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

HOÀNG MINH THỦY

NGHIÊN CỨU CÁC PHƯƠNG PHÁP TRÍCH CHỌN THÔNG TIN VÀ ỨNG DỤNG TRÍCH CHỌN THÔNG TIN DU LỊCH TRONG VĂN BẢN TIẾNG VIỆT

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

GS. VŨ ĐỨC THI

Thái Nguyên – 2015

LỜI CAM ĐOAN

Tác giả Hoàng Minh Thủy xin cam kết rằng nội dung của Luận văn này chưa được nộp cho bất kỳ một chương trình cấp bằng cao học nào cũng như bất kỳ một chương trình đào tạo cấp bằng nào khác.

Ngoài ra, tác giả cũng xin cam kết Luận văn thạc sĩ này là nỗ lực riêng của cá nhân tác giả. Các kết quả, phân tích, kết luận trong Luận văn thạc sĩ này (ngoài các phần được trích dẫn) đều là kết quả làm việc của cá nhân tác giả.

Thái Nguyên, ngày 10 tháng 11 năm 2015

Tác Giả

Hoàng Minh Thủy

LỜI CẢM ƠN

Lời đầu tiên em xin gửi lời cảm ơn chân thành đến Các quý thầy cô giáo, Tổ chuyên môn Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tận tình giảng dạy, truyền đạt những kiến thức, kinh nghiệm quý báu trong suốt thời gian em theo học tại trường. Các kiến thức, kinh nghiệm quý báu của các Quý thầy cô giáo không chỉ giúp cá nhân em hoàn thiện hệ thống kiến thức trong học tập mà còn giúp em ứng dụng các kiến thức đó trong công tác hiện tại tại đơn vị.

Đặc biệt, em xin chân thành cảm ơn thầy giáo GS. Vũ Đức Thi đã rất nhiệt tình và tâm huyết trong việc định hướng và giúp đỡ em hoàn thành luận văn này.

Ngoài ra, em cũng xin chân thành cảm ơn Ban lãnh đạo và cán bộ viên chức Trường Đại học Lâm nghiệp đã tạo điều kiện cung cấp những ý kiến quý báu và những kiến thức thực tiễn cho em thực hiện luận văn tốt nghiệp này.

Em cũng xin được bày tỏ tình cảm với gia đình, đồng nghiệp, bạn bè đã tạo điều kiện để cá nhân em có thể dành thời gian cho khóa học. Xin chân thành cảm ơn những người bạn lớp cao học CK13, trong 2 năm qua đã luôn luôn động viên, khích lệ và hỗ trợ em trong quá trình học tập.

Trong quá trình thực hiện Luận văn mặc dù đã cố gắng hết mình, song chắc chắn luận văn của em vẫn còn nhiều thiếu sót. Em rất mong nhận được sự chỉ bảo vào đóng góp tận tình của các thầy cô để luận văn của em được hoàn thiện hơn.

Thái Nguyên, ngày 10 tháng 11 năm 2015

Tác Giả

Hoàng Minh Thủy

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1.1. Sự cần thiết lựa chọn đề tài

Trích chọn thông tin (IE - Information Extraction) là một lĩnh vực nghiên cứu quan trọng trong khai phá dữ liệu văn bản [3, 4]. Trích chọn thông tin là quá trình thu thập thông tin từ các nguồn dữ liệu theo nhiều định dạng khác nhau, không đồng nhất, thậm chí không có định dạng cụ thể, sau đó chuyển thành một dạng đồng nhất. Dữ liệu sau khi trích chọn được lưu vào cơ sở dữ liệu để xử lý hay được sử dụng cho những hệ thống khai phá dữ liệu. Từ dữ liệu, thông tin được trích chọn ra có thể sử dụng các kỹ thuật phân tích, khai phá để khám phá ra các mẫu thông tin có ích, tiềm ẩn trong dữ liệu.

Ngày nay, cùng với sự phát triển của công nghệ thông tin, Tin học đã dần được ứng dụng rộng rãi trong nhiều lĩnh vực như kinh tế, du lịch, thương mại, y tế, ngân hàng và mang lại nhiều lợi ích to lớn. Nền kinh tế không ngừng phát triển, đời sống văn hoá - xã hội ngày càng được nâng cao thì du lịch đã trở thành một nhu cầu không thể thiếu trong cuộc sống của người dân, trên các trang web du lịch là hàng loạt thông tin về các tour du lịch trong nước và ngoài nước. Tuy nhiên lượng thông tin về các tour du lịch trên Internet là vô cùng lớn, gây khó khăn cho người có nhu cầu du lịch trong việc lựa chọn địa điểm tham quan, lựa chọn công ty cung cấp dịch vụ,…. Do vậy, một bài toán đặt ra là cần phải xây dựng một hệ thống tìm kiếm và tư vấn du lịch, giúp người dùng có thể lựa chọn được những tour du lịch phù hợp nhất với yêu cầu đề ra. Để có một hệ thống tìm kiếm và tư vấn tốt thì trước tiên ta phải xây dựng được tập dữ liệu có độ chính xác cao. Cùng với nó là bài toán con trích chọn thông tin du lịch trong văn bản tiếng Việt.

Để có thể tiến đến tìm hiểu được những vấn đề trên, em lựa chọn đề tài "***Nghiên cứu các phương pháp trích chọn thông tin và ứng dụng trích chọn thông tin du lịch trong văn bản Tiếng Việt***" làm luận văn tốt nghiệp Thạc sĩ của mình.

1.2. Mục tiêu đề tài

Tìm hiểu các phương phương pháp trích chọn thông tin và xây dựng mô hình giải quyết bài toán trích chọn thông tin về các tour du lịch từ các trang thông tin điện tử tiếng Việt trên Internet.

1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các phương pháp tiếp cận giải quyết bài toán trích chọn thông tin trong văn bản tiếng Việt và các trang thông tin điện tử tiếng Việt trên mạng Internet về lĩnh vực du lịch.

Phạm vi nghiên cứu của đề tài là bài toán trích chọn thông tin về các tour du lịch trên một số trang thôn tin điện tử tiếng Việt (website) trên mạng Internet.

1.4. Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài là nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

Về nghiên cứu lý thuyết, đề tài đã tổng hợp các kết quả nghiên cứu về các phương pháp trích chọn thông tin từ văn bản tiếng Việt phục vụ phân tích, thống kê, báo cáo, ra quyết định. Về nghiên cứu thực nghiệm, đề tài xây dựng và cài đặt, thử nghiệm mô hình trích chọn thông tin du lịch từ một số trang web về du lịch bằng tiếng Việt trên mạng Internet.

1.5. Cấu trúc của luận văn

Cấu trúc luận văn gồm: mở đầu, bốn chương chính, kết luận và tài liệu tham khảo.

Phần mở đầu: Lý do chọn đề tài và bố cục luận văn

Chương 1: Giới thiệu tổng quan bài toán trích chọn thông tin và một số lĩnh vực nghiên cứu liên quan.