ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ THỊ BÍCH HẢO

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP TRÍCH CHỌN ĐẶC TRƯNG TRONG KHAI PHÁ QUAN ĐIỂM VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2016

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ THỊ BÍCH HẢO

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP TRÍCH CHỌN ĐẶC TRƯNG TRONG KHAI PHÁ QUAN ĐIỂM VÀ ỨNG DỤNG

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: TS NGUYỄN VIỆT ANH

THÁI NGUYÊN – 2016

LỜI CẢM ƠN

Trước hết tôi xin bày tỏ lòng biết ơn sâu sắc và gửi lời cảm ơn đặc biệt nhất tới Thầy TS. Nguyễn Việt Anh, người đã định hướng đề tài, cung cấp cho tôi những kiến thức, những tài liệu và tận tình hướng dẫn chỉ bảo tôi trong suốt quá trình thực hiện đề tài luận văn cao học này, từ những ý tưởng trong đề cương nghiên cứu, phương pháp nghiên cứu, phương pháp giải quyết vấn đề cho đến những lần kiểm tra cuối cùng để hoàn thành luận văn này.

Tôi xin gửi lời cảm ơn chân thành tới Ban Giám hiệu Nhà trường, Phòng Đào tạo sau đại học, Đại học Công nghệ thông tin và truyền thông Thái Nguyên đã tạo điều kiện tốt nhất giúp tôi trong suốt quá trình học tập.

Cuối cùng tôi xin gửi lời cảm ơn đến gia đình, bạn bè những người đã luôn động viên khuyến khích tôi trong suốt quá trình học tập cũng như thực hiện đề tài luận văn của mình.

Thái Nguyên, ngày 6 tháng 4 năm 2016

Học viên

Lê Thị Bích Hảo

LỜI CAM ĐOAN

Tôi xin cam đoan nội dung trình bày trong luận văn này là do tôi tự nghiên cứu tìm hiểu dựa trên các tài liệu và tôi trình bày theo ý hiểu của bản thân dưới sự hướng dẫn trực tiếp của Thầy TS. Nguyễn Việt Anh. Các nội dung nghiên cứu, tìm hiểu và kết quả thực nghiệm là hoàn toàn trung thực.

Luận văn này của tôi chưa từng được ai công bố trong bất cứ công trình nào.

Trong quá trình thực hiện luận văn này tôi đã tham khảo đến các tài liệu của một số tác giả, tôi đã ghi rõ tên tài liệu, nguồn gốc tài liệu, tên tác giả và tôi đã liệt kê trong mục "DANH MỤC TÀI LIỆU THAM KHẢO" ở cuối luận văn.

Học viên

Lê Thị Bích Hảo

MỤC LỤC

Trang

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH VẼ ĐỒ THỊ

MỞ ĐẦU

Trên thế giới nói chung và ở Việt Nam nói riêng, thương mại điện tử đã trở nên phổ biến và ngày càng phát triển. Một phần quan trọng trong thương mại điện tử là bán hàng trực tuyến. Số lượng người mua hàng trực tuyến gia tăng, số lượng đánh giá, nhận xét của người dùng về các sản phẩm cũng ngày càng nhiều. Một sản phẩm thông dụng có thể có hàng trăm, hàng nghìn đánh giá. Cùng với các trang web bán hàng trực tuyến là các trang web đánh giá sản phẩm như epinions.com, dpreview.com, vnreview.vn, trustedreviews.com, tinhte.vn, ... Các trang web này là nơi người tiêu dùng viết các đánh giá của mình về một sản phẩm nào đó. Các đánh giá được đăng trên một trang web loại này cần tuân theo một số quy định do các trang web đó đưa ra và sẽ được chấm điểm bởi đông đảo người dùng của trang web căn cứ vào độ tin cậy, hợp lý và hữu dụng mà các đánh giá này mang lại. Chính bởi vậy, các bài đánh giá từ các trang web loại này được coi là nguồn tổng hợp lớn các đánh giá sản phẩm tin cậy từ khách hàng. Đây là nguồn thông tin quan trọng, cung cấp cho người mua hàng cái nhìn toàn diện hơn về một sản phẩm mà họ định mua. Còn đối với nhà sản xuất, đánh giá của khách hàng là cơ sở để tiến hành cải tiến, hoàn thiện sản phẩm của mình. Tuy nhiên, một vấn đề đặt ra là số lượng các ý kiến đánh giá rất lớn. Điều này gây khó khăn cho cả người mua hàng và nhà sản xuất. Người mua hàng sẽ gặp khó khăn trong việc tổng hợp ý kiến của những người tiêu dùng trước để đưa ra quyết định mua hay không mua một sản phẩm. Còn nhà sản xuất thì khó theo dõi, nắm bắt được tất cả phản hồi của người tiêu dùng về sản phẩm của mình. Thực tế trên làm nảy sinh yêu cầu tổng hợp tất cả nhận xét của khách hàng về các đặc trưng của sản phẩm trên một trang web đánh giá sản phẩm.

Theo cuộc khảo sát hơn 2000 người Mỹ trưởng thành cho th ấy 81% người dùng internet (chiếm tỷ lệ 60% người Mỹ) đã thực hiện việc tìm hiểu về một sản phẩm thông qua internet. Có từ 73% đến 87% số người nói rằng các nhận xét về sản phẩm có sự ảnh hưởng quan trọng đến việc lựa chọn mua sản phẩm của họ . Như vậy, quan điểm của người khác giúp chúng ta có thêm thông tin khi quyết định một

vấn đề, nó ảnh hưởng rất lớn đến hành vi của chúng ta. Tại Việt Nam theo báo cáo thương mại điện tử của Bộ công thương công bố năm 2014 [1] loại mặt hàng được mua trực tuyến là đồ công nghệ điện tử chiếm tới 61%, yếu tố được quan tâm khi mua sắm là 81% người ra rằng uy tín của người bán hàng 64% theo thương hiệu của sản phẩm; thống kê năm 2015 của Google [2] về người dùng internet có xu hướng theo lời khuyên trực tuyến 50% để mua đồ.

Việc giúp người có ý định mua có thể tham khảo tốt hơn ý kiến người dùng, hay giúp nhà cung cấp sản phẩm biết được cộng đồng đang quan tâm đến sản phẩm của mình trên những khía cạnh nào, chính là động lực để học viên nghiên cứu đề tài.

Đối với bài toán trên cũng đã có rất nhiều các công trình nghiên cứu và ứng dụng trên thế giới trong hơn một thập kỷ qua và đã đưa ra nhiều kết quả đáng chú ý được mô tả tổng hợp bởi một số nhà nghiên cứu uy tín trong ngành như Bing Liu [3] hay Moghaddam [4]… và đó là trên thế giới, trong nước đề tài này cũng đang nhận được nhiều sự chú ý quan tâm của các nhà nghiên cứu trong những năm gần đây, nổi bật có các nhóm tác giả Bảo Sơn [5] và nhóm của Hà Thụy [6], [7] đã đưa ra một số kết quả là mô hình áp dụng đối với một số bộ dữ liệu tiếng Việt và bộ từ điển miền Tiếng Việt…

Luận văn định hướng tìm hiểu các phương pháp trích chọn đặc trưng trong khai phá quan điểm để biểu diễn đối tượng được quan tâm, trên cơ sở đó đề xuất phương pháp và thử nghiệm ứng dụng hệ thống trong bài toán trích chọn đặc trưng sản phẩm cụ thể là *điện thoại di động*, từ những dữ liệu thu thập được trên website diễn dàn đánh giá sản phẩm. Với ý nghĩa thực tế có thể ứng dụng trong thị trường trong nước, học viên xin được đề xuất nghiên cứu và đưa ra mô hình ứng dụng của mình. Mô hình bao gồm các bước từ thu thập dữ liệu, tiền xử lý dữ liệu, đến ứng dụng các thuật toán mô hình lan truyền kép để trích chọn ra các đặc trưng, sử dụng phân cụm để gộp nhóm các đặc trưng. Cuối cùng là đưa ra những đánh giá đối với riêng hiệu quả thuật toán, bộ dữ liệu, kết quả đạt được và đánh giá về tính khả thi ứng dụng mô hình.

Cấu trúc của luận văn sẽ chia thành 4 phần chính:

Phần I. Mô tả tổng quan về bài toán khai phá quan điểm, trong đó nêu rõ những vấn đề nổi bật trong bài toán này tiếp tới là đi sâu hơn vào bài toán khai phá quan điểm dựa trên đặc trưng, những bài toán con cần giải quyết và phương pháp đánh giá. Những vấn đề nêu trên đều có giới thiệu các nghiên cứu trong và ngoài nước liên quan.

Phần II. Mô tả cụ thể chi tiết các phương pháp giải quyết bài toán trích trọn đặc trưng nổi bật trên thế giới, phân tích và đưa ra quyết định ứng dụng vào mô hình giải quyết bài toán của mình.

Phần III. Phát biểu bài toán và đưa ra mô hình ứng dụng đối với bài toán trích chọn đặc trưng cho miền dữ liệu tiếng Việt về sản phẩm điện thoại di động. Tiếp theo là đưa ra kết quả thực nghiệm và những phân tích chủ quan của học viên về kết quả đạt được của mô hình.

Phần IV. Kết luận tổng kết quá trình thực hiện luận văn, những khó khăn, thách thức, những kết quả đạt được và định hướng hướng nghiên cứu áp dụng tiếp theo.