ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

TRƯƠNG MẠNH HÀ

NGHIÊN CỨU MỘT SỐ KỸ THUẬT LẤY TIN TỰ ĐỘNG TRÊN INTERNET

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Người hướng dẫn khoa học: **TS. Phạm Việt Bình**

Thái Nguyên - Năm 2009

LỜI CAM ĐOAN

Tôi xin cam đoan toàn bộ nội dung bản luận văn này là do tôi tự sưu tầm, tra cứu và sắp xếp cho phù hợp với nội dung yêu cầu của đề tài.

Nội dung luận văn này chưa từng được công bố hay xuất bản dưới bất kỳ hình thức nào và cũng không được sao chép từ bất kỳ một công trình nghiên cứu nào.

Tất cả phần mã nguồn của chương trình đều do tôi tự thiết kế và xây dựng, trong đó có sử dụng một số thư viện chuẩn và các thuật toán được các tác giả xuất bản công khai và miễn phí trên mạng Internet.

Nếu sai tôi xin tôi xin hoàn toàn chịu trách nhiệm.

Thái Nguyên, ngày 11 tháng 11 năm 2009

Người cam đoan

Trương Mạnh Hà

MỞ ĐẦU

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 dây điện thoại, cáp quang, sóng radio... 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 tự động các thông tin phù hợp và 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 [2].

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 sự 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 lọc 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 để cho người dùng có thể dễ dàng tìm được thông tin mà họ cần. Đã có nhiều thuật toán phân cụm tài liệu dựa trên phân cụm ngoại tuyến toàn bộ tập tài liệu. 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 một số kỹ thuật lấy tin tự động trên internet"

Mục tiêu của đề tài: Nghiên cứu xây dựng giải pháp phát triển hệ thống phần mềm thu thập, đánh giá và phân cụm thông tin tự động trên Internet phục vụ cho việc nghiên cứu, học tập, giảng dạy.

Ngoài phần mở đầu, phần kết luận, mục lục, tài liệu tham khảo, phụ lục, luận văn gồm 3 chương:

- Chương 1: Khái quát về khai phá dữ liệu và phân cụm tài liệu Web

Giới thiệu một số khái niệm cơ bản về khai phá dữ liệu, khai phá dữ liệu web, các hướng tiếp cận, ứng dụng của khai phá dữ liệu, và nêu bài toàn phân cụm tài liệu Web.

- Chương 2: Một số thuật toán phân cụm tài liệu

Nghiên cứu một số kỹ thuật phân cụm tài liệu liên quan, tư tưởng của các thuật toán đã được nghiên cứu, nghiên cứu đề xuất phương pháp cải tiến.

- Chương 3: Ứng dụng trong lấy tin tự động

Ứng dụng xây dựng bài toán Thu thập dữ liệu về Kinh tế trên Internet.

Để hoàn thành được luận văn Cao học, em xin được gửi lời cảm ơn tới các thầy trong Viện Công nghệ thông tin, các thầy trong Khoa Công nghệ thông tin đã tận tình giảng dạy, cung cấp nguồn kiến thức quý giá trong suốt quá trình học tập.

Đặc biệt em xin chân thành cảm ơn TS. Phạm Việt Bình, đã tận tình hướng dẫn, góp ý, tạo điều kiện cho em hoàn thành luận văn này.

Xin chân thành cảm ơn các thầy cô, anh chị em đang công tác tại phòng VRLAB - Viện công nghệ thông tin - Viện khoa học và Công nghệ Việt Nam, các thầy cô đang công tác tại Viện Công nghệ thông tin - Viện khoa học và Công nghệ Việt Nam.

Cảm ơn đồng nghiệp Đỗ Văn Đại đã cung cấp những tài liệu, cùng những kinh nghiệm quý báu đã được làm trong cuốn Đồ án tốt nghiệp đại học của đồng nghiệp Đỗ Văn Đại giúp cho em trong quá trình nghiên cứu giảm bớt được những khó khăn trong việc tiếp cận vấn đề và nghiên cứu tài liệu.

Xin được cảm ơn Ban lãnh đạo Khoa Công nghệ thông tin - Đại học Thái Nguyên, lãnh đạo phòng Công nghệ thông tin - Thư viện, cùng toàn thể các đồng nghiệp trong Khoa Công nghệ thông tin - Đại học Thái Nguyên đã giúp đỡ em về thời gian, vật chất và tinh thần giúp em hoàn thành tốt nhiệm vụ học tập, công tác.

Chương 1: KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU VÀ PHÂN CỤM TÀI LIỆU WEB

1.1 Khai phá dữ liệu:

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 [4].

Mô hình khai phá dữ liệu bao gồm năm giai đoạn chính:

- Tìm hiểu nghiệp vụ và dữ liệu
- Chuẩn bị dữ liệu
- Mô hình hoá dữ liệu
- Hậu xử lý và đánh giá mô hình
- Triển khai tri thức

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 sau. Tham gia chính trong quá trình Khai phá dữ liệu là các nhà tư vấn và phát triển chuyên nghiệp trong lĩnh vực Khai phá dữ liệu.

Trong giai đoạn đầu tiên, tìm hiểu nghiệp vụ dữ liệu, nhà tư vấn nghiên cứu kiến thức về lĩnh vực sẽ áp dụng, bao gồm các tri thức cấu trúc về hệ thống và tri thức, các nguồn dữ liệu hiện hữu, ý nghĩa, vai trò và tầm quan trọng của các thực thể dữ liệu. Việc nghiên cứu này được thực hiện qua việc tiếp xúc giữa nhà từ vấn và người dùng. Khác với phương pháp giải quyết vấn đề truyền thống khi bài toán được xác định chính xác ở bước đầu tiên, nhà tư vấn tìm hiểu các yêu cầu sơ khởi của người dùng và đề nghị các bài toán tiềm năng có thể giải quyết với nguồn dữ liệu hiện hữu. Tập các bài toán tiềm năng được tinh chỉnh và làm hẹp lại trong các giai đoạn sau. Các nguồn và đặc tả dữ liệu có liên quan đến tập các bài toán tiềm năng cũng được xác định [4].

Giai đoạn chuẩn bị dữ liệu sử dụng các kỹ thuật tiền xử lý để biến đổi và cải thiện chất lượng dữ liệu để thích hợp với những yêu cầu của các giải thuật học. Phần lớn các giải thuật khai phá dữ liệu hiện nay chỉ làm việc trên một tập dữ liệu đơn và phẳng, do đó dữ liệu phải được trích xuất và biến đối từ các dạng cơ sơ dữ liệu phân bố, quan hệ hay hướng đối tượng sang dạng cơ sở dữ liệu quan hệ đơn giản với một bảng dữ liệu. Các giải thuật tiền xử lý tiêu biểu bao gồm:

(a) Xử lý dữ liệu bị thiếu/mất: các dữ liệu bị thiếu sẽ được thay thế bởi các giá trị thích hợp.

(b) Khử sự trùng lắp: các đối tượng dữ liệu trùng lắp sẽ bị loại bỏ đi. Kỹ thuật này không được sử dụng cho các tác vụ có quan tâm đến phân bố dữ liệu.

(c) Giảm nhiễu: nhiễu và các đối tượng tách rời (outlier) khỏi phân bố chung sẽ bị loại đi khỏi dữ liệu.

(d) Chuẩn hóa: miền giá trị của dữ liệu sẽ được chuẩn hóa.

(e) Rời rạc hóa: các dữ liệu số sẽ được biến đổi ra các giá trị rời rạc.

(f) Rút trích và xây dựng đặc trưng mới từ các thuộc tính đã có.

(g) Giảm chiều: các thuộc tính chứa ít thông tin sẽ được loại bỏ bớt.

Các bài toán được giải quyết trong giai đoạn Mô hình hóa dữ liệu. Các giải thuật học sử dụng các dữ liệu đã được tiền xử lý trong giai đoạn hai để tìm kiếm các qui tắc ẩn và chưa biết. Công việc quan trọng nhất trong giai đoạn này là lựa chọn kỹ thuật phù hợp để giải quyết các vấn đề đặt ra. Các bài toán được phân loại vào một trong những nhóm bài toán chính trong Khai phá dữ liệu dựa trên đặc tả của chúng [4].

Các mô hình kết quả của giai đoạn ba sẽ được hậu xử lý và đánh giá trong giai đoạn (d). Dựa trên các đánh giá của người dùng sau khi kiểm tra trên các tập thử, các mô hình sẽ được tinh chỉnh và kết hợp lại nếu cần. Chỉ các mô hình đạt được mức yêu cầu cơ bản của người dùng mới đưa ra triển khai trong thực tế. Trong giai đoạn này, các kết quả được biến đổi từ dạng học thuật sang dạng phù hợp với nghiệp vụ và dễ hiểu hơn cho người dùng.

Trong giai đoạn cuối, Triển khai tri thức, các mô hình được đưa vào những hệ thống thông tin thực tế dưới dạng các module hỗ trợ việc đưa ra quyết định.

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 [3]. 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.

1.1.1 Các dạng dữ liệu

1.1.1.1 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. Ý nghĩa của mỗi từ trong tài liệu không cố định mà tùy thuộc vào từng ngữ cảnh khác nhau sẽ mang ý nghĩa khác nhau. Các từ trong tài liệu được liên kết với nhau theo một ngôn ngữ nào 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.

Cơ sở dữ liệu Full text là một dạng cơ sở dữ liệu phi cấu trúc mà dữ liệu bao gồm các tài liệu và thuộc tính của tài liệu. Cơ sở dữ liệu Full_Text thường được tổ chức như một tổ hợp của hai thành phần: Một cơ sở dữ liệu có cấu trúc thông thường (chứa đặc điểm của các tài liệu) và các tài liệu.

1.1.1.2 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. Ví dụ khi đọc một cuốn sách người đọc không phải đọc lần lượt từng trang từ đầu đến cuối mà có thể nhảy cóc đến các đoạn sau để tham khảo về các vấn đề họ quan tâm.

Như vậy văn bản Hypertext bao gồm dạng chữ viết không liên tục, chúng được phân nhánh và cho phép người đọc có thể chọn cách đọc theo ý muốn của mình. Hiểu theo nghĩa thông thường thì Hypertext là một tập các trang chữ viết được kết nối với nhau bởi các liên kết và cho phép người đọc có thể đọc theo các cách khác nhau. Như ta đã làm quen nhiều với các trang định dạng HTML, trong các trang có những liên kết trỏ tới từng phần khác nhau của trang đó hoặc trỏ tới trang khác và người đọc sẽ đọc văn bản dựa vào những liên kết đó.

Bên cạnh đó, Hypertext cũng là một dạng văn bản Text đặc biệt nên cũng có thể bao gồm các chữ viết liên tục (là dạng phổ biến nhất của chữ viết). Do không bị hạn chế bởi tính liên tục trong Hypertext, chúng ta có thể tạo ra các dạng trình bày mới, do đó tài liệu sẽ phản ánh tốt hơn nội dung muốn diễn đạt. Hơn nữa người đọc có thể chọn cho mình một cách đọc phù hợp chẳng hạn như đi sâu vào một vấn đề mà họ quan tâm. Sáng kiến tạo ra một tập các văn bản cùng với các con trỏ tới các văn bản khác để liên kết một tập các văn bản có mối quan hệ với nhau là một cách thực sự hay và hữu ích để tổ chức thông tin. Với người viết, cách này cho phép họ có thể thoải mái loại bỏ những băn khoăn về thứ tự trình bày mà có thể tổ chức vấn đề thành những phần nhỏ rồi sử dụng kết nối để chỉ ra mối liên hệ giữa các phần nhỏ đó với nhau.

Với người đọc, cách này cho phép họ có thể đi tắt trên mạng thông tin và quyết định phần thông tin nào có liên quan đến vấn đề mà họ quan tâm để tiếp tục tìm hiểu. So sánh với cách đọc tuyến tính tức là đọc lần lượt thì Hypertext đã cung cấp cho chúng ta một giao diện để có thể tiếp xúc với nội dung thông