ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN DANH HÙNG

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG TỔNG HỢP, PHÂN LOẠI THÔNG TIN TỰ ĐỘNG TRÊN WEB

Chuyên ngành: Khoa học máy tính
Mã số : 60.48.0101

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Người hướng dẫn khoa học: PGS.TS ĐOÀN VĂN BAN

Thái nguyên – Năm 2014

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

KDD	Knowledge Discovery in Database
KPDL	Khai phá dữ liệu
IGATEC	Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese
kNN	K–Nearest Neighbor
LLSF	Linear Least Square Fit
NB	Naïve Bayes
NNet	Neural Network
LLSF	Linear Lest Square Fit
DF	Tần suất tài liệu (Document Frequency
TBL	Phương pháp giải thuật học cải biến (Transformation – based Learning
IDF	Tần suất tài liệu ngược (Inverse document frequency)
TF	Tần suất từ (Term frequency

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây cùng với sự phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ về tri thức. Kho dữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vô tận làm cho vấn đề khai thác các nguồn tri thức đó ngày càng trở nên nóng bỏng và đặt ra thách thức lớn cho nền công nghệ thông tin thế giới.

Cùng với những tiến bộ vượt bậc của công nghệ thông tin là sự phát triển mạnh mẽ của mạng thông tin toàn cầu, nguồn dữ liệu Web trở thành kho dữ liệu khổng lồ. Nhu cầu khai thác và xử lý thông tin phục vụ cho công tác quản lý, hoạt động sản xuất, kinh doanh, học tập… đã trở nên cấp thiết trong xã hội hiện đại. Do đó số lượng văn bản xuất hiện trên mạng Internet cũng tăng theo một tốc độ chóng mặt. Với lượng thông tin đồ sộ như vậy, một yêu cầu lớn đặt ra là làm sao tổ chức, tìm kiếm và có được thông tin nhanh chóng, hiệu quả nhất.

Để giải quyết vấn đề này, có một hướng giải quyết là nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu trong môi trường Web. Vì vậy tôi chọn đề tài "nghiên cứu xây dựng hệ thống tổng hợp, phân loại thông tin tự động trên web" nhằm tìm hiểu phương pháp tổng hợp tin từ nhiều website và tự động phân loại các tin được lấy về.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

Tìm hiểu về khai phá dữ liệu web, các thuật toán phân loại tài liệu và ứng dụng trong truy xuất thông tin tự động. Trên cơ sở đó, xây dựng hệ thống tổng hợp, phân loại thông tin tự động trên web.

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 tổng hợp, phân loại thông tin trực tuyến.

3. Hướng nghiên cứu của đề tài

Về lý thuyết: Nghiên cứu các giải pháp kỹ thuật trong việc thu thập thông tin tự động trên internet, ứng dụng kỹ thuật khai phá dữ liệu cho việc phân tích thông tin thu thập được theo các lĩnh vực khác nhau nhằm giúp người dung theo dõi, tìm kiếm thông tin dễ dàng, thuận tiện.

Về thực tiễn: Ứng dụng hệ thống này trong việc xây dựng hệ thống tổng hợp, phân loại thông tin việc làm tự động.

4. Những nội dung chính

Luận văn được trình bày trong 3 chương, có phần mở đầu, phần kết luận, phần mục lục, phần tài liệu tham khảo. Các nội dung cơ bản của luận văn được trình như sau:

Chương 1: Trình bày những nội dung tổng quan về khai phá dữ liệu, khai phá web, phân loại văn bản.

Chương 2: Trình bày một số phương pháp tách, phân loại từ tiếng Việt và mô hình hệ thống tổng hợp, phân loại tin tức.

Chương 3: Trình bày giải pháp xây dựng thử nghiệm hệ thống tổng hợp, phân loại thông tin việc làm tự động.

5. Phương pháp nghiên cứu

Nghiên cứu lý thuyết:

- Tìm hiểu lý thuyết về khai phá dữ liệu và khai phá dữ liệu web.

- Tìm hiểu các thuật toán phâm cụm tài liệu.

- Tìm hiểu cơ chế hoạt động của các hệ thống tìm kiếm thu thập thông tin.

Nghiên cứu thực nghiệm:

- Dựa trên lý thuyết đã nghiên cứu, tiến hành xây dựng hệ thống thu thập và phân loại thông tin từ các kênh tin được cấu hình trước.

- Thử nghiệm trên máy đơn qua localhost có kết nối internet.

6. Ý nghĩa khoa học

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

Về mặt thực tiễn: Xây dựng hệ thống tổng hợp, phân loại thông tin tự động trên web. Cho phép người dung cập nhật các thông tin mới nhất từ các website khác, lưu trữ, tìm kiếm thông tin theo các chuyên mục.