BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

TRẦN THỊ ÁI QUỲNH

ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ TRÍCH RÚT THÔNG TIN THEO CHỦ ĐỀ TỪ CÁC MẠNG XÃ HỘI

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. Hoàng Thị Thanh Hà

Phản biện 2: PGS. TS. Lê Mạnh Thạnh

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 19 tháng 11 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. Tính cấp thiết của đề tài

Trong những năm gần đây, công nghệ thông tin phát triển mạnh mẽ và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực đời sống, kinh tế xã hội đã làm cho lượng dữ liệu tăng lên nhanh chóng từ mức độ terabytes đến mức độ petabytes. Do đó, việc khai thác và chọn lọc những dữ liệu có ích từ lượng dữ liệu khổng lồ đó là việc cần thiết, đóng vai trò quyết định trong mọi hoạt động. Hiện nay, mạng xã hội có đa dạng người sử dụng, ở đó họ chia sẻ ý kiến về nhiều chủ đề khác nhau, do đó nó là nguồn dữ liệu có giá trị. Chúng ta cũng biết việc trích lọc được các ý kiến của người dùng có sức ảnh hưởng mang lại nhiều lợi ích thiết thực như mang đến những cơ hội kinh doanh, các ý kiến về các mặt hàng mà họ đã mua, tốt xấu…, có ảnh hưởng đến các cuộc bỏ phiếu chính trị, cũng như ảnh hưởng đến các cuộc thảo luận mang tính xã hội,….

Hơn một thập niên trở lại đây, khai phá dữ liệu (KPDL) đã trở thành một trong những hướng nghiên cứu quan trọng trong lĩnh vực khoa học máy tính và công nghệ tri thức. Hàng loạt nghiên cứu, đề xuất ra đời đã được thử nghiệm và ứng dụng thành công vào đời sống cùng với lịch sử cho của nó thấy rằng KPDL là một lĩnh vực nghiên cứu ổn định, có một nền tảng lý thuyết vững chắc. Ngày nay, với sự phát triển internet và nhu cầu đưa thông tin lên mạng, các trang web với dữ liệu fulltex đã trở nên phổ biến. Cùng với các kỹ thuật khai phá dữ liệu nói chung, các kỹ thuật khai phá web cũng rất được quan tâm nhằm chắt lọc, trích rút thông tin phục vụ cho một mục đích ứng dụng nào đó là rất cần thiết. Mặt khác, với mục tiêu tạo môi trường giao lưu, chia sẻ thông tin đa dạng, phong phú. Vì vậy, đề tài " *Ứng dụng khai phá dữ liệu để trích rút thông tin theo chủ đề từ các trang mạng xã hội*" là cần thiết và có ý nghĩa về mặt lý thuyết và thực tiễn.

2. Mục đích nghiên cứu

– Nghiên cứu các phương pháp, kỹ thuật khai phá văn bản.

– Nghiên cứu phương pháp tách từ, phân loại văn bản Tiếng Việt.

– Nghiên cứu phương pháp lấy dữ liệu của người dùng về các chủ đề trên mạng xã hội Twitter.

– Xây dựng hệ thống phân loại văn bản SVM theo chủ đề từ dữ liệu lấy từ mạng xã hội Twitter.

– Đưa ra định hướng và hướng phát triển đề tài.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

- Kỹ thuật, phương pháp khai phá dữ liệu.
- Phương pháp thu thập dữ liệu từ mạng xã hội.
- Các chủ đề đang được quan tâm trên mạng xã hội Twitter.

Phạm vi nghiên cứu

Ứng dụng các thuật toán của kỹ thuật rút trích thông tin để xây dựng đưa ra danh sách các ý kiến người dùng về một chủ đề đang được quan tâm trên mạng xã hội Twitter.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu lý luận

Thu thập, đọc hiểu, phân tích thông tin, dữ liệu từ các tài liệu, giáo trình, sách liên quan đến khai phá dữ liệu, rút trích thông tin.

Phương pháp nghiên cứu thực tiễn

- Tiến hành nghiên cứu kỹ thuật rút trích thông tin, ứng dụng các kỹ thuật đó để xây dựng mô hình đưa ra danh sách ý kiến người dùng theo chủ đề trên mạng xã hội.
- So sánh và đánh giá kết quả đạt được để từ đó đề xuất ra hướng phát triển tốt hơn.

5. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học

Với sự phát triển lớn mạng của Internet và lượng người dùng tham gia vào các trang mạng xã hội không ngừng tăng lên như hiện nay thì việc khai thác nguồn dữ liệu từ các trang mạng xã hội để phục vụ cho công việc kinh doanh cũng như các mục đích chính trị xã hội khác nhau đang là một trào lưu được ưu chuộng.

Dữ liệu trên các trang mạng xã hội rất đa dạng và có số lượng rất lớn. Với lượng dữ liệu khổng lồ như thế, làm thế nào để khai thác, chọn lọc dữ liệu có ích từ nguồn dữ liệu khổng lồ đó. Nhu cầu phát triển các kỹ thuật chọn lọc, thu thập, phân tích dữ liệu,trích rút thông tin một cách thông minh và hiệu quả, vì thế, được đặt ra hơn bao giờ hết. Từ đó, các kỹ thuật khai phá dữ liệu giúp tự động phân tích các tập dữ liệu rất lớn để khám phá ra các tri thức cũng như trích rút các mẫu quan trọng là rất cần thiết và có ý nghĩa thực tiễn cao.

Ý nghĩa thực tiễn

Xây dựng công cụ để trích rút thông tin chủ đề, đưa ra được danh sách ý kiến theo chủ đề của người dùng trên mạng xã hội, từ đó thống kê được ý kiến của người dùng về một chủ đề nào đó.

6. Bố cục của luận văn

Nội dung chính của luận văn được chia thành 3 chương với nội dung như sau:

+ Chương 1: Nghiên cứu tổng quan về khai phá dữ liệu.

+ Chương 2: Nghiên cứu phương pháp lấy dữ liệu từ mạng xã hội Twitter và thuật toán CONGA

+ Chương 3: Thử nghiệm và đánh giá

CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. KHÁI NIỆM VÀ QUÁ TRÌNH KHAI PHÁ DỮ LIỆU

1.1.1. Khái niệm khai phá dữ liệu

Khai phá dữ liệu là một lĩnh vực nghiên cứu ổn định, nó ra đời vào khoảng những năm cuối của của thập kỷ 1980.

KPDL là quá trình khảo sát và phân tích một lượng lớn các dữ liệu được lưu trữ trong các CSDL, kho dữ liệu,…để từ đó trích xuất ra các thông tin quan trọng, có giá trị tiềm ẩn bên trong

Khám phá tri thức trong cơ sở dữ liệu (KDD) là mục tiêu chính của KPDL, do vậy hai khái niệm khai phá dữ liệu và KDD được các nhà khoa học xem là tương đương nhau. Thế nhưng, nếu phân chia một cách chi tiết thì khai phá dữ liệu là một bước chính trong quá trình KDD.

1.1.2. Quá trình khai phá dữ liệu

Quá trình khá phá tri thức có thể chia thành 5 bước như sau [10]:

- Trích lọc dữ liệu
- Tiền xử lý dữ liệu
- Biến đổi dữ liệu
- Khai phá dữ liệu
- Đánh giá và biểu diễn tri thức

1.1.3. Những chức năng chính của khai phá dữ liệu

Hai mục tiêu chính của KPDL là mô tả và dự báo.

a. Mô tả và khái niệm

b. Phân tích sự kết hợp

c. Phân lớp và dự báo

d. Phân cụm

e. Phân tích các đối tượng ngoài cuộc

f. Phân tích sự tiến hóa

1.1.4. Các công trình khai phá và xử lý dữ liệu đã được phát triển

- 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 SVM.
- Phân loại văn bản Tiếng Việt với máy học vector hỗ trợ và cây quyết định.
- Phương pháp luật kết hợp và ứng dụng
- Ứng dụng khai phá dữ liệu để tư vấn học tập
- Nghiên cứu ứng dụng phân lớp dữ liệu trong quản lý khách hàng trên mạng
- Dự báo bùng nổ sự kiện trong mạng xã hội
- Phát hiện cộng đồng sử dụng thuật toán CONGA và khai phá quan điểm cộng đồng
- Khai phá quan điểm trên dữ liệu twitter.

1.1.5. Một số thách thức đặt ra cho việc khai phá dữ liệu

❖ Các cơ sở dữ liệu lớn

❖ Số chiều lớn

❖ Thay đổi dữ liệu và tri thức có thể làm cho các mẫu đã phát hiện không còn phù hợp.

❖ Dữ liệu bị thiếu hoặc nhiễu

❖ Quan hệ giữa các trường phức tạp

- ❖ Giao tiếp với người sử dụng và kết hợp với các tri thức đã có.
- ❖ Tích hợp với các hệ thống khác…

1.2. PHƯƠNG PHÁP VÀ KỸ THUẬT KHAI PHÁ DỮ LIỆU

1.2.1. Các kỹ thuật áp dụng trong khai phá dữ liệu

KDD là một lĩnh vực liên ngành, bao gồm: Tổ chức dữ liệu, học máy, trí tuệ nhân tạo và các khoa học khác

a. Theo quan điểm học máy

- Học có giám sát
- Học không có giám sát
- Học nửa giám sát

b. Căn cứ vào lớp các bài toán cần giải quyết

Chia làm 2 nhóm chính:

- Kỹ thuật mô tả
- Kỹ thuật dự đoán

1.2.2. So sánh các kỹ thuật khai phá dữ liệu

1.2.3. So sánh phương pháp khai phá dữ liệu với các phương pháp học máy, phương pháp hệ chuyên gia và phương pháp thống kê

1.3. KHAI PHÁ DỮ LIỆU WEB

1.3.1. Các dạng dữ liệu

1.3.2. Các loại khai phá Web

1.3.3. Một số vấn đề xử lý dữ liệu văn bản

1.4. CÁC PHƯƠNG PHÁP TÁCH TỪ TIẾNG VIỆT HIỆN NAY

1.4.1. Phương pháp Maximum Matching

1.4.2. Phương pháp giải thuật học cải biến (Transformation-based Learning, TBL)

1.4.3. Mô hình tách từ bằng WFST và mạng Neural

1.4.4. Phương pháp quy hoạch động (dynamic programming)

1.4.5. Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật toán di truyền (Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese - IGATEC)

1.4.6. So sánh các phương pháp tách từ Tiếng Việt hiện nay

1.5. KẾT LUẬN CHƯƠNG 1

Chương 1 của luận văn giới thiệu khái quát về khái niệm, quá trình, các kỹ thuật và phương pháp khai phá dữ liệu. Đồng thời, trong chương này tôi đã trình bày các phương pháp phân tách từ Tiếng Việt hiện nay, so sánh các phương pháp này với nhau để chọn ra một phương pháp tốt nhất phù hợp cho bài toán phân loại văn bản SVM theo chủ đề được đề cập ở chương 3 của luận văn.

Chương tiếp theo tôi sẽ giới thiệu về mạng xã hội Twitter, cấu trúc và tính cộng đồng của nó. Đồng thời, tôi sẽ trình bày về phương pháp thu thập dữ liệu từ mạng xã hội Twitter và thuật toán CONGA để phát hiện cộng đồng, các phương pháp phân loại văn bản hiện nay.

CHƯƠNG 2

CÁC PHƯƠNG PHÁP LẤY DỮ LIỆU TỪ MẠNG XÃ HỘI TWITTER VÀ THUẬT TOÁN CONGA

2.1. MẠNG XÃ HỘI TWITTER

2.1.1. Giới thiệu

2.1.2. Cấu trúc mạng xã hội Twitter

2.1.3. Tính cộng đồng trên mạng xã hội

a. Cộng đồng mạng xã hội

Việc phát hiện cộng đồng có ý nghĩa rất quan trọng trong việc xác định các môđun và ranh giới của chúng cho phép ta phân lớp các đỉnh dựa trên cấu trúc vị trí của chúng trong môđun [5].

Mục tiêu của việc phát hiện cộng đồng là từ các mạng xã hội cho trước, phát hiện được các cấu trúc cộng đồng nằm trong đó và tìm hiểu về mối liên hệ bên trong các cộng đồng cũng như giữa các cộng đồng với nhau, mối liên hệ đó có ảnh hưởng thế nào đến cấu trúc của toàn mạng xã hội.

b. Bài toán khai phá quan điểm người dùng mạng xã hội về một chủ đề nào đó

Đầu vào: Quan điểm người dùng về các chủ đề trên mạng xã hội

Đầu ra: Phân lớp các quan điểm theo từng chủ đề

Nghiên cứu các tính chất và trích chọn những thông tin quan trọng từ các cộng đồng trực tuyến như từ các diễn đàn (forums), blogs và mạng xã hội trực tuyến (online social networks) là một trong những hướng thu hút được sự chú ý của cộng đồng khai phá web hiện nay

Bài toán phân lớp quan điểm theo chủ đề nào đó trên mạng xã hội rất được sự quan tâm của con người trong quá trình làm việc với một tập các đối tượng. Chính vì điều này mà giúp cho việc sắp xếp, tìm kiếm các đối tượng một cách nhanh chóng hơn

c. Thuật toán Girvan-Newman

Ý tưởng thuật toán: Thuật toán này dựa trên ý tưởng khi các cộng đồng được gắn kết với nhau thì đường đi giữa cộng đồng này đến cộng đồng khác sẽ đi qua các cạnh nối giữa các cộng đồng với tần suất cao. Mục đích chính của thuật toán là tìm những cạnh nối đó [5].

Thuật toán được thực hiện theo các bước sau:

1. Tính độ đo trung gian cho tất cả các cạnh trong mạng.
2. Hủy bỏ các cạnh có độ trung gian cao nhất.
3. Tính lại độ trung gian cho tất cả các cạnh bị ảnh hưởng theo các cạnh đã loại bỏ.
4. Lặp lại từ bước 2 cho đến khi không còn các cạnh trung gian.

Ưu điểm của thuật toán: Thuật toán khá đơn giản và dễ hiểu. Toàn bộ thuật toán có thể được biểu diễn trong một dendrogram, ở đây ta có thể hiểu là thuật toán đi từ gốc đến các lá. Các nhánh của cây biểu diễn cho các phép loại bỏ cạnh để chia đồ thị thành các cộng đồng riêng rẽ.

Nhược điểm của thuật toán:

Số lượng cộng đồng hoàn toàn không kiểm soát trước được vì thuật toán Girvan-Newman sử dụng phương pháp loại trừ đến khi không có cạnh nào vượt qua ngưỡng của độ trung gian cao nhất.

Khó có thể xác định được phân vùng nào mang lại hiệu quả cao nhất.

Độ phức tạp của thuật toán khá lớn $O(m^2n)$.

Với cách phân chia của Girvan-Newman thì không giải quyết được hiện tượng chồng chéo cộng đồng bởi vì trên thực tế, mỗi đơn vị nút mạng có thể thuộc rất nhiều cộng đồng khác nhau.

Dựa trên những ưu điểm và nhược điểm trên của thuật toán Girvan-Newman, các nhà khoa học đã tìm cách để cải tiến thuật toán trên nhằm khắc phục những nhược điểm của thuật toán Girven-Newman như tìm phép phân vùng tốt nhất, giảm độ phức tạp của thuật toán, giải quyết hiện tượng chồng chéo cộng đồng. Với cách tiếp cận khác nhau, năm 2007 Gregory đề xuất thuật toán CONGA (Cluster Overlap Newman-Girvan Algorithm)

d. Thuật toán CONGA

Thuật toán CONGA được Gregory cải tiến từ thuật toán Girvan-Newman nhằm mục đích giải quyết vấn đề về chồng chéo cộng đồng [16**].**

Ý tưởng thuật toán: Dựa trên ý tưởng thuật toán Girvan-Newman, tác giả đề xuất thêm một ý tưởng mới đó là phép chia các đỉnh thành nhiều phần khác nhau, để một phần của đỉnh được chia đó có thể xuất hiện trong các cộng đồng con.

Tác giả đề ra một độ đo mới, là độ trung gian của phép phân chia, độ đo này cho phép ta có thể xác định được khi nào cần phân chia một đỉnh, thay vì loại bỏ các cạnh, đỉnh nào cần phân chia và phân chia như thế nào.

Thuật toán CONGA chia làm các bước như sau:

- Tính độ trung gian của tất cả các cạnh trong đồ thị

− Tính độ trung gian của các đỉnh trong đồ thị, dựa vào độ trung gian của các cạnh như trong công thức ở trên

− Tìm danh sách các đỉnh mà độ trung gian của đỉnh đó lớn hơn giá trị lớn nhất của các độ trung gian cạnh

− Nếu danh sách ở bước 3 không rỗng, tính các độ trung gian theo cặp của các đỉnh trong danh sách, sau đó xác định phép phân chia tối ưu nhất cho các đỉnh đó

− Thực hiện việc loại bỏ cạnh, hoặc phân chia đỉnh để chia đồ thị thành các thành phần

− Tính lại độ trung gian của các cạnh trong tất cả các thành phần vừa được chia ra

− Lặp lại bước 2 đến khi không còn cạnh nào.

Ưu điểm của thuật toán: Giải quyết được vấn đề chồng chéo cộng đồng bằng cách đặt ra phép phân chia đỉnh, ngoài ra nội dung thuật toán tương đối dễ hiểu và xác định được phép phân chia tối ưu nhất trong các trường hợp.

Nhược điểm của thuật toán: Thời gian tính toán, với độ phức tạp tính toán lên tới $O(m^3)$ với m là số cạnh.

2.2. PHƯƠNG PHÁP THU THẬP DỮ LIỆU VÀ PHÁT HIỆN CỘNG ĐỒNG TỪ MẠNG XÃ HỘI TWITTER

Quá trình thực hiện như sau:

Thu thập dữ liệu: Mạng xã hội Twitter cung cấp một API giúp người sử dụng có thể lấy được các thông tin về các người dùng trong mạng xã hội đó, chính từ nguồn dữ liệu của Twitter được cung cấp qua Twitter API, ta sẽ tiến hành thu thập các dữ liệu về người sử dụng như ID, tên truy cập, và các thông tin cá nhân của người dùng, các bình luận,.... Bộ thư viện mã nguồn mở Twitter4j được thiết kếvới mục đích giúp người sử dụng có khả năng tương tác với Twitter API qua Java và lấy được dữ liệu từ Twitter. Bộ thư viện do Yusuke Yamamoto, một lập trình viên người Nhật và các cộng sự phát triển năm 2009 [17].

Tiền xử lý dữ liệu: Từ dữ liệu thu thập được, tôi tiến hành tiền xử lý như loại bỏ những thông tin người dùng thiếu về thông tin sử dụng, thiếu kết nối với các đỉnh khác trong mạng. Như vậy trong bước này, dữ liệu thu thập về đã được chuẩn hóa phù hợp với mô hình cần xây dựng.

Xây dựng mô hình mạng xã hội: Từ tập dữ liệu đã được chuẩn hóa sẽ tiến hành xây dựng đồ thị mô tả mạng xã hội. Trong đó với các đỉnh là những người sử dụng thu về được và dựa vào danh sách friends và danh sách followers của mỗi người dùng để đưa ra danh sách mối liên kết của các đỉnh đó với nhau. Do thuật toán yêu cầu đầu vào của thuật toán CONGA là đồ thị vô hướng, không có trọng số nên kết quả đầu ra được lưu vào một file.txt, trong đó mỗi hàng sẽ đưa ra một cạnh liên kết trong đồ thị, bao gồm hai đỉnh đầu vào cuối của cạnh đó.

Áp dụng thuật toán CONGA: Từ mạng xã hội vừa xây dựng được ở bước 3, cho qua CONGA để phát hiện cộng đồng mạng xã hội. Dựa trên đồ thị vừa xây dựng được, chúng tôi tiến hành cài đặt thuật toán CONGA cho đồ thị đó, dựa trên bộ thư viện mà tác giả thuật toán cung cấp. Đầu vào của chương trình là tập tin văn bản biểu diễn đồ thị xây dựng được ở bước trên. Đầu ra của chương trình là tập cộng đồng phân cách phân chia mang lại hiệu quả cao nhất.

2.3. CÁC PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN HIỆN NAY

2.3.1. Máy vector hỗ trợ (SVM)

2.3.2. K lân cận (kNN)

2.3.3. Xác suất Naïve Bayes (NB).

2.3.4. Mạng Nơron (NNet)

2.3.5. Tuyến tính bình phương tối thiểu (LLSF)

2.3.6. Vector trọng tâm (Centroid- based vector)

2.3.7. So sánh các phương pháp phân loại văn bản

Các thuật toán phân loại trên từ thuật toán phân loại 2 lớp (SVM) đến các thuật toán phân loại đa lớp (kNN) đều có điểm chung là yêu cầu văn bản phải được biểu diễn dưới dạng vector đặc trưng. Ngoài ra các thuật toán như kNN,NB,LLSF đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất (hơn 10000 chiều) trong khi đó chỉ là 2000 đối với NB, 2415 cho kNN và LLSF, 1000 cho Nnet [6]. Thời gian huấn luyện cũng khác nhau đối với từng phương pháp, Nnet (sử dụng mỗi mạng tương ứng một chủ đề) và SVM là hai phương pháp có thời gian huấn luyện lâu nhất trong khi đó kNN,NB,LLSF và Centroid là các phương pháp có tốc độ (thời gian huấn luyện, phân loại) nhanh và cài đặt dễ dàng.

Về hiệu suất, dựa vào thử nghiệm của Yang trên tập dữ liệu Reuter-21578 với hơn 90 chủ đề và trên 7769 văn bản, ta có thể sắp xếp các phương pháp phân loại văn bản theo thứ tự như sau SVM > kNN >> {LLSF,NB,Nnet}[6]. Tuy nhiên kết quả trên có thể không còn đúng khi áp dụng thử nghiệm phân loại trên Tiếng Việt.

2.4. KẾT LUẬN CHƯƠNG 2

Trong chương 2, khóa luận đã giới thiệu mạng xã hội Twitter, cấu trúc và tính cộng đồng của nó. Chính nhờ vào cấu trúc, tính cộng đồng của mạng xã hội, áp dụng thuật toán CONGA để phát hiện cộng đồng và thu thập dữ liệu từ mạng xã hội Twitter. Đồng thời, trong chương này tôi đã trình bày cụ thể phương pháp thu thập dữ liệu từ mạng xã hội Twitter, các phương pháp phân loại văn bản hiện nay và so sánh các phương pháp này với nhau nhằm để tìm ra phương pháp phân loại văn bản tốt nhất phù hợp cho quá trình thực nghiệm phân loại văn bản theo chủ đề mà dữ liệu được thu thập từ mạng xã hội Twitter.

Chương tiếp theo tôi sẽ trình bày cụ thể phương pháp, thuật toán SVM để áp dụng cho bài toán phân lớp ý kiến người dùng theo từng chủ đề và mô hình cũng như giải pháp cho bài toán, kết quả thực nghiệm và đánh giá.

CHƯƠNG 3

THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. ÁP DỤNG PHƯƠNG PHÁP SVM CHO BÀI TOÁN PHÂN LỚP Ý KIẾN NGƯỜI DÙNG THEO TỪNG CHỦ ĐỀ

3.1.1. Lý do chọn phương pháp SVM

Chúng ta có thể thấy từ các thuật toán phân lớp hai lớp như SVM đến các thuật toán phân lớp đa lớp đều có đặc điểm chung là yêu cầu văn bản phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp thì SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất (hơn 10.000 chiều) trong khi đó các phương pháp khác có số chiều bé hơn nhiều (như Naïve Bayes là 2000, k-Nearest Neighbors là 2415…).

So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể [3].

3.1.2. Thuật toán SVM

Đặc trưng cơ bản quyết định khả năng phân loại của một bộ phân loại là hiệu suất tổng quát hóa. 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). Kích thước VC là một khái niệm quan trọng đối với một họ hàm phân tách (hay là bộ phân loại). Đại lượng này được xác định bằng số điểm cực đại mà họ hàm có thể phân tách hoàn toàn trong không gian đối tượng. Một bộ phân loại tốt là bộ phân loại có năng lực thấp nhất (có nghĩa là đơn giản nhất) và đảm bảo sai số huấn luyện nhỏ.

Tập phân lớp SVM là *mặt siêu phẳng phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại*, trong đó *độ chênh lệch – còn gọi là Lề* (margin) xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất (Hình 2.5). Mặt siêu phẳng này được gọi là *mặt siêu phẳng lề tối ưu.*.

Máy học SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số w và b. Mục tiêu của phương pháp SVM là ước lượng w và b để cực đại hóa lề giữa các lớp dữ liệu dương và âm. Các giá trị khác nhau của lề cho ta các họ mặt siêu phẳng khác nhau, và lề càng lớn thì năng lực của máy học càng giảm. Như vậy, cực đại hóa lề thực chất là việc tìm một máy học có năng lực nhỏ nhất. Quá trình phân loại là tối ưu khi sai số phân loại là cực tiểu.

Ta phải giải phương trình sau:

$$\min(w,b,\xi) \quad C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i\,[w \,.\, x_i - b] + \xi_i \geq 1 \qquad (3.2)$$

$$\xi_i \geq 0,\ i = 1, \ldots, N$$

Tìm ra được vector trọng số w và sai số của mỗi điểm trong tập huấn luyện là ξ_i, với C là tham số cho trước, từ đó ta có phương trình tổng quát của siêu phẳng tìm ra được bởi thuật toán SVM là:

$$(x_1, x_2, \ldots, x_n) = C + \sum w_i\, x_i$$

Với i = 1,…, n. Trong đó n là số dữ liệu huấn luyện.

Sau khi đã tìm được phương trình của siêu phẳng bằng thuật toán SVM, sử dụng công thức này để tìm ra nhãn lớp cho các dữ liệu mới.

3.1.3. Huấn luyện SVM

SVM là bộ phân loại tốt vì được huấn luyện với nhiều đặc trưng nhất. Điều này làm cho SVM trở thành một phương pháp thích hợp cho phân loại văn bản, bởi vì giải thuật SVM có khả năng điều chỉnh năng lực phân loại tự động đảm bảo hiệu suất tổng quát hóa tốt, thậm chí cả trong không gian dữ liệu có số chiều cao (số đặc trưng rất lớn) và lượng tài liệu mẫu là có hạn.

3.1.4. Áp dụng SVM cho bài toán phân lớp ý kiến người dùng theo từng chủ đề

Quy trình thực hiện như sau:

Bước 1: Thu thập những câu câu đánh giá, nhận xét về các sự kiện được nhắc đến dựa vào công cụ Twitter4j.

Bước 2: Tiền xử lý dữ liệu. Sau khi làm sạch, dữ liệu sẽ được đưa qua module tách câu. Mỗi câu được biểu diễn trên một dòng. Module tách câu sẽ lọc dữ liệu, loại bỏ những câu cảm thán, những câu không có nghĩa. Như đã trình bày 1.4, phương pháp tách từ tiếng Việt tôi áp dụng cho quá trình thực nghiệm là phương pháp khớp tối đa (Maximum Matching).

Bước 3: Trích xuất tập từ đặc trưng và xây xựng vector đặc trưng văn bản được tiến hành lựa chọn đặc trưng và 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. Để xây dựng bộ vector đặc trưng, tôi sẽ chọn phương pháp lựa chọn tần suất nghịch đảo từ TFxIDF và đo lượng tin tương hỗ.

- ❖ Phương pháp tần suất từ TF
- ❖ Phương pháp tần suất nghịch đảo từ TFxIDF

$$IDF = \log(N/DF) + 1 \quad (3.4)$$

❖ Đo lượng tin tương hỗ. Lượng tin tương hỗ giữa từ t và lớp c được tính như sau:

$$MI(t,c) = \sum_{t\in\{0,1\}}\sum_{c\in\{0,1\}} P(t,c)\log\frac{P(t,c)}{P(t)P(c)} \quad (3.5)$$

❖ Độ đo MI toàn cục (tính trên toàn bộ tập tài liệu huấn luyện) cho từ t được tính như sau:

$$MI_{avg}(t) = \sum_i P_{(c_i)} MI(t, c_i) \quad (3.6)$$

Bước 4: Chọn ra tập dữ liệu học, qua bộ phân lớp nhị phân, từ đó cho ra mô hình huấn luyện. Tại bộ phân lớp nhị phân, vector đặc trưng của tập dữ liệu học sẽ được sử dụng để tính toán cho ra mô hình huấn luyện. Trong đó, mỗi đặc trưng trong vector sẽ được xem xét và phân lớp thuộc Iphone hay Bana Hill.

Bước 5: Tập dữ liệu kiểm tra, cho qua mô hình huấn luyện, ta được kết quả của đánh giá cộng đồng trên mạng xã hội. Dựa vào mô hình huấn luyện được hình thành tại bước 4, ta phân lớp cho từng câu trong tập dữ liệu kiểm tra (với đầu vào là các vector đặc trưng).

3.2. MÔ HÌNH VÀ GIẢI PHÁP CHO BÀI TOÁN

3.2.1. Đề xuất giải quyết bài toán

Thông tin người dùng Twitter cùng follow sẽ được lấy về, xây dựng lại mạng xã hội và được cho qua bộ CONGA để phát hiện cộng đồng. Từ những cộng đồng đó, ta có thể xây dựng dữ liệu về đánh giá của từng nhóm người dùng về một sự kiện, hiện tượng chung nào đó. Với dữ liệu lấy về là Tiếng Việt, tôi sử dụng bộ phân lớp SVM để phân tách các nhận định người dùng theo 2 chủ đề là sản phẩm Iphone hoặc dịch vụ du lịch tại Bana Hill, để từ đó đưa ra được những đánh giá chung về sự kiện, hiện tượng nào đó, và phần này thì 2 người cùng nhóm hướng dẫn của thầy TS. Huỳnh Công Pháp là bạn Nguyễn Hải Minh và Phùng Hữu Đoàn thực hiện.

Đầu vào: Tập người dùng mạng xã hội, các liên kết tương ứng, và các nhận xét, đánh giá của người dùng về sự kiện, hiện tượng.

Đầu ra: Phân lớp theo chủ đề của từng nhóm cộng đồng về tất cả các ý kiến, đánh giá, nhận xét

Phát biểu bài toán: Coi mỗi người dùng là một nút mạng, xây dựng mạng xã hội và phân chia thành các nhóm (cộng đồng) dựa trên những liên kết của các nút mạng. Đưa ra danh sách quan điểm về sự kiện, hiện tượng của từng cộng đồng vừa được xây dựng theo chủ đề đã chọn.

Như đã trình bày ở chương 2, và phần 3.1 thì tôi sẽ chọn thuật toán CONGA trong phát hiện cộng đồng, bộ phân lớp SVM để giải quyết bài toán của mình.

3.2.2. Mô tả thực nghiệm

Như đã trình bày, mô hình ở phần 2.2, mô hình đề xuất cho bài toán thì Khóa luận tập trung chủ yếu vào việc đánh giá kết quả thực nghiệm trên 2 pha chính: phân nhóm cộng đồng CONGA và bộ phân lớp SVM.

a. Mô tả dữ liệu

Dữ liệu được thu thập theo 2 phần:

Phần 1: Ta có thể thu thập được thông tin về những người sử dụng trực tuyến của Twitter, như ID, tên truy cập, danh sách bạn bè, các follower và following, các status, những mẫu tin Tweet mà người sử dụng gửi từ một API mà Twitter cung cấp cho người sử dụng để tương tác với cơ sở dữ liệu của Twitter.

Phần 2: Thu thập dữ liệu về những tweet mà những người dùng trong mạng xã hội vừa xây dựng đề cập đến những sự kiện, hiện tượng. Từ đó phân chia dữ liệu đó đến từng cộng đồng trong mạng xã hội của từng cộng đồng về mỗi sự kiện. Số lượng tin tức được retweet và tweet từ tương đối lớn, đủ để phục vụ cho việc học và kiểm tra của bộ phân lớp theo các cộng đồng khác nhau.

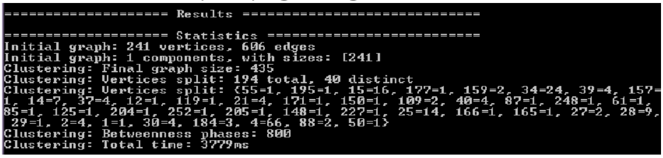
b. Môi trường thực nghiệm

c. Các công cụ và phần mềm sử dụng

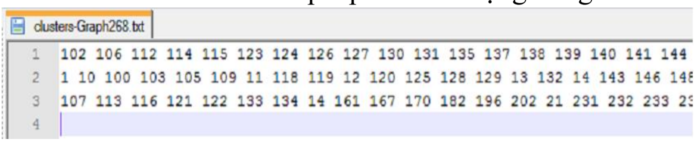
3.3. Kết quả thực nghiệm và đánh giá

a. Kết quả thực nghiệm

❖ **Phần 1: Phát hiện cộng đồng**



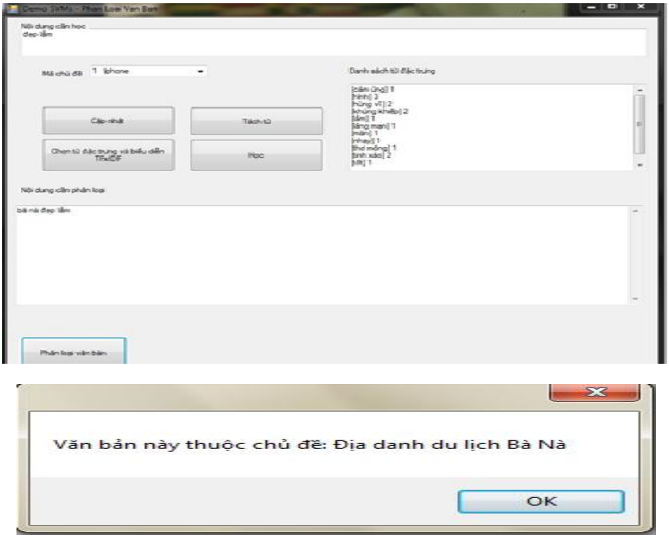
Hình 3.2. Kết quả phân chia cộng đồng



Hình 3.3. Cấu trúc đồ thị chia thành 3 cộng đồng

❖ **Phần 2: Phân loại văn bản SVM**

Giao diện chính của chương trình



Hình 3.5. Kết quả phân loại văn bản

Tập dữ liệu đầu vào từ người dùng được chia theo các nhóm cộng đồng đầu ra của CONGA, sau khi qua bước tiền xử lý cho ra tổng cộng 3053 câu quan điểm để xây dựng máy học và kiểm chứng hiệu quả. Sau khi tách từ và loại bỏ stopword, số từ còn lại là 19937 từ. Sau khi mô hình hóa, mỗi văn bản là một vector trọng số các từ, trong đó các trọng số là chỉ số TF*IDF như đã trình bày ở trên. Như vậy tập ngữ liệu được mô hình hóa như là một ma trận chứa TF*IDF của các từ và có kích thước 19937*3053 phần tử.

Kết quả bước đầu, chương trình đã phân lớp theo từng chủ để của văn bản đầu vào khá chính xác dựa trên những dữ liệu đã học được, đạt 78,08% độ chính xác.

b. Đánh giá

+ Kết quả đánh giá phát hiện cộng đồng mạng sử dụng CONGA đạt 86,9 % độ chính xác.

+ Kết quả đánh giá bộ phân lớp SVM đạt 78,08% độ chính xác.

Nhận xét: Dựa vào kết quả đánh giá, có thể nhận thấy phương pháp phân lớp các quan điểm cộng đồng theo chủ đề sử dụng thuật toán CONGA và bộ vector đặc trưng SVM mang lại kết quả hợp lý.

3.4. KẾT LUẬN CHƯƠNG 3

Trong chương này, tôi đã trình bày lý do lựa chọn thuật toán SVM, thuật toán của nó. Tôi đã tiến hành thực nghiệm, xem xét và đánh giá kết quả của quá trình thực nghiệm mô hình gồm phát hiện và khai phá các quan điểm cộng đồng trên mạng xã hội Twitter với miền tiếng Việt sử dụng phương pháp phát hiện cộng đồng CONGA và phân lớp văn bản tho chủ đề bằng máy vector hỗ trợ SVM. Qua đánh giá cho thấy kết quả khá khả quan.

KẾT LUẬN

1. Nội dung nghiên cứu và kết quả đạt được

Trong luận văn này, tôi đã xây dựng được mô hình phát hiện cộng đồng trên mạng xã hội và thực nghiệm trên mạng xã hội Twitter bằng cách sử dụng thuật toán phát hiện cộng đồng CONGA. Phương pháp này đã đem lại kết quả khá tốt trong việc phát hiện được các cộng đồng chồng chéo nhau trong mạng xã hội. Đồng thời áp dụng phương pháp và xây dựng được mô hình phân lớp SVM về quan điểm của người dùng theo chủ đề.

Tôi đã tiến hành cài đặt thử nghiệm trên tập người dùng Twitter cho kết quả khá khả quan, mô hình phân nhóm cộng đồng khá chuẩn xác, phân lớp quan điểm người dùng theo từng chủ đề đạt độ chính xác 78,08%.

2. Hướng phát triển

Mở rộng cài đặt thử nghiệm với các thuật toán phân loại văn bản khác như kNN, Naïve Bayes, ... sẽ đem nhiều kết quả hơn trong lĩnh vực này.

Mở rộng cài đặt thử nghiệm khai phá dữ liệu để rút trích thông tin đối với các mạng xã hội khác như Facebook,…

Áp dụng cho vùng dữ liệu lớn và tổng quát hơn.