BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN HẢI MINH

KHAI PHÁ DỮ LIỆU TỪ CÁC MẠNG XÃ HỘI ĐỂ KHẢO SÁT Ý KIẾN CỦA KHÁCH HÀNG ĐỐI VỚI MỘT SẢN PHẨM THƯƠNG MẠI ĐIỆN TỬ

Chuyên ngành: Khoa học máy tính
Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: TS. HUỲNH CÔNG PHÁP

Phản biện 1: TS. Nguyễn Thanh Bình

Phản biện 2: PGS.TS. Trương Công Tuấn

Luận văn đã được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp Thạc sĩ kỹ thuật họp tại Đại Học Đà Nẵng vào ngày 16 tháng 10 năm 2013.

Có thể tìm hiểu Luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời buổi kinh tế thị trường ngày hôm nay, một doanh nghiệp muốn tồn tại và phát triển thì cần phải khai thác và thu thập được các ý kiến phản hồi của người dùng về sản phẩm hay dịch vụ của mình từ đó đưa ra những định hướng và điều chỉnh về hoạt động sản xuất kinh doanh phù hợp hơn.

Cùng với sự ra đời của internet, sự xuất hiện và phát triển không ngừng của lĩnh vực thương mại điện tử khiến cho việc xúc tiến các hoạt động kinh doanh, buôn bán, quảng bá sản phẩm, dịch vụ diễn ra trên khắp các kênh thông tin xã hội đặc biệt là trên mạng internet. Điều này vô hình dung tạo nên cầu nối giữa người dùng và nhà cung cấp, và từ cầu nối này người dùng có thể đưa ra ý kiến của họ đối với sản phẩm hay dịch vụ mà nhà cung cấp mang lại.

Như chúng ta đã biết ngày nay mọi thông tin đều được đưa lên các trang mạng xã hội dưới dạng các posts và rất nhiều người dùng để lại các các nhận xét của mình về các posts này dưới dạng các comments, ta nhận thấy đây là kho thông tin khổng lồ mà từ đó nếu chúng ta có thể khai phá và trích rút tất cả các comments của người dùng, sau đó phân tích và phân loại dữ liệu ấy, chúng ta có thể thu được các kết quả khảo sát cần thiết phục vụ cho hoạt động sản xuất kinh doanh. Kết quả khảo sát ấy có thể là tỉ lệ người dùng thích, không thích hay không có ý kiến đối với sản phẩm hay dịch vụ mà họ quan tâm.

Từ việc nhìn thấy kho dữ liệu khổng lồ có thể trích rút được từ các trang mạng xã hội, kết hợp với niềm cảm hứng về một dự án khảo sát ý kiến của người tiêu dùng đối với các sản phẩm trong điều kiện phát triển mạnh mẽ của lĩnh vực thương mại điện tử, tôi quyết

định xây dựng đề tài ***"Khai phá dữ liệu từ các mạng xã hội để khảo sát ý kiến của khách hàng đối với một sản phẩm thương mại điện tử".***

2. Mục tiêu và nhiệm vụ nghiên cứu

Nghiên cứu tổng quan về khai phá dữ liệu và các kỹ thuật khai phá dữ liệu.

Nghiên cứu các kĩ thuật phân loại văn bản tiếng Việt.

Nghiên cứu các kĩ thuật tách từ tiếng Việt.

Nghiên cứu các phương pháp phân loại ý kiến đã và đang được phát triển ngày nay.

Nghiên cứu phương pháp phân loại ý kiến dựa vào phân lớp văn bản, áp dụng kĩ thuật máy học vector hỗ trợ SVM.

Xây dựng một công cụ mà với đầu vào là tập hợp các ý kiến nhận xét của người dùng về một sản phẩm thương mại điện tử được trích rút từ các trang mạng xã hội thì đầu ra sẽ là thống kê ý kiến phản hồi của người dùng về sản phẩm đó, từ đó biết được số lượng ý kiến tích cực, tiêu cực và chưa xác định.

3. Đối tượng và phạm vi nghiên cứu

– Đối tượng nghiên cứu: các nhận xét của người dùng về một sản phẩm thương mại điện tử trên các trang mạng xã hội như facebook, twister, yahoo...

– Phạm vi nghiên cứu

❖ Về lý thuyết:

o Cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên, trí tuệ nhân tạo.

o Tìm hiểu tổng quan về các kĩ thuật khai phá dữ liệu.

o Tìm hiểu tổng quan về các kĩ thuật phân loại văn bản tiếng Việt.

o Tìm hiểu tổng quan về các kĩ thuật tách từ tiếng Việt.

o Tìm hiểu tổng quan về các kĩ thuật các phương pháp phân loại ý kiến hiện nay.

❖ Về mặt thực nghiệm:

o Trình bày và ứng dụng phương pháp phân loại SVM để phân loại ý kiến của khách hàng đối với một sản phẩm thương mại điện tử**.** Áp dụng trên miền sản phẩm điện thoại Iphone5.

o Chỉ xử lý đối với văn bản tiếng Việt có dấu.

o Có nhiều tiêu chí để phân loại ý kiến, trong đề tài tôi chỉ xét ba tiêu chi cơ bản đó là tích cực, tiêu cực và không xác định.

4. Phương pháp nghiên cứu

– Tìm hiểu các các kĩ thuật khai phá dữ liệu.

– Tìm hiểu các kĩ thuật phân loại văn bản tiếng Việt.

– Tìm hiểu các kỹ thuật tách từ tiếng Việt

– Tìm hiểu các phương pháp phân loại ý kiến hiện nay.

– Phân tích thiết kế hệ thống chương trình ứng dụng.

– Xây dựng kho dữ liệu huấn luyện thể hiện quan điểm của người dùng đối với một sản phẩm thương mại điện tử, mà trong phạm vi đề tài là sản phẩm điện thoại Iphone5 của hãng Apple.

5. Ý nghĩa khoa học và thực tiễn

– *Ý nghĩa khoa học:* Nghiên cứu và tìm hiểu các kĩ thuật trích rút thông tin, xử lý ngôn ngữ tự nhiên, xử lý văn bản tiếng việt và các phương pháp phân loại ý kiến hiện nay.Tạo tiền đề cho những nghiên cứu tiếp theo trong tương lai.

– *Ý nghĩa thực tiễn:* Xây dựng giải pháp cơ bản về khảo sát ý kiến của khách hàng đối với một sản phẩm thương mại điện tử.

6. Cấu trúc luận văn

Ngoài phần mở đầu và kết luận, luận văn gồm có 3 chương:

Chương 1 Tổng quan về các phương pháp khai phá dữ liệu: chương này trình bày lý thuyết về khai phá dữ liệu và các kĩ thuật khai phá dữ liệu.

Chương 2 Các phương pháp khảo sát ý kiến của khách hàng đối với một sản phẩm thương mại điện tử: trong chương này trình bày các phương pháp khảo sát ý kiến khách hàng hiện nay, các vấn đề liên quan đến phân loại ý kiến, hướng tiếp cận bài toán phân loại ý kiến và các giải pháp phân loại ý kiến hiện nay.

Chương 3 Đề xuất giải pháp và xây dựng chương trình thực nghiệm: chương này trình bày phương pháp phân loại SVM và áp dụng vào bài toán phân loại ý kiến khách hàng đối với một sản phẩm thương mại điện tử. Sau đó trình bày đề xuất hướng cải tiến bài toán hiệu quả hơn.

CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

Chương này trình bày tổng quan về khai phá dữ liệu, quá trình khai phá dữ liệu, và các phương pháp và kĩ thuật dùng trong khai phá dữ liệu như các kĩ thuật phân lớp, các kĩ thuật tách từ tiếng Việt hiện nay.

1.1. VÀI NÉT KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU

1.1.1. Khái niệm khai phá dữ liệu

Khai phá dữ liệu là một lĩnh vực khoa học mới xuất hiện, nhằm tự động hóa khai thác những thông tin, tri thức hữu ích, tiềm ẩn trong các CSDL cho các tổ chức, doanh nghiệp,... từ đó thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh của tổ chức, doanh nghiệp này.

Khai phá dữ liệu là quá trình tìm kiếm, phát hiện các tri thức mới, hữu ích tiềm ẩn trong cơ sở dữ liệu lớn.

1.1.2. Các bước khai phá dữ liệu

1.2. MỘT SỐ NGHIÊN CỨU GẦN ĐÂY VỀ KHAI PHÁ DỮ LIỆU

– Khai phá dữ liệu website bằng kĩ thuật phân cụm.

– Lựa chọn thuộc tính trong khai phá dữ liệu.

– Nghiên cứu ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét ngữ nghĩa.

– Phân loại văn bản tiếng Việt với bộ vector hỗ trợ và cây quyết định.

– Ứng dụng khai phá dữ liệu để tư vấn học tập.

1.3. CÁC KĨ THUẬT KHAI PHÁ VÀ XỬ LÝ DỮ LIỆU HIỆN NAY.

1.3.1. Các kĩ thuật khai phá dữ liệu

– Đứng trên quan điểm của học máy, thì các kỹ thuật trong KPDL bao gồm:

❖ Học có giám sát

❖ Học không có giám sát

❖ Học nửa giám sát

– Nếu căn cứ vào lớp các bài toán cần giải quyết, thì KPDL bao gồm các kỹ thuật áp dụng sau:

❖ Phân lớp và dự đoán

❖ Phân cụm

❖ Luật kết hợp

❖ Phân tích hồi quy

❖ Phân tích các mẫu theo thời gian

❖ Mô tả khái niệm

1.3.2. So sánh khai phá dữ liệu với các phương pháp khác

1.3.3. Các phương pháp phân lớp văn bản

a. Support Vector Machine (SVM)

SVM là phương pháp tiếp cận phân loại rất hiệu quả được Vapnik giới thiệu năm 1995 để giải quyết vấn đề nhận dạng mẫu 2 lớp sử dụng nguyên lý cực tiểu hóa rủi ro có cấu trúc (Structural Risk Minimization).

Ý tưởng của thuật toán bắt đầu từ việc cho trước một tập huấn luyện được biểu diễn trong không gian vector trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu mặt phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp + và lớp –. Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm được khoảng cách biên lớn nhất.

b. K–Nearest Neighbor (kNN)

kNN là phương pháp truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trong nhận dạng mẫu hơn bốn thập kỷ qua. kNN được đánh giá là một trong những phương pháp tốt nhất (áp dụng trên tập dữ liệu Reuters phiên bản 21450), được sử dụng từ những thời kỳ đầu của việc phân loại văn bản.

c. Naïve Bayes (NB)

NB là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961 sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm, các bộ lọc mail...

d. Neural Network (NNet)

Nnet được nghiên cứu mạnh trong hướng trí tuệ nhân tạo. Wiener là người đã sử dụng Nnet để phân loại văn bản, sử dụng 2 hướng tiếp cận: kiến trúc phẳng (không sử dụng lớp ẩn) và mạng nơron 3 lớp (bao gồm một lớp ẩn) Cả hai hệ thống trên đều sử dụng một mạng nơron riêng rẽ cho từng chủ đề, NNet học cách ánh xạ phi tuyến tính những yếu tố đầu vào như từ, hay mô hình vector của một văn bản vào một chủ đề cụ thể.

e. Linear Least Square Fit (LLSF)

LLSF là một cách tiếp cận ánh xạ được phát triển bởi Yang và Chute vào năm1992. Đầu tiên, LLSF được Yang và Chute thử nghiệmtrong lĩnh vực xác định từ đồng nghĩa sau đó sử dụng trong phân loại vào năm 1994. Các thử nghiệm của Ỵang cho thấy hiệu suất phân loại của LLSF có thể ngang bằng với phương pháp kNN kinh điển.

f. Centroid- based vector

Là một phương pháp phân loại đơn giản, dễ cài đặt và tốc độ nhanh do có độ phức tạp tuyến tính O(n) .

Mỗi lớp trong dữ liệu huấn luyện sẽ được biểu diễn bởi một vector trọng tâm. Việc xác định lớp của một văn bản thử bất kì sẽ thông qua viêc tìm vector trọng tâm nào gần với vector biểu diễn văn bản thử nhất. Lớp của văn bản thử chính là lớp mà vector trọng tâm đại diện. Khoảng cách được tính theo độ đo cosine.

1.3.4. Nhận xét về các phương pháp phân lớp văn bản

1.3.5. Một số phương pháp tách từ tiếng Việt hiện nay

a. Phương pháp Maximum Matching

b. Phương pháp giải thuật học cải biến

c. Mô hình tách từ bằng WFST và mạng Neural

d. Phương pháp quy hoạch động

e. Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật toán di truyền

1.3.6. Đánh giá các phương pháp tách từ tiếng Việt hiện nay

CHƯƠNG 2
CÁC PHƯƠNG PHÁP KHẢO SÁT VÀ PHÂN LOẠI Ý KIẾN CỦA KHÁCH HÀNG ĐỐI VỚI MỘT SẢN PHẨM TMĐT

Chương 2 tập trung trình bày các vấn đề liên quan đến các phương pháp khảo sát và phân loại ý kiến của khách hang đối với một sản phẩm TMĐT như tìm hiểu về khái niệm sản phẩm TMĐT, sự cần thiết của việc lấy ý kiến khách hàng, các phương pháp khảo sát ý kiến khách hàng hiện nay, các vấn đề liên quan đến phân loại ý kiến, hướng tiếp cận bài toán phân loại ý kiến và cuối chương là trình bày một số phương pháp phân loại ý kiến hiện nay.

2.1. TÌM HIỂU CHUNG VỀ SẢN PHẨM TMĐT

Trước khi đi vào tìm hiểu về các phương pháp khảo sát và phân loại ý kiến của khách hàng đối với một sản phẩm thương mại điện tử, chúng ta nên tìm hiểu về lĩnh vực thương mại điện tử và các sản phẩm thương mại điện tử ngày nay. Vì thương mại điện tử và sản phẩm thương mại điện tử sẽ là môi trường và đối tượng cần thiết để từ đó chúng ta có thể khai thác được kho dữ liệu khổng lồ về ý kiến

của người dùng nhằm phục vụ cho mục đích nghiên cứu trong luận văn.

2.1.1. Khái niệm về thương mại điện tử

Thương mại điện tử, hay còn gọi là e-commerce, e-comm hay EC, là sự mua bán sản phẩm hay dịch vụ trên các hệ thống điện tử như internet và các mạng máy tính.

Thương mại điện tử thông thường được xem ở các khía cạnh của kinh doanh điện tử (e-business). Nó cũng bao gồm việc trao đổi dữ liệu tạo điều kiện thuận lợi cho các nguồn tài chính và các khía cạnh thanh toán của việc giao dịch kinh doanh.

2.1.2. Sự hình thành thương mại điện tử

Về nguồn gốc, thương mại điện tử được xem như là điều kiện thuận lợi của các giao dịch thương mại điện tử, sử dụng công nghệ như EDI và EFT. Cả hai công nghệ này đều được giới thiệu thập niên 70, cho phép các doanh nghiệp gửi các hợp đồng điện tử như đơn đặt hàng hay hóa đơn điện tử. Sự phát triển và chấp nhận của thẻ tín dụng, máy rút tiền tự động (ATM) và ngân hàng điện thoại vào thập niên 80 cũng đã hình thành nên thương mại điện tử. Một dạng thương mại điện tử khác là hệ thống đặt vé máy bay bởi Sabre ở Mỹ và Travicom ở Anh.

Vào thập niên 90, thương mại điện tử bao gồm các hệ thống hoạch định tài nguyên doanh nghiệp (ERP), khai thác dữ liệu và kho dữ liệu.

Năm 1990, internet ra đời, con người bắt đầu có mối liên hệ với từ "ecommerce" với quyền trao đổi các loại hàng hóa khác nhau thông qua internet dùng các giao thức bảo mật và dịch vụ thanh toán điện tử.

2.1.3. Sản phẩm thương mại điện tử

Sản phẩm thương mại điện tử là các sản phẩm được buôn bán, giao dịch trong môi trường thương mại điện tử.

2.2. VÌ SAO PHẢI LẤY Ý KIẾN KHÁCH HÀNG

Khảo sát ý kiến của khách hàng là một cách tuyệt vời để tìm hiểu xem khách hàng của chúng ta cảm thấy như thế nào về sản phẩm mới, dịch vụ, địa điểm, chính sách hoặc bất cứ điều gì quan trọng đối với công việc kinh doanh của chúng ta.

Thông qua cuộc khảo sát chúng ta sẽ biết được những điều khách hàng đang mong đợi, và từ đó có những định hướng chuyển biến phù hợp trong hoạt động sản xuất kinh doanh.

2.3. CÁC PHƯƠNG PHÁP KHẢO SÁT Ý KIẾN KHÁCH HÀNG

2.3.1. Khảo sát ý kiến khách hàng bằng các phương pháp thủ công

Trong lĩnh vực điều tra khảo sát ý kiến khách hàng, có nhiều phương pháp giúp người thu thập thông tin có được cái nhìn toàn diện nhất về cuộc khảo sát của mình, các phương pháp này có thể quy về 2 phương pháp chính đó là phương pháp phỏng vấn và phương pháp dùng phiếu thăm dò ý kiến khách hàng.

a. Phương pháp phỏng vấn

b. Dùng phiếu thăm dò ý kiến khách hàng

c. Các phương pháp khác

2.3.2. Khảo sát ý kiến khách hàng bằng phương pháp tự động

a. Sự cần thiết của việc khảo sát ý kiến khách hàng theo hướng tự động

b. Các công trình nghiên cứu và ứng dụng khảo sát ý kiến

của khách hàng

– Khai phá quan điểm trên dữ liệu twister

– Phát hiện cộng đồng sử dụng thuật toán CONGA và khai phá quan điểm cộng đồng trên mạng xã hội.

– Dự báo thị trường chứng khoán dựa trên khai phá dữ liệu Twitter.

– Khai phá quan điểm của các Blog để dự đoán việc bán sản phẩm.

2.4. CÁC VẤN ĐỀ LIÊN QUAN ĐẾN PHÂN LOẠI Ý KIẾN

2.4.1. Khái quát về phân loại ý kiến

– Phân loại ý kiến đang là một lĩnh vực mới và hiện đang thu hút được sự quan tâm bởi nhiều nhà khoa học, các nhà sản xuất và rất nhiều công ty doanh nghiệp. Việc phân loại ý kiến có ý nghĩa rất quan trọng trong việc nhìn nhận quyết định một vấn đề.

– Phân loại ý kiến áp dụng nhiều kết quả nghiên cứu của lĩnh vực xử lý ngôn ngữ tự nhiên, học máy và khai phá văn bản.

– Phân loại ý kiến bắt đầu bằng việc xác định các từ thể hiện quan điểm như "tốt", "xấu", "tuyệt vời"..., từ đó xác định xu hướng quan điểm của một từ, một cụm từ, một câu, một đoạn văn bản, hoặc một đặc trưng.

2.4.2. Các khái niệm thường dùng trong phân loại ý kiến

2.4.3. Các bài toán trong phân loại ý kiến

Phân loại ý kiến còn gọi là khai phá quan điểm hay phân lớp nhận định, nó có ba bài toán điển hình đó là:

– Phân lớp ý kiến.

– Khai phá và tổng hợp quan điểm dựa trên đặc trưng.

– Khai phá quan hệ (so sánh).

2.5. HƯỚNG TIẾP CẬN BÀI TOÁN PHÂN LOẠI Ý KIẾN

2.5.1. Xu hướng của các nghiên cứu gần đây về phân loại ý kiến

a. Xác định từ, cụm từ thể hiện quan điểm

b. Xác định chiều hướng của từ, cụm từ thể hiện quan điểm

c. Phân lớp câu, tài liệu chỉ quan điểm

2.5.2. Những thách thức của bài toán phân loại ý kiến hiện nay

Những vấn đề thách thức chính trong đánh giá quan điểm còn tồn tại trong việc sử dụng các từ loại, việc xây dựng các từ ngữ chỉ quan điểm, sự nhập nhằng trong câu phủ định, mức độ của tình cảm (như excellent thì hơn good), các câu hay văn bản phức tạp, từ ngữ trong văn cảnh khác nhau,…

a. Các từ loại khác

b. Thuật ngữ chỉ quan điểm

c. Tính phủ định

d. Cấp độ quan điểm

e. Sự phức tạp của câu, tài liệu

f. Quan điểm theo ngữ cảnh

g. Tài liệu không đồng nhất

2.6. MỘT SỐ PHƯƠNG PHÁP PHÂN LOẠI Ý KIẾN

Hiện nay có ba phương pháp phân loại ý kiến được sử dụng phổ biến đó là: Phân loại ý kiến dựa vào cụm từ thể hiện quan điểm, phân loại ý kiến dựa vào hàm tính điểm số, và phân loại ý kiến dựa vào phương pháp phân lớp văn bản.

2.6.1. Phân loại ý kiến dựa vào cụm từ thể hiện quan điểm

Phương pháp phân loại dựa vào từ thể hiện quan điểm tích cực hay tiêu cực trong mỗi văn bản đánh giá. Thuật toán này sử dụng kỹ thuật xử lí ngôn ngữ tự nhiên gọi là gán nhãn từ loại (part-of-speech). Đánh dấu cho một từ được xác định bởi cú pháp ngữ nghĩa của nó.

2.6.2. Phân loại ý kiến dựa vào hàm tính điểm số

Phương pháp này sẽ dựa vào các từ thể hiện quan điểm để tính điểm số cho từng văn bản, sau đó dựa vào điểm số này để xác định văn bản cần phân loại thuộc lớp nào.

2.6.3. Phân loại ý kiến dựa vào phương pháp phân lớp văn bản

– Đây là phương pháp đơn giản nhất để giải quyết các bài toán phân lớp quan điểm dựa vào chủ đề. Sau đó, có thể áp dụng bất kì kỹ thuật học máy nào để phân lớp như Bayesian, SVM, KNN…

– Ý tưởng chính của phương pháp là đưa bài toán phân loại ý kiến về bài toán phân lớp văn bản để giải quyết. Khi đó mỗi ý kiến được xem như là một văn bản. Ý kiến được chia làm nhiều loại, mỗi loại ý kiến xem như là một chủ đề.

CHƯƠNG 3
ĐỀ XUẤT GIẢI PHÁP VÀ XÂY DỰNG CHƯƠNG TRÌNH THỰC NGHIỆM

Sau khi tìm hiểu tổng quan về lĩnh vực khai phá dữ liệu ở chương 1 và đến chương 2 luận văn cũng đã phân tích và trình bày một cách cơ bản về các phương pháp khảo sát ý kiến của khách hàng, hướng tiếp cận bài toán phân loại ý kiến và các phương pháp

phân loại ý kiến hiện nay. Từ những tiền đề lý thuyết ban đầu ấy, tôi quyết định chọn phương pháp phân loại ý kiến dựa vào phân loại văn bản để giải quyết bài toán phân loại ý kiến khách hàng vì tính đơn giản và hiệu quả cao của nó. Hiện nay có rất nhiều các phương pháp phân loại văn bản, vấn đề đặt ra là phải tìm được phương pháp phân loại tối ưu nhất có thể áp dụng vào bài toán hiện tại. Bằng việc so sánh đối chiếu các phương pháp phân lớp văn bản đã nêu ở cuối chương 1, tôi nhận thấy SVM là phương pháp phân loại phù hợp nhất, cho nên việc áp dụng nó để giải quyết bài toán phân loại ý kiến của khách hàng đối với một sản phẩm thương mại điện tử sẽ cho chúng ta những kết quả nghiên cứu khả quan cả về mặt lý thuyết và thực nghiệm.

3.1. PHƯƠNG PHÁP MÁY HỌC VECTOR HỖ TRỢ SVM

3.1.1. Lý do sử dụng SVM

– SVM có khả năng phân loại khá tốt trong bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...

– SVM có nhiều đặc tính nổi bật trong cả lý thuyết và thực thi so với các phương pháp khác trong lĩnh vực phân lớp văn bản. Ưu điểm chính của SVM so với các phương pháp khác là cách giải quyết vấn đề mang tính tổng quát trong khi các phương pháp khác có thể mang tính cục bộ.

– Việc mở rộng nghiên cứu SVM trên các ứng dụng khác nhau cho thấy tính nhất quán giữa lý thuyết và thực hành tạo nên tính thuyết phục cho phương pháp SVM

3.1.2. Phương pháp SVM

a. Định nghĩa

– Máy học vectơ hỗ trợ (SVM - viết tắt tên tiếng Anh

support vector machine) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy.

– SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó SVM là một thuật toán phân loại nhị phân. Với một bộ các ví dụ luyện tập thuộc hai thể loại cho trước, thuật toán luyện tập SVM xây dựng một mô hình SVM để phân loại các ví dụ khác vào hai thể loại đó.

b. Nguồn gốc ra đời của thuật toán

– Thuật toán SVM ban đầu được tìm ra bởi Vladimir N. Vapnik và dạng chuẩn hiện nay sử dụng lề mềm được tìm ra bởi Vapnik và Corinna Cortes năm 1995.

c. Thuật toán SVM

d. Huấn luyện SVM

3.2. MÔ TẢ BÀI TOÁN

– Yêu cầu của bài toán đặt ra là chúng ta phải khai phá dữ liệu từ các trang mạng xã hội để lấy tất cả các ý kiến phản hồi, các bình luận của người dùng về một sản phẩm thương mại điện tử, từ đó phân tích các quan điểm ấy dựa vào tập tiêu chí để xem quan điểm nào thuộc tiêu chí nào.

– Trong phạm vi của đề tài việc khai phá dữ liệu từ mạng xã hội để lấy về tất cả các ý kiến bình luận đã được thực hiện bởi học viên Trần Thị Ái Quỳnh, là học viên cùng nhóm do giảng viên T.S Huỳnh Công Pháp hướng dẫn. Cho nên từ dữ liệu đầu vào đã có ấy tôi sẽ đi phân loại các ý kiến dựa vào tập tiêu chí đưa ra.

– Có nhiều tiêu chí để đánh giá ý kiến của khách hàng nhưng trong phạm vi của đề tài tôi xin đưa ra 3 tiêu chí: tích cực, tiêu cực, không xác định.

– Về sản phẩm thương mại điện tử thì rất là đa dạng, trong đề tài tôi chỉ chú trọng đến dòng sản phẩm được nhiều người ưa thích và chú ý đó là dòng điện thoại Iphone5 của Apple.

– Có nhiều phương pháp để phân loại các ý kiến của khách hàng vào từng nhóm tiêu chí cụ thể, tôi quyết định chọn phương pháp phương pháp phân loại sử dụng máy học vector hỗ trợ SVM để nghiên cứu và xây dựng ứng dụng mô tả cho lý thuyết nghiên cứu.

3.3. ĐỀ XUẤT GIẢI PHÁP

3.3.1. Giải pháp cho bài toán

– Giải pháp: Việc khảo sát ý kiến của khách hàng đối với một sản phẩm thương mại điện tử chính là việc đi giải bài toán phân loại ý kiến của khách hàng. Với đầu vào của bài toán là dữ liệu đã khai phá được từ mạng xã hội về sản phẩm thương mại điện tử IPhone5, ta giải bài toán phân loại ý kiến của khách hàng dựa vào phương pháp phân lớp văn bản, sử dụng phương pháp máy học vector hỗ trợ SVM.

– Một số cải tiến:

o Giảm kích thước của không gian đặc trưng đến miền các từ vựng chỉ quan điểm bằng cách:

▪ Từ từ điển Tiếng Việt đã có ta trích xuất tất cả các từ vựng thể hiện quan điểm có thể xảy ra trong ngữ cảnh người dùng sử dụng nó để nhận xét về sản phẩm điện thoại Iphone5.

▪ Bổ sung thêm các từ ngữ, thuật ngữ thể hiện quan điểm.

o Cho phép học có giám sát: Điều này có nghĩa là hệ thống phải cho phép người dùng can thiệp để kiểm tra độ chính xác của quá trình phân loại, từ đó có thể ghi nhận và thay đổi các ý kiến phân loại không đúng nhằm làm cho hệ thống tối ưu hơn.

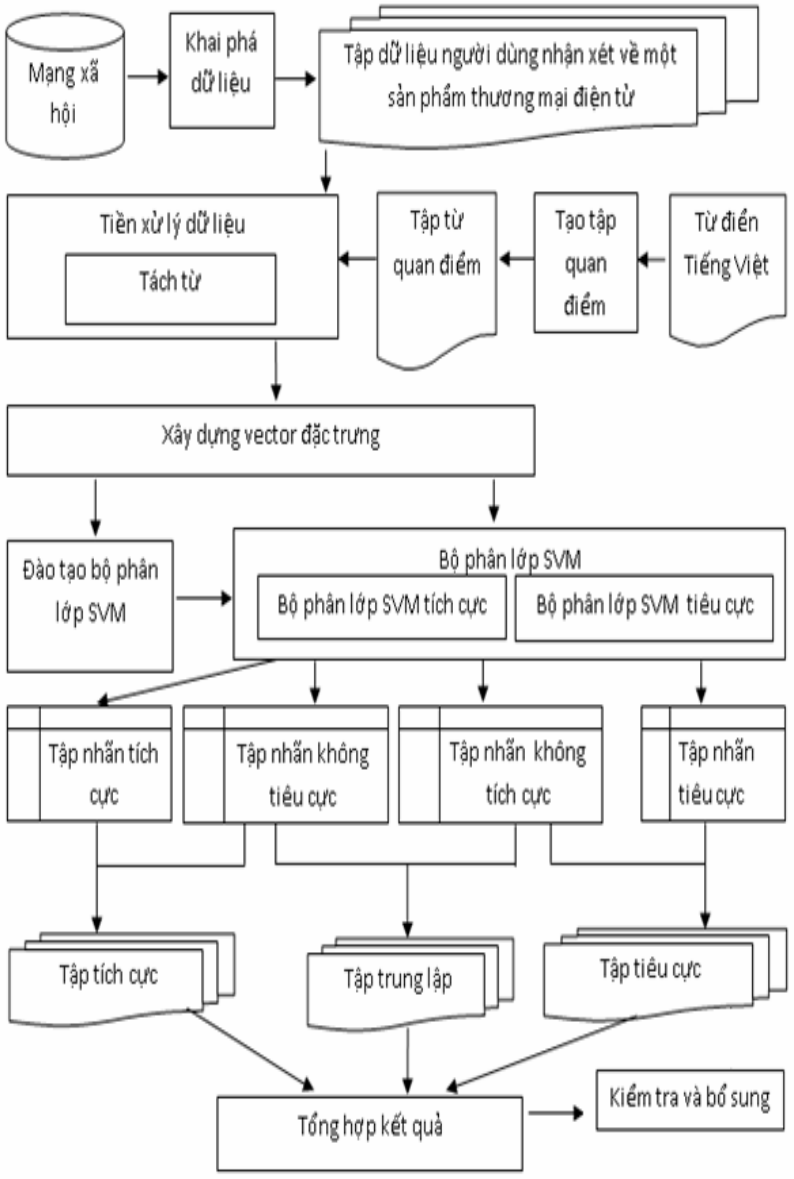
o Tối ưu hóa ngữ nghĩa: Những từ ngữ thể hiện quan điểm nhập nhằn và chưa có trong dữ liệu học nhưng được phát hiện sau quá trình phân loại thuộc nhóm ý kiến chưa xác định, hệ thống cần tự động cảnh báo và cho phép người dùng bổ sung vào tập dữ liệu huấn luyện.

o Xử lý vấn đề tính phủ định: Tính phủ định của từ cũng ảnh hưởng đến ngữ nghĩa của nó, đối với những từ ngữ thể hiện quan điểm điều này càng được thể hiện rõ nét hơn.

– Đầu vào: dữ liệu bình luận đánh giá của người dùng về một sản phẩm thương mại điện tử trên các trang mạng xã hội được trích rút lấy về. Sau đó dữ liệu này được xử lý loại bỏ thông tin dư thừa và làm đầu vào cho bộ phân lớp SVM.

– Đầu ra: là ba tập dữ liệu gồm tập các ý kiến tích cực, tập các ý kiến tiêu cực và tập các ý kiến trung lập.

3.3.2. Mô hình giải pháp



Mô hình phân loại ý kiến sử dụng kĩ thuật SVM

3.3.3. Các bước thực hiện

a. Bước 1: Khai phá và thu thập dữ liệu

– Việc trích rút thông tin từ các trang mạng xã hội được thực hiện bằng cách sử dụng các wrapper. Một wrapper có thể được xem như là một thủ tục được thiết kế để có thể rút trích các thông tin cần quan tâm. Đã có nhiều công trình khác nhau trên thế giới sử dụng nhiều phương pháp tạo wrapper khác nhau để thực hiện trích rút thông tin trên web. Các wrapper này được xây dựng bằng tay hoặc phát sinh tự động trên các vùng thông tin người dùng xác định trước trên các trang web mẫu. Wrapper xây dựng theo phương pháp này có nhược điểm là phải cập nhật lại khi có sự thay đổi về quy cách trình bày trên trang web.

– Dữ liệu đầu vào sau khi được khai phá từ các trang mạng xã hôi, được tổng hợp thành một file text.

– Mỗi dòng của file text này là nội dung của một comment của người dùng bình luận về sản phẩm điện tử Iphone 5.

b. Bước 2: Tạo tập từ vựng quan điểm

– Tập từ vựng quan điểm được trích xuất bằng tay từ từ điển Tiếng Việt, chủ yếu gồm những từ thể hiện quan điểm, tình cảm, thái độ của người dùng trong ngữ cảnh muốn nhận xétvề một sản phẩm điện thoại Iphone5.

– Tập từ này còn được bổ sung từ từ điển quan điểm VietSentiWordNet trên miền dữ liệu Tiếng Việt.

– Tập từ cũng được bổ sung thêm nhiều từ mang quan điểm được dùng thường xuyên trong xu thế bây giờ.

c. Bước 3: Tiền xử lý dữ liệu

– Dữ liệu cần phải được tiền xử lý để loại bỏ các kí tự dư thừa. Do phần lớn các comments được người dùng viết nhanh, vội,

vắn tắt nên thường nhập nhằng và có nhiều lỗi về cú pháp, lỗi ngữ nghĩa. Chính vì vậy chúng ta cần loại bỏ các kí tự, kí hiệu dư thừa và điều chỉnh ngữ nghĩa của các từ viết tắt để chúng trở nên có nghĩa và tìm thấy được trong từ điển Tiếng Việt.

– Dữ liệu này được chia thành 2 tập: một tập dữ liệu huấn luyện và một tập dữ liệu kiểm tra.

– Sau các bước tiền xử lý thủ công đơn giản ta tiến hành tách câu và tách từ cho mỗi văn bản dữ liệu.

– Sau khi tách từ ta tiến hành các bước tối ưu hóa ngữ nghĩa.

– Tiến hành xử lý các từ có tính phủ định.

d. Bước 4: Xây dựng vector đặc trưng và biểu diễn TFxIDF

– Dữ liệu ta đang xét là tập hợp các văn bản. Sau khi tiền xử lý, tách từ, đối với dữ liệu dùng để huấn luyện ta trích xuất tập từ đặc trưng và xây xựng vector đặc trưng văn bản. Khi đó tập dữ liệu huấn luyện sẽ được biểu diễn như là tập các vector đặc trưng.

– Mỗi từ trong văn bản sẽ được tính trọng số TFxIDF và sẽ được đưa vào vector đặc trưng.Vector đặc trưng này sẽ là đầu vào cho quá trình huấn luyện SVM ở bước tiếp theo.

e. Bước 5: Huấn luyện bộ phân lớp SVM

– Đặc trưng cơ bản quyết định khả năng phân loại của một bộ phân loại là hiệu suất tổng quát hóa, hay là khả năng phân loại những dữ liệu mới dựa vào những tri thức đã tích lũy được trong quá trình huấn luyện.

– Thuật toán huấn luyện được đánh giá là tốt nếu sau quá trình huấn luyện, hiệu suất tổng quát hóa của bộ phân loại nhận được

cao. Hiệu suất tổng quát hóa phụ thuộc vào hai tham số là sai số huấn luyện và năng lực của máy học. Trong đó sai số huấn luyện là tỷ lệ lỗi phân loại trên tập dữ liệu huấn luyện. Còn năng lực của máy học được xác định bằng kích thước Vapnik- Chervonenkis (kích thước VC).

f. Bước 6: Phân lớp dữ liệu

– Dữ liệu đầu vào sau khi được xử lý và biểu diễn dưới dạng các vector đặc trưng, sẽ được đưa qua bộ phân lớp đã được tạo ra ở bước trước để tính F.

– Từ giá trị max của F ta tìm được loại ý kiến tương ứng.

g. Bước 7: Tổng hợp kết quả

– Tổng hợp số ý kiến tích cực, tiêu cực và không xác định.

– Kết quả tổng hợp được chính là mục tiêu cần khảo sát

h. Bước 8: Kiểm tra và bổ sung từ quan điểm

– Nếu số ý kiến chưa xác định tồn tại thì hệ thống sẽ cho phép ta xem, kiểm tra và bổ sung các từ vựng mang quan điểm cần thiết vào tập từ vựng quan điểm và bổ sung ý kiến này vào tập dữ liệu học với nhãn chính xác.

– Công việc này sẽ giúp cho quá trình phân loại sẽ ngày càng chính xác hơn, hoàn thiện hơn.

3.3.4. Cài đặt và thử nghiệm

a. Công cụ

– Visual Studial 2010

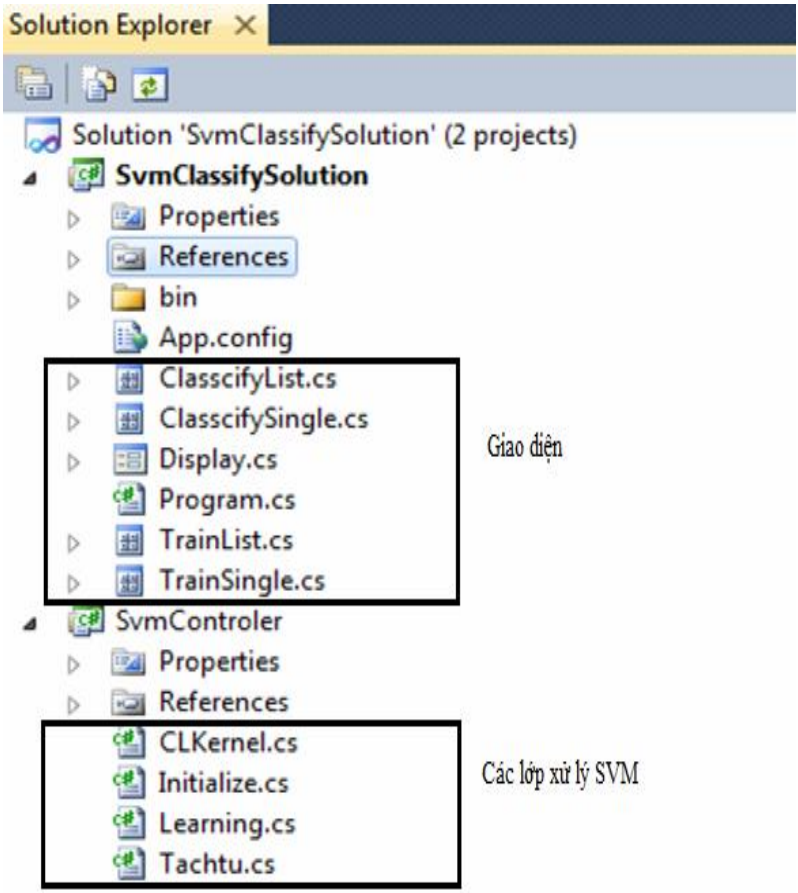
– Sql Server 2008

b. Tổ chức chương trình

– Solution gồm 2 project:

❖ SvmClassifySolution : chứa các lớp xử lý hiển thị kết quả

❖ SvmControler : Chứa các lớp xử lý SVM



Tổ chức chương trình

c. Chức năng hệ thống

– Cập nhật dữ liệu huấn luyện

❖Cập nhật một nội dung ý kiến cần huấn luyện

❖Cập nhật nhiều nội dung huấn luyện từ file text

– Phân loại ý kiến

❖Phân loại một ý kiến

❖Phân loại nhiều ý kiến từ file text

3.3.5. Kết quả thực nghiệm

– Sau khi xây dựng công cụ minh họa cho lý thuyết nghiên cứu. Để tiến hành kiểm chứng kết quả của việc áp dụng phương pháp SVM vào phân loại ý kiến khách hàng, tôi đã sử dụng dữ liệu mẫu gồm 976 nhận xét của khách hàng về sản điện thoại Iphone5 để làm dữ liệu huấn luyện. Dữ liệu này được gán nhãn bằng tay với 488 nhãn tích cực và 488 nhãn tiêu cực.

– Tập dữ liệu kiểm tra gồm 468 nhận xét của khách hàng về sản phẩm điện thoại Iphone5. Sau khi tiến hành kiểm thử tôi thu được kết quả gồm 242 ý kiến tích cực và 190 ý kiến tiêu cực và 36 ý kiến chưa xác định.

– Để kiểm chứng độ chính xác của phương pháp tôi tiến hành kiểm thử nó trên tập dữ liệu huấn luyện thu được 392 ý kiến tích cực và 584 ý kiến tiêu cực.

– Như vậy độ chính xác của thuật toán là 80,3%.

– Ta nhận thấy các ý kiến phân loại không đúng là do trong tập đặc trưng có nhiều từ không dùng để thể hiện quan điểm, sự xuất hiện của những từ này trong quá trình tính toán trọng số sẽ ảnh hưởng đến độ chính xác của việc phân loại. Trong khi đó việc phân loại một ý kiến hầu hết đều thể hiện qua từ hoặc nhóm từ thể hiện quan điểm trong ý kiến đó. Bên cạnh đó việc phân loại ý kiến không đúng còn do sự phức tạp về ngữ nghĩa của câu. Các ý kiến không xác định là các ý kiến mà không phải tích cực cũng không phải tiêu cực. Các ý kiến này cần xem xét và quay lại bổ sung cho quá trình huấn luyện.

KẾT LUẬN

Những kết quả đạt được của luận văn:

– Trình bày khái quát về các kĩ thuật khai phá dữ liệu.

– Nêu lên các phương pháp phân loại văn bản đặc biệt là phương pháp phân loại sử dụng máy học vector hỗ trợ SVM.

– Trình bày các phương pháp phân loại ý kiến hiện nay, áp dụng phương pháp phân loại văn bản vào bài toán phân loại ý kiến sử dụng phương pháp SVM.

– Đề xuất cải tiến hệ thống nhằm nâng cao tính hiệu quả của việc sử dụng phương pháp SVM vào phân loại ý kiến.

Bên cạnh những kết quả đạt được, dù đã rất cố gắng nhưng do sự hữu hạn về thời gian và kiến thức, luận văn vẫn còn một số hạn chế:

– Hiệu quả phân loại còn phụ thuộc vào sự phức tạp của ngữ nghĩa, nếu ý kiến có ngữ nghĩa phức tạp thì khi áp dụng bài toán phân loại văn bản vào phân loại ý kiến thì hiệu quả sẽ không cao.

– Cấp độ của quan điểm trong ý kiến chỉ còn hạn chế ở hai mức tích cực và tiêu cực.

– Mức phân lớp chỉ dừng lại ở mức tài liệu, chưa sâu đến mức đặc trưng.

Định hướng nghiên cứu trong tương lai:

– Nâng cao hiệu quả phân loại trong trường hợp các ý kiến có ngữ nghĩa phức tạp.

– Cấp độ của quan điểm cần phải cao hơn, không nên chỉ giới hạn ở hai mức là tích cực và tiêu cực.

– Hướng phân lớp đến mức đặc trưng chứ không chỉ dừng lại ở mức tài liệu.