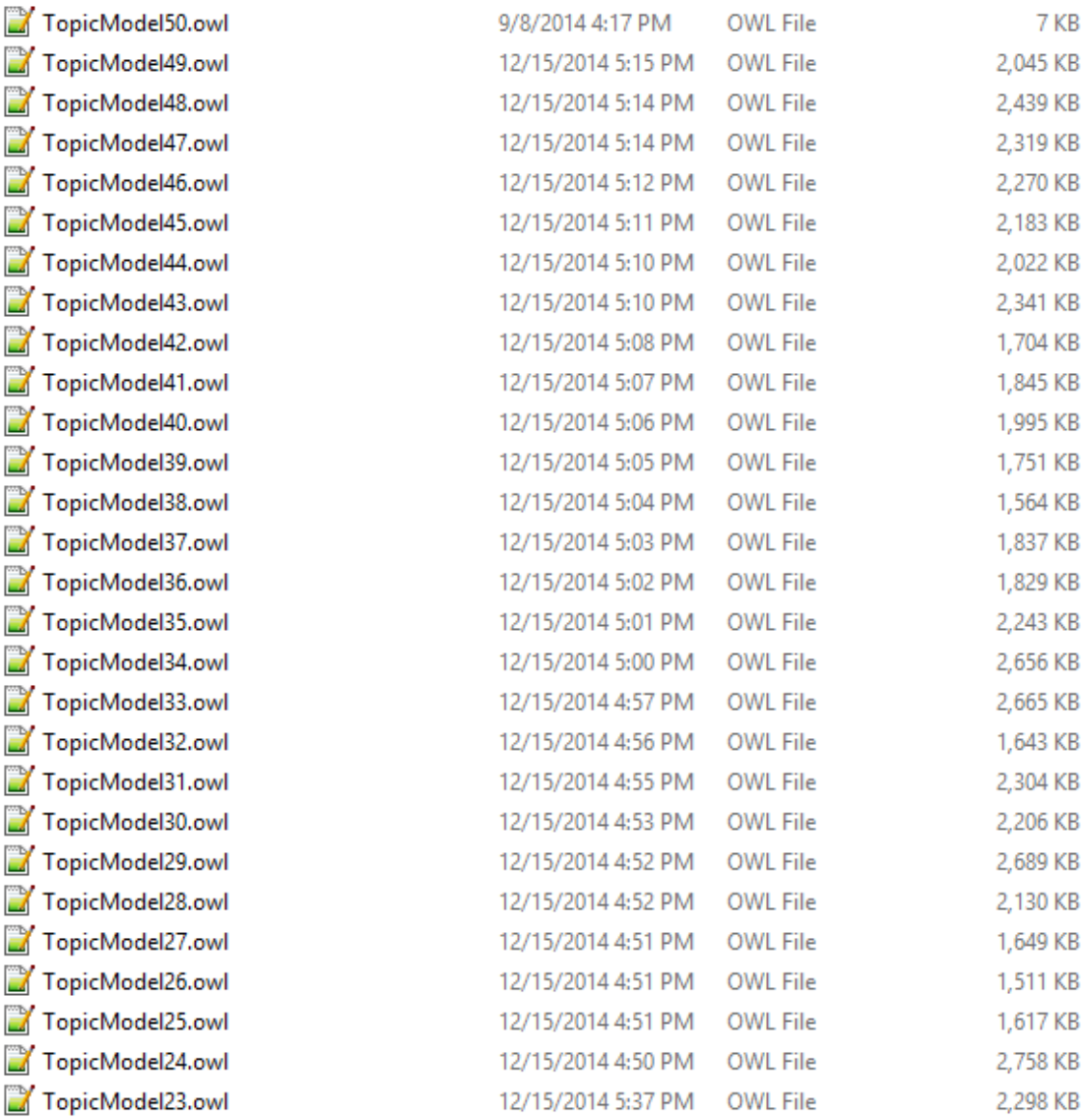
Với các vấn đề trên tác giả đã đưa ra phương pháp để khắc phục là phân tán tập tin Ontology ra thành nhiều tập tin nhỏ để tăng tốc độ ghi tập tin và giải quyết được vấn đề không ghi được vào tập tin dung lượng lớn của C#.

Để tiến hành xây dựng tập tin theo phương pháp phân tán tác giả đã tiến hành thí nghiệm 2 phương pháp ghi tập tin phân tán theo chiều rộng và ghi tập tin phân tán theo chiều sâucả hai đều trả về kết quả tốt và tăng tốc độ đáng kể với cấu hình như trên tác giả chỉ mất 6 giờ để tạo tập tin Ontology với dung lượng 100MB.

4.2.1. Phương pháp ghi tập tin phân tán theo chiều rộng:

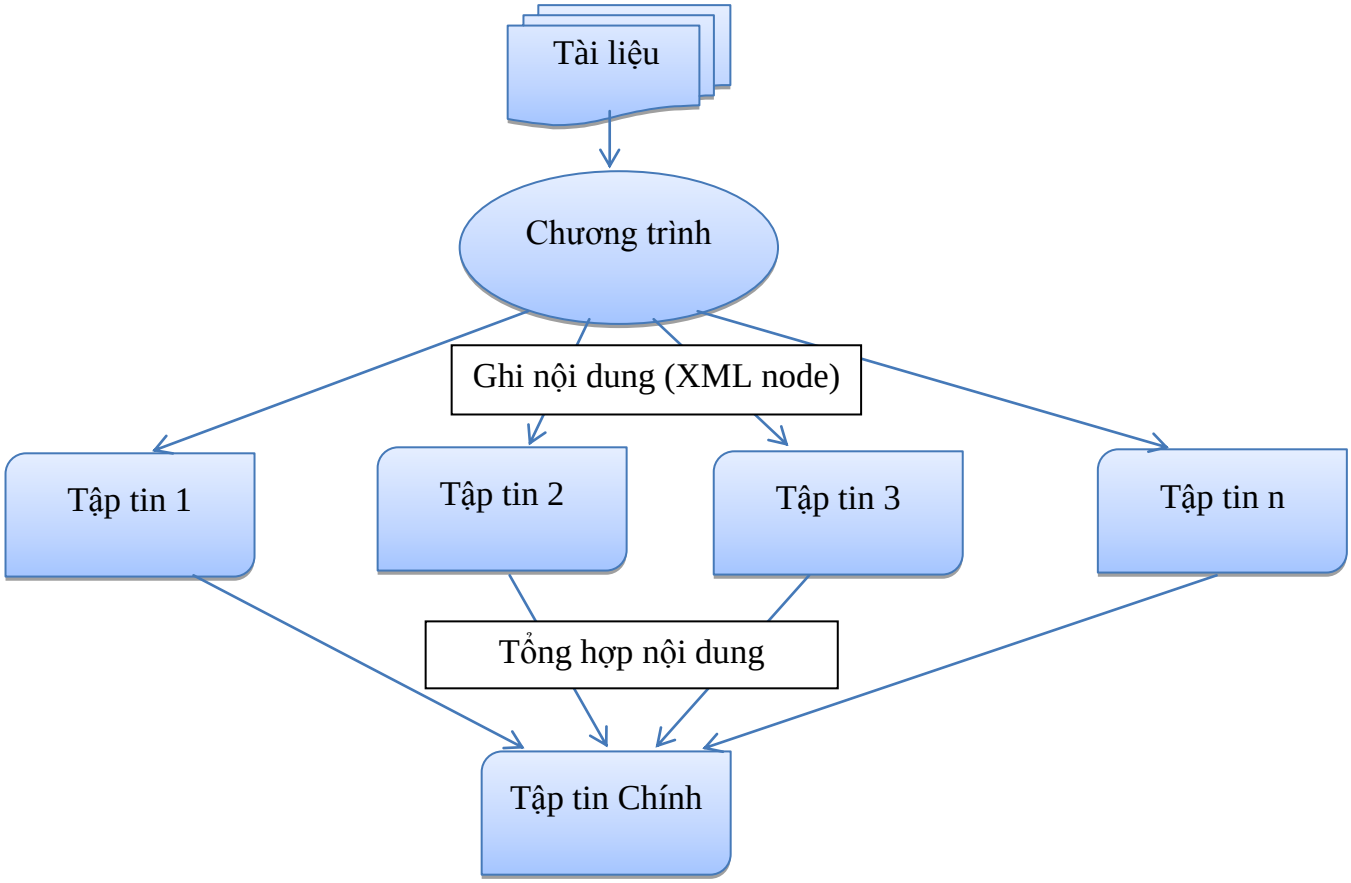
Thay vì chúng ta ghi vào một tập tin duy nhất ta sẽ chia tập tin đó ra thành số lượng tập tin nhất định tùy vào cấu hình máy và số lượng tập tin ta xây dựng lớn bao nhiêu,trong luận văn này với 660 tài liệu 20 chủ đề và 700 từ tác giả đã tạo ra 50 tập tin con để tăng tốc cho quá trình thí nghiệm.

Quá trình phân tán tập tin được thể hiện như hình sau:



Hình 4.5. Thực nghiệm việc phân tán tập tin

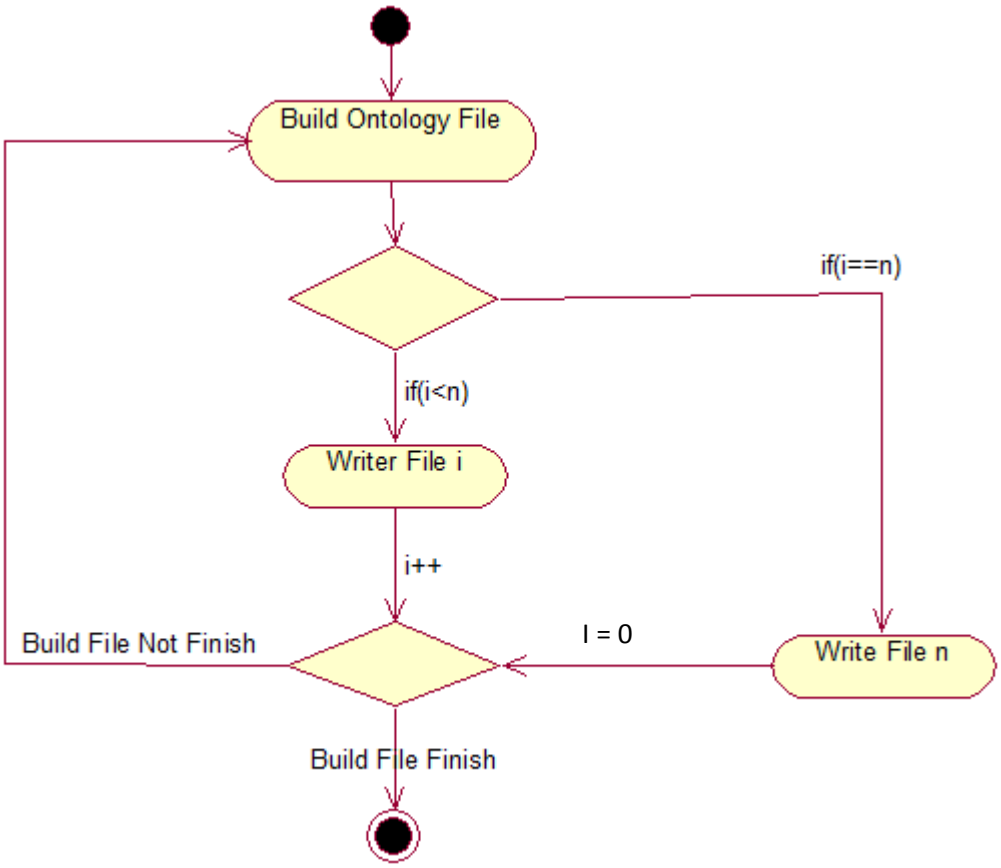
Việc phân tán ghi tập tin theo chiều rộngthực hiện theo cấu trúc sau:



Hình 4.6. Mô hình ghi tập tin phân tán theo chiều rộng

Thực hiện ghi tập tin theo mô hình này mỗi tập tin sẽ được ghi dữ liệu vào theo thứ tự như sau:

Nếu như theo cách bình thường thì dữ liệu nội dung của tập tin ontology mô hình các chủ đề sẽ được ghi vào một tập tin chính, tuy nhiên theo phương pháp này thì dữ liệu của nội dung tập tin ontology sẽ được chia đều ra nhiều tập tin.Tập tin đầu tiên sẽ được ghinội dungtrước rồi đến tập tin thứ 2 cứ thế ghi theo thứ tự đến tập tin cuối cùng rồi lại tiếp tục ghi nội dung vào tập tin đầu tiên và lặp lại cho đến khi tất cả nội dung xây dựng tập tin ontology mô hình các chủ đề kết thúc. Cuối cùng ta sẽ tập hợp tất cả các dữ liệu đó vào một tập tin duy nhất đó chính là tập tin ontology mô hình các chủ đề. Chương trình sẽ chạy theo qui tắc sau:



Hình 4.7. Phương pháp ghi tập tin theo chiều rộng

4.2.2. Phương pháp ghi tập tin phân tán theo chiều sâu:

Trong phương pháp này ta vẫn chia tập tin chính ra nhiều tập tin nhỏ tuy nhiên ta phải xác định dung lượng cần ghi cho mỗi tập tin và tùy vào ước lượng dung lượng nội dunglớn hay không mà chúng ta có thể xác định số lượng tập tin phân rã cho phù hợp cho phù hợp. Trong luận văn này mỗi tập tin được xác định là 10MB, sau khi tập tin đầu tiên ghi đủ dung lượng thì sẽ chuyển qua ghi tiếp tập tin tiếp theo nếu như chưa đến tập tin cuối cùng mà đã hoàn tất xây dựng tập tin Ontology thì sẽ kết thúc, ngược lại nếu đến tập tin cuối cùng mà vẫn chưa kết thúc việc xây dựng tập tin Ontology thì chúng ta sẽ tiếp tục quy trình bằng cách tiếp tục ghi vào tập tin đầu tiên và tiếp tục quy trình ,tuy nhiên lần này dung lượng quy định cho mỗi tập tin sẽ tăng lên để tiếp tục ghi tập tin với một hạn định mới để chuyển sang ghi tập tin kế tiếp. Dung lượng tăng thêm sẽ được định nghĩa trong chương trình ,trong luận

văn này dung lượng sẽ tăng lên 3MB nếu như đến tập tin cuối cùng mà vẫn chưa hoàn thành việc xây dựng tập tin Ontology.

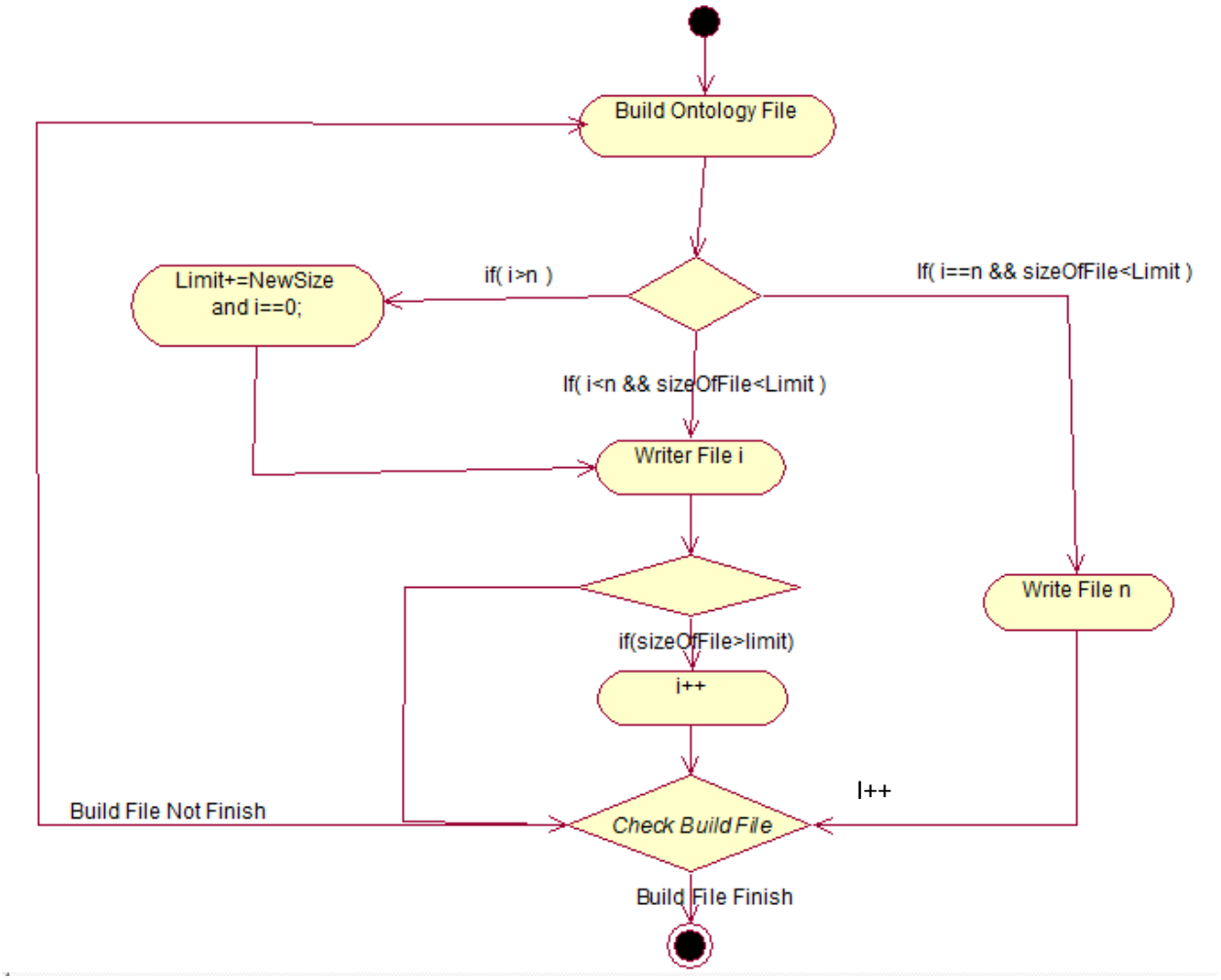
Phương pháp này có khuyết điểm so với phương pháp trên là:

Thời gian ghi tập tin sẽ lâu hơn do lúc đầu ta định nghĩa dung lượng mỗi tập tin khá lớn.

Phương pháp này có ưu điểm so với phương pháp trên là:

Số lượng tập tin ban đầu sẽ ít hơn phù hợp giúp cho việc gom nội dung tập tin sau khi hoàn thành vào một tập tin thống nhất nhanh hơn.

Quá trình phân tán tập tin được thể hiện như hình sau:



Hình 4.8. Phương pháp ghi tập tin theo chiều sâu

4.3. Xây dựng chương trình tìm kiếm theo ngữ nghĩa:

Sau khi xây dựng thành công tập tin ontologymô hình các chủ đề, tác giả tiến hành xây dựng chương trình áp dụng mô hình trên cho việc tìm kiếm theo ngữ nghĩa.

Chương trình được xây dựng trên ngôn ngữ Java và dùng thư viện Jena của Apache [17] phục vụ cho việc xử lý ontology, về ngôn ngữ truy vấn dữ liệu chương trình dùng ngôn ngữ truy vấn SPARDQL để thực hiện chương trình.

Công cụ tìm kiếm theo ngữ nghĩa hoạt động theo trình tự sau:

- Sau khi người dùng nhập vào đoạn văn cần tìm kiếm chương trình sẽ gỡ bỏ những stopword trong đoạn văn để tăng hiệu năng tìm kiếm đồng thời tiến hành bóc tách đoạn văn thành những từ và cụm từ có ý nghĩa.
- Sau khi bóc tách đoạn văn để tăng hiệu năng tìm kiếm công cụ tiến hành kiểm tra và loại bỏ các từ và cụm từ không tồn tại trong tập tin ontology vì tốc độ kiểm tra các từ tồn tại trong tập tin hay không sẽ nhanh hơn rất nhiều so với tốc độ tìm kiếm các từ và cụm từ đó.
- Sau đó công cụ sẽ tiến hành tìm kiếm từngtừ và cụm từ đó, sau đó kết hợp các kết quả lại với nhau theo hướng nhóm các tựa đề các tài liệu có chứa các từ và cụm từ vừa tìm đượcđồng thời cộng dồn trọng số của các từ và cụm từ đó và cộng dồn trọng số của các tài liệu và chủ đề(bao gồm trọng số của từ và cụm từ xuất hiện trong chủ đề và trọng số của các chủ đề xuất hiện trong các tài liệu).Thuật toán được thể hiện theo công thức sau:

$$f_t = \sum_{i=1}^{n} R_{(w_i \in t)}$$

Trong đó:

f_t : Trọng số của một tài liệu vừa tìm được

n: Số từ hoặc cụm từ tìm được thuộc tài liệu đó

w_i : Từ hoặc cụm từ tìm được thuộc tài liệu đó

t: Tài liệu

R: Trọng số của các từ hoặc cụm từ tìm được.

Sau khi tính được f tác giả tiến hành sắp xếp các tiêu đề theo trọng số f rồi đến các ký tự tiêu đề theo hướng giảm dần rồi hiển thị kết quả cho người dùng.

- Cuối cùng công cụ sẽ tiến hành sắp xếp các tài liệu theo các trọng số và tựa đề cho người dùng, trong quá trình sắp xếp công cụ sẽ ưu tiên trọng số của các từ và cụm từ trong các chủ đề trước sau đó đến trọng số của các chủ đề trong tài liệu và cuối cùng là sắp xếp các tiêu đề giảm dần để tiện việc tìm kiếm.
- Do lượng dữ liệu nhiều nên công cụ chỉ lấy top 100 dữ liệu theo trọng số của từng từ và cụm từ sau đó kết hợp lại. Đó cũng là lý do ta tìm theo từng từ và cụm từ rồi kết hợp lại mà không tìm một lần tất cả các từ rồi lấy kết quả cuối cùng về xử lý(do cách tìm một lần tất cả các từ rồi lấy kết quả cuối cùng về xử lý ta không thể lấy top trọng số theo từng từ cụm từ sẽ dẫn đến dữ liệu quá lớn làm ảnh hưởng đến tốc độ xử lý và lấy lên nhiều tài liệu có độ chính xác không cao)

Tuy nhiên do lượng dữ liệu lớn mà trong quá trình tìm kiếm chúng ta lại phải sắp xếp giảm dần theo trọng số để lấy top từ trên xuống làm tốc độ xử lý chậm do sắp xếp một lượng lớn dữ liệu. Để khắc phục vấn đề trên đồng thời tận dụng xử lý đa luồng của CPU đa nhân tác giả đã lấy tất cả các từ và cụm từ sau khi bóc tách đoạn văn đưa vào các Thread trong Java để tiến hành xử lý tìm kiếm đồng thời cùng một lúc, sau khi tất cả các Thread đã hoàn tất xử lý sẽ tiến hành kết hợp kết quả và tính trọng số theo các bước như trên.

4.3.1.Sesame Sever:

Do dung lượng tập tin ontology phục vụ chương trình quá lớn nên để quản lý các tập tin tốt và tăng tốc độ truy cập chương trình dùng mã nguồn mở Sesame sever do Apache phát triển và hỗ trợ tích hợp tốtvới Jena và ngôn ngữ truy vấn SPARQL rất tốt và đặt biệt Sesame hỗ trợ phân tán dữ liệu rất tốt và tăng tốc độ truy vấn dữ liệu,….

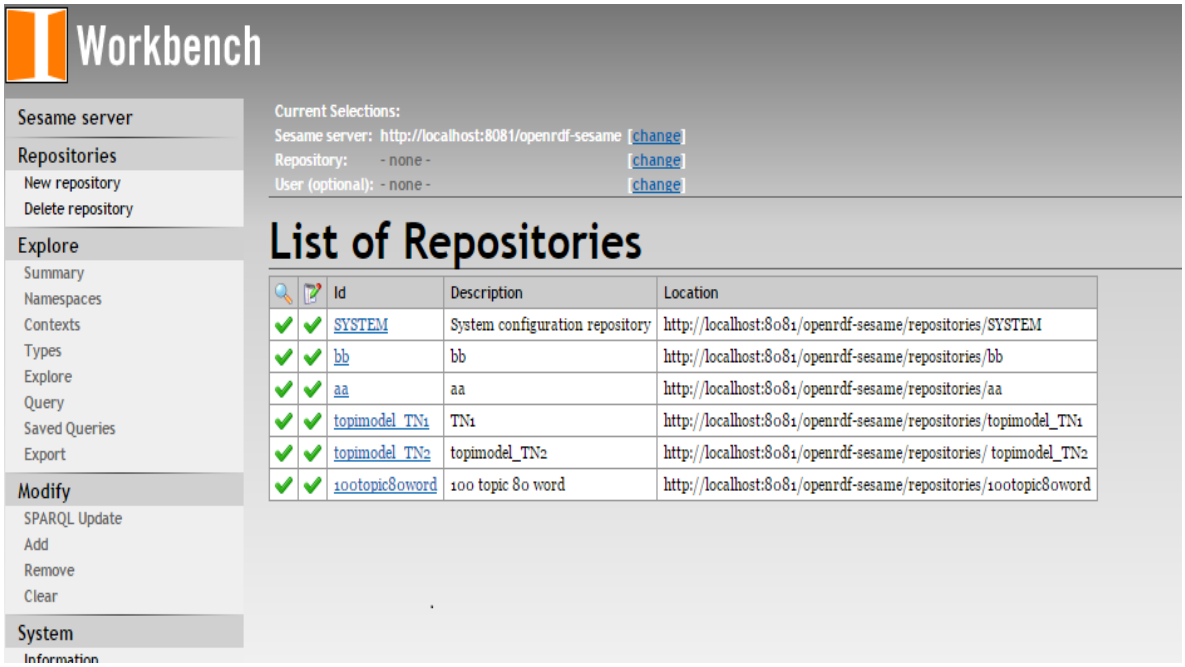
Sesame [18]hỗ trợ Webserver với giao diện đơn giản giúp người dùng upload các tập tin và quản lý các tập tin đó theo các repository, ngoài ra Sesame cũng hỗ trợ các lệnh ở CMD của window và linux để quản lý và tải tập tin lên máy chủ như:

Ví dụ kết nối và tải tập tin lên server bằng CMD của window:

connect http://localhost:8081/openrdf-sesame.

open TN1.

load d:\topicmodel.owl.



Hình 4.9. Giao diện sử dụng của Sesame

4.3.2. Jena Framework và ngôn ngữ truy vấn dữ liệu SPARQL:

Jena Framework[17] được phát triển bởi Apache giúp cho việc xử lý ontology trên java và tích hợp tốt với Sesame sever, đặc biệt Jena hỗ trợ việc truy vấn dữ liệu ontology bằng ngôn ngữ SPARQL rất tốt. Trong luận văn tác giả đã sử dụng Jena và SPARQL để truy vấn dữ liệu lưu trên Sesame để thực hiện việc truy vấn dữ liệu hỗ trợ cho việc tìm kiếm theo ngữ nghĩa. Sau đây là câu truy vấn SPARQL mà tác giả đã dùng cho chương trình:

```
PREFIXe: <http://www.semanticweb.org/thuong/ontologies/2014/4/untitled-ontology-16#>
SELECT distinct *
where {
?TopicWord e:TopicWordHasWord ?TopicWordHasWord.
?TopicWordHasWord e:WordName ?WordName.
?TopicWord e:TopicWordRate ?TopicWordRate.
?TopicWord e:TopicWordHasTopic ?TopicWordHasTopic.
?TopicWordHasTopic e:TopicName ?TopicName.
FILTER (REGEX(STR(?WordName), '"+strSearch+"', 'i'))
{
  select ?Title ?Link ?WordName ("1" as ?DocumentTopicRate) where{
  ?DocumentWordDetail e:HasDocument ?HasDocument.
  ?HasDocument e:Title ?Title.
  ?HasDocument e:Link ?Link.
  ?DocumentWordDetail e:HasWord ?HasWord.
  ?HasWord e:WordName ?WordName.
  FILTER (REGEX(STR(?WordName), '"+strSearch+"', "i"))
  }
  order byDESC(?Rate) ?Title
}
union
{select * where{
?DocumentTopic e:DocumentTopicHasTopic ?DocumentTopicHasTopic.
?DocumentTopicHasTopic e:TopicName ?TopicName.
?DocumentTopic e:DocumentTopicRate ?DocumentTopicRate.
?DocumentTopic e:DocumentTopicHasDocument
?DocumentTopicHasDocument.
?DocumentTopicHasDocumente:Title ?Title.
```

?DocumentTopicHasDocument e:Link ?Link.

} order byDESC(?DocumentTopicRate) ?Title limit 100

}}order byDESC(?TopicWordRate)

- Trong câu truy vấn trên thực hiện từng bước như sau:
- Đầu tiên ta sẽ tìm những chủ đề chứa các từ mà người dùng cần tìm
- Sau đó ta tiến hành tìm những tài liệu chứa các từ trong chủ đề đó, nếu tài liệu nào chứa những từ đó sẽ được ưu tiên trọng số là 1
- Kế đến là tìm những tài liệu chứa các chủ đề trên có cùng tựa đề và liên kết trang với các tài liệu có chứa các từ mà người dùng nhập vào, và đồng thời để hạn chế số lượng các tài liệu đó bằng cách lấy giới hạn tối đa chỉ 100 tài liệu không chứa từ nhập vào nhưng chứa các chủ đề có chứa các từ đó.
- Sau cùng ta sắp xếp các trọng số của chúng giảm dần để hỗ trợ người xem ưu tiên những kết quả chính xác nhất. Tuy nhiên với cách sắp xếp trọng số giảm dần như trên thì ta đã ưu tiên trọng số của các từ trong chủ đề rồi mới ưu tiên trọng số của các tài liệutrong chủ đềđó .

4.3.3. Xử lý dữ liệu tìm kiếm:

Sau khi dữ liệu được lấy về để tăng cường tính chính xác cho chương trình tác giả cần thêm một số việc lọc dữ liệu lại như sau:

- Đối với những tài liệu mà chứa đựng các từ người dùng nhập vào sẽ được ưu tiên về trọng số bằng cách ta kiểm tra xem trọng số từ đó có bé hơn 1 không, nếu bé hơn 1 thì ta sẽ cộng thêm 1 vào trọng số đó.
- Đối với tài liệunào càng chứa nhiều từ hoặc cụm từ mà người dùng nhập vào thì các trọng số của các từ và trọng số các tài liệu có trong chủ đề sẽ được cộng dồn vào các trọng số của tài liệu đó.
- Trường hợp mà tài liệu đó chứa các từ của chủ đề khác thì các trọng số trong nó vẫn được cộng dồn lên với các trọng số mới.
- Cuối cùng ta sắp xếp lại dữ liệu và theo thứ tự trọng số giảm dần và hiển thị dữ liệu ra ngoài.

CHƯƠNG 5: ĐÁNH GIÁ KẾT QUẢ THỰC NGHIỆM

5.1Kết quả thực nghiệm:

5.1.2.Môi trường thực nghiệm:

Bảng 5.1. Môi trường thực nghiệm

Thành Phần	Chỉ số
CPU	**Core I5 1.6Ghz**
RAM	**12Gb**
HDD	**500Gb**
OS	**Window 8.1**

5.1.3. Công cụ:

Công cụ mã nguồn mở được dùng cho chương trình:

Bảng 5.2. Công cụ mã nguồn mở sử dụng

Tên công cụ	Công dụng và nguồn
Crawler4j	Công cụ thu thập tài liệu http://www.winwebcrawler.com/download.htm
JvnSegmenter	Công cụ bóc tách dữ liệu http://jvnsegmenter.sourceforge.net/
JGibbLDA-v.1.0	Công cụ phân tích chủ đề ẩn http://jgibblda.sourceforge.net/
Sesame Server	Công cụ lưu trữ và phân tán tập tin Ontology http://sourceforge.net/projects/sesame/files/Sesame%202/
Jena	Mã nguồn mở giúp xử lý Ontology https://jena.apache.org/documentation/inference/#OWLintro

Chương trình xây dựng trên nền JDK 1.8.0 với công cụ hỗ trợ lập trình NetBean 8.1[19]

Ngoài các công cụ trên tác giả còn xây dựng các công cụ xử lý sau:

- Chương trình xây dựng tập tin Ontology từ nguồn dữ liệu sẵn có.
- Chương trình tìm kiếm theo ngữ nghĩa.

5.1.4.Dữ liệu:

Dữ liệu bao gồm 660 tài liệu từ trang Web docbao.vn phục vụ cho quá trình tìm kiếm theo ngữ nghĩa.

Trong luận văn này tác giả đã xây dựng tập tin ontology theo mô hình trên với 3 trường hợp sau:

- 20 chủ đề và 700 từ. Sau khi dùngchương trình xây dựng tập tin Ontology tự động thì tạo ra được tập tin ontology với 3.249.199 Statements và dung lượng tập tin là 356MB.
- 10 chủ đề và 400 từ. Sau khi dùngchương trình xây dựng tập tin Ontology tự động thì tạo ra được tập tin ontology với 1.583.749Statements và dung lượng tập tin là 173MB.
- 10 chủ đề và 700 từ. Sau khi dùngchương trình xây dựng tập tin Ontology tự động thì tạo ra được tập tin ontology với 1.971.847Statements và dung lượng tập tin là 215MB.

5.1.5.Kết quả đạt được:

Tiến hành thực nghiệm với 20 chủ đề 700 ký tự tác giả thu được kết quả sau:

Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Cô giáo dốc sức leo, bám vách đá đi dạy*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Cô giáo","dốc","sức le"," bám"," vách đá"," dạy",*Các từ có trong dữ liệu là:" *bám"," vách đá, "dạy*".

repositories 20 topic 700 word

Key Word: Cô giáo dốc sức leo, bám vách đá đi dạy Search Keyword search: "bám","vách đá","dạy"

Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName
Cô giáo dốc sức leo, bám vách đá đi dạy	http://docbao.vn/tin-t...	vách đá , dạy , bám	3.00169724059173...	3.0	Topic_19th , Topic...
13 tuổi thay mẹ nuôi em	http://docbao.vn/tin-t...	dạy , bám	2.001474783252001	2.0	Topic_6th , Topic_...
"Tôi không biết tâm sự cùng ai về sex"	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
ATM mini đựng được gần 70 triệu đồng hút khách	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Cứ bỏ học rồi sẽ thành tỷ phú?	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Hoàng Bách: "Đại gia, chân dài cũng có lúc khổ sở vì tiền"	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Học trò tả cô giáo dáng như siêu mẫu, tóc màu nâu đỏ	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Thiếu nữ gặp họa vì mãi làm điệu bên cột ATM	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Tình dục tuổi teen: Góc khuất của nữ sinh làm mẹ ở tuổi 14	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
UNESCO công nhận dân ca Ví, Giặm là di sản đại diện của n...	http://docbao.vn/tin-t...	dạy	1.00099987181130...	1	Topic_6th
Bạn gái xinh đẹp của Việt Anh "Chạy án"	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Chuyện bi hài của trai trẻ thích tán gái nơi công cộng	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Ngọc Trinh đi xe 8 tỷ mới sắm trước ngày lên đường dự Victo...	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Phim Tết Việt 2015 "ồ ạt" ra rạp	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Rượt đuổi, nổ súng bắt kẻ cướp giật iPhone 6 giữa Sài Gòn	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Taylor Swift phải thuê vệ sĩ vì bị "fan cuồng" dọa giết	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Tàu Trung Quốc manh động tấn công tàu cá Việt Nam	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th
Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp	http://docbao.vn/tin-t...	bám	1.00047491144069...	1	Topic_16th

Hình 5.1. Kết quả thực nghiệm 1 của 20 chủ đề 700 ký tự

Kết quả đạt được chính xác cao tìm ra được nội dung cần tìm và đề xuất các nội dung khác tương tự nội dung cần tìm.

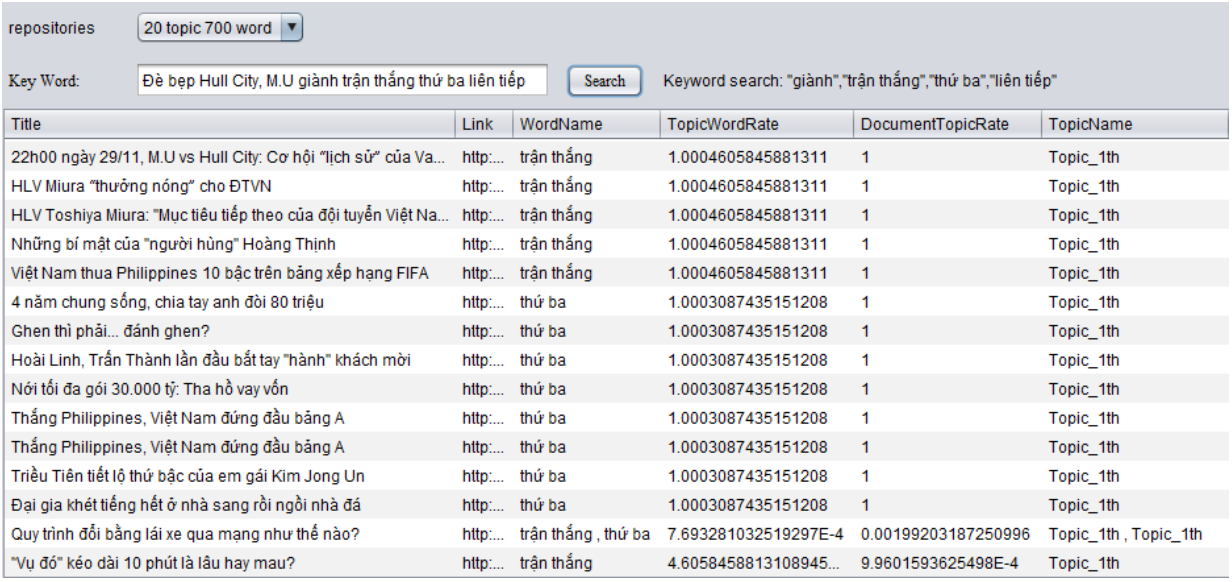
Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Đè bẹp","Hull City", "M.U","giành","trận thắng","thứ ba","liên tiếp",* Các từ có trong dữ liệu là:*"giành","trận thắng","thứ ba","liên tiếp".*

repositories 20 topic 700 word

Key Word: Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp Search Keyword search: "giành","trận thắng","thứ ba","liên tiếp"

Title	Link	WordName	TopicWordRate	DocumentTopic...	TopicName
Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp	http://docbao....	trận thắng , liên tiếp , thứ ba , giành	4.002173567681284	4.0	Topic_1th , Topi...
Cảnh sát Hồng Kông dụng độ dữ dội với người biểu t...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Giá xăng tiếp tục giảm hơn 500 đồng/lít	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Indonesia 5-1 Lào: Chiến thắng danh dự	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Nghi ngờ có chỉ điểm trong vụ chặt tay cướp 1,5 tỷ đồ...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Thủ môn Inter lập kỷ lục cản được 6 quả penalty liên t...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Tìm con gái nghi bị bán sang nước ngoài, bố bị đánh...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
U19 HA.GL-Arsenal JMG 1-2 U21 Sydney: Thua sút v...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Xe chở học sinh đâm xe tải, 11 trẻ em thiệt mạng	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
Đại gia Việt mua siêu du thuyền hơn 20 tỷ tặng khách...	http://docbao....	liên tiếp	1.000814880425155	1	Topic_1th
"Đập cánh giữa không trung" giành giải xuất sắc tại L...	http://docbao....	giành	1.0005893591528772	1	Topic_3th
00h30 ngày 30/11, Sunderland vs Chelsea: The Blue...	http://docbao....	giành	1.0005893591528772	1	Topic_3th
1001 lý do thí sinh bỏ thi hoa hậu	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Bánh xe container phát nổ, hất văng nam thanh niên	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Hoa hậu mất điểm vì những trang phục cắt xẻo kỳ dị	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Quán quân "The Voice Kids" tái xuất "Cặp đôi hoàn h...	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Sao rực rỡ trên thảm đỏ bế mạc LHP	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Thí sinh Hoa khôi Áo dài có tướng ngồi "bá đạo"	http://docbao....	giành	1.0005893591528772	1	Topic_3th
Tàu Trung Quốc manh động tấn công tàu cá Việt Nam	http://docbao....	giành	1.0005893591528772	1	Topic_3th
22h00 ngày 29/11, M.U vs Hull City: Cơ hội "lịch sử" c...	http://docbao....	trận thắng	1.0004605845881311	1	Topic_1th
HLV Miura "thưởng nóng" cho ĐTVN	http://docbao....	trận thắng	1.0004605845881311	1	Topic_1th
HLV Toshiya Miura: "Mục tiêu tiếp theo của đội tuyển ...	http://docbao....	trận thắng	1.0004605845881311	1	Topic_1th
Những bí mật của "người hùng" Hoàng Thịnh	http://docbao....	trận thắng	1.0004605845881311	1	Topic_1th
Việt Nam thua Philippines 10 bậc trên bảng xếp hạng...	http://docbao....	trận thắng	1.0004605845881311	1	Topic_1th

Hình 5.2. Kết quả thực nghiệm 2 của 20 chủ đề 700 ký tự



repositories: 20 topic 700 word

Key Word: Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp — Search — Keyword search: "giành","trận thắng","thứ ba","liên tiếp"

Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName
22h00 ngày 29/11, M.U vs Hull City: Cơ hội "lịch sử" của Va...	http:...	trận thắng	1.0004605845881311	1	Topic_1th
HLV Miura "thưởng nóng" cho ĐTVN	http:...	trận thắng	1.0004605845881311	1	Topic_1th
HLV Toshiya Miura: "Mục tiêu tiếp theo của đội tuyển Việt Na...	http:...	trận thắng	1.0004605845881311	1	Topic_1th
Những bí mật của "người hùng" Hoàng Thịnh	http:...	trận thắng	1.0004605845881311	1	Topic_1th
Việt Nam thua Philippines 10 bậc trên bảng xếp hạng FIFA	http:...	trận thắng	1.0004605845881311	1	Topic_1th
4 năm chung sống, chia tay anh đòi 80 triệu	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Ghen thì phải... đánh ghen?	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Hoài Linh, Trấn Thành lần đầu bắt tay "hành" khách mời	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Nới tối đa gói 30.000 tỷ: Tha hồ vay vốn	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Thắng Philippines, Việt Nam đứng đầu bảng A	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Thắng Philippines, Việt Nam đứng đầu bảng A	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Triều Tiên tiết lộ thứ bậc của em gái Kim Jong Un	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Đại gia khét tiếng hết ở nhà sang rồi ngồi nhà đá	http:...	thứ ba	1.0003087435151208	1	Topic_1th
Quy trình đổi bằng lái xe qua mạng như thế nào?	http:...	trận thắng , thứ ba	7.693281032519297E-4	0.00199203187250996	Topic_1th , Topic_1th
"Vụ đó" kéo dài 10 phút là lâu hay mau?	http:...	trận thắng	4.6058458813108945...	9.9601593625498E-4	Topic_1th

Hình 5.3. Kết quả thực nghiệm 2 của 20 chủ đề 700 ký tự

Theo hình trên thì hai tài liệu cuối cùng là có cùng chủ đề với các tài liệu chứa các từ tìm kiếm nhưng không chứa các từ tìm kiếm nên mặc dù được chương trình tìm ra nhưng trọng số nhỏ nên được đặt ở vị trí cuối.

Tiến hành thực nghiệm với 10 chủ đề 700 ký tự tác giả thu được kết quả sau:

Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Cô giáo dốc sức leo, bám vách đá đi dạy*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Cô giáo","dốc","sức leo","bám","vách đá","dạy*" ,Các từ có trong dữ liệu là:" *bám","dốc", "dạy*".

repositories 10 topic 700 word

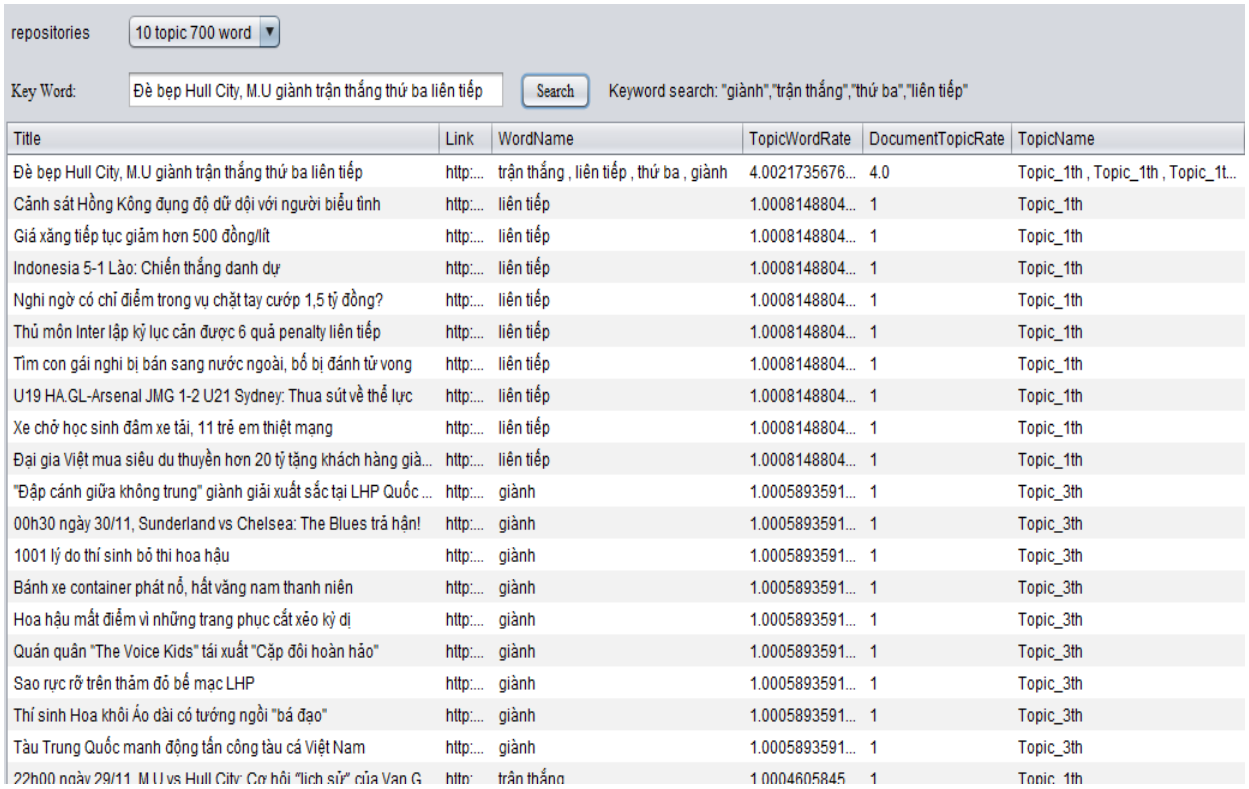
Key Word: Cô giáo dốc sức leo, bám vách đá đi dạy Search Keyword search: "dốc","bám","dạy"

Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName
13 tuổi thay mẹ nuôi em	http://d...	bám , dạy	2.001474783252001	2.0	Topic_16th , Topic_6th
Cô giáo dốc sức leo, bám vách đá đi dạy	http://d...	bám , dạy	2.001474783252001	2.0	Topic_16th , Topic_6th
"Tôi không biết tâm sự cùng ai về sex"	http://d...	dạy	1.0009998718113062	1	Topic_6th
ATM mini đựng được gần 70 triệu đồng hút khách	http://d...	dạy	1.0009998718113062	1	Topic_6th
Cứ bỏ học rồi sẽ thành tỷ phú?	http://d...	dạy	1.0009998718113062	1	Topic_6th
Hoàng Bách: "Đại gia, chân dài cũng có lúc khổ sở vì tiền"	http://d...	dạy	1.0009998718113062	1	Topic_6th
Học trò tả cô giáo dáng như siêu mẫu, tóc màu nâu đỏ	http://d...	dạy	1.0009998718113062	1	Topic_6th
Thiếu nữ gặp họa vì mải làm điệu bên cột ATM	http://d...	dạy	1.0009998718113062	1	Topic_6th
Tình dục tuổi teen: Góc khuất của nữ sinh làm mẹ ở tuổi 14	http://d...	dạy	1.0009998718113062	1	Topic_6th
UNESCO công nhận dân ca Ví, Giặm là di sản đại diện của nh...	http://d...	dạy	1.0009998718113062	1	Topic_6th
Bạn gái xinh đẹp của Việt Anh "Chạy án"	http://d...	bám	1.0004749114406946	1	Topic_16th
Chuyện bi hài của trai trẻ thích tán gái nơi công cộng	http://d...	bám	1.0004749114406946	1	Topic_16th
Ngọc Trinh đi xe 8 tỷ mới sắm trước ngày lên đường dự Victori...	http://d...	bám	1.0004749114406946	1	Topic_16th
Phim Tết Việt 2015 "ồ ạt" ra rạp	http://d...	bám	1.0004749114406946	1	Topic_16th
Rượt đuổi, nổ súng bắt kẻ cướp giật iPhone 6 giữa Sài Gòn	http://d...	bám	1.0004749114406946	1	Topic_16th
Taylor Swift phải thuê vệ sĩ vì bị "fan cuồng" dọa giết	http://d...	bám	1.0004749114406946	1	Topic_16th
Tàu Trung Quốc manh động tấn công tàu cá Việt Nam	http://d...	bám	1.0004749114406946	1	Topic_16th
Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp	http://d...	bám	1.0004749114406946	1	Topic_16th

Hình 5.4. Kết quả thực nghiệm 1 của 10 chủ đề 700 ký tự

Kết quả trên cho chúng ta thấy được với cách chia chủ đề không phù hợp sẽ dẫn đến kết quả tìm kiếm giảm đi độ chính xác.

Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Đè bẹp*"," *Hull City*", "*M.U*"," *giành*"," *trận thắng*"," *thứ ba*"," *liên tiếp*", Các từ có trong dữ liệu là:"*giành*"," *trận thắng*"," *thứ ba*"," *liên tiếp*".



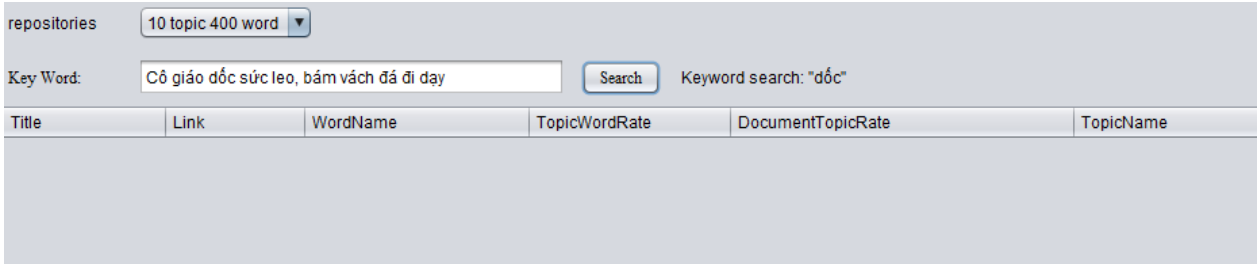
Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName
Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp	http:...	trận thắng , liên tiếp , thứ ba , giành	4.0021735676...	4.0	Topic_1th , Topic_1th , Topic_1t...
Cảnh sát Hồng Kông đụng độ dữ dội với người biểu tình	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Giá xăng tiếp tục giảm hơn 500 đồng/lít	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Indonesia 5-1 Lào: Chiến thắng danh dự	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Nghi ngờ có chỉ điểm trong vụ chặt tay cướp 1,5 tỷ đồng?	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Thủ môn Inter lập kỷ lục cản được 6 quả penalty liên tiếp	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Tìm con gái nghi bị bán sang nước ngoài, bố bị đánh tử vong	http:...	liên tiếp	1.0008148804...	1	Topic_1th
U19 HA.GL-Arsenal JMG 1-2 U21 Sydney: Thua sút về thể lực	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Xe chở học sinh đâm xe tải, 11 trẻ em thiệt mạng	http:...	liên tiếp	1.0008148804...	1	Topic_1th
Đại gia Việt mua siêu du thuyền hơn 20 tỷ tặng khách hàng già...	http:...	liên tiếp	1.0008148804...	1	Topic_1th
"Đập cánh giữa không trung" giành giải xuất sắc tại LHP Quốc ...	http:...	giành	1.0005893591...	1	Topic_3th
00h30 ngày 30/11, Sunderland vs Chelsea: The Blues trả hận!	http:...	giành	1.0005893591...	1	Topic_3th
1001 lý do thí sinh bỏ thi hoa hậu	http:...	giành	1.0005893591...	1	Topic_3th
Bánh xe container phát nổ, hất văng nam thanh niên	http:...	giành	1.0005893591...	1	Topic_3th
Hoa hậu mất điểm vì những trang phục cắt xẻo kỳ dị	http:...	giành	1.0005893591...	1	Topic_3th
Quán quân "The Voice Kids" tái xuất "Cặp đôi hoàn hảo"	http:...	giành	1.0005893591...	1	Topic_3th
Sao rực rỡ trên thảm đỏ bế mạc LHP	http:...	giành	1.0005893591...	1	Topic_3th
Thí sinh Hoa khôi Áo dài có tướng ngồi "bá đạo"	http:...	giành	1.0005893591...	1	Topic_3th
Tàu Trung Quốc manh động tấn công tàu cá Việt Nam	http:...	giành	1.0005893591...	1	Topic_3th
22h00 ngày 29/11, M.U vs Hull City: Cơ hội "lịch sử" của Van G	http:	trận thắng	1.0004605845	1	Topic_1th

Hình 5.5. Kết quả thực nghiệm 2 của 10 chủ đề 700 ký tự

Với kết quả đạt được từ thí nghiệm trên chúng ta thấy được với cách chia chủ đề và số từ hợp lý chương trình sẽ có độ chính xác cao.

Tiến hành thực nghiệm với 10 chủ đề 400 ký tự tác giả thu được kết quả sau:

Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Cô giáo dốc sức leo, bám vách đá đi dạy*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Cô giáo","dốc","sức leo","bám","vách đá","dạy"* , Các từ có trong dữ liệu là:" *dốc"*.

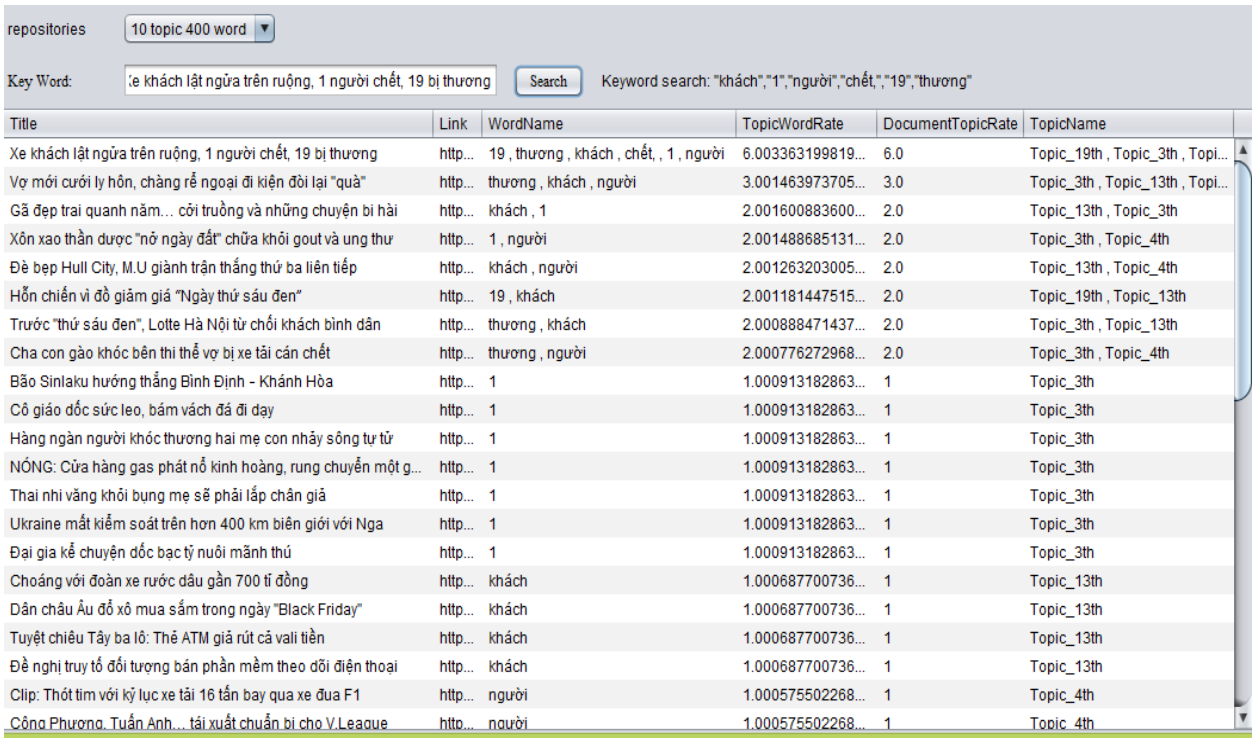


Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName

Hình 5.6. Kết quả thực nghiệm 1 của 10 chủ đề 400 ký tự

Với kết quả trên cho ta thấy được với số ký tự xét cho mỏi chủ đề quá ít sẽ làm mất đi những từ quan trọng trong việc tìm kiếm dẫn đến kết quả tìm kiếm thiếu chính xác.

Tiến hành tìm kiếm thử nghiệm với đoạn văn "*Xe khách lật ngửa trên ruộng, 1 người chết, 19 bị thương*" chương trình loại bỏ stopword và bóc tách được các từ khóa cần tìm là: "*Xe", "khách"," lật ngửa" ," ruộng"," 1"," người","chết"," 19"," thương"* , Các từ có trong dữ liệu là:
"khách","1","người","chết","19","thương".



Title	Link	WordName	TopicWordRate	DocumentTopicRate	TopicName
Xe khách lật ngửa trên ruộng, 1 người chết, 19 bị thương	http...	19 , thương , khách , chết, , 1 , người	6.003363199819...	6.0	Topic_19th , Topic_3th , Topi...
Vợ mới cưới ly hôn, chàng rể ngoại đi kiện đòi lại "quà"	http...	thương , khách , người	3.001463973705...	3.0	Topic_3th , Topic_13th , Topi...
Gã đẹp trai quanh năm... cởi truồng và những chuyện bi hài	http...	khách , 1	2.001600883600...	2.0	Topic_13th , Topic_3th
Xôn xao thần dược "nở ngày đất" chữa khỏi gout và ung thư	http...	1 , người	2.001488685131...	2.0	Topic_3th , Topic_4th
Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp	http...	khách , người	2.001263203005...	2.0	Topic_13th , Topic_4th
Hỗn chiến vì đồ giảm giá "Ngày thứ sáu đen"	http...	19 , khách	2.001181447515...	2.0	Topic_19th , Topic_13th
Trước "thứ sáu đen", Lotte Hà Nội từ chối khách bình dân	http...	thương , khách	2.000888471437...	2.0	Topic_3th , Topic_13th
Cha con gào khóc bên thi thể vợ bị xe tải cán chết	http...	thương , người	2.000776272968...	2.0	Topic_3th , Topic_4th
Bão Sinlaku hướng thẳng Bình Định - Khánh Hòa	http...	1	1.000913182863...	1	Topic_3th
Cô giáo dốc sức leo, bám vách đá đi dạy	http...	1	1.000913182863...	1	Topic_3th
Hàng ngàn người khóc thương hai mẹ con nhảy sông tự tử	http...	1	1.000913182863...	1	Topic_3th
NÓNG: Cửa hàng gas phát nổ kinh hoàng, rung chuyển một g...	http...	1	1.000913182863...	1	Topic_3th
Thai nhi văng khỏi bụng mẹ sẽ phải lắp chân giả	http...	1	1.000913182863...	1	Topic_3th
Ukraine mất kiểm soát trên hơn 400 km biên giới với Nga	http...	1	1.000913182863...	1	Topic_3th
Đại gia kể chuyện dốc bạc tỷ nuôi mãnh thú	http...	1	1.000913182863...	1	Topic_3th
Choáng với đoàn xe rước dâu gần 700 tỉ đồng	http...	khách	1.000687700736...	1	Topic_13th
Dân châu Âu đổ xô mua sắm trong ngày "Black Friday"	http...	khách	1.000687700736...	1	Topic_13th
Tuyệt chiêu Tây ba lô: Thẻ ATM giả rút cả vali tiền	http...	khách	1.000687700736...	1	Topic_13th
Đề nghị truy tố đối tượng bán phần mềm theo dõi điện thoại	http...	khách	1.000687700736...	1	Topic_13th
Clip: Thót tim với kỷ lục xe tải 16 tấn bay qua xe đua F1	http...	người	1.000575502268...	1	Topic_4th
Công Phượng, Tuấn Anh... tái xuất chuẩn bị cho V.League	http...	người	1.000575502268...	1	Topic_4th

Hình 5.7. Kết quả thực nghiệm 1 của 10 chủ đề 400 ký tự

Từ thực nghiệm trên tác giả rút ra rằng để ứng dụng hoạt động với độ chính xác cao và tốc độ truy vấn tốt cần có một sự lựa chọn tốt về số chủ đề và số lượng từ của mỗi chủ đề. Nếu chọn số chủ đề và số lượng từ quá lớn sẽ tạo ra tập tin ontology lớn ảnh hưởng đến tốc độ truy vấn. Nếu chọn ngược lại sẽ ảnh hưởng đến độ chính xác của việc tìm kiếm. Trong quá trình thí nghiệm với 100 trường hợp tìm kiếm ngẫu nhiên và số lượng chủ đề và số lượng từ cho chủ đề khác nhau tác giả thấy rằng chương trình tìm kiếm với độ chính xác cao vào khoảng 92% và giảm

đáng kể số lượng dữ liệu lưu trữ so với việc lưu trữ dữ liệu vào cơ sở dữ liệu SQL do chương trình chỉ lưu những từ quan trọng phục vụ cho việc tìm kiếm. Tuy nhiên do các cơ sở dữ liệu được các nhà đầu tư phát triển nhiều thuật toán hỗ trợ việc lưu trữ xử lý số lượng lớn dữ liệu, tìm kiếm giúp cải thiện đáng kể tốc độ tìm kiếm và sắp xếp dữ liệu. Nếu dùng cơ sở dữ liệu để triển khai mô hình trên sẽ được cải thiệng đáng kể về tốc độ tìm kiếm tuy nhiên ontology lại được một số ưu điểm mà các cơ sở dữ liệu không có như tổ chức dữ liệu kiểu tri thức, hỗ trợ web 4.0, .v.v.

5.2. Đánh giá chương trình:

Tác giả tiến hành thí nghiệm tìm kiếm ngẫu nhiênvới 660 tài liệu, 20 chủ đề, 700 từ để tiến hành đánh giá chương trình.

Chương trình được dánh giá theo phương pháp đánh giá của LSP[7]

5.2.1.Thời giantìm kiếm của chương trình:

Để đo lường thời gian tìm kiếm của chương trình, luận văn tiến hành tìm kiếm theo phương pháp sau:

Khi người dùng điền đầy đủ thông tin và nhấn tìm kiếm thì sẽ lưu lại thời gian tại thời điểm đó ta có được t_1.

Sau khi chương trình tiến hành tìm kiếm và hiển thị kết quả ra ta có được t_2.

Ta tính được thời gian tìm kiếm của chương trình bằng công thức sau:

$$t_i = t_2 - t_1$$

Để tính tốc độ trung bình của chương trình ta cộng tất cả thời gian tìm kiếm của từng lần thực hiện và chia cho tổng số lần thực hiện, ta có công thức như sau:

$$T_{tb} = \frac{\sum_{i=1}^{n} T_i}{\mathrm{n}}$$

Trong đó:

T_{tb}:là thời gian tìm kiếm trung bình của chương trình

T_i: là thời gian tìm kiếm trong mỏi lần thực hiện

n: là tổng số lần thực hiện tìm kiếm.

Tác giả tiến hành thí nghiệm tính thời gian tìm kiếm của các câu sau:

Bảng 5.3. Thí nghiệm độ chính xác của chương trình

Nội dung tìm kiếm	Thời gian tìm kiếm
Thiếu nữ gặp họa vì mải làm điệu bên cột ATM	15s
Việt Nam - Ethiopia thúc đẩy hợp tác kinh tế	17s
Hà Nội không lấy phiếu tín nhiệm 3 Phó chủ tịch	18s
Bão Sinlaku hướng thẳng Bình Định – Khánh Hòa	19s
Mỹ thả nhầm vũ khí vào tay phiến quân IS	18s
Gặp họa vì thử lòng chồng	17s
Sao rực rỡ trên thảm đỏ bế mạc LHP	17s
Lộ diện bạn trai của Vũ Ngọc Anh	15s
Những sao Việt dùng xe sang rước dâu	18s
Lam Trường trao nhẫn cưới cho Yến Phương	18s

Áp dụng công thức trên ta được:

T_{tb} =(15+17+18+19+18+17+17+15+18+18)/10=17,2s

Vậy ta có thể ước lượng thời gian tìm kiếm trung bình của chương trình cho tập tin ontology 356MB, 660 tài liệu 20 chủ đề và 700 từ vào khoảng 17,2 giây.

Về tốc của chương trình trong quá trình tìm kiếm tác giả chia thành nhiều luồn xử lý cùng lúc nên đối với CPU nhiều nhân sẽ cho tốc độ vượt trội.

Trong quá trình tìm kiếm phải tìm ra những kết quả có trọng số cao nhất xếp từ trên xuống, nên quá trình tìm kiếm buộc phải sắp xếp dữ liệu theo trọng số đối với dữ liệu lớn thì việc sắp xếp sẽ ảnh hưởng đến tốc độ tìm kiếm rất nhiều. Nên trong thực tế có tốc độ tốt nhất cần đưa thêm những giảm pháp về thuật toán để hạn chế ảnh hưởng của việc tìm kiếm như là ta tính ra một cái ngưỡng cho trọng số đó rồi tiến hành lấy dữ liệu có trọng số trên ngưỡng đó sẽ hạn chế được quá trình sắp xếp dữ liệu,…

Tuy nhiên với lượng dữ liệu như trên và cấu hình máy như thực nghiệm trên thì chương trình đạt được tốc độ tìm kiếm trung bình khoảng 17,2 giây cho một câu tìm kiếm.

Về thực tế để cải thiện tốc độ tìm kiếm ta có áp dụng những biện pháp để hạn chế việc sắp xếp dữ liệu khi tìm kiếm kết hợp sử dụng cache cho chương trình,… sẽ tăng đáng kể tốc độ tìm kiếm của chương trình.

5.2.2. Độ chính xác của chương trình:

Để đo lường độ chính xác của chương trình luận văn áp dụng theo phương pháp sau:

Ta có D là các tài liệu liên quan với nội dung tìm kiếm và A là các tài liệu mà sau khi tìm kiếm chương trình trả về.

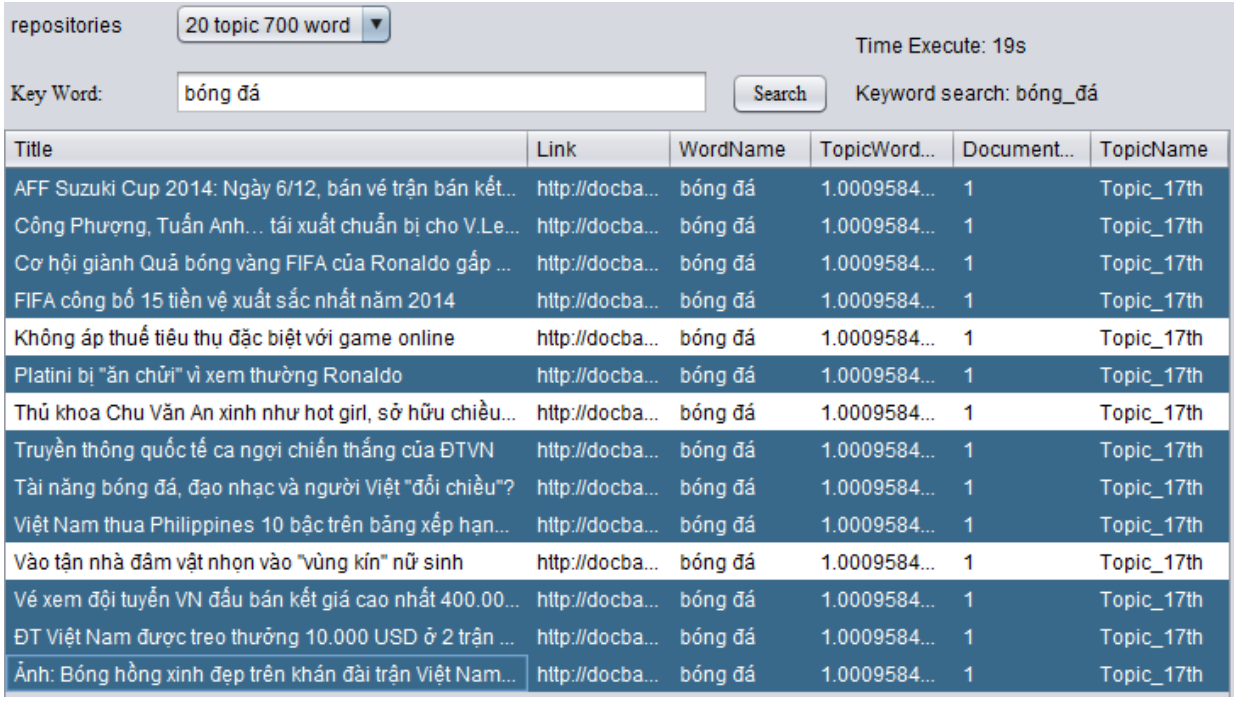
Tính độ chính xác của chương trình ta có công thức sau:

$$R = \frac{|D \cap A|}{|A|}$$

Trong đó:

R là độ chính xác của chương trình.

Trong luận văn này tác giả tiến hành tìm kiếm với từ “bóng đá” chương trình sẽ trả về những kết quả liên quan với bóng đá như sau:



Title	Link	WordName	TopicWord...	Document...	TopicName
AFF Suzuki Cup 2014: Ngày 6/12, bán vé trận bán kết...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Công Phượng, Tuấn Anh… tái xuất chuẩn bị cho V.Le...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Cơ hội giành Quả bóng vàng FIFA của Ronaldo gấp ...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
FIFA công bố 15 tiền vệ xuất sắc nhất năm 2014	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Không áp thuế tiêu thụ đặc biệt với game online	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Platini bị "ăn chửi" vì xem thường Ronaldo	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Thủ khoa Chu Văn An xinh như hot girl, sở hữu chiều...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Truyền thông quốc tế ca ngợi chiến thắng của ĐTVN	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Tài năng bóng đá, đạo nhạc và người Việt "đối chiều"?	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Việt Nam thua Philippines 10 bậc trên bảng xếp hạn...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Vào tận nhà đâm vật nhọn vào "vùng kín" nữ sinh	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Vé xem đội tuyển VN đấu bán kết giá cao nhất 400.00...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
ĐT Việt Nam được treo thưởng 10.000 USD ở 2 trận ...	http://docba...	bóng đá	1.0009584...	1	Topic_17th
Ảnh: Bóng hồng xinh đẹp trên khán đài trận Việt Nam...	http://docba...	bóng đá	1.0009584...	1	Topic_17th

Hình 5.8 Kết quả tìm kiếm của từ khóa "bóng đá"

Trong tìm kiếm trên chương trình trả về tất cả 14 tài liệu bao gồm: 11 tài liệu liên quan đến bóng đávà 3 tài liệu có chứa từ "bóng đá" nhưng không nói nhiều về lĩnh vực bóng đá.

Ví dụ trong tìm kiếm trên trong tài liệu "Không áp thuế tiêu thụ đặc biệt với Game online " có đoạn chứa từ bóng đá như sau"*Trong thời gian vừa qua, UBTVQH đã cho ý kiến để Chính phủ ban hành Nghị định kinh doanh đặt cược đua ngựa, đua chó và bóng đá quốc tế. Theo đó, hình thức kinh doanh đặt cược được phép chỉ gồm 03 loại hình nêu trên.*"

Tuy nhiên trong nguồn dữ liệu của luận văn còn có được những tài liệu liên quan đến bóng đá nhưng không được tìm ra như:

- Thắng U21 Việt Nam 4-3 bằng thi đá luân lưu 11m, U19 HA.GL vào chung kết
- U19 HA.GL-Arsenal JMG 1-2 U21 Sydney: Thua sút về thể lực
- Đè bẹp Hull City, M.U giành trận thắng thứ ba liên tiếp

Áp dụng công thức trên ta tính ra độ chính xác của chương trình như sau:

$$R = 11/14 = 0{,}7857$$

Tiến hành thí nghiệm tương tự với từ khóa “kinh tế” ta có được kết quả như sau:

repositories: 20 topic 700 word — Time Execute: 11s

Key Word: kinh tế — Search — Keyword search: kinh_tế

Title	Link	WordName	TopicWord...	DocumentT...	TopicName
7 ứng viên sáng giá cho ngôi vị Hoa Hậu thế giới 2014	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Bán "lúa non": Món hời triệu đô, đại gia xót lòng	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Chủ đầu tư sửng sốt với giá đất đền bù tăng 100%	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Doanh nghiệp lo trữ cả trăm tỷ hàng Tết	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Giá dầu hạ chóng mặt sau quyết định của OPEC	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Hai người Việt Nam được trao Huân chương Quốc c...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Hà Nội không lấy phiếu tín nhiệm 3 Phó chủ tịch	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Nga sẽ công nhận kết quả bầu cử quốc hội Ukraine	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Ngày mai, xét xử phúc thẩm Nguyễn Đức Kiên cùng đ...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Ngân hàng Nhà nước cam kết bán ngoại tệ	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Nhà họ Lee dần lấy lại quyền lực tại Samsung	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Nỗi buồn sau ngày cưới của cặp đôi đồng tính nữ ở ...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Tiệc sinh nhật nhiều nụ cười, nước mắt của Duy Nhân	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Trúng thưởng casino không bị đánh thuế thu nhập	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Từ mai 29-10, đồng loạt hạ trần lãi suất tiền gửi xuốn...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Xét xử phúc thẩm Nguyễn Đức Kiên và đồng phạm tr...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Đua nhau xây trụ sở nghìn tỷ: Hải Dương muốn chứn...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Đại gia khét tiếng hết ở nhà sang rồi ngồi nhà đá	http://docba...	kinh tế	1.0006877...	1	Topic_13th
Đại gia Việt mua siêu du thuyền hơn 20 tỷ tặng khách...	http://docba...	kinh tế	1.0006877...	1	Topic_13th
“Bóng ma” nhập siêu quay lại	http://docba...	kinh tế	1.0006877...	1	Topic_13th

Hình 5.9 Kết quả tìm kiếm của từ khóa “kinh tế”

Trong tìm kiếm trên chương trình trả về tất cả 20 tài liệu bao gồm: 14 tài liệu liên quan đến kinh tế và 6 tài liệu có chứa từ “kinh tế” nhưng không nói nhiều về lĩnh vực kinh tế.

Tuy nhiên trong nguồn dữ liệu lại có 19 tài liệu liên quan đến lĩnh vực kinh tế.

Áp dụng công thức trên ta tính ra độ chính xác của chương trình như sau:

$$R = 14/20 = 0{,}7$$

Bảng 5.4. Kết quả tìm kiếm ngẫu nhiên của 5 từ khóa

Từ khóa tìm kiếm	Độ chính xác(R_i)
bóngđá	0,7857
Kinh tế	0,7
Văn hóa	0,8
Ngân hàng	0,71428
Lãnh đạo	0,8

Từ bảng trên ta có độ chính xác của chương trình:

$$R_{tb} = 100\% \frac{\sum_1^5 R_i}{5} = 76\%$$

Vậy với 5 lĩnh vực tìm kiếm ngẫu nhiên ta có được độ chính xác trung bình của chương trình vào khoảng 76%.

Do chương trình tìm kiếm theo ngữ nghĩa nên ngoài việc tìm kiếm các dữ liệu người dùng nhập vào chương trình còn đề xuất các nội dung tương tự nội dung cần tìm nhằm hỗ trợ người dùng có được kết quả tìm kiếm phong phú. Hiện trang web www.docbao.vn chưa hỗ trợ tính năng tìm kiếm vì trang web chỉ tập trung vào những nội dung mới nhất mỗi ngày cho người xem không quan tầm nhiều đến nội dung cũ. So với công cụ tìm kiếm trên mạng như google hoặc yahoo,.v.v. thì chương trình hỗ trợ việc tìm kiếm tập trung hơn về lĩnh vực tin tức và tài liệu do tác giả xây dựng.

5.2.3.Độ phản hồi của chương trình:

Độ phản hồi của chương trình là dùng để đo lường các tài liệu liên quan đến tìm kiếm trả về của chương trình. Độ phản hồi dùng để đánh giá tỉ lệ tương đối về mức độ chính xác tìm kiếm chương trình. Để tính độ phản hồi của chương trình ta áp dụng công thức sau:

$$C = \frac{|D \cap A|}{|D|}$$

Trong đó:

C: Độ phản hồi của chương trình

D: Số tài liệu liên quan đến tìm kiếm.

A: Số tàiliệu chương trình trả về trong quá trình tìm kiếm

Với các thông tin đo độ chính xác của chương trình ở trên ta có thể dùng để áp dụng cho phần tính độ phản hồi như sau:

Đối với từ khóa "bóngđá" trong kết quả tìm kiếm trên chương trình trả về tất cả 14 tài liệu bao gồm: 11 tài liệu liên quan đến bóng đá và 3 tài liệu có chứa từ "bóng đá" nhưng không nói nhiều về lĩnh vực bóng đá.

Trong nguồn dữ liệu thì số tài liệu liên quan đến bóng đá là 14 tài liệu. Áp dụng công thức trên ta có được độ phản như sau:

$$C = 11/14 = 0{,}78571$$

Đối với từ khóa "kinh tế" trong kết quả tìm kiếm trên chương trình trả về tất cả 20 tài liệu bao gồm: 14 tài liệu liên quan đến kinh tế và 6 tài liệu có chứa từ "kinh tế" nhưng không nói nhiều về lĩnh vực kinh tế.

Trong nguồn dữ liệu thì số tài liệu liên quan đến kinh tếlà 19 tài liệu. Áp dụng công thức trên ta có được độ phản như sau:

$$C = 14/19 = 0{,}73684$$

Bảng 5.5. Kết quả tìm kiếm đo độ phản hồi

Từ tìm kiếm	Độ phản hồi (C_i)
bóngđá	0,78571
Kinh tế	0,73684
Văn hóa	0,88235
Ngân hàng	0,84615
Lãnh đạo	0,8

$$C_{tb} = 100\% \frac{\sum_1^5 C_i}{5} = 81\%$$

Vậy với 5 lĩnh vực tìm kiếm ngẫu nhiên ta có được độ phản hồi trung bình của chương trình vào khoảng 81%.

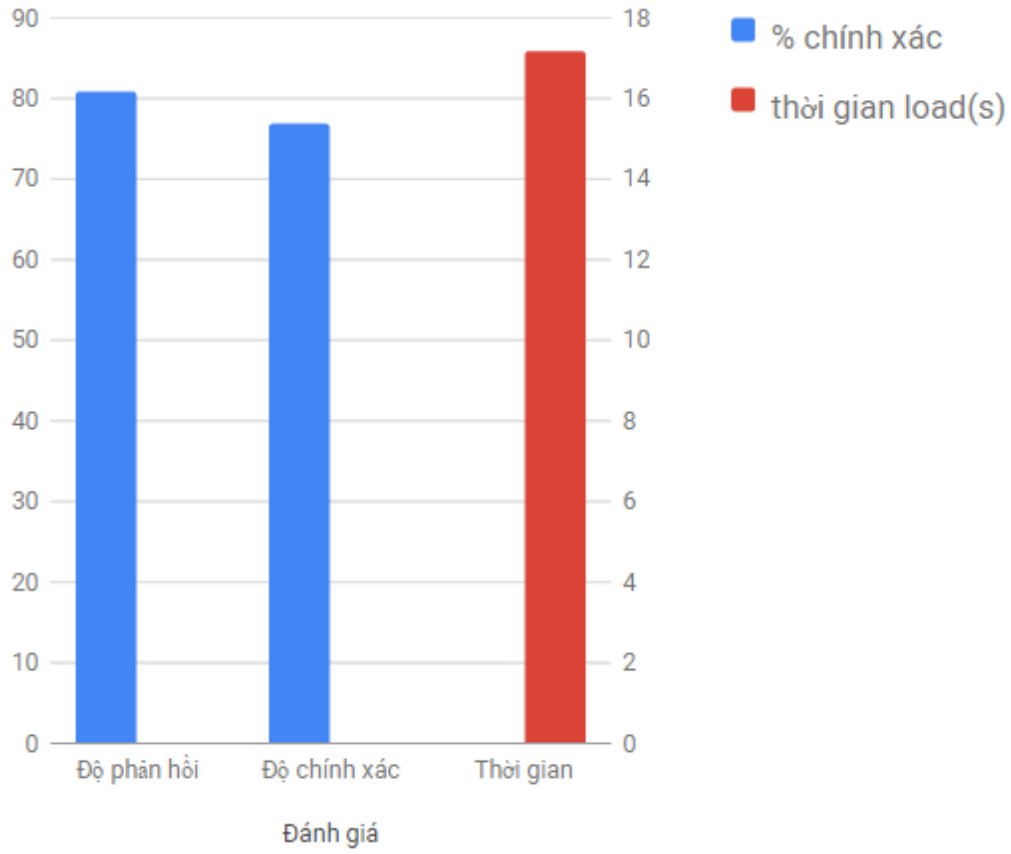
Với các trang web dùng SQL cho việc tìm kiếm thì độ chính xác cao hơn do truy vấn tất cả dữ liệu trong cơ sở dữ liệu tuy nhiên tốc độ lại chậm hơn, tuy nhiên dùng truy vấn SQL để tìm kiếm sẽ không tìm được các tài liệu liên quan do tìm kiếm theo SQL không thể tìm kiếm theo ngữ nghĩa. Với mô hình này nội dung hỗ trợ tìm kiếm được lưu trữ nhỏ hơn việc tìm kiếm chỉ thực hiện trên một dữ liệu nhỏ để lưu trữ các từ quan trọng và trọng số của nó nên nó có tốc độ tìm kiếm tốt hơn các ngôn ngữ truy vấn.

5.2.4. Độtổng quát của chương trình:

Chương trình hỗ trợ tìm kiếm bằng ngôn ngữ tiếng Việt với tất cả các thể loại và lĩnh vực. Với WebCrawler chương trình có thể thu thập các tài liệu từ các trang Web khác nhau trên internet sau đó tiến hành xây dựng mô hình các chủ đề cho việc tìm kiếm giúp cho nguồn tiềm kiếm trở nên phong phú chương trình có thể hỗ trợ tìm kiếm cho tất cả các trang web tiếng Việt và tất cả các chủ đề trên mạng cũng như các tài liệu nội bộ lưu trữ trong các tập tin hoặc cơ sở dữ liệu điều có thể xây dựng được mô hình các chủ đề .

5.2.5. Kết luận:

Chương trình hoạt động với độ chính xác và độ phản hồi cao tuy nhiên tốc độ tìm kiếm cần phải áp dụng thêm những thuật giải hoặc công nghệ khác để giúp tăng tốc độ tìm kiếm của chương trình.



Biểu đồ 5.1 Kết quả đánh giá chương trình

5.2.6. Các vấn đề rút ra được từ thí nghiệm trên:

Từ thí nghiệm trên tác giả rút ra được một số vấn đề như sau:

Khi tác giả chia chủ đề cho tài liệu để tăng độ chính xác và giảm số lượng các đối tượng trong tập tin ontology chúng ta chúng ta có thể tính ra số lượng các chủ đề dựa trên công thức sau:

$$N_{topic} = \frac{N_{document}}{K} (1)$$

N_{topic}: Số lượng chủ đề dùng cho tìm kiếm theo ngữ nghĩa.

$N_{document}$: Số lượng tài liệu dùng cho việc tìm kiếm

K: Hằng số, trong quá trình thực nghiệm tác giả chọn K=20

Công thức trên sẽ trả về số lượng các chủ đề tương ứng với số lượng tài liệu thu thập được trong quá trình thử nghiệm tác giả rút ra được hằng số công thức trên trả về kết quả khá tốt.

Khi tác giả chọn số từ cho quá trình tìm kiếm theo ngữ nghĩa để tăng độ chính xác cho quá trình tìm kiếm và tăng tốc độ cho chương trình, trong quá trình thực nghiệm tác giả đề xuất chọn số từ theo công thức sau:

$$N_{word}=K \cdot N_{document} \quad (2)$$

N_{word} : Số lượng từ dùng cho chương trình

$N_{document}$: Số lượng tài liệu dùng cho việc tìm kiếm

K: Hằng số, trong quá trình thực nghiệm tác giả chọnK=1.1.K càng lớn thì tập tin ontology càng lớn và độ chính xác càng cao

PHẦN KẾT LUẬN

❖ **Kết quả đạt được của luận văn:**

Luận văn tiến hành nghiên cứu xây dựng mô hình tìm kiếm theo ngữ nghĩa phục vụ cho lĩnh vực tìm kiếm hiện nay. Luận văn cũng đạt được những thành tựu như:

Xây dựng được mô hình các chủ đề phục vụ cho việc tìm kiếm theo ngữ nghĩa

Xây dựng chương trình hiện thực việc tìm kiếm.

Các quy trình như thu thập dữ liệu xây dựng mô hình đều được thực hiện một cách tự động hoá .

Mô hình các chủ đề hỗ trợ tìm kiếm theo ngữ nghĩa đưa ra các nội dung cần tìm và đề xuất các nội dungtương tự nội dung cần tìm cho người dùng.

Tuy nhiên để đạt được những hiệu quả tốt nhất cần phải khắc phục một số vấn đề quan trọng như: Tìm cách tăng tốc quá trình tìm kiếm trong trường hợp dữ liệu lớn, giảm thời gian xây dựng tập tin ontology trong trường hợp dữ liệu lớn.

TÀI LIỆU THAM KHẢO

1. Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, Padhraic Smyth (2004). *The Author-Topic Model for Authors and Documents. Dept. of Computer Science UC Irvine, Dept*. of Psychology Stanford University, Dept. of Cognitive Sciences UC Irvine, Dept. of Computer Science UC Irvine

2.David Newman, Arthur Asuncion, Padhraic Smyth, Max Welling (2009) . *Distributed Algorithms for Topic Models*. Department of Computer Science University of California, Irvine Irvine, CA 92697, USA

3.Yuening Hu • Jordan Boyd-Graber , Brianna Satinoff (2011). *Interactive Topic Modeling*. Computer Science University of Maryland, iSchool and UMIACS University of Maryland

4.Gautam Pant, Padmini Srinivasan and Filippo Menczer (2004). *Crawling the Web*. The University of Iowa, Iowa City IA 52242, USA, The University of Iowa, Iowa City IA 52242, USA, 3 School of Informatics Indiana University, Bloomington, IN 47408, USA

5.Cam-Tu Nguyen, Trung-Kien Nguyen & Xuan-Hieu Phan & Le-Minh Nguyen & Quang-Thuy Ha (2008). *Vietnamese Word Segmentation with CRFs and SVMs: An Investigation*. College of Technology, Vietnam National University, Hanoi School of Information Science, Japan Advanced Institute of Science and Technology

6.David M. Blei & Andrew Y. Ng & Michael I. Jordan (2003). *Latent Dirichlet Allocation*. Computer Science Division and Department of Statistics, University of California, Berkeley, CA

7.Nguyen Cam Tu (2008). *Hidden topic discovery toward classification and clustering in VietNamese web documents*. Viet Nam national university, Ha Noi college of technology

8.Jozo Dujmović và Haishi Bai (2006). *Evaluation and Comparison of Search Engines Using the LSP Method*. Department of Computer Science San Francisco State University

Internet:

9.Công cụ tạo các mô hình chủ đề https://code.google.com/p/topic-modeling-tool/

10.Công cụ phân tích chủ đề ẩn, http://jgibblda.sourceforge.net/

11.Công cụ thu thập dữ liệu từ Internet có tính phí http://www.winwebcrawler.com/download.htm

12.Công cụ mã nguồn mở dùng để thu thập dữ liệu từ internet không tính phí https://code.google.com/p/crawler4j/

13.Công cụ tách từ http://jvnsegmenter.sourceforge.net/

14.Công cụ bóc tách từ http://mim.hus.vnu.edu.vn/phuonglh/projects

15.Giới thiệu web ngữ nghĩa http://www.academia.edu/7476371/SW_hay

16.Công cụ soan thảo Ontology http://protege.stanford.edu/download/protege/4.3/installanywhere/Web_Installers/

17.Công cụ lập trình Ontology http://jena.apache.org/documentation/query/

18.Công cụ quản lý tập tin Ontology mã nguồn mở Sesame sever https://jena.apache.org/documentation/inference/#OWLintro

19.Công cụ lập trình Java https://netbeans.org/downloads/

Phụ Lục

Danh sách các Stopword:

Tuy	ai_nấy	bất_chợt	bển
bị	ái	bất_cứ	bệt
Các	ái_chà	bất_giác	biết_bao
Đi	ái_dà	bất_kể	biết_chừng_nào
Đó	alô	bất_kì	biết_đâu
rất	amen	bất_kỳ	biết_đâu_chừng
vào	áng	bất_luận	biết_đâu_đấy
này	ào	bất_nhược	biết_mấy
giữa	ắt	bất_quá	bộ
cho	ắt_hẳn	bất_thình_lình	bội_phần
là	ắt_là	bất_tử	bông
của	âu_là	bây_bẩy	bỗng
và	ầu_ơ	bây_chừ	bỗng_chốc
một	ấy	bây_giờ	bỗng_dưng
không	bài	bây_giờ	bỗng_đâu
lại	bản	bây_nhiêu	bỗng_không
nói	bao_giờ	bấy	bỗng_nhiên
với	bao_lâu	bấy_giờ	bỏ_mẹ
qua	bao_nả	bấy_chầy	bớ
a_ha	bao_nhiêu	bấy_chừ	bởi
a-lô	bay_biến	bấy_giờ	bởi_chưng
à_ơi	bằng	bấy_lâu	bởi_nhưng
á	bằng_ấy	bấy_lâu_nay	bởi_thế
à	bằng_không	bấy_nay	bởi_vậy
à	bằng_nấy	bấy_nhiêu	bởi_vì
ạ	bắt_đầu_từ	bèn	bức cả

ạ_ơi	bập_bà_bập_bõm	béng	cả_thảy
ai	bập_bõm	biết_bao_nhiêu	
ai_ai	bất_đồ	chú	
cái	chết_nỗi	chu_cha	coi_mòi
các	chết_tiệt	chú_mày	con
cả_thảy	chết_thật	chú_mình	còn
cả_thể	chí_chết	chui_cha	cô
càng	chỉn	chùn_chùn	cô_mình
căn	chính	chùn_chũn	cổ_lai
căn_cắt	chính_là	chủn	công_nhiên
cật_lực	chính_thị	chung_cục	cơ
cật_sức	chỉ	chung_qui	cơ_chừng
cây	chỉ_do	chung_quy	cơ_hồ
cha_	chỉ_là	chung_quy_lại	cơ_mà
cha_chả	chỉ_tại	chúng_mình	cơn
chành_chạnh	chỉ_vì	chúng_ta	cu_cậu
chao_ôi	chiếc	chúng_tôi	của
chắc	cho_đến	chứ	cùng
chắc_hẳn	cho_đến_khi	chứ_lị	cùng_cực
chăn_chắn	cho_nên	có_chăng_là	cùng_nhau
chăng	cho_tới	có_dễ	cùng_với
chẳng_lẽ	cho_tới_khi	có_vẻ	cũng
chẳng_những	choa chốc_chốc	cóc_khô	cũng_như
chẳng_nữa	chớ	coi_bộ	cũng_vậy
chẳng_phải	chớ_chi	nghe_chừng	cũng_vậy_thôi
chậc	chợt	nghe_đâu	cứ
chầm_chập	đáng_lí	nghen	cứ_việc
cực_lực	đáng_lý	nghiễm_nhiên	cực_kì
cuộc	đành_đạch	nghỉm	cực_kỳ

cuốn	đánh_đùng	ngõ_hầu	nhất_quyết
dào	đáo_để	ngoải	nhất_sinh
dạ	nấy	ngoài	nhất_tâm
dần_dà	nên_chi	ngôi	nhất_tề
dần_dần	nền	ngọn	nhất_thiết
dầu_sao	nếu	ngọt	nhé
dẫu	nếu_như	ngộ_nhỡ	nhỉ
dẫu_sao	ngay	ngươi	nhiên_hậu
dễ_sợ	ngay_cả	nhau	nhiệt_liệt
dễ_thường	ngay_lập_tức	nhân_dịp	nhón_nhén
do	ngay_lúc	nhân_tiện	nhỡ_ra
do_vì	ngay_khi	nhất	nhung_nhăng
do_đó	ngay_từ	nhất_đán	như
do_vậy	ngay_tức_khắc	nhất_định	như_chơi
dở_chừng	ngày_càng	nhất_loạt	như_không
dù_cho	ngày_ngày	nhất_luật	như_quả
dù_rằng	ngày_xưa	nhất_mực	như_thể
duy	ngày_xửa	nhất_nhất	như_tuồng
dữ	ngăn_ngắt	quá	như_vậy
dưới	những	quá_chừng	nhưng
đã	những_ai	quá_độ	nhưng_mà
đại_để	những_như	quá_đỗi	ren_rén
đại_loại	nhược_bằng	quá_lắm	rén
đại_nhân	nó	quá_sá	rích
đại_phàm	nóc	quá_thể	riệt
đang	nọ	quá_trời	riu_ríu
đáng_lẽ	nổi	quá_ư	rón_rén
phải_chi	nớ	quá_xá	rồi
phải_chăng	nữa	quý_hồ	rốt_cục

phăn_phắt	nức_nở	quyển	rốt_cuộc
phắt	oai_oái	quyết	rút_cục
phè	oái	quyết_nhiên	rứa
phỉ_phui	ô_hay	ra	sa_sả
pho	ô_hô	ra_phết	
phóc	ô_kê	ra_trò	
phỏng	ô_kìa	ráo	sạch
phỏng_như	ồ	ráo_trọi	sao
phót	ôi_chao	rày	sau_chót
phốc	ôi_thôi	răng	sau_cùng
phụt	ối_dào	rằng	sau_cuối
phương_chi	ối_giời	rằng_là	sau_đó
phứt	ối_giời_ơi	rất	sắp
qua_quít	ôkê	rất_chi_là	sất
qua_quýt	ồng	rất_đỗi	sẽ
quả	ơ	rất_mực	sì
quả_đúng	ơ_hay	thái_quá	song_le
quả_là	ơ_kìa	than_ôi	số_là
quả_tang	ờ	thanh	sốt_sột
quả_thật	ớ	thành_ra	sở_dĩ
quả_tình	ơi	thành_thử	suýt
quả_vậy	phải	thảo_hèn	sự
tỏ_vẻ	thỉnh_thoảng	thảo_nào	tà_tà
tò_te	thoạt	thậm	tại
toà	thoạt_nhiên	thậm_chí	tại_vì
toé_khói	thoắt	thật_lực	tấm
toẹt	thỏm	thật_vậy	tấn
tọt	thọt	thật_ra	tự_vì
tốc_tả	thốc	thẩy	tanh

tôi	thốc_tháo	thế	tăm_tắp
tối_ư	thộc	thế_à	tắp
tông_tốc	thôi	thế_là	tắp_lự
tột	thốt	thế_mà	tất_cả
tràn_cung_mây	thốt_nhiên	thế_nào	tất_tần_tật
trên	thuần	thế_nên	tất_tật
trển	thục_mạng	thế_ra	tất_thảy
trệt	thúng_thắng	thế_thì	tênh
trếu_tráo	thửa	thếch	tha_hồ
trệu_trạo	thực_ra	thi_thoảng	thà
trong	thực_vậy	thì	thà_là
trỏng	thương_ôi	thình_lình	thà_rằng
trời_đất_ơi	tiện_thể	trước_tiên	tuốt_tuột
trước	tiếp_đó	trừ_phi	tuy
trước_đây	tiếp_theo	tù_tì	tuy_nhiên
trước_đó	tít_mù	tuần_tự	tuy_rằng
trước_kia	tỏ_ra,	tuốt_luốt	tuy_thế,
trước_nay,	…….	tuốt_tuồn_tuột	……
……..		……	