BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN HỒ HIẾU

ỨNG DỤNG KỸ THUẬT THU THẬP THÔNG TIN TRÊN WEB ĐỂ XÂY DỰNG HỆ THỐNG TỔNG HỢP THÔNG TIN KINH TẾ XÃ HỘI

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

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Người hướng dẫn khoa học PGS.TS. VÕ TRUNG HÙNG

ĐÀ NẴNG 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Võ Trung Hùng**

Phản biện 1: **PGS. TSKH. Trần Quốc Chiến**

Phản biện 2: **TS. Trương Công Tuấn**

Luận văn sẽ được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật ngành Khoa học máy tính họp tại Đại học Đà Nẵng vào ngày 15 tháng 10 năm 2011

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin-Học liệu, Đại học Đà Nẵng
- Thư viện Trường Đại học Bách khoa, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Công tác điều hành, quản lý nhà nước trên lĩnh vực kinh tế - văn hóa – xã hội đòi hỏi người lãnh đạo phải thường xuyên nắm bắt, tổng hợp thông tin tình hình thực tiễn trên các báo, internet, các báo cáo của cấp dưới, ... để từ đó có cơ sở cho việc ra các quyết định phù hợp. Hằng ngày, tại Văn phòng UBND đều có cán bộ tổng hợp thông tin phục vụ lãnh đạo. Các thông tin được trích lọc từ các báo, website, từ thông tin trong nước, quốc tế, đặc biệt là thông tin trong tỉnh. Việc tổng hợp thủ công vừa tốn thời gian công sức, vừa không đầy đủ thông tin. Đặc biệt, thông tin trên interrnet hiện nay rất đa dạng, phong phú, nếu không có sự kiểm soát thông tin chặt chẽ sẽ xuất hiện những thông tin không đúng sự thật, gây ảnh hướng xấu đến hình ảnh của tỉnh.

Chính vì vậy, việc xây dựng hệ thống website thông tin kinh tế chính trị xã hội phục vụ điều hành lãnh đạo là hết sức cần thiết, trên cơ sở tự động tổng hợp thông tin từ các website trên internet theo tiêu chí chọn trước. Hiện nay, có nhiều phương pháp tự động tìm kiếm thông tin khác nhau, nhưng nhìn chung là các cách tiếp cận đều dựa vào các trọng số trang Web (Chỉ số quan trọng của trang trong tập kết quả), như: Page Bank, HITS và ứng dụng kỹ thuật khai phá dữ liệu. Trong đó Khai phá dữ liệu (Data Mining) là một lĩnh vực khoa học liên ngành mới xuất hiện gần đây nhằm đáp ứng nhu cầu này. Các kết quả nghiên cứu cùng với những ứng dụng thành công trong khai phá dữ liệu, khám phá tri thức cho thấy khai phá dữ liệu là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống.

Chính vì vậy, sau khi nghiên cứu các tài liệu và được sự đồng ý, hướng dẫn, động viên tận tình của TS. Võ Trung Hùng tôi đã chọn đề tài: ***"Ứng dụng kỹ thuật thu thập thông tin trên web xây dựng hệ thống tổng hợp thông tin kinh tế xã hội"*** làm đề tài nghiên cứu cho luận văn cao học của mình.

2. Mục tiêu và nhiệm vụ

Đề tài này nhằm mục đích xây dựng hệ thống tự động tổng hợp thông tin trực tuyến từ các website phục vụ cho công tác theo dõi, quản lý, chỉ đạo của lãnh đạo bằng cách sử dụng kỹ thuật khai phá dữ liệu web. Hệ thống cho phép:

- Tự động trích xuất các tin tức từ các website theo các chủ đề được chọn.
- Cho phép quản lý các chuyên mục tin.
- Quản lý các kênh tin tức.
- Quản lý thông tin lưu trữ.
- Tìm kiếm thông tin đã lưu trữ.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Đề tài này nhằm mục đích tìm hiểu về khai phá dữ liệu web, các thuật toán phân cụm (cluster) tài liệu và ứng dụng trong truy xuất thông tin tự động (information retrieval). Trên cơ sở đó, xây dựng hệ thống tự động tổng hợp, phân loại thông tin từ các website trên internet nhằm xây dựng hệ thống thông tin tổng hợp kinh tế - chính trị - xã hội.

Phạm vi nghiên cứu

- Khai phá dữ liệu web.
- Các giải thuật phân cụm tài liệu.
- Các kỹ thuật và công nghệ hỗ trợ trích xuất thông tin tự động.
- Kết hợp các yếu tố trên để xây dựng hệ thống tự động tổng hợp tin tức trực tuyến.

4. Phương pháp nghiên cứu

- ❖ Nghiên cứu lý thuyết
  - o Tìm hiểu lý thuyết về khai phá dữ liệu và khai phá dữ liệu web.
  - o Tìm hiểu các thuật toán phân cụm tài liệu.
  - o Tìm hiểu cơ chế hoạt động của các hệ thống tìm kiếm thu thập thông tin.

- Ứng dụng các công cụ để xây dựng hệ thống thu thập thông tin: RSS, Xpath, dotnetnuke, …

❖ Nghiên cứu thực nghiệm

- Dựa trên lý thuyết đã nghiên cứu, tiến hành xây dựng hệ thống thu thập thông tin từ các kênh tin cấu hình trước.
- Thử nghiệm trên máy đơn qua localhost có kết nối internet.

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

Về mặt lý thuyết: Giới thiệu tổng quan, và ứng dụng của khai phá dữ liệu web, các thuật toán phân cụm tài liệu và cơ chế của hệ thống thu thập tin.

Về mặc thực tiễn: Xây dựng hệ thống tổng hợp thông tin kinh tế chính trị xã hội phục vụ công tác quản lý chỉ đạo điều hành của lãnh đạo các cấp. Website cho phép người sử dụng cập nhật các thông tin mới nhất từ các website tin tức, lưu trữ, tìm kiếm thông tin theo các chuyên mục.

6. Bố cục của luận văn

Báo cáo của luận văn được được tổ chức thành ba chương chính.

Chương 1, dành để trình bày những nghiên cứu tổng quan về khai phá dữ liệu, thu thập thông tin từ internet.

Chương 2, dành để trình bày quá trình phân tích và thiết kế hệ thống thu thập thông tin;

Chương 3, dành để trình bày giải pháp xây dựng thử nghiệm hệ thống.

CHƯƠNG 1. TỔNG QUAN

Trong chương này chúng tôi trình bày một số khái niệm, định nghĩa liên quan đến Khai phá dữ liệu; các mô hình, các giai đoạn của quá trình khai phá dữ liệu, các dạng dữ liệu liên quan, các bài toán thông dụng và phạm vi ứng dụng của khai phá dữ liệu. Tiếp theo là giới thiệu về Kỹ thuật phân cụm tài liệu, các biểu diễn tài liệu trong mô hình không gian vector, các thuật toán ứng dụng trong phân cụm tài liệu. Sau đó giới thiệu về các quá trình thu thập thông tin, các kỹ thuật thu thập thông tin trên web. Cuối cùng là giới thiệu một số phần mềm tổng hợp thông tin tự động.

1.1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1.1. Giới thiệu

Trong thời đại ngày nay, với sự phát triển vượt bậc của công nghệ thông tin, các hệ thống thông tin có thể lưu trữ một khối lượng lớn dữ liệu về hoạt động hàng ngày. Từ khối dữ liệu này, các kỹ thuật trong Khai phá dữ liệu và Máy học có thể dùng để trích xuất những thông tin hữu ích mà chúng ta chưa biết. Các tri thức vừa học được có thể vận dụng để cải thiện hiệu quả hoạt động của hệ thống thông tin ban đầu.

Giáo sư Tom Mitchell đã đưa ra định nghĩa của Khai phá dữ liệu như sau: "Khai phá dữ liệu 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: "Khai phá dữ liệu, 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". Nói tóm lại, Khai phá dữ liệu là một quá trình học tri thức mới từ những dữ liệu đã thu thập được.

Quá trình này có thể được lặp lại nhiều lần một hay nhiều giai đoạn dựa trên phản hồi từ kết quả của các giai đoạn. Mối quan hệ chặt chẽ giữa các giai đoạn trong quá trình Khai phá dữ liệu là rất quan trọng cho việc nghiên cứu trong Khai phá dữ liệu. Một giải thuật trong Khai phá dữ liệu không thể được phát triển độc lập, không quan tâm đến bối cảnh áp dụng mà thường được xây dựng để giải quyết một mục tiêu cụ thể. Do đó, sự hiểu biết bối cảnh vận dụng là rất cần thiết. Thêm vào đó, các kỹ thuật được sử dụng trong các giai đoạn trước có thể ảnh hưởng đến hiệu quả của các giải thuật sử dụng trong các giai đoạn tiếp theo.

1.1.2. Các dạng dữ liệu

Full text

Dữ liệu dạng Full text là một dạng dữ liệu phi cấu trúc với thông tin chỉ gồm các tài liệu dạng text. Mỗi tài liệu chứa thông tin về một vấn đề nào đó thể hiện qua nội dung của tất cả các từ cấu thành tài liệu đó.

Trong các dữ liệu hiện nay thì văn bản là một trong những dữ liệu phổ biến nhất, nó có mặt khắp mọi nơi và chúng ta thường xuyên bắt gặp do đó các bài toán về xử lý văn bản đã được đặt ra khá lâu và hiện nay vẫn là một trong những vấn đề trong khai phá dữ liệu Text,

trong đó có những bài toán đáng chú ý như tìm kiếm văn bản, phân loại văn bản, phân cụm văn bản hoặc dẫn đường văn bản.

Hypertext

Theo từ điển của Đại Học Oxford (Oxford English Dictionary Additions Series) thì Hypertext được định nghĩa như sau: Đó là loại Text không phải đọc theo dạng liên tục đơn, nó có thể được đọc theo các thứ tự khác nhau, đặc biệt là Text và ảnh đồ họa (Graphic) là các dạng có mối liên kết với nhau theo cách mà người đọc có thể không cần đọc một cách liên tục.

Có hai khái niệm về Hypertext cần quan tâm: Hypertext Document (Tài liệu siêu văn bản) và Hypertext Link (Liên kết siêu văn bản)

1.1.3. Các bài toán thông dụng trong khai phá dữ liệu

1.1.3.1. Phân lớp (Classification).

Với một tập các dữ liệu huấn luyện cho trước và sự huấn luyện của con người, các giải thuật phân loại sẽ học ra bộ phân loại (classifier) dùng để phân các dữ liệu mới vào một trong những lớp (còn gọi là loại) đã được xác định trước. Nhận dạng cũng là một bài toán thuộc kiểu phân loại.

1.1.3.2. Dự đoán (Prediction).

Với mô hình học tương tự như bài toán Phân loại, lớp bài toán Dự đoán (Prediction) sẽ học ra các bộ dự đoán. Khi có dữ liệu mới đến, bộ dự đoán sẽ dựa trên thông tin đang có để đưa ra một giá trị số học cho hàm cần dự đoán. Bài toán tiêu biểu trong nhóm này là dự đoán giá sản phẩm để lập kế hoạch trong kinh doanh.

1.1.3.3. Tìm luật liên kết (Association Rule)

Các giải thuật Tìm luật liên kết (Association Rule) tìm kiếm các mối liên kết giữa các phần tử dữ liệu, ví dụ như nhóm các món hàng thường được mua kèm với nhau trong siêu thị.

1.1.3.4. Phân cụm (Clustering)

Các kỹ thuật Phân cụm (Clustering) sẽ nhóm các đối tượng dữ liệu có tính chất giống nhau vào cùng một nhóm. Có nhiều cách tiếp cận với những mục tiêu khác nhau trong phân loại. Các kỹ thuật trong bài toán này thường được vận dụng trong vấn đề phân hoạch dữ liệu tiếp thị hay khảo sát sơ bộ các dữ liệu.

1.1.4. Ứng dụng của khai phá dữ liệu

Khai phá dữ liệu được vận dụng trong nhiều lĩnh vực khác nhau nhằm khai thác nguồn dữ liệu phong phú được lưu trữ trong các hệ thống thông tin. Tùy theo bản chất của từng lĩnh vực, việc vận dụng Khai phá dữ liệu có những cách tiếp cận khác nhau. Khai phá dữ liệu cũng được vận dụng hiệu quả để giải quyết các bài toán phức tạp trong các ngành đòi hỏi kỹ thuật cao như tìm kiếm mỏ dầu từ ảnh viễn thám, xác định các vùng gãy trong ảnh địa chất để dự đoán thiên tai, cảnh báo hỏng hóc trong các hệ thống sản xuất,… Các bài toán này đã được giải quyết từ khá lâu bằng các kỹ thuật nhận dạng hay xác suất nhưng được giải quyết với yêu cầu cao hơn bởi các kỹ thuật của Khai phá dữ liệu. Phân nhóm và dự đoán là những công cụ rất cần thiết cho việc qui hoạch và phát triển các hệ thống quản lý và sản xuất trong thực tế.

1.2. PHÂN CỤM TÀI LIỆU

1.2.1. Phân cụm tài liệu

Phân cụm (Clustering) là quá trình nhóm một tập các đối tượng vật lý hoặc trừu tượng thành các nhóm hay các lớp đối tượng tương tự nhau. Một cụm (cluster) là một tập các đối tượng giống nhau hay là tương tự nhau, chúng khác hoặc ít tương tự so với các đối tượng thuộc lớp khác. Không giống như quá trình phân loại, ta thường biết trước tính chất hay đặc điểm của các đối tượng trong cùng một lớp và dựa vào đó để ấn định một đối tượng vào lớp của nó, trong quá trình chia lớp ta không hề biết trước tính chất của các lớp và thường dựa vào mối quan hệ của các đối tượng để tìm ra sự giống nhau giữa các đối tượng dựa vào một độ đo nào đó đặc trưng cho mỗi lớp.

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...

1.2.2. Biểu diễn tài liệu trong mô hình không gian vector

1.2.2.1. Khái niệm

Mô hình không gian vector (Vector space model- VSM) là một cách biểu diễn một tài liệu như một vector. Đây là khái niệm quan trọng trong Information Retrieval-IR, được sử dụng để lượng hóa những đối tượng khó quản lý như tài liệu, khái niệm, câu truy vấn ,….

Tập hợp toàn bộ các tài liệu mà ta xem xét tương ứng với một không gian vector. Tài liệu được xem là một vector với các thành phần là trọng số tính trên các khái niệm xuất hiện trong nó (term), thông thường người ta xem các term này chính là các từ vựng xuất hiện trong tài liệu.

Dữ liệu web về bản chất chính là văn bản, do đó có thể áp dụng các kỹ thuật phân cụm văn bản cho việc xây dựng hệ thống tìm kiếm và phân loại thông tin trên web.

1.2.2.2. Hàm tương tự giữa hai vector tài liệu trong không gian

Để tiến hành các thao tác xử lý tài liệu như tìm kiếm, so sánh, phân lớp, phân cụm, … cần thiết phải có công cụ để so sánh các tài liệu với nhau. Khi đã xây dựng được không gian vector, một cách tự nhiên người ta muốn xây dựng hàm tương tự giữa hai vector. Điều này phục vụ việc tính toán độ tương tự giữa hai tài liệu trong việc phân cụm tài liệu ,hay độ phù hợp của một tài liệu với một câu truy vấn khi tìm kiếm. Bản chất của quá trình này là chúng ta xem xét xem thế nào là hai vector giống nhau, hay tương tự nhau.

1.2.3. Các thuật toán ứng dụng trong phân cụm tài liệu

1.2.3.1. Phân cụm dữ liệu không gian và các tiếp cận

Các kỹ thuật áp dụng để giải quyết vấn đề phân cụm dữ liệu đều hướng tới hai mục tiêu chung: Chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ phân cụm dữ liệu có thể phân loại theo các cách tiếp cận chính như: Phân cụm phân hoạch, Phân cụm dữ liệu phân cấp, Phân cụm dữ liệu dựa trên mật độ, Phân cụm dữ liệu dựa trên lưới, Phân cụm dữ liệu dựa trên mô hình, Phân cụm dữ liệu có ràng buộc,

1.2.3.2. Phân cụm dữ liệu dựa vào thuật toán K-means

Tư tưởng thuật toán

K-means là một trong số những phương pháp học không có giám sát cơbản nhất thường được áp dụng trong việc giải các bài toán về phân cụm dữliệu. 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

$$Xi = (xi1, xi2, \ldots, xid) \ (i = \overline{1,n})$$

sao cho hàm tiêu chuẩn:

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

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

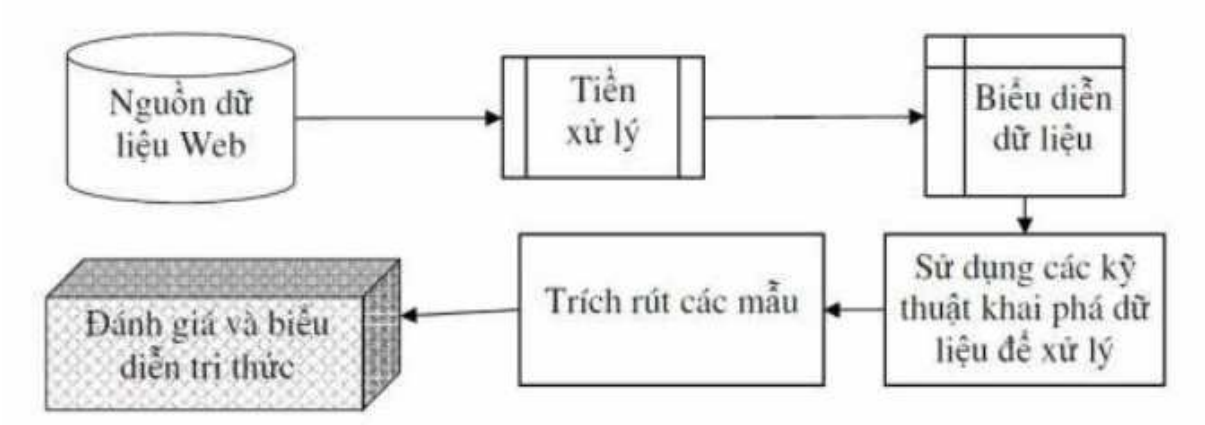
Trọng tâm của một cụm là một véc tơ, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng véc tơ dữ liệu trong cụm đang xét. Tham số đầu vào của thuật toán là số cụm k, và tham số đầu ra của thuật toán là các trọng tâm của các cụm dữ liệu. Độ đo khoảng cách d giữa các đối tượng dữ liệu thường được sử dụng là khoảng cách Euclide, bởi vì đây là mô hình khoảng cách dễ để lấy đạo hàm và xác định các cực trị tối thiểu. Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tuỳ vào ứng dụng hoặc các quan điểm của người dùng.

1.3. THU THẬP THÔNG TIN TRÊN WEB

1.3.1. Giới thiệu tổng quan về thu thập thông tin trên web

Thu thập thông tin (Information Retrieval - IR) trên web tập trung vào việc khám phá một cách tự động nguồn thông tin có giá trị trực tuyến. Nội dung web có thể được tiếp cận theo 2 cách khác nhau: Tìm kiếm thông tin và khai phá dữ liệu trong cơ sở dữ liệu lớn. Khai phá dữ liệu đa phương tiện là một phần của khai phá nội dung Web, nó hứa hẹn việc khai thác được các thông tin và tri thức ở mức cao từ nguồn đa phương tiện trực tuyến rộng lớn.

Khai phá văn bản Web là việc sử dụng kỹ thuật khai phá dữ liệu đối với các tập văn bản để tìm ra tri thức có ý nghĩa tiềm ẩm trong nó. Dữ liệu của nó có là dữ liệu có cấu trúc hoặc không cấu trúc. Kết quả khai phá không chỉ là trạng thái chung của mỗi tài liệu văn bản mà còn là sự phân loại, phân cụm các tập văn bản phục vụ cho mục đích nào đó.



1.3.2. Quá trình thu thập thông tin trên web

Nắm bắt những đặc tính của người dùng Web là việc rất quan trọng đối với người thiết kế Website. Thông qua việc khai phá lịch sử các mẫu truy xuất của người dùng Web, không chỉ thông tin về Web được sử dụng như thế nào mà còn nhiều đặc tính khác như các hành vi của người dùng có thể được xác định. Sự điều hướng đường dẫn người dùng Web mang lại giá trị thông tin về mức độ quan tâm của người dùng đến các Website đó. Khai phá Web theo sử dụng Web là khai phá truy cập Web để khám phá các mẫu người dùng truy cập vào Website.

1.3.3. Các kỹ thuật crawling và indexing

Một Web thu thập thông tin (Web Crawler) là một chương trình máy tính có thể "duyệt web" một cách tự động và theo một phương thức nào đó được xác định trước. Vì là một chương trình nên quá trình "duyệt web" của các web crawler không hoàn toàn giống với quá trình duyệt web của con người (web crawler phải sử dụng các phương thức dựa trên HTTP trực tiếp chứ không thông qua web browser như con người). Các web crawler thường bắt đầu với một danh sách URL của các web page để ghé thăm đầu tiên. Khi ghé thăm một URL, crawler sẽ đọc nội dung web page, tìm tất cả các hyperlink có trong web page đó và đưa các URL được trỏ tới bởi các hyperlink đó vào danh sách URL. Dựa vào danh sách URL này, Crawler lại tiếp tục quá trình duyệt đệ quy để ghé thăm tất cả các URL chưa được duyệt đến. Quá trình này được gọi là web crawling hoặc là web spidering, các web crawler còn được gọi là các robot (bot) hoặc nhện web (web spider).

Về bản chất, web crawling chính là quá trình duyệt đệ quy một đồ thị cây có các node là các web page.

1.4. KHẢO SÁT MỘT SỐ PHẦN MỀM TỔNG HỢP TIN

1.4.1. Google Reader

Google Reader là công cụ tổng hợp tin hữu ích của Google. Việc dùng Google Reader khá đơn giản, chỉ cần thêm địa chỉ URL của feed/rss của nguồn tin muốn theo dõi, mỗi khi nguồn tin có thay đổi, Google Reader sẽ lấy tin về tự động.

Google Reader còn có nhiều tiện ích như:

- Chia sẻ trực tiếp các tin đọc trong Google Reader cho bạn bè (bấm vào nút Share), thông tin này sẽ được hiển thị trên Google Buzz hoặc dùng nút Send To để gửi đến các dịch vụ khác như Twitter, Facebook, Blogger. Chia sẻ các danh sách nguồn tin mà bạn thấy hữu ích cho bạn bè.

- Kiểm tra sự cập nhật của các trang web, không nhất thiết ở dưới định dạng feed bằng cách thêm URL của trang web cần lấy vào Google Reader.

1.4.2. iGoogle

iGoogle là dịch vụ trang chủ tìm kiếm cá nhân hoá (Personalized Homepage) với các tính năng mới như "Gadget Maker" và khả năng hiển thị kết quả tìm kiếm dựa trên từng vùng. iGoogle cho phép người dùng có thể tạo lập một trang chủ tìm kiếm hoàn toàn theo ý thích. Tại trang chủ này, người dùng có thể đặt các "gadget" (tiện ích nhỏ) chứa các thông tin quan tâm như thời tiết, chứng khoán, tin tức, và thậm chí là cả ngày tháng hiện tại. Ngoài ra iGoogle cung cấp nhiều tiện ích khác như: xem RSS tin tức từ các site khác, To do list, đếm ngược thời gian, khung tìm kiếm của Wikipedia …

1.4.3. Yahoo

Yahoo hiện đang thử nghiệm dịch vụ tổng hợp thông tin tự động tại địa chỉ. Yahoo!Pipes (http://pipes.yahoo.com/).

Đây là công cụ tương tác qua web hỗ trợ xử lý và tổng hợp các nguồn tin từ internet cho phép người dùng thu thập thông tin từ các nguồn khác nhau, lọc và xem tin tùy theo lĩnh vực quan tâm. Yahoo Pipe hỗ trợ nhiều nguồn tin khác nhau như Data, Page, Url, Rss, yahoo Search, … và nhiều công cụ cho phép người dùng xác định từ khóa tin cần lấy.

CHƯƠNG 2. THIẾT KẾ GIẢI PHÁP XÂY DỰNG HỆ THỐNG THU THẬP THÔNG TIN KINH TẾ XÃ HỘI

Chương này tập trung vào phân tích và xác định các yêu cầu xây dựng Hệ thống thu thập thông tin kinh tế xã hội. Tiếp theo là giới thiệu mô hình kiến trúc, các thành phần của hệ thống thu thập thông tin. Sau đó là trình bày các giải pháp, các công cụ sử dụng và cuối cùng là phân tích và thiết kế hệ thống.

2.1. PHÂN TÍCH VÀ XÁC ĐỊNH YÊU CẦU

2.1.1. Đặt vấn đề

Trong thời đại bùng nổ thông tin như hiện nay thì việc khai thác, thu thập và chia sẻ thông tin đóng một vai trò quan trọng. Với một dữ liệu khổng lồ trên mạng, làm sao ta có thể nắm bắt được thông tin mới nhất, nhanh chóng nhất mà không phải tốn thời gian xem từng website để đọc và tìm kiếm thông tin.

Trên cơ sở này, hệ thống bóc tách thông tin được xây dựng nhằm phục vụ cho việc trích xuất thông tin từ các website, rồi tất cả thông tin được hiển thị trên một website, giúp cho người đọc có thể nắm bắt được thông tin một cách xúc tích, nhanh chóng và tiết kiệm thời gian.

Đối tượng sử dụng hệ thống là tất cả cộng đồng người sử dụng mạng. Quản trị viên có thể quản lý tài khoản người dùng, quản lý các đường dẫn (link).

Khảo sát, phân tích và đánh giá yêu cầu

Khảo sát một số chương trình hỗ trợ đọc tin tức RSS

2.1.2. Xác định yêu cầu của Hệ thống

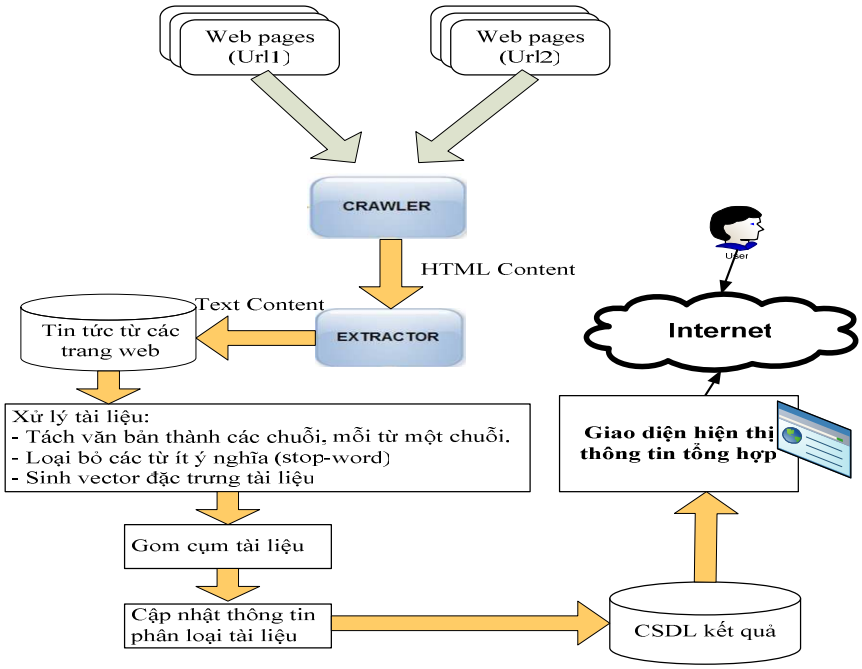
Mục tiêu của đề tài là xây dựng nên một hệ thống hỗ trợ người dùng chọn kênh tin tức, thu thập tin tức, quản lý các kênh tin, tạo ra một website tin tức cho chính người dùng mà không phải lướt từng website để đọc tin tức.

Thông qua việc khảo sát một số phần mềm đọc tin tức trong và ngoài nước, và yêu cầu từ phía người dùng, có thể tóm tắt yêu cầu của người dùng đối với hệ thống bóc tách thông tin.

2.2. MÔ HÌNH HỆ THỐNG

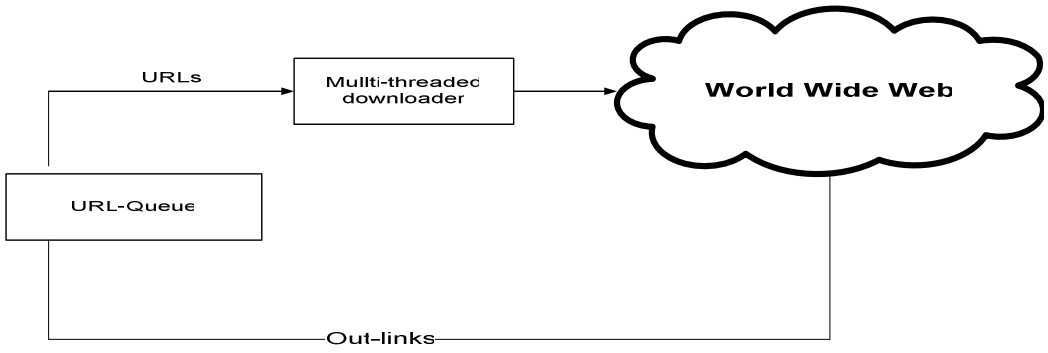
2.2.1. Kiến trúc chung

Hệ thống khai thác và tổng hợp nội dung có nhiệm vụ khai thác, tổng hợp, lưu trữ rồi phát hành lại tới người dùng. Crawler nhận cấu hình đầu vào của một website (tin tức) tiến hành bóc tách, tổng hợp chủ đề liên quan, lưu trữ trong database và phát hành lại trên trang tin tổng hợp. Giải pháp đề xuất dựa trên mô hình trích xuất dữ liệu đặc tả của nội dung (còn gọi là meta data - cung cấp các thông tin cơ bản bao gồm : tên tin bài, ngày phát hành, sơ lược nội dung, người viết,...). Nội dung được bóc tách toàn vẹn, sạch sẽ và được tổng hợp từ nhiều nguồn khác nhau giúp người đọc có thể theo dõi, kiểm soát, tìm kiếm, biên soạn, lưu trữ một cách hiệu quả. Sau đó những đặc tả dữ liệu (meta data) được xây dựng tự động trên nền nội dung đã bóc tách. Sau quy trình khai thác, nội dung sẽ trở thành độc lập với website nguồn, được lưu trữ và tái sử dụng cho những mục đích khác nhau.



2.2.2. Thành phần web Crawler

Crawler là thành phần quan trọng của hệ thống có nhiệm vụ dò tìm của Url và tải nội dung từ các Url. Kiến trúc và hoạt động của một Crawler đơn giản như sau:



Hình 2-1: Mô hình hệ thống crawler.

Hoạt động của hệ thống có thể được mô tả như sau:

- Bước 1: URL-Queue sẽ chọn ra một tập các URLs cần download,gửi cho Multi-threaded downloader
- Bước 2: Downloader tiến hành download các tài liệu này, phân tích chúng, trích ra các đường link xuất hiện bên trong các tài liệu, rồi gửi cho URL-Queue. Lặp lại bước 1.

Quá trình này dừng lại khi thỏa mãn một số điều kiện dừng nào đó.

2.2.3. Thành phần web Extractor

Tài liệu trên Web là những văn bản được lưu trữ trong các máy tính kết nối với Internet. Để xem các tài liệu này, người dùng dùng một trình duyệt Web (Web Browser) mở và hiển thị chúng.

2.2.4. Xử lý tài liệu

Thông thường một tài liệu, trước khi được lưu trữ và lập chỉ mục trong các hệ thống tìm kiếm bao giờ cũng phải trải qua những bước tiền xử lý .Mục đích của nó là đưa tài liệu về một dạng mang nhiều thông tin hơn, đơn giản hơn, tiện cho các quá trình xử lý sau này. Tài liệu ở đây là các tin tức được tải tự động từ các trang web. Vì nội dung tin tức có thể rất dài, chứa hàng ngàn từ, do đó để giảm kích thước xử lý, chúng ta chỉ xử lý đối với phần tóm tắt của tin tức. Phần này thường chỉ gồm 1-5 câu, khái quát được chủ đề của tin tức, do đó có thể đại diện cho tin tức.

2.2.5. Gom cụm tài liệu

Việc gom cụm tài liệu sẽ được thực hiện dựa vào mô hình không gian vector (phần I.2.2) dựa vào trọng số của các từ đặc trưng trong tài liệu.

2.3. GIẢI PHÁP CÔNG NGHỆ SỬ DỤNG

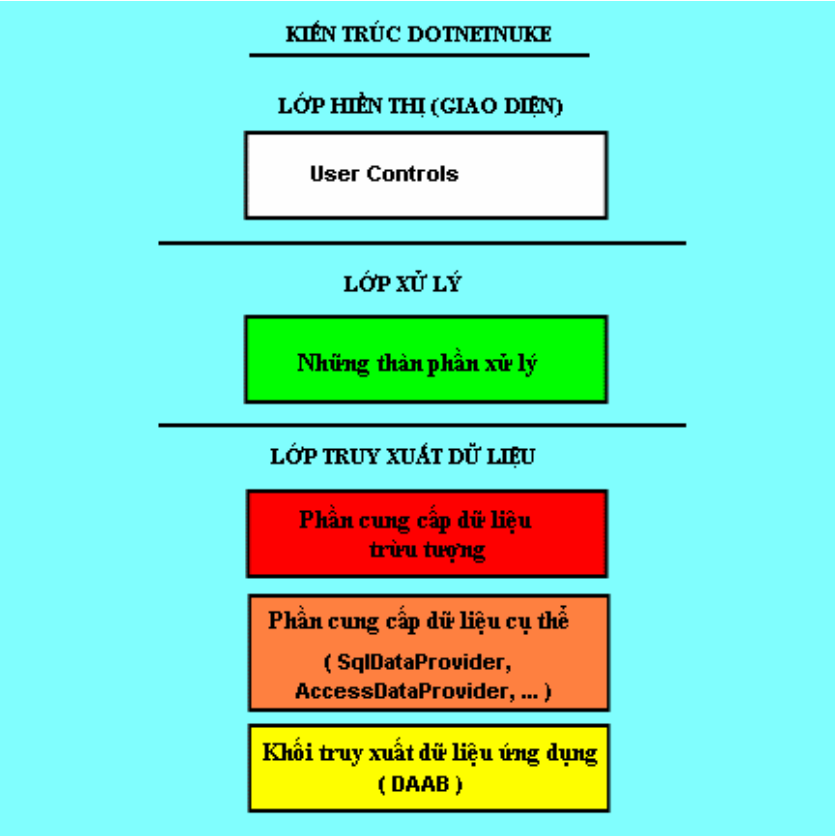
2.3.1. Công cụ phân tích dữ liệu XPath

Xpath – XML Path – là một ngôn ngữ truy vấn được định nghĩa bởi W3C, sử dụng để truy vấn các node hoặc tính toán các giá trị lấy trong một tài liệu XML [1]. Một biểu thức XPath (Xpath expression) có thể chọn một node hoặc một tập hợp các node, hoặc nó có thể trả lại một giá trị dữ liệu dựa trên một hoặc nhiều node trong tài liệu. XPath hiện có 2 phiên bản là XPath 1.0 và XPath 2.0.

2.3.2. Công nghệ Portal Dotnetnuke

Kiến trúc mà DotNetNuke xây dựng là kiến trúc đa cổng (multi portal). Khái niệm cổng được gọi là portal trong DotNetNuke. DotNetNuke hỗ trợ nhiều portal cùng chạy trên một cơ sở dữ liệu và một mã nguồn duy nhất.

DotNetNuke được thiết kế theo mô hình ba lớp hoàn chỉnh. Vì vậy, nó tạo ra rất nhiều tiện lợi cho người lập trình. Không những thế, khả năng hỗ trợ rất tốt và dễ dùng lại trong việc truy xuất dữ liệu chính là một trong những thế mạnh của DotNetNuke. Mô hình ba lớp của DotNetNuke được mô tả trong mô hình sau :



Hình 2-2: Mô hình kiến trúc công nghệ dotnetnuke portal

DotNetNuke sử dụng đối tượng DataReader để chuyển những dữ liệu có được từ Lớp Truy xuất Dữ liệu lên Lớp Xử lý.

Lớp hiển thị (Giao diện)

Lớp hiển thị sử dụng những dịch vụ của Lớp xử lý cung cấp. Lớp giao diện chính là những UserControl

Lớp Xử lý

Những hàm xử lý của cùng một đối tượng xử lý được lưu chung vào một tập tin có phần mở rộng (*.vb). Lớp này sử dụng những hàm do lớp truy xuất dữ liệu cung cấp.

Lớp Truy xuất dữ liệu

Lớp này là lớp cuối cùng, thực hiện nhiệm vụ truy xuất dữ liệu. Một hàm quan trọng của lớp này là hàm SQLGenerator..

2.4. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

2.4.1. Mô tả chức năng hệ thống

2.4.1.1. Phân hệ thu thập và xử lý tin tức

Đây là phân hệ quan trọng của hệ thống có chức năng tự động lấy tin tức từ các báo điện tử trên mạng và lưu vào CSDL. Gồm các phân hệ con: crawler, extractor và xử lý dữ liệu.

Tin tức do phân hệ này sẽ cung cấp cho Cổng thông tin điện tử để người quản trị tin có thể duyệt/xuất bản tin.

2.4.1.2. Phân hệ Cổng thông tin điện tử kinh tế xã hội tổng hợp

Phân hệ tin tức được chia thành 2 mảng chức năng tương ứng 2 đối tượng sử dụng: mảng chức năng đối với người dùng (user) và mảng chức năng quản trị (admin)

Chức năng người dùng:

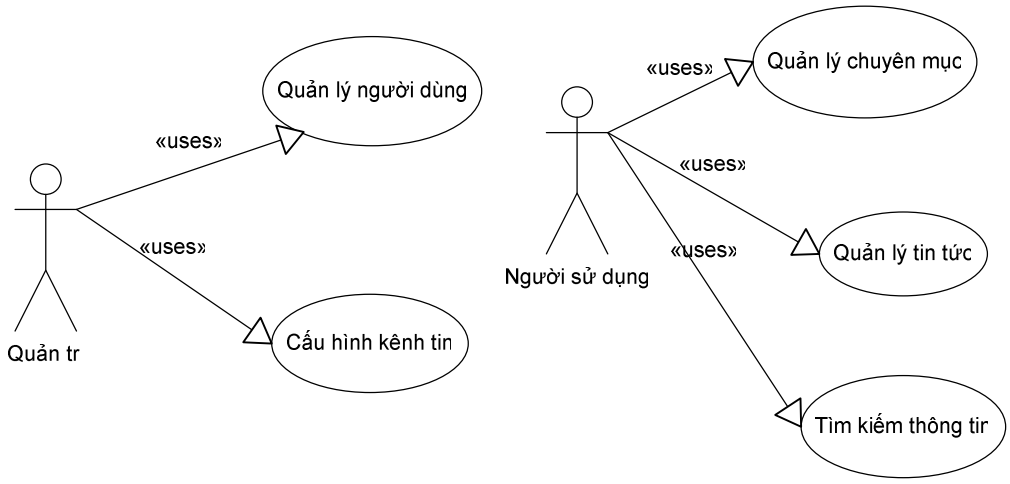
- Xem tin: Cho phép người dùng xem chi tiết một tin. Giống với một trang báo thông thường.
- Tìm kiếm: Cho phép người dùng tìm kiếm tin bài một cách nhanh chóng, thuận tiện.

Chức năng quản trị (admin)

- Quản trị các chuyên mục tin tức: Chức năng này cho phép người quản trị tổ chức các tin thành các chuyên mục.
- Quản trị tin tức theo chuyên mục: Các tin tức được liệt kê theo từng chuyên mục, chỉ những người được phân quyền quản trị đối với chủ đề này mới được phép xem danh sách này.
- Cập nhật tin tức: Người được cấp quyền đối với một chủ đề có thể thêm mới, sửa, hay xóa một tin.
- Phân quyền quản trị tin tức: Đây là chức năng quản trị quyền trong phân hệ quản trị tin tức. Các quyền được phân cho từng đối tượng người dùng theo từng chủ đề.

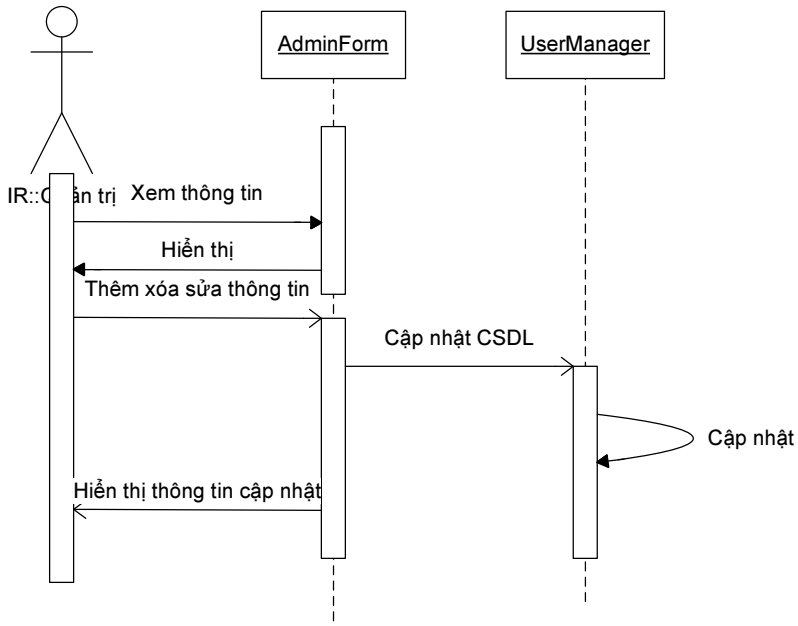
2.4.2. Phân tích thiết kế hệ thống

2.4.2.1. Danh sách User case và Actor



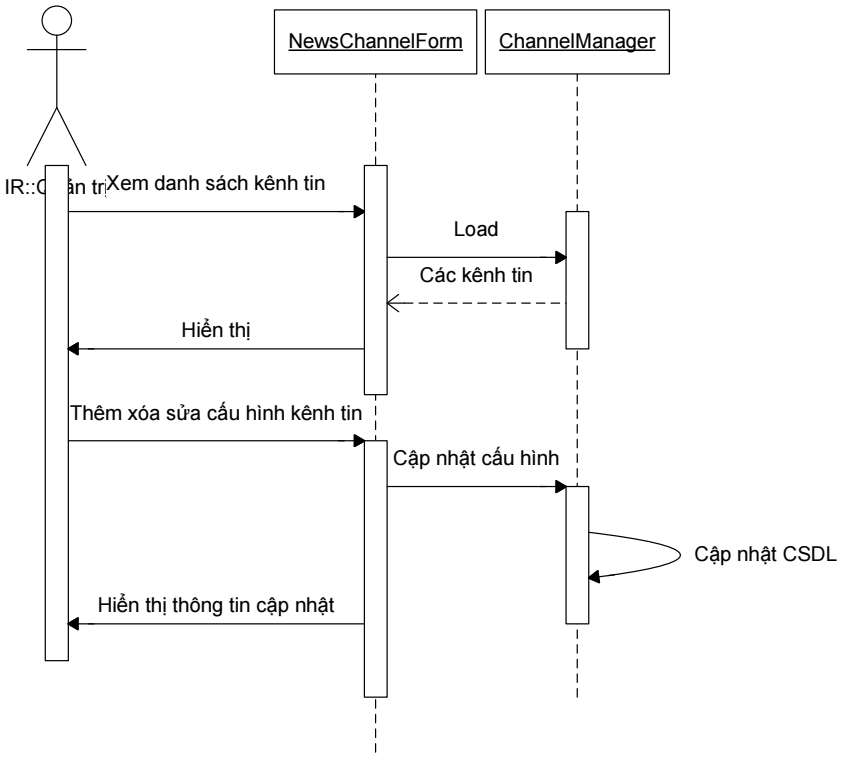
2.4.2.2. Biểu đồ tuần tự

Biểu đồ tuần tự của thao tác quản lý người dùng:



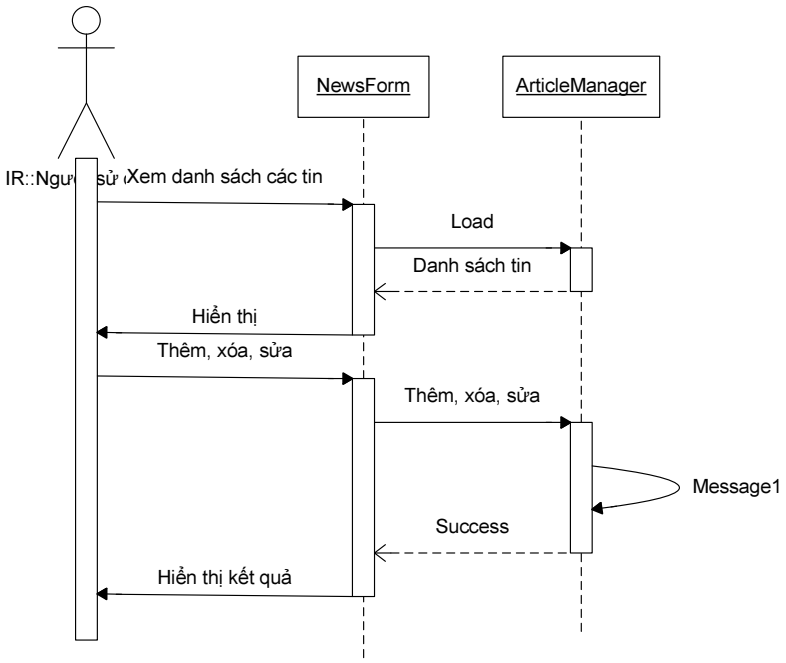
Hình 2-3: Biểu đồ tuần tự - quản lý người dùng

Biểu đồ tuần tự của quá trình quản lý cấu hình kênh tin:



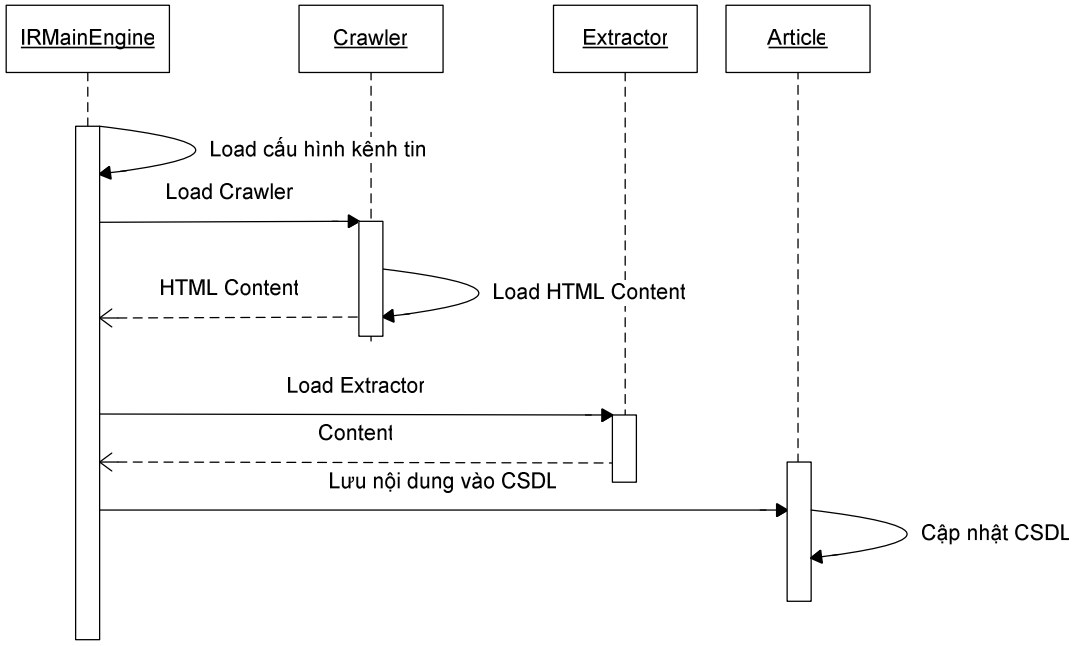
Hình 2-4: Biểu đồ tuần tự - quản lý kênh tin

Biểu đồ tuần tự của quá trình quản lý tin:



Hình 2-5: Biểu đồ tuần tự - quản lý tin

Biểu đồ tuần tự của quá trình lấy tin:



Hình 2-6: Biểu đồ tuần tự - Lấy thông tin từ internet

CHƯƠNG 3. XÂY DỰNG HỆ THỐNG TỔNG HỢP THÔNG TIN

Chương này tập trung trình bày về cài đặt cơ sở dự liệu, phát triển chương trình ứng dụng thử nghiệm và đánh giá kết quả thử nghiệm hệ thống

3.1. CÔNG CỤ SỬ DỤNG

Hệ điều hành: Microsoft Windows Server, Windows XP, Windows 7.

Hệ quản trị CSDL: MS SQL Server 2005.

Web Server: IIS (Internet Information Services).

Công nghệ lập trình: C#, ASPX, Javascript, DHTML, XML, CSS.

3.2. CÀI ĐẶT CSDL

Cơ sở dữ liệu tin tức tổng hợp được dùng chung cho chương trình chính (dạng winform) và cổng thông tin điện tử (portal) nhằm phục vụ cho việc duyệt tin từ xa thông qua giao diện web.

3.3. PHÁT TRIỂN CHƯƠNG TRÌNH

3.3.1. Xây dựng Phân hệ Crawler

WebCrawler được xây dựng trong hệ thống là các robot thu thập thông tin tự động từ các kênh tin được cấu hình sẵn trong hệ thống. Khi chương trình xem/quản lý tin tức chính được khởi động, nó sẽ load danh sách các kênh tin trong CSDL và ứng với mỗi kênh tin sẽ tạo ra một crawler để tải các tin từ kênh đó về. Việc khởi tạo và chạy nhiều crawler sẽ khiến chương trình chính bị chậm lại, ảnh hưởng đến việc duyệt các tin đã lưu của người dùng. Do đó, các crawler được tạo ra sẽ chạy ở chế độ nền, theo một tiến trình (thread) khác với chương trình chính. Do đó chương trình chính sẽ không bị ảnh hưởng.

3.3.2. Xây dựng phân hệ Extractor:

Tài liệu do crawler tải về ở dạng HTML trong đó chứa nội TEXT và các thẻ (tag) HTML. Đặc thù của file HTML là định dạng trang web bằng các thẻ. Mỗi thẻ sẽ có các thuộc tính và giá trị, các thẻ cũng có thể lồng nhau. Do đó cần phải bóc tách các thẻ để lấy nội dung thông tin. Việc bóc tách nội dung được thực hiện cụ thể tùy theo từng kênh tin.

3.3.3. Xây dựng phân hệ xử lý dữ liệu

Phân hệ này có chức năng xử lý các tin tức thu thập được nhằm mục đích phân loại chuyên mục cho tin tức. Các bước xử lý bao gồm: Loại bỏ dấu câu, tách từ, tính toán ma trận trọng số TFIDF của tập tin tức, so sánh độ tương tự giữa tin mới và các tin có sẵn trong chuyên mục, xác định chuyên mục cho tin mới cập nhật.

3.3.4. Xây dựng Cổng thông tin tổng hợp (portal)

Cổng thông tin điện tử được xây dựng trên nền tảng Dotnetnuke portal. Các phân hệ tin tức được xây dựng thành 02 module chính trên dotnetnuke: module tin tức và module chuyên mục. Module tin tức có nhiệm vụ lấy và hiển thị tin trên trang chủ, quản lý tin (sửa, xóa, duyệt, ...), hiển thị tin theo chuyên mục, tìm kiếm, ... Module chuyên mục có chức năng quản lý chuyên mục (nhóm) tin, cho phép thêm, xóa, sửa nhóm tin, gán các tin được tải về tự động vào các chuyên mục nếu hệ thống phân loại sai.

3.4. KẾT QUẢ THỬ NGHIỆM HỆ THỐNG

Chương trình được cài đặt trên 2 máy trong mạng LAN. Các máy có cấu hình Intel Core 2 Duo, 3 GHz, RAM 1G.

Máy chủ	
Hệ điều hành	Microsoft Windows
Dung lượng ổ đĩa trống	500 MB
Cơ sở dữ liệu	Microsoft SQL Server 2005
Webserver	IIS
Server Application	ASP. NET
Máy trạm (phía người dùng)	
Hệ điều hành	Windows 98, 2000, XP hoặc Linux
Trình duyệt	IE, Netscape, Mozilla, Opera, FireFox…

Đánh giá kết quả:

Phân hệ Crawler và Extractor: hoạt động tốt và đúng theo yêu cầu đề ra, cho phép tải tin tức về từ các kênh cấu hình sẵn. Kết quả bóc tách nội dung tốt, không có sai sót, tuy nhiên phần xử lý tải hình ảnh có liên quan chưa được thực hiện.

Phân hệ xử lý dữ liệu và phân loại: kết quả phân loại tương đối chính xác, tuy nhiên do số tin thử nghiệm chưa nhiều do đó chưa có số liệu về tỉ lệ sai sót. Thời gian xử lý gom cụm tương đối chậm, do phải tính toán trên toàn bộ dữ liệu.

Các phân hệ quản lý hệ thống khác: vận hành tốt theo đúng thiết kế.

KẾT LUẬN

Đánh giá kết quả đề tài

Đề tài đã tìm hiểu được kiến thức tổng quan về khai phá dữ liệu, ứng dụng của phân cụm dữ liệu trong khai phá dữ liệu web, các thuật toán phân cụm tài liệu và cơ chế của hệ thống thu thập tin. Đồng thời ứng dựng xây dựng hệ thống tổng hợp thông tin kinh tế - chính trị - xã hội phục vụ công tác quản lý, chỉ đạo điều hành của lãnh đạo.

Đề tài đã thực hiện các nội dung sau:

- Tìm hiểu tổng quan về khai phá dữ liệu, các bài toán trong khai phá dữ liệu và ứng dụng.
- Tìm hiểu các kỹ thuật phân cụm tài liệu, mô hình không gian vector biểu diễn tài liệu.
- Tìm hiểu các kỹ thuật thu thập thông tin tự động trên internet và quá trình khai phá dữ liệu web.
- Đề xuất giải pháp kỹ thuật thu thập thông tin trên internet và phân cụm tin thu thập được.
- Xây dựng phần mềm thu thập tổng hợp thông tin, cổng thông tin (portal) và cài đặt, thử nghiệm hệ thống.

Hạn chế

- Về xử lý dữ liệu: chưa nghiên cứu các giải pháp tách từ tiếng Việt đầy đủ, do đó ảnh hưởng đến độ chính xác của việc phân cụm tài liệu.
- Hệ thống Crawler được xây dựng còn đơn giản chưa hỗ trợ duyệt các Url trên internet ở các cấp mức độ khác nhau.

Phạm vi áp dụng của đề tài:

Về lý thuyết: Qua nghiên cứu đề tài đã bước đầu đề cập đến các giải pháp kỹ thuật trong việc thu thập thông tin tự động trên internet, ứng dụng kỹ thuật khai phá dữ liệu phục vụ cho việc phân tích thông tin thu thập được theo các lĩnh vực, chủ đề khác nhau nhằm giúp cho người dùng theo dõi thông tin một cách thuận tiện, dễ dàng.

Về thực tiễn: Với việc triển khai hệ thống thử nghiệm cho thấy khả năng ứng dụng kết quả này trong việc xây dựng Hệ thống thông tin tổng hợp tự động cho phép kết nối nhiều

nguồn tin khác nhau và có thể ứng dụng phục vụ trong các cơ quan nhà nước, trong bối cảnh nhiều cơ quan ban ngành, địa phương đã và đang xây dựng các website riêng và cung cấp nhiều thông tin trên website của mình, do đó cần thiết phải có hệ thống kết nối và tổng hợp thông tin nhằm chia sẻ dữ liệu của các ban ngành khác trên địa bàn tỉnh để phục vụ tốt công tác quản lý nhà nước của địa phương.

Hướng phát triển

Mặc dù đã thực hiện các nội dung cơ bản và xây dựng vận hành thành công. Tuy nhiên, để có thể hoàn thiện tốt hơn, đề tài cần nghiên cứu bổ sung thêm các nội dung sau:

- Cải thiện chức năng của phân hệ bóc tách dữ liệu Text từ nội dung HTML một cách linh động hơn thay vì chỉ dựa trên cấu hình có sẵn.
- Nghiên cứu ứng dụng các giải thuật phân cụm nhằm tăng cường hiệu năng và độ chính xác của việc phân loại thông tin.