BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRẦN HỮU PHÚ

XÂY DỰNG HỆ THỐNG THU THẬP THÔNG TIN TỰ ĐỘNG PHỤC VỤ CẬP NHẬT NỘI DUNG CHO TRANG WEB

Chuyên ngành : **KHOA HỌC MÁY TÍNH**
Mã số : **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 1: **PGS.TSKH. TRẦN QUỐC CHIẾN**

Phản biện 2: **PGS.TS. LÊ MẠNH THẠNH**

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 16 tháng 10 năm 2011

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

Sự phát triển nhanh chóng của mạng Internet kèm theo khối lượng dữ liệu khổng lồ, đa dạng và tăng trưởng không ngừng. Đối với mọi cá nhân, tổ chức, việc cập nhật thường xuyên các nguồn thông tin trên mạng Internet là rất quan trọng, quyết định đến hiệu quả, thành công, trong lĩnh vực hoạt động của mình. Tuy nhiên, việc tìm kiếm được các thông tin phù hợp và có giá trị đối với người truy cập từ mạng Internet sẽ tốn kém thời gian do dữ liệu nằm phân tán trên mạng và không được sắp xếp, phân loại như mong muốn. Do đó, việc tìm kiếm, trích lọc và thu thập các thông tin có ý nghĩa từ Internet về một điểm truy cập tập trung phục vụ nhu cầu người khai thác là một bài toán cần thiết được giải quyết.

Nhu cầu thu thập và phát lại các thông tin cần thiết từ internet đối với trang TTĐT Quảng Nam là rất lớn. Là một cán bộ đang công tác tại Sở Thông Tin & Truyền Thông Quảng Nam, đơn vị quản lý cổng TTĐT này, tôi thiết nghĩ cần thiết phải đưa ra một giải pháp xây dựng hệ thống thu thập thông tin tự động phục vụ cập nhật nội dung cho trang TTĐT .

Từ những lý do như trên nên tôi chọn đề tài:

"Xây dựng hệ thống thu thập thông tin tự động phục vụ cập nhật nội dung cho trang web".

Các nội dung chính nghiên cứu trong luận văn :

- *Tìm hiểu tổng quan kỹ thuật thu thập thông tin trên Internet, tổng quan về khai phá dữ liệu, các thuật toán phân cụm dữ liệu.*

- *Tiếp cận bài toán Tìm kiếm và phân cụm tài liều web ứng dụng thuật toán K-means và các kỹ thuật tiền xử lý và biểu diễn dữ liệu.*
- *Áp dụng Bài toán Tìm kiếm và phân cụm tài liệu web vào việc Xây dựng hệ thống thu thập tin tự động hỗ trợ thu thập và biên tập các tin tức từ các nguồn trên Internet, phục vụ nhu cầu người truy cập một cách tập trung các tin tức liên quan đến chủ đề cần thu thập trên Trang TTĐT Quảng Nam.*

2. Mục tiêu và nhiệm vụ

Nắm vững cơ sở lý thuyết về khai phá dữ liệu và các kỹ thuật phân cụm tài liệu web, qua đó xây dựng hệ thống thu thập thông tin tự động phục vụ cập nhật nội dung trang TTĐT Quảng Nam, kết quả thực nghiệm đáp ứng yêu cầu đề ra..

3. Đối tượng và phạm vi nghiên cứu

Khai phá dữ liệu là một lĩnh vực rộng lớn trong ngành khoa học máy tính, phân cụm tài liệu web là một trong những lĩnh vực ứng dụng điển hình của khai phá dữ liệu, tuy nhiên có rất nhiều kỹ thuật thông qua rất nhiều thuật toán cho bài toán phân cụm dữ liệu, trong phạm vi của đề tài này, chủ yếu tập trung đi vào nghiên cứu lý thuyết về phân cụm tài liệu web và các thuật toán, trọng tâm đi vào phân tích, ứng dụng thuật toán K-Means để tiến hành cài đặt ứng dụng thực nghiệm.

4. Phương pháp nghiên cứu

Trong đề tài này sử dụng phương pháp nghiên cứu lý thuyết kết hợp với phát triển ứng dụng thực nghiệm. Trên cơ sở lý thuyết về khai phá dữ liệu, và cụ thể hơn nữa là lý thuyết về phân cụm dữ liệu và các thuật toán phân cụm tài liệu, tiến hành cài đặt và phân tích tối

ưu các thuật toán, đi đến chọn lựa thuật toán phù hợp cho việc triển khai xây dựng ứng dụng thực nghiệm.

Tiến hành đánh giá kết quả thực nghiệm để đưa ra hướng phát triển mở rộng của đề tài để đáp ứng những yêu cầu triển khai thực tế.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Về mặt lý thuyết: đề tài tổng hợp các cơ sở lý thuyết về khai phá dữ liệu, phân cụm tài liệu, phân tích các phương pháp phân cụm, cài đặt và đánh giá hiệu quả của các thuật toán phân cụm và từ đó chọn thuật toán tối ưu nhất để triển khai thực nghiệm.

Về mặt thực tiễn: với việc phát triển và triển khai thực nghiệm ứng dụng thu thập tin tự động trên Internet, đề tài này có thể ứng dụng vào thực tế là hỗ trợ cho việc thu thập và biên tập tin tức cho Trang thông tin điện tử tỉnh Quảng Nam, đem lại hiệu quả kinh tế nhờ tiết kiệm thời gian và chi phí.

6. Cấu trúc luận văn

Ngoài phần mở đầu, phần kết luận, mục lục, danh mục hình vẽ, danh mục bảng biểu, tài liệu tham khảo, phụ lục, phần chính của luận văn gồm 3 chương như sau :

Chương 1: Nguyên cứu tổng quan

Chương 2 : Phân tích thiết kế hệ thống

Chương 3 : Xây dựng và triển khai hệ thống.

Chương 1: NGHIÊN CỨU TỔNG QUAN

1.1 Tổng quan về kỹ thuật thu thập thông tin trên Internet

Có nhiều hình thái về thu thập và bóc tách thông tin đã được nghiên cứu và phát triển. Chúng ta có một loạt khái niệm như Robot, Search, Web Crawler, Data Wrapper, Web Spider, Web Clipping, Semantic Web,... để mô tả về những hình thái khai thác nội dung thông tin trên Internet. Xin lấy mô hình tìm kiếm là một ví dụ: Nội dung sau khi khai thác có thể được lưu trữ trong các hệ thống database và phát hành lại tới người dùng trực tiếp thông qua hệ thống tích hợp, tìm kiếm, lọc, chia sẻ đặt tả,...hay sử dụng cho một mục đích chuyên biệt nào đó. Google là minh chứng cụ thể cho giải pháp đó, các Website tồn tại trên Internet sẽ được Google Crawler ghé thăm và thu thập lại toàn bộ, sau đó nội dung được lưu trữ trong cơ sở dữ liệu, được đánh chỉ mục,... và được tìm kiếm mỗi khi có yêu cầu từ phía người dùng. Một sản phẩn khác là GoogleNews lại có nhiệm vụ tổng hợp tất cả các tin tức diễn ra hàng ngày trên Internet. Ở Việt nam, ta có thể tìm kiếm những mô hình tương tự như Baomoi.com hay Thegioitin.com, VietSpider, InewsCrawler.

Có nhiều giải pháp khác nhau như RSS, phân tích cây DOM, web clustering (phân cụm tài liệu web)... Trong khóa luận này ta sẽ chọn giải pháp web clustering.

1.2 Tổng quan về Khai phá dữ liệu

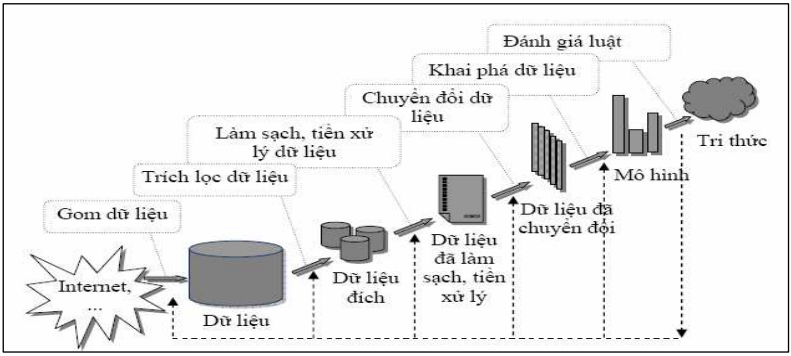
1.2.1 Khái niệm Khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là một khái niệm ra đời vào những năm cuối của thập kỷ 1980. Nó là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong lượng lớn dữ liệu được lưu trữ trong các CSDL, kho dữ liệu... Đây là giai đoạn quan trọng nhất trong tiến trình Phát hiện tri thức từ cơ sở dữ liệu, các tri thức này hỗ

trợ trong việc ra quyết định trong khoa học và kinh doanh và các hoạt động khác.

1.2.2 Quá trình phát hiện tri thức

Quá trình Phát hiện trị thức được tiến hành qua 6 giai đoạn như hình 1.1:



Hình 1.1 : Quá trình phát hiện tri thức

Bắt đầu của quá trình là kho dữ liệu thô và kết thúc với tri thức được chiết xuất ra. Về lý thuyết thì có vẻ rất đơn giản nhưng thực sự đây là một quá trình rất khó khăn gặp phải rất nhiều vướng mắc như: quản lý các tập dữ liệu, phải lặp đi lặp lại toàn bộ quá trình, v.v... Quá trình gồm 6 bước:

(1) Gom dữ liệu

(2) Trích lọc dữ liệu

3) Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu

4) Chuyển đổi dữ liệu

(5) Khai phá dữ liệu

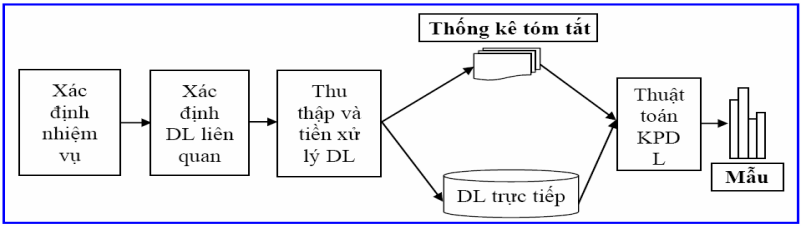
(6) Đánh giá các luật và biểu diễn tri thức

1.2.3 Quá trình khai phá dữ liệu

Khai phá dữ liệu là một giai đoạn quan trọng trong quá trình phát hiện tri thức. Về bản chất, nó là giai đoạn duy nhất tìm ra được

thông tin mới, thông tin tiềm ẩn có trong CSDL chủ yếu phục vụ cho mô tả và dự đoán.

Quá trình Khai phá dữ liệu bao gồm các bước chính được thể hiện như Hình 1.2 sau:



Hình 1.2: Quá trình Khai phá dữ liệu

- *Xác định nhiệm vụ:* Xác định chính xác các vấn đề cần giải quyết.
- *Xác định các dữ liệu liên quan:* Dùng để xây dựng giải pháp.
- *Thu thập và tiền xử lý dữ liệu:* Thu thập các dữ liệu liên quan và tiền xử lý chúng sao cho thuật toán KPDL có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều các vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), v.v..
- *Thuật toán khai phá dữ liệu:* Lựa chọn thuật toán KPDL và thực hiện việc KPDL để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định... tương ứng với ý nghĩa của nó.

1.2.4 Các phương pháp khai phá dữ liệu

Với hai mục đích khai phá dư liệu là Mô tả và Dự đoán, người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- *Luật kết hợp (association rules)*
- *Phân lớp (Classfication)*
- *Hồi qui (Regression)*
- *Trực quan hóa (Visualiztion)*
- *Phân cụm (Clustering)*
- *Tổng hợp (Summarization)*
- *Mô hình ràng buộc (Dependency modeling)*
- *Biểu diễn mô hình (Model Evaluation)*
- *Phân tích sự phát triển và độ lệch (Evolution and deviation analyst)*
- *Phương pháp tìm kiếm (Search Method)*

Có nhiều phương pháp khai phá dữ liệu được nghiên cứu ở trên, trong đó có 3 phương pháp được các nhà nghiên cứu sử dụng nhiều nhất đó là: Luật kết hợp, Phân lớp dữ liệu và Phân cụm dữ liệu.

1.2.5 Các bài toán thông dụng trong Khai phá dữ liệu

Trong Khai phá dữ liệu, các bài toán có thể phân thành 4 loại chính: Phân lớp dữ liệu, Dự đoán dữ liệu, Tìm luật liên kết (Association Rule), Phân cụm dữ liệu.

1.3 Phân cụm dữ liệu

1.3.1 Khái niệm Phân cụm dữ liệu

Phân cụm dữ liệu là một kỹ thuật trong Data Mining, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định.

Trong học máy, phân cụm dữ liệu được xem là vấn đề học không có giám sát, vì nó phải đi giải quyết vấn đề tìm một cấu trúc trong tập hợp các dữ liệu chưa biết trước các thông tin về lớp hay các thông tin về tập ví dụ huấn luyện.

Trong lĩnh vực khai thác dữ liệu, các vấn đề nghiên cứu trong phân cụm chủ yếu tập trung vào tìm kiếm các phương pháp phân cụm hiệu quả và tin cậy trong cơ sở dữ liệu lớn.

Trong lĩnh vực khai phá dữ liệu Web, phân cụm có thể khám phá ra các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu...

1.3.2 Ứng dụng của Phân cụm dữ liệu

Phân cụm dữ liệu có thể được ứng dụng trong nhiều lĩnh vực như: thương mại, sinh học, thư viện, bảo hiểm, quy hoạch đô thị, nghiên cứu trái đất, WWW…

1.3.3 Các tiêu chuẩn của Phân cụm dữ liệu

Phân cụm là một thách thức trong lĩnh vực nghiên cứu ở chỗ những ứng dụng tiềm năng của chúng được đưa ra ngay chính trong những yêu cầu đặc biệt của chúng. Sau đây là những yêu cầu cơ bản của phân cụm trong KPDL:

- Có khả năng mở rộng
- Khả năng thích nghi với các kiểu thuộc tính khác nhau
- Khám phá các cụm với hình dạng bất kỳ
- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào Khả năng thích nghi với dữ liệu nhiễu
- Ít nhạy cảm với thứ tự của các dữ liệu vào
- Số chiều lớn
- Phân cụm có tính ràng buộc

- Dễ hiểu và dễ sử dụng:

1.3.4 Các phương pháp Phân cụm dữ liệu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các cách tiếp cận chính sau :

1.3.4.1 Phân cụm phân hoạch

1.3.4.2 Phân cụm dữ liệu phân cấp

1.3.4.3 Phân cụm dữ liệu dựa trên mật độ

1.3.4.4 Phân cụm dữ liệu dựa trên lưới

1.3.4.5 Phân cụm dữ liệu dựa trên mô hình

1.3.4.6 Phân cụm dữ liệu có ràng buộc

1.3.5 Các đặc tính của thuật toán phân cụm

1.3.5.1 Mô hình dữ liệu

- *Mô hình dữ liệu tài liệu*
- *Mô hình dữ liệu số*
- *Mô hình phân loại dữ liệu*
- *Mô hình dữ liệu kết hợp*

1.3.5.2 Độ đo sự tương tự

Để có thể nhóm các đối tượng dữ liệu, một ma trận xấp xỉ đã được sử dụng để tìm kiếm những đối tượng (hoặc phân cụm) tương tự nhau.

1.3.6 Thuật toán K-means

K-means là một trong số những phương pháp học không có giám sát cơ bản nhất thường được áp dụng trong việc giải các bài toán về phân cụm dữ liệu. Mục đích của thuật toán k-means là sinh ra k cụm dữ liệu {C1, C2,…,Ck} từ một tập dữ liệu chứa n đối tượng

trong không gian d chiều Xi =(xi1,xi2,…xid)(i=$\overline{1,n}$) sao cho hàm tiêu chuẩn:

$$\sum_{i=1}^{k}\sum_{x\in C_i}|x-m_i|^2$$

đạt giá trị tối thiểu. Trong đó: mi là trọng tâm của cụm Ci, là khoảng cách giữa hai đối tượng.

1.4 Đề xuất giải pháp

1.4.1 Đặt vấn đề

Máy tìm kiếm có thể giúp chúng ta tìm kiếm các thông tin cần thiết phân tán trên mạng internet, mặc dù danh sách tài liệu trả về theo truy vấn đã được xác định thứ hạng quan trọng của nó, nhưng thông thường người dùng khó đưa ra quyết định chính xác đối với các tài liệu vì khả năng gây nhập nhằn của danh sách trả về cũng như người dùng không đủ kiên nhẫn để duyệt qua tất cả các tài liệu. Để thu thập các thông tin có ý nghĩa chúng ta có thể đưa ra giải pháp là: phân cụm các tài liệu trả về từ máy tìm kiếm để chọn ra cụm tài liệu phù hợp nhất phục vụ cho mục đích sử dụng. Như vậy, giải pháp được đưa ra đồng nghĩa với việc chúng ta đi giải quyết bài toán *tìm kiếm và phân cụm tài liệu web.* Trên cơ sở áp dụng các lý thuyết về khai phá dữ liệu, chúng ta sẽ đi giải quyết bài toán này.

1.4.2 Các yêu cầu

- Tính phù hợp
- Tính đa hình
- Sử dụng các mẩu thông tin
- Tốc độ
- Tính gia tăng.

1.4.3 Hướng tiếp cận

Thay vì dựa vào liên kết trang để xác định trọng số cho trang, ta có thể tiếp cận theo một hướng khác đó là dựa vào nội dung của các tài liệu để xác định trọng số, nếu các tài liệu "gần nhau" về nội dung thì sẽ quan trọng tương đương và sẽ thuộc về cùng một nhóm, nhóm nào gần với câu truy vấn hơn sẽ quan trọng hơn.

Cách tiếp cận giải quyết được các vấn đề sau:

+ Kết quả tìm kiếm sẽ được phân thành các cụm chủ đề khác nhau, tùy vào yêu cầu cụ thể mà người dùng sẽ xác định chủ đề mà họ cần.

+ Quá trình tìm kiếm và xác định trọng số cho các trang chủ yếu tập trung vào nội dung của trang hơn là dựa vào các liên kết trang.

+ Giải quyết được vấn đề từ/cụm từ đồng nghĩa trong câu truy vấn của người dùng.

+ Có thể kết hợp phương pháp phân cụm trong lĩnh vực khai phá dữ liệu với các phương pháp tìm kiếm đã có.

1.4.4 Quá trình tìm kiếm và phân cụm tài liệu

Quá trình bao gồm các bước sau:

1.4.4.1 Tìm kiếm dữ liệu trên web

Nhiệm vụ chủ yếu của giai đoạn này là dựa vào tập từ khóa tìm kiếm để tìm kiếm và trả về tập gồm toàn văn tài liệu, tiêu đề, mô tả tóm tắt tài liệu, URL,… tương ứng với các trang đó. Dữ liệu được lưu trữ vào CSDL để tiếp tục được xử lý.

1.4.4.2 Tiền xử lý và biểu diễn dữ liệu

Quá trình làm sạch dữ liệu và chuyển dịch các tài liệu thành các dạng biểu diễn thích hợp bao gồm các bước:

- *Chuẩn hóa văn bản*
- *Xóa bỏ từ dừng*
- *Kết hợp các từ có cùng gốc*

- *Xây dựng từ điển*
- *Tách từ, số hóa văn bản và biểu diễn tài liệu*

1.4.4.3 Phân cụm tài liệu:

Sau khi đã tìm kiếm, trích rút dữ liệu và tiền xử lý, sử dụng kỹ thuật phân cụm để phân cụm tài liệu bằng thuật toán K-means như đã nêu.

1.4.5 Ứng dụng

Với hướng tiếp cận như trên, bài toán Tìm kiếm và Phân cụm tài liệu web có thể áp dụng trong việc xây dựng hệ thống thu thập tin tự động. Việc tìm kiếm thông tin trên internet đã được tận dụng thế mạnh của các Search Engine trên Internet hiện nay, việc phân cụm các kết quả tìm kiếm bằng thuật toán K-means có thể đem lại các cụm tài liệu với độ tương tự của các tài liệu trong cụm là rất cao và từ đó hỗ trợ người dùng ra quyết định trong việc chọn lựa một trong các cụm tài liệu để phục vụ cho mục đích nào đó của mình .

Chương 2: PHÂN TÍCH THIẾT KẾ HỆ THỐNG

2.1 Hiện trạng và nhu cầu

Xây dựng hệ thống thu thập thông tin tự động phục vụ cập nhật nội dung cho trang TTĐT là việc làm hết sức cần thiết.

Trang TTĐT Quảng nam có số lượng truy cập rất lớn và nhu cầu tìm kiếm thông tin trên đó là rất cao. Hiện nay chủ đề "Xây dựng nông thôn mới" là chủ đề đang được quan tâm nhất, các thông tin về chủ đề này được đăng rất nhiều trên các báo bộ, ngành, địa phương và cần được thu thập về ngay trên trang TTĐT Quảng Nam để phục vụ nhu cầu của nhân dân trong tỉnh.

Các thông tin thu thập về và đăng tải lại trên trang TTĐT Quảng Nam phải có nội dung thật sự phù hợp với chủ đề và các thông tin là chính thống, không lấy từ các nguồn báo không rõ ràng.

2.2 Yêu cầu của hệ thống

2.2.1 Cơ sở lý thuyết áp dụng

- Hệ thống được xây dựng trên cơ sở áp dụng phương pháp phân cụm các tài liệu web trả về của máy tìm kiếm.

- Thuật toán phân cụm được áp dụng là thuật toán K-means (với số cụm tùy chọn)

- Các lý thuyết hỗ trợ như độ đo độ tương tự, chuẩn hóa, tách từ, biểu diễn dữ liệu theo vecto không gian cũng được áp dụng.

2.2.2 Xác định các yêu cầu của hệ thống

2.2.2.1 Yêu cầu phi chức năng

- Hệ thống được phát triển để tích hợp phục vụ cho trang TTĐT Quảng Nam do đó nó phải được thiết kế tuân theo mô hình của Portal đang sử dụng (Liferay).

- Đảm bảo yếu tố tốc độ trong quá trình xử lý thu thập và phân cụm tài liệu.

- Hệ thống được xây dựng với các module chức năng chuyên trách và giao diện dễ sử dụng, tạo điều kiện dễ dàng cho người biên tập tin bài.

2.2.2.2 Yêu cầu về chức năng

Đối với các thành viên của Ban biên tập:

- Hệ thống cho phép quản lý cấu hình hệ thống
- Có thể xem kết quả của tập tài liệu đã tìm kiếm theo từ khóa được trả về từ máy chủ Google
- Có thể xem được kết quả phân cụm
- Có thể xuất bản tài liệu hoặc cụm tài liệu lên trang chủ

Đối với người truy cập vào Trang TTĐT:

- Có thể xem tin tức được thu thập từ Internet trên trang chủ
- Tin tức được hiển thị bao gồm tiêu đề và trích dẫn, để xem chi tiết tin bài, người dùng kích chuột vào tiêu đề bài viết trích dẫn.

2.3 Mô hình hoạt động của hệ thống

Quá trình hoạt động của hệ thống được thực hiện qua 4 giai đoạn sau đây:

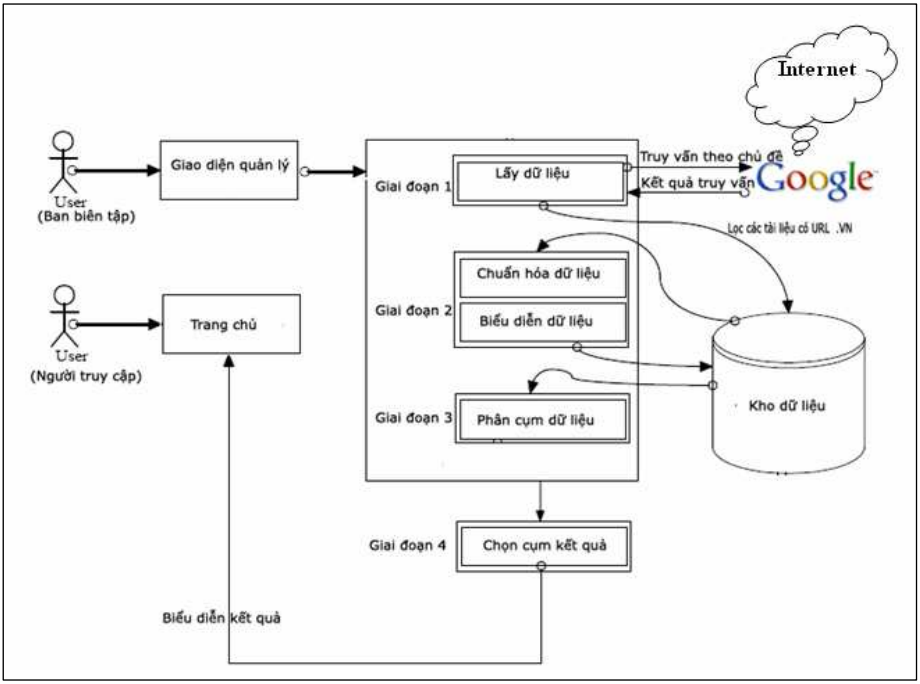
Giai đoạn 1: Lấy dữ liệu trả về từ máy tìm kiếm theo nội dung truy vấn. Để lấy được dữ liệu trên danh sách trả về từ máy tìm kiếm, chức năng Crawler sẽ thực hiện download các tài liệu về và lưu trữ vào cơ sở dữ liệu.

Giai đoạn 2: đây là giai đoạn chuẩn bị dữ liệu bao gồm tiền xử lý, chuẩn hóa và biểu diễn dữ liệu trước khi thực hiện phân cụm .

Giai đoạn 3: chức năng phân cụm tài liệu sẽ tiến hành phân cụm dữ liệu đã thu thập thành các cụm với độ tương tự của các tài liệu trong cụm là gần nhau nhất.

Giai đoạn 4: đánh giá và lựa chọn cụm tài liệu để phát hành lên trang chủ website.

Hình dưới đây minh họa mô hình hoạt động của hệ thống:



Hình 2.2: Mô hình hoạt động của hệ thống thu thập tin tự động

2.4 Chức năng của hệ thống

Dựa trên mô hình hoạt động của hệ thống ta có thể thiết kế các thành phần chức năng như sau:

- **Quản lý hệ thống:** quản lý các cấu hình hệ thống
- **Lập từ điển:**Xây dựng bộ từ điển để phục vụ cho việc tách từ và vecto hóa tài liệu chuẩn bị cho quá trình phân cụm tài liệu.
- **Lấy dữ liệu:**Thành phần Crawler trong hệ thống sẽ download tập các tài liệu từ danh sách trả về của máy tìm kiếm và sau đó lưu vào CSDL để tiếp tục tiền xử lý trước

khi phân cụm.

- **Xử lý dữ liệu và phân cụm:**Hệ thống tiến hành tiền xử lý các dữ liệu trả về từ máy chủ tìm kiếm và thực hiện phân cụm. Đầu ra là các cụm dữ liệu được gom theo các chủ đề nhỏ với mức độ tương đồng của các tài liệu trong cụm.
- **Đánh giá và chọn kết quả xuất bản:**Đây là bước người biên tập đưa ra quyết định chọn cụm tài liệu cần xuất bản lên trang chủ. Quá trình này cũng có thể thiết lập tự động dựa vào một tiêu chí đánh giá độ tương tự của cụm với chủ đề theo một tiêu chuẩn đánh giá định trước.
- **Biểu diễn tài liệu trên trang chủ:** dữ liệu được phát hành lên trang chủ phục vụ nhu cầu truy cập.

2.5 Phân tích và thiết kế hệ thống

2.5.1 Xác định Actor

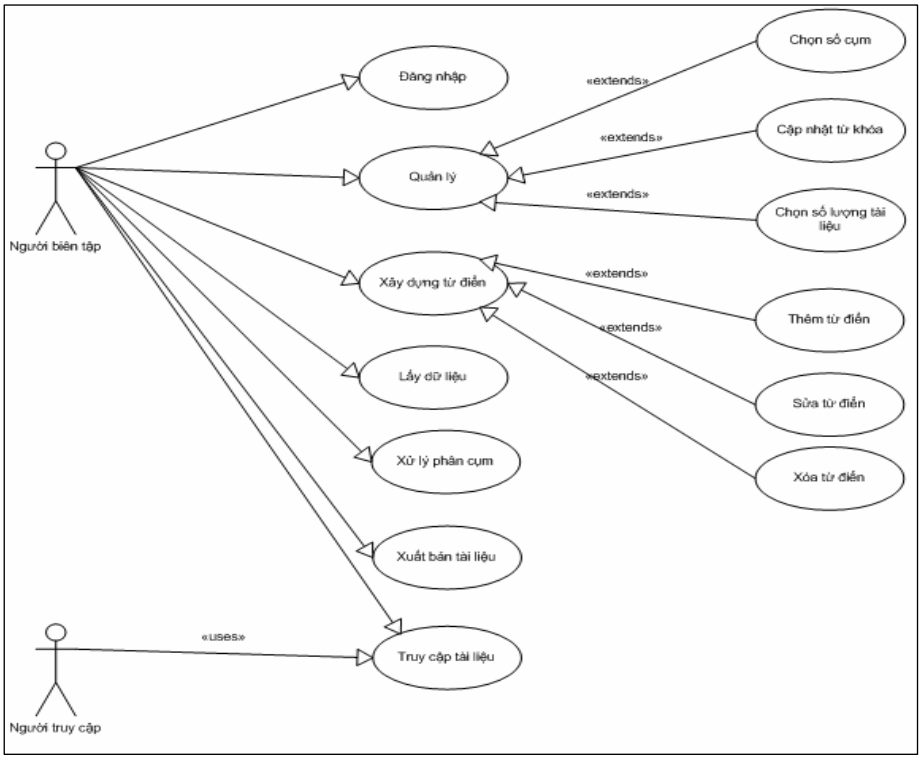
- Người biên tập: quản lý quá trình thu thập, xử lý, phân cụm và xuất bản tài liệu

- Người truy cập: Xem tài liệu được xuất bản trên trang chủ

2.5.2 Xác định Use Case

Ta xác định được các use case sau đây: Đăng nhập, Quản lý hệ thống, Lấy dữ liệu, Xây dựng từ điển, Xử lý phân cụm, Xuất bản tài liệu, Truy cập tài liệu.

2.5.3 Sơ đồ Use Case



Hình 2.3 : Sơ đồ Use case của hệ thống thu thập tin tự động

2.5.4 Đặc tả Use Case

Bao gồm 7 ca sử dụng được đặc tả với các thông tin : tác nhân, mô tả, tiền điều kiện, hậu điều kiện.

Các use case bao gồm: Xây dựng từ điển, Lấy dữ liệu, Xử lý phân cụm, Xuất bản tài liệu, Truy cập tài liệu

2.5.5 Biểu đồ tuần tự

Chúng ta có các biểu đồ tuần tự sau: Đăng nhập, Quản lý, Xây dựng từ điển, Lấy dữ liệu, Xử lý phân cụm, Xuất bản tài liệu, Truy cập tài liệu

2.5.6 Biểu đồ hoạt động

Xây dựng biểu đồ hoạt động cho ca sử dụng Lấy dữ liệu

2.5.7 Biểu đồ lớp

Dựa vào mô tả hệ thống và Use case, ta xác định các lớp chính của hệ thống thu thập tin tự động như sau:

Lớp Dictionary : lưu trữ thông tin của từ điển

Lớp Document : lưu trữ các tài liệu được lấy về từ internet

Lớp Cluster: lưu trữ các thông tin về các cụm dữ liệu sau khi phân cụm

Lớp DocumentIndex: Lưu trữ các thông tin trong quá trình làm sạch dữ liệu và tách từ

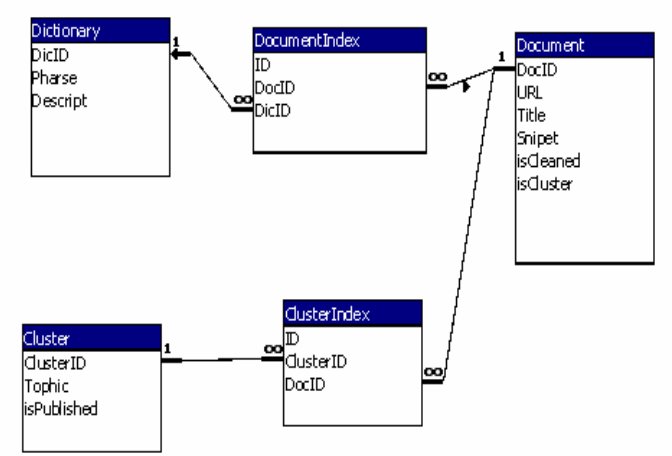
Lớp ClusterIndex: Lưu trữ các kết quả phân cụm

2.5.8 Thiết kế cơ sở dữ liệu

2.5.8.1 Các bảng dữ liệu

Document, Dictionary, Cluster, DocumentIndex, ClusterIndex

2.5.8.1 Mô hình cơ sở dữ liệu quan hệ



Hình 2.13: Mô hình cơ sở dữ liệu quan hệ

Chương 3: XÂY DỰNG VÀ TRIỂN KHAI HỆ THỐNG

3.1 Giải pháp kỹ thuật công nghệ

3.1.1 Tìm hiểu công nghệ Liferay Portal

Tìm hiểu về nền tảng công nghệ Portal Liferay và mô hình phát triển tích hợp các thành phần mở rộng

3.1.2 Thiết lập môi trường phát triển

- *Công cụ phát triển ứng dụng Java*
- *Cơ sở dữ liệu MySQL*
- *Máy chủ ứng dụng TomCat*
- *Môi trường phát triển tích hợp Eclipse IDE*
- *Triển khai Ext*
- *Thiết lập môi trường công cụ phát triển bổ sung (Plugin SDK)*

3.2 Xây dựng ứng dụng

Ứng dụng được xây dựng bởi các module cơ bản như sau:

- Module Lập từ điển dữ liệu
- Module Lấy dữ liệu
- Module Xử lý và phân cụm
- Module quản lý hệ thống
- Module hiển thị tin trên trang chủ

3.3 Triển khai ứng dụng

- Các module sau khi lập trình được đóng gói thành dạng Portlet và cài đặt vào hệ thống Portal

- Hệ thống Portal được cài đặt trên máy chủ thực thi web server Apache Tomcat.

- Hệ điều hành máy chủ MS Window 2003 Server

- Cấu hình máy chủ tối thiểu (thử nghiệm): CPU Intel core 2 duo, DDR 2 Gb

3.3 Thử nghiệm hệ thống

3.3.1 Dữ liệu

- Truy vấn vào máy chủ Google với từ khóa “Nông thôn mới”

- Chỉ lấy về 200 tài liệu đầu tiên tư danh sách trả về của máy tìm kiếm để phục vụ việc xử lý và phân cụm.

3.3.2 Kịch bản sử dụng

- Để tiến hành thu thập thông tin cho trang TTĐT Quảng Nam:

- Người biên tập cung cấp từ khóa theo chủ đề đã định trước, ở đây từ khóa là “Nông thôn mới” và ra lệnh tìm kiếm.

- Hệ thống tự động chuyển truy vấn đến máy chủ Google và kết quả trả về được hiển thị cho người sử dụng xem ngay trong màn hình hệ thống.

- Hệ thống đồng thời tiến hành việc trích lọc các tài liệu trả về từ Google có địa chỉ ở Việt Nam và lưu vào cơ sở dữ liệu.

- Quá trình làm sạch dữ liệu được tiến hành tự động

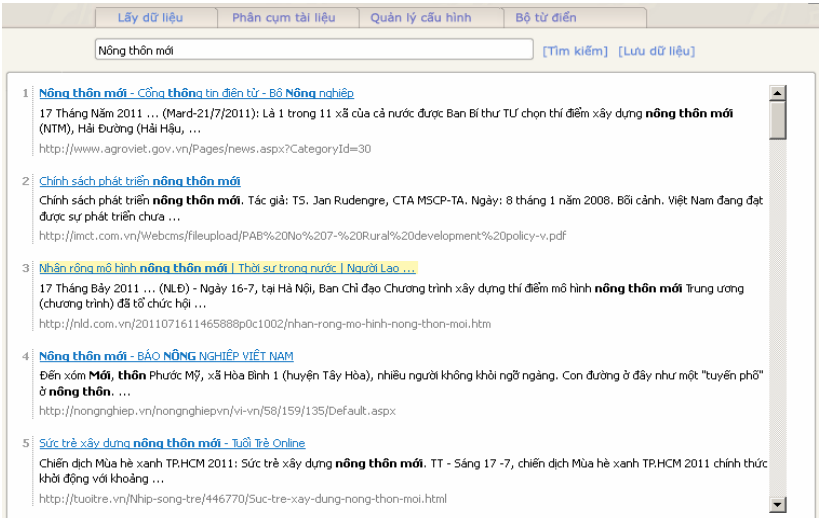
- Người dùng ra lệnh phân cụm tập dữ liệu và xem kết quả phân cụm

- Với kết quả phân cụm nhận được người dùng có thể cho xuất bản hoặc không xuất bản một hoặc nhiều cụm.

- Sau khi xuất bản, tin tức được hiển thị lên trang chủ thuộc chuyên mục của chủ đề cần thu thập dưới dạng tiêu đề và trích lượt.

- Người truy cập khi xem tin, hệ thống sẽ chuyển hướng trang sang phần xem chi tiết ngay trên web nguồn, tuy nhiên vẫn hiển thị trong phạm vi của trang TTĐT Quảng Nam.

3.4 Quá trình chạy thử nghiệm



Hình 3.15: Màn hình lấy dữ liệu

Hình 3.16 : Màn hình phân cụm dữ liệu

Hình 3.17 : Kết quả xuất bản tin tức về Nông thôn mới lên trang chủ website

3.5 Đánh giá kết quả thử nghiệm

Kết quả thử nghiệm hệ thống đáp ứng yêu cầu cơ bản đề ra về chất lượng phân cụm, tốc độ xử lý phân cụm.

KẾT LUẬN VÀ KIẾN NGHỊ

Các vấn đề đã được nghiên cứu, tìm hiểu trong luận văn:

Nghiên cứu tổng quan về Data Mining và các ứng dụng của Data Mining trong đó chủ yếu nghiên cứu kỹ thuật phân cụm dữ liệu. Trọng tâm đi vào tìm hiểu và cài đặt thuật toán K-means, ứng dụng thuật toán K-means tiếp cận bài toán Tìm kiếm và phân cụm tài liệu Web, bài toán là cơ sở để áp dụng xây dựng hệ thống thu thập tin tự động trên Internet.

Đã tìm hiểu các kỹ thuật xử lý, chuẩn hóa và biểu diễn tài liệu. Đây là kỹ thuật khá quan trọng trong lĩnh vực khai phá văn bản web.

Đã xây dựng thử nghiệm hệ thống thu thập tin tự động cho trang TTĐT tỉnh Quảng Nam dựa trên cơ sở lý thuyết đã tìm hiểu, nghiên cứu. Kết quả thử nghiệm hệ thống đáp ứng cơ bản yêu cầu đề ra.

Hạn chế của đề tài:

Do thời gian và khả năng kiến thức, khóa luận còn những hạn chế sau:

-Chưa đi vào nghiên cứu kỹ các hướng tiếp cận trong phân cụm dữ liệu, phân tích, so sánh các thuật toán để đánh giá thực chất về chất lượng phân cụm. Từ đó lựa chọn giải pháp tối ưu hơn.

- Vấn đề xử lý tài liệu tiếng Việt có ảnh hưởng rất lớn đến chất lượng phân cụm, truy nhiên khóa luận chưa đi sâu vào vấn đề này.

- Ứng dụng được xây dựng chỉ ở mức độ thử nghiệm nhằm thực nghiệm lý thuyết đã tìm hiểu, để triển khai thực tế cần phát triển hoàn chỉnh các tính năng trong đó quá trình thu thập và phân cụm có thể thiết lập tự động theo định kỳ và việc xuất bản cụm chủ đề sẽ tự động dựa vào tiêu chuẩn định trước.

Hướng nghiên cứu tiếp theo

Tiếp tục nghiên cứu các kỹ thuật phân cụm dữ liệu, trong đó nhấn mạnh đến kỹ thuật phân cụm K-Means mở rộng, thời gian tuyến tính đáp ứng được các yêu cầu của bài toán phân cụm tài liệu Web. Ngoài ra, cần nghiên cứu kỹ hơn các các kỹ thuật xử lý tiếng Việt, đây là kỹ thuật quan trọng trong việc tiền xử lý và Vectơ hóa tài liệu, có ảnh hưởng lớn đến chất lượng phân cụm tài liệu.

Phát triển hệ thống với đầy đủ các tính năng, đáp ứng việc triển khai sử dụng thực tế, đem lại hiệu quả kinh tế nhờ tiết kiệm thời gian, công sức và chi phí cho việc sưu tầm và xuất bản lại tin tức của Ban biên tập trang TTĐT tỉnh Quảng Nam.