ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN DUY DŨNG

Các thuật toán phân lớp dữ liệu và ứng dụng xây dựng hệ thống hỏi đáp tự động về một số bệnh thường gặp

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là kết quả nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong luận văn là trung thực. Được các tác giả cho phép tham khảo và sử dụng các tài liệu đăng tải trên các tác phẩm, tạp chí và các trang web theo danh mục tài liệu tham khảo của luận văn.

LỜI CẢM ƠN

Tôi xin được gửi lời cảm ơn trân trọng và sâu sắc nhất đến thầy giáo ***PGS.TS. Đoàn Văn Ban*** *– thầy đã tận tình giúp đỡ, hướng dẫn cho tôi trong suốt quá trình học tập và nghiên cứu, thực hiện đề tài này.*

Tôi cũng xin gửi lời biết ơn chân thành đến quý Thầy giáo, cô giáo Viện Công nghệ thông tin và quý Thầy cô trường Đại học Công nghệ thông tin & truyền thông Đại học Thái Nguyên đã tận tình giảng dạy, trang bị cho tôi những kiến thức quý báu trong suốt quá trình học tập tại trường.

Tôi cũng xin gửi lời biết ơn chân thành đến Ban giám hiệu, các phòng ban trường Cao đẳng Y tế Thanh Hóa đã tạo điều kiện cho tôi tham gia lớp học này.

Tôi cũng xin gửi lời biết ơn chân thành đến cơ quan Bắc Trung Bộ đã giúp đỡ hỗ trợ cho tôi tham gia khóa học này.

Tôi xin được cảm ơn, chia sẻ niềm vui này với gia đình, bạn bè đồng nghiệp và các y bác sĩ cùng anh chị em lớp Cao học K12G trường Đại học Công nghệ thông tin & truyền thông Đại học Thái Nguyên, những người đã luôn ở bên tôi, giúp đỡ và tạo điều kiện thuận lợi để cho tôi được học tập, nghiên cứu, hoàn thành luận văn.

MỤC LỤC

DANH MỤC CÁC CHỮ VIẾT TẮT

Chữ viết tắt	Tiếng anh	Tiếng việt
Q&A	Question Answering	Hỏi - đáp
CSDL		Cơ sở dữ liệu

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC HÌNH

ĐẶT VẤN ĐỀ

Ngày nay với sự phát triển mạnh mẽ của khoa học kỹ thuật từ lý thuyết đến ứng dụng, người ta đang cố gắng đưa công nghệ thông tin vào các ngành nghề như: khoa học kỹ thuật, giáo dục, y tế, v.v, trong đó lĩnh vực y tế ngày càng được nhiều người quan tâm. Các nhà nghiên cứu về hệ thống hỏi đáp cũng bắt đầu khai thác web như là một nguồn dữ liệu cho việc tìm kiếm câu trả lời.

Phân tích câu hỏi là phần đầu tiên trong kiến trúc chung của một hệ thống hỏi đáp, có nhiệm vụ tìm ra các thông tin cần thiết làm đầu vào cho quá trình xử lý của các phần sau (trích chọn tài liệu, trích xuất câu trả lời, v.v). Vì vậy,việc phân tích câu hỏi có vai trò hết sức quan trọng, ảnh hưởng trực tiếp đến hoạt động của toàn bộ hệ thống. Nếu phân tích câu hỏi không tốt thì sẽ không thể tìm ra được câu trả lời.

Hệ thống hỏi - đáp tự động là một công cụ hữu hiệu phục vụ cho nhu cầu tìm kiếm trao đổi thông tin ngày càng cao của con người, trong hệ thống hỏi đáp có rất nhiều dạng câu hỏi như: Câu hỏi dạng định nghĩa (What), câu hỏi về nơi chốn (Where), câu hỏi như thế nào (How), câu hỏi đúng/sai (Yes/No). Nhưng hệ thống hỏi - đáp (Yes/No) lại mới chỉ được quan tâm trong vài năm gần đây. Như vậy,việc xây dựng một hệ thống hỏi - đáp (Yes/No) là một nhu cầu cần thiết. Hướng tới mục tiêu này, chúng tôi muốn xây dựng một mô hình hệ thống hỏi - đáp tự động (Yes/No) nhằm phục vụ cho một lĩnh vực cụ thể là hỗ trợ việc chẩn đoán và khuyến nghị điều trị các bệnh lý thông thường trong cuộc sống.

Trong cuộc sống hằng ngày, có rất nhiều các loại bệnh thường xuyên đe dọa đến sức khỏe của con người chúng ta. Thường các loại bệnh này xuất phát từ các triệu chứng, nhưng không phải ai cũng biết. Cho nên yêu cầu của con người chúng ta cần có một hệ thống hỏi - đáp giúp họ chẩn đoán được các bệnh và giúp họ hướng giải quyết để đảm bảo được sức khỏe cho chính mình.

Như vậy, mục tiêu của đề tài này là tìm hiểu các tri thức cơ bản của y khoa về các loại bệnh thông thường, thu thập tri thức để xây dựng một hệ thống hỏi đáp

nhằm hỗ trợ chẩn đoán và phân loại các bệnh thường gặp, cho người sử dụng những lời khuyên hữu ích trong việc phòng và điều trị bệnh.

Nhận thấy tính thiết thực của vấn đề này và được sự gợi ý của giảng viên hướng dẫn, tôi đã chọn đề tài "*Các thuật toán phân lớp dữ liệu và ứng dụng xây dựng hệ thống hỏi đáp tự động về một số bệnh thường gặp".*

1. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

- Nghiên cứu kỹ thuật phân lớp dữ liệu trong khai phá dữ liệu.

- Tìm hiểu về các bệnh thường gặp và xây dựng cơ sở tri thức về các biểu hiện của bệnh dựa trên cơ sở dữ liệu thu thập được tại Bệnh viện Đa khoa tỉnh Thanh Hóa để phân lớp các loại bệnh.

2. PHƯƠNG PHÁP NGHIÊN CỨU

- ***Phương pháp nghiên cứu lý thuyết***: Nghiên cứu qua các tài liệu, sách, sách điện tử, các bài báo, thông tin tài liệu trên các website và các tài liệu liên quan và công nghệ liên quan, tổng hợp các tài liệu, phân tích và thiết kế hệ thống thông tin theo quy trình xây dựng ứng dụng phần mềm.

- ***Phương pháp nghiên cứu thực nghiệm***: Phân tích hiện trạng và yêu cầu thực tế của bài toán và xây dựng các bước phân tích hệ thống để hỗ trợ việc lập trình, xây dựng ứng dụng,vận dụng các vấn đề nghiên cứu về mã hóa thông tin trong tiến trình xây dựng hệ thống, đánh giá kết quả đạt được.

3. HƯỚNG NGHIÊN CỨU CỦA ĐỀ TÀI

- Nghiên cứu phương pháp phân lớp dữ liệu trong KPDL, các thuật toán liên quan đến quy nạp cây quyết định, tìm hiểu các ngôn ngữ mã lệnh siêu tìm kiếm.

- Tìm hiểu hệ thống hỏi đáp tự động, ứng dụng công nghệ tri thức hỗ trợ phục vụ chẩn đoán và đưa ra khuyến nghị điều trị một số bệnh thường gặp.

4. BỐ CỤC LUẬN VĂN

Sau phần mở đầu, nội dung chính của luận văn được chia thành 3 chương: