BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN ĐÌNH BÌNH

NGHIÊN CỨU KHAI PHÁ DỮ LIỆU WEB VÀ ỨNG DỤNG TÌM KIẾM TRÍCH CHỌN THÔNG TIN THEO CHỦ ĐỀ

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. Lê Văn Sơn

Phản biện 1: PGS.TS. Võ Trung Hùng

Phản biện 2: GS.TS. Nguyễn Thanh Thủy

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp Thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 19 tháng 01 năm 2013.

**** Có thể tìm hiểu Luận văn tại:***

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Hơn bốn thập niên kể từ khi Internet ra đời cho đến nay, nó mang lại rất nhiều tiện ích hữu dụng cho người sử dụng như: hệ thống thư điện tử (Email), trò chơi (Game), trò chuyện trực tuyến (Chat), máy truy vấn dữ liệu (Search engine), các dịch vụ thương mại, y tế và giáo dục… Sự phát triển nhanh chóng của mạng Internet đã sinh ra một khối lượng khổng lồ các dữ liệu dạng siêu văn bản (dữ liệu Web). Các tài liệu siêu văn bản chứa đựng văn bản và thường nhúng các liên kết đến các tài liệu khác phân bố trên Web. Ngày nay, Web bao gồm hàng tỉ tài liệu của hàng triệu tác giả được tạo ra và được phân tán qua hàng triệu máy tính được kết nối qua đường hữu tuyến (dây điện thoại, cáp quang) và đường vô tuyến (sóng radio, bức xạ hồng ngoại hay sóng truyền qua vệ tinh) . Web đang ngày càng được sử dụng phổ biến trong nhiều lĩnh vực như báo chí, phát thanh, truyền hình, hệ thống bưu điện, trường học, các tổ chức thương mại, chính phủ…Chính vì vậy lĩnh vực Web mining hay tìm kiếm các thông tin phù hợp có giá trị trên Web là một chủ đề quan trọng trong Data Mining và là vấn đề quan trọng của mỗi đơn vị, tổ chức có nhu cầu thu thập và tìm kiếm thông tin trên Internet. Các hệ thống tìm kiếm thông tin hay nói ngắn gọn là các máy tìm kiếm Web thông thường trả lại một danh sách các tài liệu được phân hạng mà người dùng sẽ phải tốn công chọn lọc trong một danh sách rất dài để có được những tài liệu phù hợp. Ngoài ra các thông tin đó thường rất phong phú, đa dạng và liên quan đến nhiều đối tượng khác nhau. Điều này tạo nên sự nhập nhằng gây khó khăn cho người sử dụng trong việc lấy được các thông tin cần thiết.

Có nhiều hướng tiếp cận khác nhau để giải quyết vấn đề này, các hướng này thường chú ý giảm sự nhập nhằng bằng các phương

pháp tìm kiếm trích chọn thông tin hay thêm các tùy chọn để cắt bớt thông tin và hướng biểu diễn các thông tin trả về bởi các máy tìm kiếm thành từng cụm, lớp để cho người dùng có thể dễ dàng tìm được thông tin mà họ cần. Đã có nhiều thuật toán phân cụm, phân lớp để tìm kiếm thông tin. Tuy nhiên việc tập hợp tài liệu của các máy tìm kiếm là quá lớn và luôn thay đổi để có thể phân cụm ngoại tuyến. Do đó, việc phân cụm phải được ứng dụng trên tập các tài liệu nhỏ hơn được trả về từ các truy vấn và thay vì trả về một danh sách rất dài các thông tin gây nhập nhằng cho người sử dụng cần có một phương pháp tổ chức lại các kết quả tìm kiếm một cách hợp lý. Do những vấn đề cấp thiết được đề cập ở trên nên em chọn đề tài: *"Nghiên cứu khai phá dữ liệu Web và Ứng dụng tìm kiếm trích chọn thông tin theo chủ đề"* .

2. Mục tiêu và nhiệm vụ nghiên cứu

Mục đích của đề tài là nghiên cứu áp dụng tìm kiếm và trích chọn mẫu mới, hữu ích, hiểu được, tiềm ẩn trong Web. Những thông tin theo chủ đề nhanh, chính xác và đầy đủ, thông tin tiềm ẩn bên trong nội dung trang Web đó và những thông tin quan trọng hay những luồng thông tin tốt nhất trên trang Web tìm kiếm trả về kết quả phù hợp với yêu cầu người dùng.

Mục tiêu cụ thể như sau:

Nghiên cứu tìm kiếm

Nghiên cứu kỹ thuật tìm kiếm trên Web.

Hiệu quả tìm kiếm một cách nhanh chóng và chính xác trên Web.

Thông tin tìm kiếm trên Web đầy đủ nguyên vẹn, cô động.

Nghiên cứu về trích chọn

Những thông tin cần khai thác còn tìm ẩn trong một câu, một vùng văn bản và một phân vùng của trang Web .

Những vấn đề khó khăn khi thực hiện về việc trích chọn thông tin chủ đề ẩn trên trang Web.

Đưa ra những luồng thông tin theo chủ đề tốt nhất để đáp ứng yêu cầu người sử dụng.

Ứng dụng thực tế

Sử dụng quy trình khai phá dữ liệu Web trong việc tìm kiếm trích chọn thông tin theo chủ đề trên những trang Web vào thực tế để đáp ứng theo yêu cầu người dùng.

Lấy được những thông tin quí giá tìm ẩn bên trong trang Web đó, để đáp ứng được nhu cầu tìm kiếm tối ưu cho người dùng.

Tìm kiếm trích chọn các mẫu hoặc tri thức hấp dẫn (không tầm thường, ẩn, chưa biết và hữu dụng tiềm năng) từ một tập hợp lớn dữ liệu. để kết quả đạt được đáp ứng yêu cầu xã hội hiện nay.

3. Đối tượng và phạm vi nghiên cứu:

Đối tượng dữ liệu là khai phá kho dữ liệu Web.

Cấu trúc đối tượng là CSDL quan hệ, CSDL đa phương tiện, Dữ liệu dạng Text và dữ liệu Web.

Phạm vi nghiên cứu luận văn này, tôi chỉ áp dụng thuật toán Viterbi, Crawling, Markov, Apriori …

Công cụ hỗ trợ dữ liệu với ngôn ngữ Java trong hệ quản trị cơ sở dữ liệu MySQL, máy tìm kiếm Google, Yahoo….

Đề xuất khai phá dữ liệu Web dựa trên lý thuyết xác suất (điển hình là mô hình xác suất Bayes, mô hình Markov ẩn, mô hình trường ngẫu nhiên có điều kiện…) trong việc tìm kiếm, trích chọn và thử nghiệm thực tế với các một cơ sở dữ liệu có sẵn trên Web.

Đề tài thuộc loại hình khai phá dữ liệu.

4. Phương pháp nghiên cứu

Phương pháp thống kê - phân tích.

Phương pháp lịch sử.

Phương pháp so sánh - đối chiếu.

Phương pháp cấu trúc - hệ thống.

Thu thập và phân tích các tài liệu và thông tin liên quan đến đề tài.

Thảo luận, lựa chọn phương hướng giải quyết vấn đề.

Triển khai xây dựng khai phá dữ liệu.

Kiểm tra, thử nghiệm và đánh giá kết quả trong quá trình khai phá.

5. Bố cục luận văn

Sau phần mở đầu, giới thiệu…, nội dung chính của luận văn được chia thành 3 chương như sau:

Chương 1, Tổng quan về khai phá dữ liệu Web, trình bày cơ sở lý thuyết làm nền tảng để xây dựng ứng dụng, bao gồm: Khai phá dữ liệu và phá hiện tri thức, các mô hình toán học thường dùng trong các bài toán khai phá dữ liệu Web.

Chương 2, Hệ thống tìm kiếm và trích chọn thông tin trên Web, tìm hiểu, giới thiệu và phân tích hệ thống máy tìm kiếm Vietseek, kiến trúc Google ở mức cao và hệ thống trích chọn thông tin dự trên mô hình phân cụm, gán nhãn,CRFs, LDA và thuật toán Viterbi, nêu những vấn đề hạn chế và đề xuất giải pháp khắc phục, đó là giải pháp ứng dụng tìm kiếm trích chọn thông tin theo chủ đề nhằm giải quyết bài toán đặt ra.

Chương 3, trình bày chi tiết về mô hình kiến trúc tổng thể của hệ thống và phương pháp xây dựng ứng dụng. Tiến hành kịch bản thử nghiệm trên số liệu thực tế, sau đó đánh giá kết quả đạt được và khả năng triển khai ứng dụng trên toàn hệ thống.

Cuối cùng là phần đánh giá, kết luận và hướng phát triển của đề tài.

CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU WEB

1.1. KHAI PHÁ DỮ LIỆU VÀ PHÁT HIỆN TRI THỨC

1.1.1. Tại sao lại khai phá dữ liệu

1.1.2. Định nghĩa khai phá dữ liệu

Định nghĩa 1: (Frawley, Piatetski – Shapiro và Matheus)

Phát hiện tri thức trong cơ sở dữ liệu (đôi khi còn được gọi là khai phá dữ liệu) là một quá trình không tầm thường nhận ra những mẫu có giá trị, mới, hữu ích tiềm năng và hiểu được trong dữ liệu.

Định nghĩa 2*: Khai phá dữ liệu (datamining)*

Khai phá dữ liệu là quá trình trích ra những thông tin dùng được, đúng và chưa biết trước từ cơ sở dữ liệu lớn, rồi dùng thông tin này để ra các quyết định.

Giáo sư Tom Mitchell đã đưa ra định nghĩa của KPDL như sau: *"KPDL là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai."*

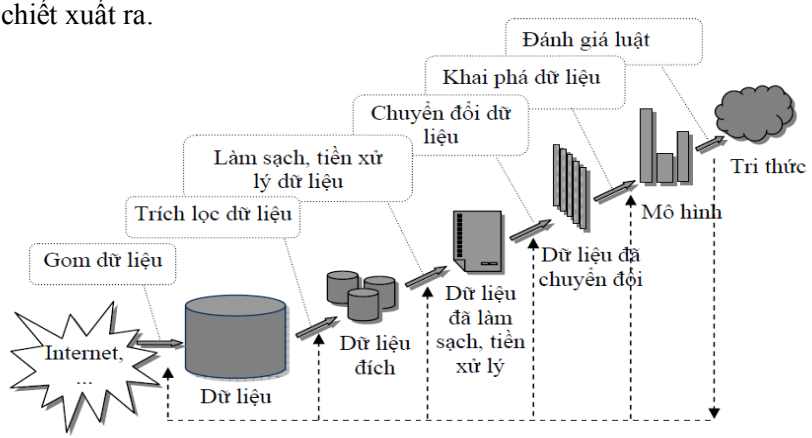
Với một cách tiếp cận ứng dụng hơn, Tiến sĩ Fayyad đã phát biểu: *"KPDL, thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu."*

Ngoài ra theo tài liệu của Weldon năm 1996, khai phá dữ liệu là việc phát hiện tri thức nhờ các công cụ hoàn thiện sử dụng thống kê truyền thống, trí tuệ nhân tạo và đồ họa máy tính. Nói tóm lại, KPDL là một quá trình học tri thức mới từ những dữ liệu đã thu thập được.

1.1.3. Quá trình khai phá tri thức (KDD)

Quá trình khai phá dữ liệu sẽ tiến hành qua 6 giai đoạn như hình 1.1,

Bắt đầu của quá trình là kho dữ liệu thô và kết thúc với tri thức được chiết xuất ra.



Hình 1.1: Quá trình phát hiện tri thức

1.1.4. Các hướng tiếp cận và các kỹ thuật áp dụng trong khai phá dữ liệu

1.1.5. Phân loại các hệ thống khai phá dữ liệu

1.1.6. Những vấn đề chú trọng và ứng dụng trong khai phá dữ liệu

1.2. CƠ SỞ DỮ LIỆU FULLTEXT VÀ HYPERTEXT

1.2.1. Cơ sở dữ liệu Fulltext

1.2.2. Cơ sở dữ liệu HyperText

1.2.3. So sánh đặc điểm của dữ liệu Fulltext và dữ liệu trang web

1.3. KHAI PHÁ DỮ LIỆU VĂN BẢN (TEXTMINING) VÀ KHAI PHÁ DỮ LIỆU WEB (WEBMINING)

1.3.1. Khai phá dữ liệu văn bản

1.3.2. Khai phá dữ liệu Web

Khai phá Web như là việc trích chọn ra các thành phần được quan tâm hay được đánh giá là có ích cùng các thông tin tiềm năng từ các tài nguyên hoặc các hoạt động liên quan tới World Wide Web

Chương 2: HỆ THỐNG TÌM KIẾM VÀ TRÍCH CHỌN THÔNG TIN TRÊN WEB

2.1. HỆ THỐNG TÌM KIẾM

2.1.1. Nhu cầu

2.1.2.Máy tìm kiếm

2.1.3 Module Crawler trong các máy tìm kiếm

2.1.4. Các thuật toán crawling

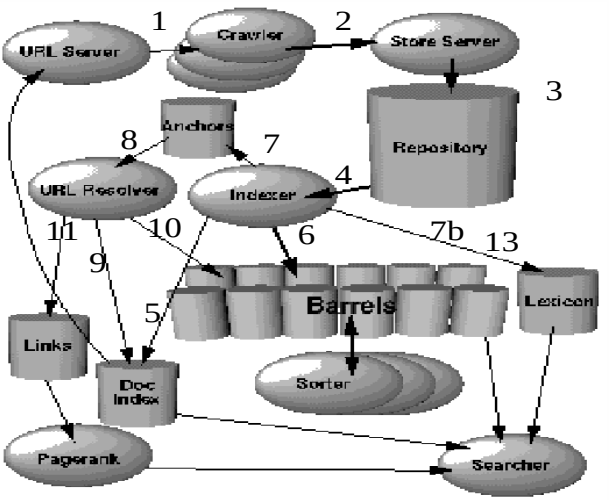
2.1.5. Phân tích và đánh chỉ số

Theo ông Sergey Brin và Lawrence Page đã trình bày cụ thể về quan điểm của nhà thiết kế máy tìm kiếm Google:

- URLserver: gửi danh sách URL Webpage sẽ đưa về cho các crawler phân tán.
- Các crawler: Tải nội dung Webpage về gửi cho StoreServer.
- StoreServer: nén và lưu Webpage lên đĩa (vào kho chứa).
- Indexer có các chức năng:
  - ✓ Đọc tài liệu từ kho chứa
  - ✓ Giải nén
  - ✓ Gọi Parser để phân tích cú pháp đưa trang Web.
- Index cùng Sorter: gán DocID cho Web page (DocID được gán mỗi khi Parser phát hiện một URL mới).
- Mỗi tài liệu
  - ✓ Được biến đổi thành tập các xuất hiện của các từ khóa (gọi là hit)
  - ✓ Hit: từ khóa, vị trí trong tài liệu, font (cỡ, ...), hoa/thường. Indexer
  - ✓ Phân bố các hit thành tập các "barrel" lưu trữ các chỉ số đã được sắp xếp.
- Indexer:
  - ✓ Phân tích các siêu liên kết

✓ Lưu các thông tin quan trọng trong file "anchor" cho phép xác định

- Nguồn, đích của siêu liên kết
- Nội dung văn bản trong siêu liên kết.



Hình 2.6 Kiến trúc Google ở mức cao

- Sinh từ điển tra cứu từ khóa: Văn bản trong siêu liên kết:

✓ Nhiều hệ chỉ gắn vào trang nguồn

✓ Google gắn vào cả trang đích⇒lợi ích

- Cho thông tin chính xác hơn, thậm chí chính trang web
  - "tóm tắt"
  - "qua chuyên gia xử lý"

✓ Index cho trang web

- "Không văn bản" (ảnh, chương trình, CSDL ...)
- Xử trí trường hợp trang web chưa tồn tại
- Lấy văn bản anchor làm "nội dung"!

- ✓ Tư tưởng này có trong WWW Worm (1994) và có trong Google
  - Kết quả chất lượng hơn.
  - Chú ý: crawling 24 triệu trang có tới 259 triệu anchor.

- URLsolver
  - ✓ Đọc file anchor.
  - ✓ Biến đổi URL tương đối thành URL tuyệt đối.
- URLsolver cập nhật lại theo chỉ số DocID
- URLsolver đưa text anchor vào *index thuận* (hướng trỏ anchor).
- URLsolver sinh CSDL liên kết gồm các cặp liên kết ($DocID_1$, $DocID_2$) được dùng để tính PageRank.
- Sorter
  - ✓ Đọc các Barrel (xếp theo DocID) để sắp lại theo WordID tạo ra các index ngược.
  - ✓ Sinh ra danh sách các wordID và gia số trong index ngược.
- DumpLexicon
  - ✓ Lấy từ lexicon + danh sách wordID
  - ✓ Sinh ra lexicon mới.
- Searcher
  - ✓ Chạy do webserver trả lời câu hỏi
  - ✓ Dựa trên lexicon mới PageRank, index ngược

2.2.TRÍCH CHỌN THÔNG TIN TRÊN WEB

2.2.1. Trích chọn thông tin

a. ***Khái niệm***

Trích chọn thông tin (IE – Information Extraction) là quá trình

lấy thông tin từ các nguồn ở những định dạng không đồng nhất và chuyển thành một dạng đồng nhất. Dữ liệu sau khi trích chọn được sử dụng, trình bày trực tiếp cho người dùng, lưu vào cơ sở dữ liệu để xử lý sau đó hay sử dụng cho những hệ thống tìm kiếm thông tin như một dữ liệu đã qua bước tiền xử lý.

b. *Phân loại hệ thống trích chọn thông tin từ web*

Ngày nay, có rất nhiều hệ thống trích chọn thông tin từ web được các nhà phát triển nghiên cứu và xây dựng. Các tiêu chí để phân loại một hệ thống trích chọn thông tin từ web như sau:

Dựa vào mức độ can thiệp của con người trong quá trình trích chọn thông tin: các hệ thống trích chọn thông tin có thể được chia ra làm 4 loại: thủ công, có giám sát, bán giám sát và không giám sát. Trong đó, các hệ thống hoàn toàn tự động, không có sự can thiệp của con người đang được các nhà nghiên cứu quan tâm nhất.

Dựa vào tầng dữ liệu được trích chọn: một trang web sẽ có nhiều trang HTML, một trang HTML sẽ có nhiều record và một record sẽ có nhiều thuộc tính. Do đó, dựa vào kết quả thông tin trích chọn được ở tầng nào, các hệ thống trích chọn được chia ra làm 4 loại: tầng thuộc tính (attribute), tầng record, tầng trang HTML (page) và tầng trang web (site). Hiện tại các hệ thống xử lý ở tầng thuộc tính và record chiếm đa số. Và cho đến nay, vẫn chưa thấy xuất hiện các hệ thống trích chọn thông tin ở tầng site.

Dựa vào các phương pháp trích chọn thông tin: Các hệ thống trích chọn thông tin cũng được chia thành 3 dạng:

• Các hệ thống dựa trên các phương pháp thủ công: sử dụng các phương pháp gán nhãn, các cách lấy thông tin trực tiếp từ cơ sở dữ liệu hoặc từ các dịch vụ web (web service).

• Các hệ thống dựa trên các phương pháp heuristic: Các phương pháp thống kê, tập luật, sử dụng các mẫu thông tin, dựa vào cấu trúc
cây,… được sử dụng để trích chọn thông tin.

• Các hệ thống dựa trên các phương pháp học: Sử dụng các phương pháp mô hình Markov, CRFs, ngữ nghĩa, học trên cấu trúc cây,… để giúp cho các hệ thống hiểu và trích chọn thông tin chính xác hơn.

2.2.2. Khuynh hướng phát triển của khai phá dữ liệu Web theo chủ đề

Bài toán mà ông Rich Caruana và cộng sự giải quyết được mô tả sơ bộ như sau: Cho trước một tập hợp (khoảng 300000) tài liệu khoa học cần phát hiện ra các chủ đề khoa học chủ chốt và qua đó dự báo được xu hướng nghiên cứu, phát triển các chủ đề khoa học mới thuộc lĩnh vực khoa học máy tính. Giải pháp tiến hành không cần khai thác các chỉ dẫn của các công trình mà chỉ cần sử dụng nội dung các công trình, hình sau mô tả kết quả nghiên cứu phát hiện ra 13 cụm chủ đề và cung cấp ý tưởng về xu hướng phát triển của 13 cụm chủ đề. Trong nghiên cứu của mình về bài toán trên, GS John E. Hopcroft một chuyên gia hàng đầu của nước Mỹ về lĩnh vực CNTT đã trình bày hướng phát triển của khoa học máy tính. Ông đề cập tới một số yếu tố nổi bật trong tương lai tác động tới sự chuyển biến của khoa học máy tính. Từ nội dung văn bản của mỗi công trình nghiên cứu, chúng ta nhận được tên các tác giả, các tài liệu tham khảo, tên tạp chí, hội thảo…

Ông Rich Caruana và cộng sự đặt ra các mục tiêu cơ bản cần hướng tới:

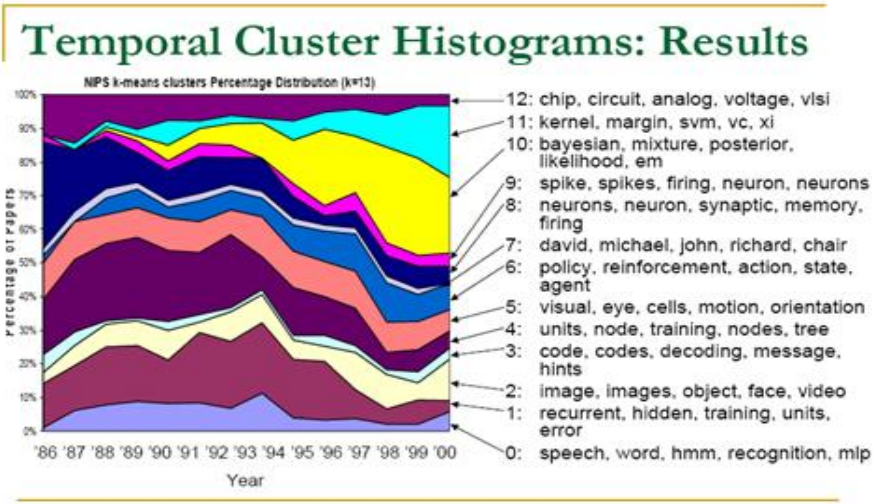
• Tìm ra diễn biến quá trình phát triển theo thời gian của các chủ đề khoa học theo một số tiêu chí như tỷ lệ các tài liệu theo

chủ đề, các chủ đề nổi bật mới, thời điểm một chủ đề cụ thể đạt đỉnh cao nhất, chủ đề nào đang tàn lụi …để tìm ra được các chủ đề có vai trò chủ chốt trong tập hợp các chủ đề.

• Nhận biết được các tài liệu có uy thế là tài liệu giới thiệu các ý tưởng mới và có chỉ số ảnh hưởng lớn.

• Nhận biết được tác giả có uy thế là tác giả có ảnh hưởng lớn đối với sự phát triển của các chủ đề.

Nhìn vào biểu đồ hình 2.8 cho thấy:



Hình 2.8. Tình hình phát triển một số nhóm chủ đề trong khoa máy qua phân cụm tài liệu khoa học

+ Một số nhóm chủ đề nghiên cứu hiện đang trong giai đoạn phát triển tốt như nhóm 10 (Bayesian, mixture, posterior, likelihood, em), nhóm 9 (Spike, spikes, firing, neuron, neurons) và nhóm 2 (Image, images, object, face, video).

+ Một số nhóm chủ đề nghiên cứu hiện đang phát triển song đang có xu hướng chững lại như nhóm 12 (chip, circuit, analog, voltage, vlsi), nhóm 4 (units, node, training, nodes, tree)

+ Các nhóm còn lại đang phát triển bình thường.

Đặc biệt nhóm chủ đề 12 chững lại song vẫn có số lượng lớn công trình nghiên cứu được công bố.

2.2.3. Thuật toán Viterbi

Thuật toán Viterbi mang tên tác giả Andrew Viterbi, là thuật toán quy hoạch động nhằm tìm dãy tương tự nhất của các trạng thái ẩn, được ứng dụng khá phổ biến để giải quyết bài toán giải mã. Khi sử dụng phương pháp máy trạng thái hữu hạn, đặc biệt đối với bài toán trích chọn thông tin trên Web. Nội dung thuật toán có sự kết hợp các nội dung của đồ thị và xác suất.

Thuật toán Viterbi được coi như tìm đường đi ngắn nhất dọc theo đồ thị là:

Input: $Z=z_1, z_2, \ldots, z_n$ // dãy quan sát đầu vào

Khởi tạo:

$K \leftarrow 1$ // chỉ số lặp

$S(c_1) \leftarrow c_1$

$L(c_1) \leftarrow 0$ // Biến chứa tổng độ dài, khởi tạo là 0

Đệ quy:

Repeat

For $\forall$ bộ chuyển $t_k=(c_k, c_{k+1})$

$L(c_k, c_{k+1}) \leftarrow L(c_k) + L[t_k=(c_k, c_{k+1})]$ theo $\forall c_k$

Tìm $L(c_{k+1})= \min L(c_k, c_{k+1})$

For mỗi c_{k+1}

Lưu $L(c_{k+1})$ và vết $S(c_{k+1})$ tương ứng

$k \leftarrow k +1$

Until $k = n$

2.2.4. Mô hình trường ngẫu nhiên (Conditional Random Fields – CRFs)

2.2.5. Mô hình phân cụm và gán nhãn cụm với chủ đề ẩn

a. *Độ tương đồng câu và các phương pháp*

- *Độ tương đồng câu*
- *Các phương pháp tính độ tương đồng câu*
- *Phương pháp tính độ tương đồng câu sử dụng độ đo Cosine*
- *Phương pháp tính độ tương đồng câu dựa vào chủ đề ẩn*

Mỗi câu có thể có nhiều phân phối xác suất topic. Với hai câu thứ i và j, chúng ta sử dụng độ đô cosine để tính độ tương đồng giữa hai câu đã được làm giàu với chủ đề ẩn.

$$sim_{i,j}(topic-parts) = \frac{\prod_{k=1}^{K} t_{i,k} \times t_{j,k}}{\sqrt{\sum_{k=1}^{K} t_{i,k}^2}\sqrt{\sum_{k=1}^{K} t_{j,k}^2}}$$

$$sim_{i,j}(word-parts) = \frac{\prod_{t=1}^{|V|} w_{i,t} \times w_{j,t}}{\sqrt{\sum_{t=1}^{|V|} w_{i,t}^2}\sqrt{\sum_{t=1}^{|V|} w_{j,t}^2}}$$

Cuối cùng, tổ hợp hai độ đo trên để ra độ tương đồng giữa hai câu:

$Sim(s_i , s_j) = \lambda x Sim(\text{topic - parts}) + (1 - \lambda) x Sim(\text{word - parts})$

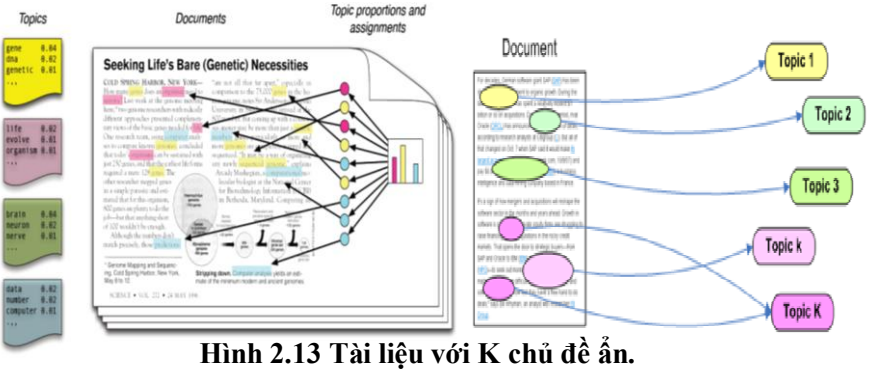
Trong công thức trên, λ là hằng số trộn, thường nằm trong đoạn [0,1]. Nó quyết định việc đóng góp giữa 2 độ đo tương đồng. Nếu λ = 0 , độ tương đồng giữa hai câu không có chủ đề ẩn. Nếu λ = 1, đo độ tương đồng giữa hai câu chỉ tính với chủ đề ẩn

2.2.6. Mô Hình Latent Dirichlet Allocation (LDA)

a. *Phân tích thông tin chủ đề dựa trên mô hình chủ đề LDA*

Phân tích chủ đề cho văn bản nói riêng và cho dữ liệu Web nói chung có vai trò quan trọng trong việc "hiểu" và định hướng thông tin trên Web. Khi ta hiểu một trang Web có chứa những chủ đề hay thông tin gì thì dễ dàng hơn cho việc xếp loại, sắp xếp, và tóm tắt nội

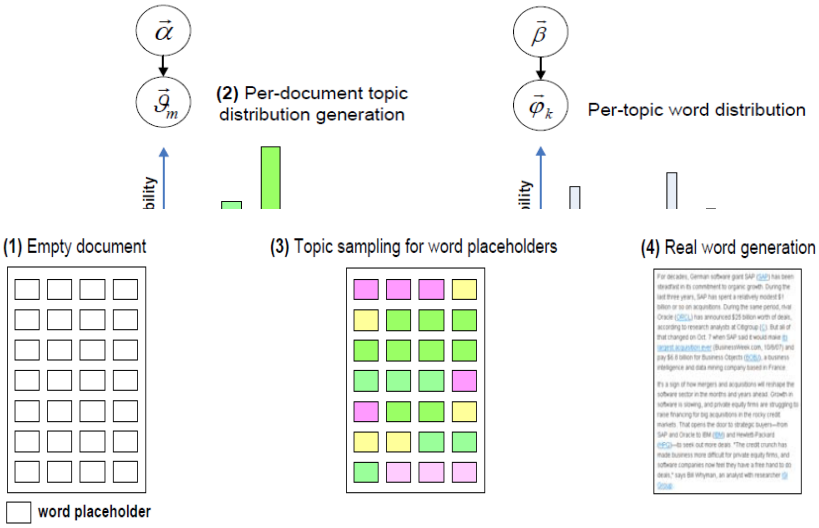
dung của trang Web đó. Trong phân lớp văn bản, mỗi văn bản thường được xếp vào một lớp cụ thể nào đó. Trong phân tích chủ đề, chúng ta giả sử mỗi văn bản đề cập đến nhiều hơn một chủ đề (*K* chủ đề) và mức độ liên quan đến chủ đề được biểu diễn bằng phân phối xác suất của của tài liệu đó trên các chủ đề.



Hình 2.13 Tài liệu với K chủ đề ẩn.

Mô hình sinh trong LDA

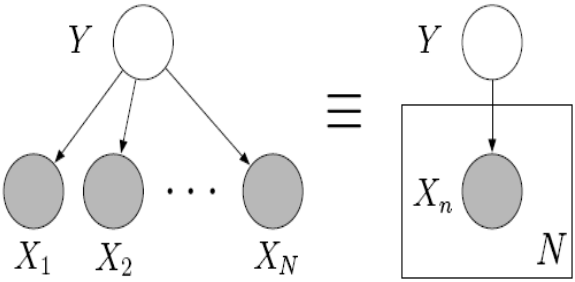
Theo Blei, Ng [8], dù pLSA một bước tiến trong việc mô hình hóa text theo xác suất nhưng nó chưa hoàn thiện. Lí do là pLSA chưa phải là một mô hình xác suất được xác định rõ ràng ở mức văn bản (document). Hệ quả là nó gặp vấn đề khi xác định xác suất với những văn bản nằm ngoài tập huấn luyện (trainning set). Hơn nữa, nó còn dẫn tới việc tăng tuyến tính số tham số của mô hình so với độ lớn của tập văn bản (corpus). LDA là mô hình phân tích chủ đề có thể xử lý được những vấn đề đó. Vì thế tôi đã chọn LDA để sử dụng trong khóa luận. Hình 2.14 giới thiệu những bước cơ bản trong tiến trình sinh của LDA.



Hình 2.14. Tiến trình sinh văn bản LDA

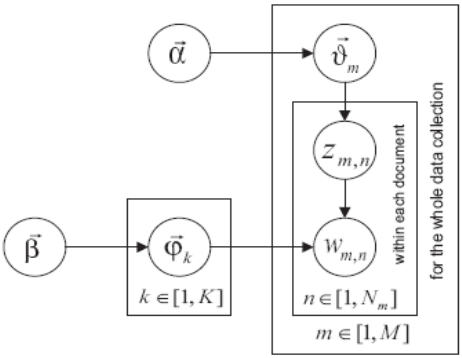
Phân phối Dirichlet ẩn (Latent Dirichlet Allocation)

LDA là mô hình sinh văn bản được giới thiệu bởi Blei, Ng và cộng sự [8] với pLSA về ý tưởng cơ bản là dựa trên việc coi văn bản là sự pha trộn của các chủ đề. Nhưng LDA là một mô hình Bayes ba mức: mức corpus, mức văn bản (document), mức từ (word). Hình 2.15 & 2.16 mô tả tiến trình sinh văn bản bằng phương pháp LDA:



Hình 2.15. Kí hiệu khối lặp lại

Cho một corpus của M tài liệu biểu diễn bởi D={d_1,d_2, …, d_M}, trong đó, mỗi tài liệu m trong corpus bao gồm Nm từ wi rút từ một tập Vocabulary của các term {t_1, …, t_v}, V là số từ. LDA cung cấp một mô hình sinh đầy đủ chỉ ra kết quả tốt hơn các phương pháp trước. Quá trình sinh ra document như sau:



Hình 2.16. Mô hình biểu diễn của LDA

Các kí hiệu:

Các khối hình vuông hình 18 biểu diễn các quá trình lặp.

Tham số đầu vào: α và β (tham số mức corpus).

α : Dirichlet prior on $\vec{\vartheta}_m$.

β : Dirichlet prior on $\vec{\varphi}_k$.

M : số văn bản trong corpus: D = {d_1 ,d_2 ,...,d_M }.

K : số chủ đề ẩn.

V : số từ trong tập từ vựng

N_m : Số lượng các từ trong tài liệu thứ m (hay còn gọi là độ dài của văn bản d_m).

$z_{m,n}$: chủ đề của từ w_n trong văn bản d_m (hay chỉ số chủ đề).

$w_{m,n}$: từ thứ n trong văn bản d_m chỉ bởi $z_{m,n}$.

$$\underline{\Theta} = \{\vec{\varphi}_m\}_{k=1}^{k} \text{ (KxV matrix)}$$

$\vec{\vartheta}_m$: Phân phối của topic trong document thứ m, $\vec{\vartheta}_m$ biểu diễn tham số cho p(z|d=m), thành phần trộn topic cho tài liệu m. Một tỷ lệ cho mỗi tài liệu

$$\underline{\Theta} = \{\vec{\vartheta}_m\}_{m=1}^{M} \text{ (MxK matrix)}$$

$\vec{\varphi}_m$: phân phối của các từ được sinh từ chủ đề $z_{m,n}$. $\vec{\varphi}_m$ biểu diễn tham số cho p(t|z=k), thành phần trộn của topic k, một tỷ lệ cho mỗi topic.

LDA sinh một tập các từ $w_{m,n}$ cho các văn bản $\vec{d}_m$ bằng cách:

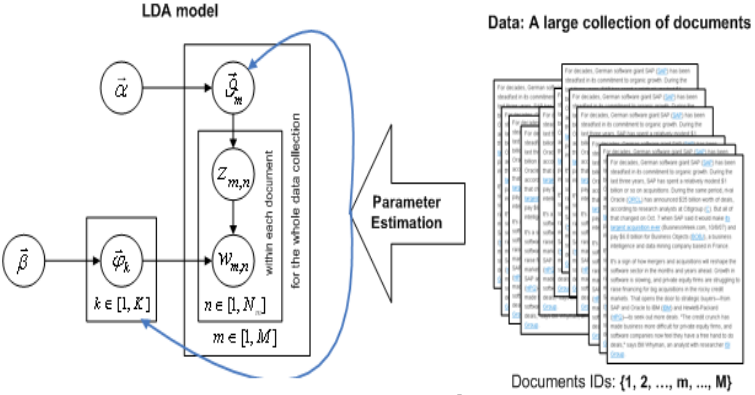
• Với mỗi văn bản m, sinh ra phân phối topic $\vec{\vartheta}_m$ cho văn bản theo Dir(α).

• Với mỗi từ, $z_{m,n}$ được lấy mẫu dựa vào phân phối topic Mult($\vec{\vartheta}_m$).

• Với mỗi topic index $z_{m,n}$, dựa vào phân phối từ $\vec{\varphi}_k$, $w_{m,n}$, được sinh ra.

Ước lượng giá trị tham số và inference thông qua Gibbs Sampling cho mô hình LDA.

Ước lượng tham số cho mô hình LDA bằng phương pháp cực đại hóa hàm likelihood trực tiếp và một cách chính xác có độ phức tạp thời gian rất cao và không khả thi trong thực tế.



Hình 2.18. Ước lượng tham số tập dữ liệu văn bản.

Người ta thường sử dụng các phương pháp xấp xỉ như Variational Methods và Gibbs Sampling . Gibbs Sampling được xem là một thuật toán nhanh, đơn giản, và hiệu quả để huấn luyện LDA.

Cho trước một tập các văn bản, tìm xem topic model nào đã sinh ra tập các văn bản trên. Bao gồm:

- Tìm phân phối xác suất trên tập từ đối với mỗi topic $\vec{\varphi}_m$.
- Tìm phân phối topic của mỗi tài liệu $\vec{\vartheta}_m$.

CHƯƠNG 3
ỨNG DỤNG VÀ THỰC NGHIỆM

3.1. ỨNG DỤNG

3.1.1. Ứng dụng tìm kiếm trích chọn theo chủ đề được lưu kho dữ liệu

Trong kho CSDL chứa các chủ đề ẩn và xác suất của các chủ đề được xác định theo mật độ ưu tiên.

• Bộ tách từ làm nhiệm vụ khi nhập vào một câu truy vấn bộ này sẽ phân tích trong câu truy vấn thuộc chủ đề ẩn nào

• Chủ đề ẩn của câu truy vấn có nhiệm vụ phân tích trong câu truy vấn thuộc loại chủ đề nào có mật độ xác xuất cao để ưu tiên theo thứ tư tăng dần, hiển thị danh sách theo chủ đề có sự trích chọn

3.1.2. Ứng dụng tìm kiếm trích chọn theo chủ đề được lưu kho CSDL trên Internet

Pha tương tác với các máy tìm kiếm Google

Pha tiền xử lý dữ liệu

Pha sắp xếp văn bản và câu theo độ quan trọng

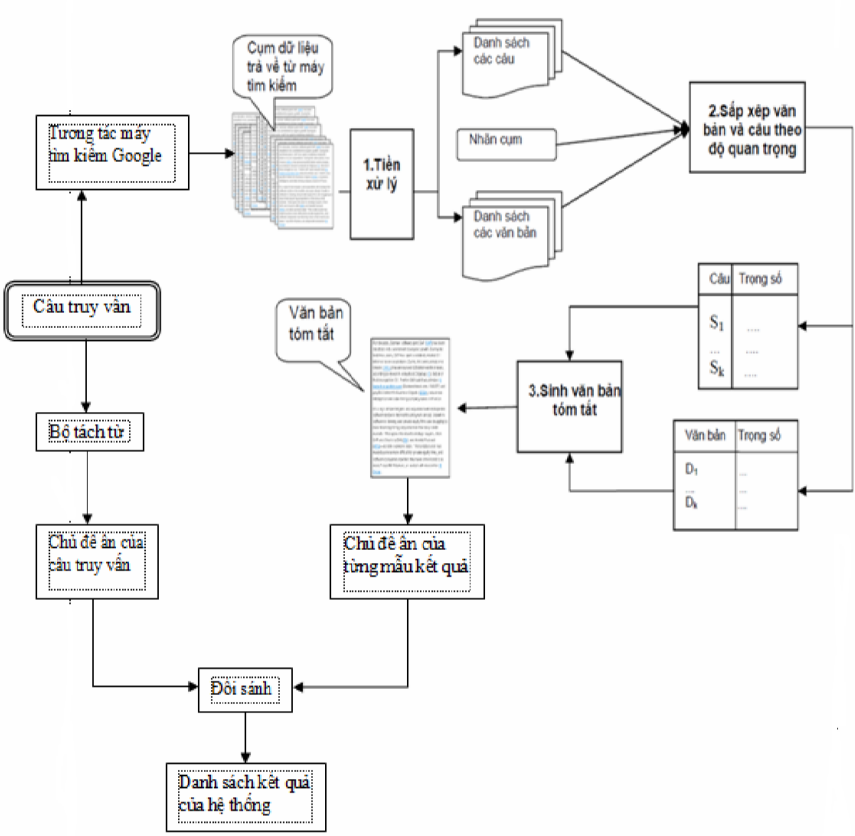
Pha sinh văn bản tóm tắt

Trong pha sinh văn bản tóm tắt, các câu được sắp xếp đã được sắp xếp ở pha trên sẽ được sắp xếp lại. Trọng số độ quan trọng của

câu sẽ được bổ sung thêm trọng số của văn bản chưa câu đấy, việc này sẽ giúp văn bản tóm tắt không có sự chồng chéo về mặt nội dung. ScoreTotal là công thức tính lại độ quan trọng của câu:

$$ScoreTotal(s_k)=(\lambda*Score(s_k)+(1- \lambda)* Score(D_i))$$

- **S_k:** là câu cần tính độ quan trọng.
- **D_i**: là văn bản chưa s_k.
- **Score(s_k), Score(D_i)**: là trọng số độ quan trọng của s_k và D_i được tính ở pha trước.
- **λ**: là các hằng số trộn nằm trong ngưỡng [0,1] thể hiện sự đóng góp của hai độ đo Score(s_k) và Score(D_i) (Các hằng số này sẽ được ước lượng trong quá trình thực nghiệm).



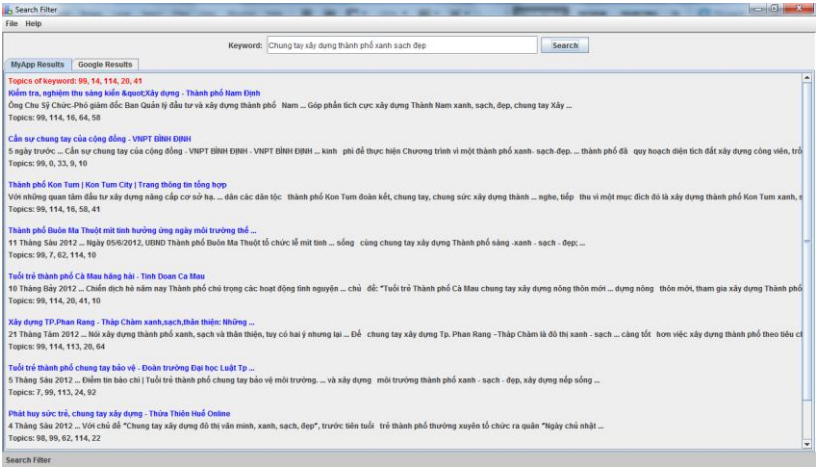
Hình 3.2. Mô hình hệ thống tìm kiếm trích chọn theo chủ đề

3.2. THỰC NGHIỆM

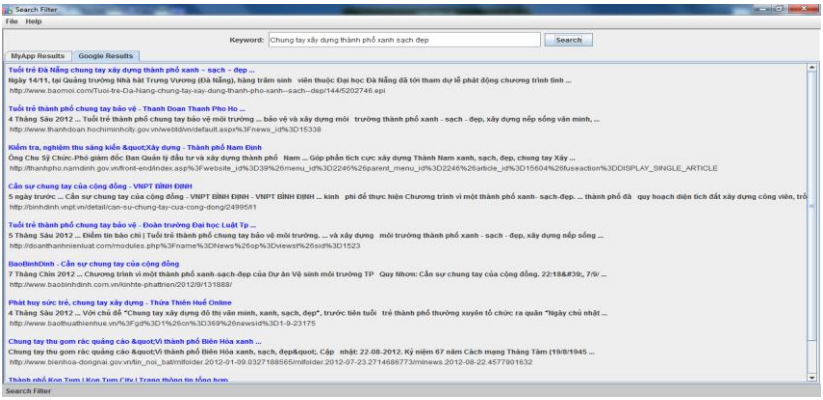
3.2.1. Môi trường thực nghiệm

3.2.2. Một số giao diện chương trình

1. Công cụ tìm kiếm trích chọn thông tin theo chủ đề "Chung tay xây dựng thành phố xanh sạch đẹp" lấy từ tài liệu trên Internet và lưu trữ vào kho dữ liệu theo chủ đề ẩn .



2. Công cụ tìm kiếm trích chọn chủ đề "Chung tay xây dựng thành phố xanh sạch đẹp" trên Internet.



KẾT LUẬN

1. KẾT QUẢ ĐẠT ĐƯỢC

Về mặt khoa học

Luận văn đã tiến hành phân tích, tìm hiểu được quy trình khai phá dữ liệu Web. Phát hiện ra những vấn đề còn hạn chế để đề xuất đưa ra giải pháp nhằm có những phương án khắc phục để nâng cao hiệu quả trong công việc tìm kiếm trích chọn thông tin theo chủ đề nhanh và chính xác hơn.

Nắm được các phương pháp và các mô hình toán học như đồ thị, xác suất Bayes và mô hình biểu diễn dữ liệu văn bản, CRFs, LDA… Áp dụng để giải quyết yêu cầu luận văn đã đặt ra.

Nghiên cứu và vận dụng giải thuật crawl, k-means, Viterbi… để xây dựng mô hình khai phá dữ liệu Web theo chủ đề .

Về mặt thực tiễn

Luận văn đã nêu được giải pháp kỹ thuật để xây dựng hệ thống trợ giúp quyết định nắm bắt được những luồng thông tin tốt trong công tác quản lí và kinh danh.

Tìm kiếm trích chọn thông tin trên Web theo chủ đề giúp chúng ta có một cái nhìn tổng thể, biết được những gì nổi bật trong quá khứ, đâu là xu hướng thông tin hiện tại và đâu là những hướng sẽ nỗi lên trong tương lai gần. Tổng hợp thông tin hướng chủ đề trên Web cũng giúp chúng ta sắp xếp lại thông tin và theo dõi các luồng thông tin tốt hơn.

Xây dựng được ứng dụng có khả năng phân tích tốt các dữ liệu về nhà trường trong những năm qua về một chủ đề nào đó.

Tìm ra diễn biến quá trình phát triển theo thời gian của các chủ đề nào đó, theo một số tiêu chí như tỷ lệ các tài liệu theo chủ đề, các chủ đề nổi bật mới, thời điểm một chủ đề cụ thể đạt đỉnh cao nhất, chủ đề nào đang tàn lụi …để tìm ra được các chủ đề có vai trò chủ

chốt trong tập hợp các chủ đề.

Hệ thống có thể giúp cho tìm kiếm trích chọn thông tin nhanh chính xác, giúp cho ban giám hiệu nhà trường và lãnh đạo các đơn vị liên kết ra quyết định một cách kịp thời, khoa học, tránh được các tình huống quyết định theo cảm tính nhằm hạn chế các trường hợp đưa ra quyết định sai không hiệu quả dẫn đến thiệt hại về kinh tế, lãng phí thời gian và tiền bạc của người học.

Có thể nói, đây là một công cụ hữu ích nhằm cung cấp cho đơn vị nắm được những chủ đề thời sự nổi bật, có thêm một giải pháp hỗ trợ về công tác quản lí sau này.

2. HẠN CHẾ

Hệ thống hiện tại chỉ tương tác dữ liệu được lưu trữ kho dữ liệu Google, chưa kết nối và truy xuất dữ liệu trực tiếp đến cơ sở dữ liệu của Yahoo, MSN, Altavista... Do đó cần một khoảng thời gian để khai phá kho dữ liệu này.

3. HƯỚNG PHÁT TRIỂN

Nghiên cứu cải tiến hệ thống thông qua giải pháp thu nhận đánh giá phản hồi của người dùng đối với chất lượng tìm kiếm trích chọn thông tin theo chủ đề để chất lượng tìm kiếm định hướng hơn tới người dùng.

Cải tiến quá trình lưu trữ và đánh chỉ mục để tăng tốc cho các việc tìm kiếm trích chọn thông tin, qua đó tăng tốc độ trả lời câu hỏi cho mô hình hỏi đáp tiếng Việt, Xây dựng và triển khai hệ thống hỏi đáp tiếng Việt cho người sử dụng.

Tự động phân lớp các trang web tiếng Việt bổ sung thêm vào cây chủ đề.

Tìm kiếm trích chọn thông tin trên Web theo chủ đề giúp chúng ta có một cái nhìn tổng thể, biết được những gì nổi bật trong quá khứ, đâu là xu hướng thông tin hiện tại và đâu là những hướng sẽ

nỗi lên trong tương lai gần. Tổng hợp thông tin hướng chủ đề trên Web cũng giúp chúng ta sắp xếp lại thông tin và theo dõi các luồng thông tin tốt hơn, giúp cho nhà quản lý đưa ra quyết định và nhà kinh tế dự báo trước những rủi ro xảy ra.

Mô hình LDA hướng phát triển lên mô hình SAM để tăng hiệu quả, đầy đủ và khái quát hơn cho việc thực hiện phân tích từ các tập dữ liệu văn bản giám sát hoặc hoàn toàn phi giám sát.