**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**---------------------------------------**



**LÊ TRUNG HIẾU**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET**

**DỰA TRÊN LỊCH SỬ TRUY CẬP**

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

HÀ NỘI - 2017

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**---------------------------------------**



**LÊ TRUNG HIẾU**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET**

**DỰA TRÊN LỊCH SỬ TRUY CẬP**

**CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN**

**MÃ SỐ: 0 60.48.01.04**

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. TỪ MINH PHƯƠNG

HÀ NỘI - 2017

**LỜI CAM ĐOAN**

Luận văn này là thành quả của quá trình học tập nghiên cứu của tôi cùng sự giúp đỡ, khuyến khích của các quý thầy cô sau 2 năm tôi theo học chương trình đào tạo Thạc sĩ, chuyên ngành Hệ thống thông tin của trường Học viện Công nghệ Bưu chính Viễn thông.

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Nội dung của luận văn có tham khảo và sử dụng một số thông tin, tài liệu từ các nguồn sách, tạp chí được liệt kê trong danh mục các tài liệu tham khảo và được trích dẫn hợp pháp.

|  |  |
| --- | --- |
|  | TÁC GIẢ  **Lê Trung Hiếu** |

**LỜI CÁM ƠN**

Tôi xin gửi lời cảm ơn và tri ân tới các thầy cô giáo, cán bộ của Học viện Công nghệ Bưu chính Viễn thông đã giúp đỡ, tạo điều kiện tốt cho tôi trong quá trình học tập và nghiên cứu để hoàn thành chương trình Thạc sĩ.

Tôi xin gửi lời cảm ơn sâu sắc tới PGS.TS Từ Minh Phương đã tận tình hướng dẫn, giúp đỡ và động viên tôi để hoàn thành tốt nhất Luận văn “DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET DỰA TRÊN LỊCH SỬ TRUY CẬP”.

Do vốn kiến thức lý luận và kinh nghiệm thực tiễn còn ít nên luận văn không tránh khỏi những thiếu sót nhất định. Tôi xin trân trọng tiếp thu các ý kiến của các thầy, cô để luận văn được hoàn thiện.

Trân trọng cám ơn.

Tác giả.

**MỤC LỤC**

[MỞ ĐẦU 1](#_Toc481607228)

[1.1 Tính cấp thiết của đề tài 1](#_Toc481607229)

[1.2 Tổng quan về vấn đề nghiên cứu 1](#_Toc481607230)

[1.3 Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu 2](#_Toc481607231)

[1.4 Cấu trúc luận văn 4](#_Toc481607232)

[Chương 1 – TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET 7](#_Toc481607233)

[1.1. Bài toán xác định giới tính và ứng dụng của bài toán vào thực tiễn 7](#_Toc481607234)

[***1.1.1.*** ***Mở đầu*** 7](#_Toc481607235)

[***1.1.2.*** ***Bài toán xác định giới tính*** 8](#_Toc481607236)

[***1.1.3.*** ***Ứng dụng của bài toán vào thực tiễn*** 11](#_Toc481607237)

[1.2. Các dạng dữ liệu lịch sử có thể dự đoán 12](#_Toc481607238)

[***1.2.1.*** ***Thuật toán K–Nearest Neighbor (kNN)*** 12](#_Toc481607239)

[***1.2.2.*** ***Thuật toán cây quyết định (Decision tree)*** 13](#_Toc481607240)

[***1.2.3.*** ***Thuật toán SVM (Support vector machine)*** 15](#_Toc481607241)

[***1.2.4.*** ***Thuật toán Naïve Bayes*** 16](#_Toc481607242)

[1.3. Các phương pháp xác định giới tính đã có 18](#_Toc481607243)

[***1.3.1.*** ***Phương pháp xác định giới tính sử dụng bài viết từ blog*** 18](#_Toc481607244)

[***1.3.2.*** ***Phương pháp phân loại giới tính dựa trên hành vi và môi trường*** 19](#_Toc481607245)

[***1.3.3.*** ***Phương pháp xác định giới tính sử dụng dữ liệu thông tin di động liên lạc hàng ngày*** 21](#_Toc481607246)

[***1.3.4.*** ***Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy*** 22](#_Toc481607247)

[1.4. Kết luận chương 22](#_Toc481607248)

[Chương 2 – DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET SỬ DỤNG LỊCH SỬ TRUY CẬP 24](#_Toc481607249)

[2.1. Giới thiệu về phương pháp học máy SVM 24](#_Toc481607252)

[***2.1.1.*** ***Giới thiệu về SVM*** 24](#_Toc481607253)

[***2.1.2.*** ***Bài toán phân 2 lớp với SVM*** 25](#_Toc481607254)

[***2.1.3.*** ***Bài toán phân đa lớp với SVM*** 29](#_Toc481607255)

[***2.1.4.*** ***Các bước chính của phương pháp SVM*** 30](#_Toc481607256)

[***2.1.5.*** ***Ưu điểm phương pháp SVM trong phân lớp dữ liệu*** 30](#_Toc481607257)

[2.2. Giới thiệu về dữ liệu sử dụng 31](#_Toc481607258)

[***2.2.1.*** ***Giới thiệu thuật toán*** 31](#_Toc481607259)

[***2.2.2.*** ***Ý tưởng thuật toán*** 32](#_Toc481607260)

[***2.2.3.*** ***Thuật toán*** 32](#_Toc481607261)

[1.5 Phân loại đa lớp nội dung văn bản với SVM 41](#_Toc481607262)

[Chương 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ 46](#_Toc481607263)

[3.1. Xây dựng bộ dữ liệu và tiền xử lý văn bản 46](#_Toc481607267)

[3.2. Giai đoạn phân lớp 49](#_Toc481607268)

[***3.2.1.*** ***URL phục vụ cho ứng dụng*** 49](#_Toc481607269)

[***3.2.2.*** ***Triển khai ứng dụng*** 49](#_Toc481607270)

[KẾT LUẬN 51](#_Toc481607271)

[DANH MỤC TÀI LIỆU THAM KHẢO 53](#_Toc481607272)

[PHỤ LỤC 55](#_Toc481607273)

**DANH MỤC TỪ VIẾT TẮT**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Ý nghĩa tiếng Anh** | **Ý nghĩa tiếng Việt** |
| 1 | SVM | Support vector machine | Máy vector hỗ trợ | |
| 2 | NB | Naïve Bayes |  | |
| 3 | kNN | K–Nearest Neighbor | K – Láng giềng gần nhất | |
| 4 | TF | Term Frequency | Tần số xuất hiện của 1 từ | |
| 5 | IDF | Inverse Document Frequency | Tần số nghịch của 1 từ trong tập văn bản | |
| 6 | RSS | Really Simple Syndication | Định dạng tài liệu | |

**DANH MỤC CÁC BẢNG BIỂU**

**DANH MỤC CÁC HÌNH VẼ**

# MỞ ĐẦU

## Tính cấp thiết của đề tài

Ngày nay, người ta thường dành một lượng lớn thời gian trong ngày để truy cập internet. Internet được người dùng sử dụng cho việc tìm kiếm thông tin, đọc tin tức, mua sắm, chơi trò chơi vv. Và các nhà quảng cáo không thể bỏ lỡ cơ hội để tiếp thị trực tuyến đến với khách hàng của họ nhằm cung cấp các dịch vụ phù hợp với nhu cầu của tổ chức, cá nhân sử dụng mạng internet. Tuy nhiên, hiện nay các nhà quảng cáo đang cung cấp toàn bộ thông tin của mình đến tất cả khách hàng họ có. Chính vì vậy người dùng thường phải đối mặt với số lượng lớn các thông tin không phù hợp ví dụ như không phù hợp về độ tuổi, về nghề nghiệp, về văn hóa và giới tính.

Tình trạng quá tải thông tin không đến đích này dẫn đến sự sụt giảm đáng kể trong việc tiếp thị trực tuyến. Từ đó việc phân loại người dùng internet để đưa ra các số liệu thống kê, kế hoạch quảng cáo giúp hệ thống tiếp cận cung cấp thông tin phù hợp, hữu ích cho từng đối tượng tương đối quan trọng. Xuất phát từ thực trạng đang xảy ra, luận văn sẽ trình bày về phương pháp xác định giới tính để phân loại người dùng internet được thực hiện bằng kỹ thuật học máy, sử dụng thông tin người dùng đã biết giới tính và các thông tin về lịch sử truy cập web của họ để huấn luyện máy nhận biết giới tính của những người dùng khác khi ta chỉ biết lịch sử truy cập các trang web và dữ liệu danh mục mà người đó quan tâm.

## Tổng quan về vấn đề nghiên cứu

Phương pháp dự đoán giới tính đã được nghiên cứu trong một thời gian dài. Ở giai đoạn đầu, hầu hết các nghiên cứu về lĩnh vực này tập trung vào việc nghiên cứu phân tích các đoạn văn bản được tạo ra từ người dùng mạng để xác định giới tính người viết là nam hay nữ. Các phương pháp nghiên cứu chủ yếu dựa trên các đặc trưng chẳng hạn như từ vựng, cú pháp hoặc các đặc trưng về nội dung với dữ liệu được thu thập từ email, blog, ý kiến vv. Gần đây, do sự phát triển của Internet và các kênh truyền thông trực tuyến, phương pháp dự đoán giới tính đã được mở rộng hơn thay vì phân tích văn bản các nhà nghiên cứu điều tra việc sử dụng các hành vi duyệt web của họ. Hu et al. (2007) đã đề xuất một phương pháp để giải quyết vấn đề của dự đoán giới tính và tuổi tác của người dùng Internet dựa trên hành vi truy cập web. Kabbur et al. (2010) cũng nghiên cứu phương pháp học máy học tập để dự đoán các thuộc tính nhân khẩu của các trang web bằng cách sử dụng thông tin từ nội dung và cấu trúc siêu liên kết. Các nghiên cứu của Dong et al. (2013) nhằm Suy luận nhân khẩu học của người dùng dựa trên các mẫu thông tin di động hàng ngày của họ. Nghiên cứu của họ được tiến hành trên một mạng di động lớn thực thế giới của hơn 7.000.000 người dùng và hơn 1.000.000.000 bản ghi thông tin liên lạc.

Trong luận văn này sẽ giải quyết các vấn đề của dự đoán các thông tin cá nhân của người dùng dựa trên dữ liệu xem danh mục của họ chẳng hạn như thời gian xem /sản phẩm xem / loại, vv. Luận văn sử dụng phương pháp học tập phổ biến SVM (Support Vector Machine) để huấn luyện và thử nghiệm trên các tập dữ liệu lịch sử truy cập người dùng internet được cung cấp bởi Công ty Cổ phần FPT PAKDD'15.

## Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu

***Mục đích nghiên cứu:***

Bằng nhận thức khi sử dụng internet ta nhận thấy việc đưa thông tin tiếp thị, tin tức phù hợp đến một người dùng nào đó thật hữu ích khi mà người đó quan tâm hoặc sẵn sàng để truy cập dịch vụ. Mặc dù trên thế giới đã có khá nhiều phương pháp xác định giới tính người dùng internet, thông thường mỗi phương pháp lại dựa trên một số đặc trưng, yếu tố nhất định với kết quả và dữ liệu thu thập tuy nhiên những nghiên cứu này vẫn chưa được bám sát được mục tiêu, mức độ nghiên cứu vẫn chỉ trong phạm vi hẹp. Trên cơ sở đó luận văn tập trung vào nghiên cứu chính là tìm hiểu và thử nghiệm phương pháp dự đoán giới tính người dùng internet dựa trên lịch sử truy cập bằng cách sử dụng kỹ thuật học máy. Nghiên cứu này nhằm mục đích thực hiện một tổng quan về các phương pháp trước đó và hiệu suất của các mô hình dự đoán đạt được để so sánh sự khác biệt về độ chính xác của các mô hình từ đó đưa ra phân tích và thảo luận với mô hình dự đoán giới tính và đưa ra các tỉ lệ, độ chính xác bên cạnh đó còn cung cấp các kiến thức khoa học hiện tại khi dự đoán giới tính và khả năng hiệu quả trong việc dự đoán. Mục tiêu cụ thể được trình bày trong luận văn như sau:

* Tìm hiểu về bài toán xác định giới tính người dùng internet.
* Các phương pháp xác định giới tính đã có.
* Hiểu được kỹ thuật học máy và học máy SVM.
* Giới thiệu về bộ dữ liệu sử dụng trong nghiên cứu này.
* Áp dụng và thử nghiệm kỹ thuật học máy cho bộ dữ liệu sử dụng.

***Đối tượng nghiên cứu:***

* Dữ liệu lịch sử truy cập của người dùng internet
* Kỹ thuật học máy SVM trong việc phân loại giới tính

***Phạm vi nghiên cứu***

* Nghiên cứu tổng quan về bài toán dự đoán giới tính
* Sau đó, luận văn đi sâu tìm hiểu ứng dụng kỹ thuật học máy SVM trong việc dự đoán giới tính dựa trên lịch sử truy cập với tập dữ liệu được cung cấp bởi Công ty Cổ phần FPT PAKDD'15.

***Phương pháp nghiên cứu:***

* **Phương pháp nghiên cứu lý thuyết**
* Tìm hiểu các phương pháp dự đoán giới tính hiện này đang có để xác định những điểm mạnh và hạn chế của các phương pháp đó.
* Tìm hiểu các kỹ thuật học máy hiện nay, những ưu nhược điểm của các phương pháp học máy đó trong việc dự đoán giới tính.
* Tìm kiểu kỹ thuật học máy SVM và ứng dụng hiện nay của phương pháp này.
* **Phương pháp nghiên cứu thực nghiệm**
* Nghiên cứu các đặc trưng dữ liệu lịch sử truy cập của người dùng trong phạm vi nghiên cứu. Xác định những yếu tốt ảnh hưởng đến giới tính người dùng và các cách để lấy dữ liệu để nghiên cứu.
* Phân tích, chuẩn hóa dữ liệu sử dụng kỹ thuật học máy SVM đưa ra kết quả.
* So sánh kết quả thực nghiệm với các phương pháp khác.

## Cấu trúc luận văn

Với mục tiêu đặt ra như vậy, nội dung và kết quả của luận văn được trình bày qua 3 chương như sau:

**CHƯƠNG I:** **TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET**

Giới thiệu chương:

Giới thiệu về dữ liệu truy cập của người dùng internet thông qua thống kê, các khái niệm và đặc trưng trong tập dữ liệu này, bao gồm các mối quan hệ giữa các trang thông tin và người dùng mạng, những hành vi của người dùng khi truy cập internet, cách thức truy cập, tìm kiếm thông tin. Giới thiệu những phương pháp nhắm mục tiêu theo hành vi hiện nay được áp dụng cho người dùng internet và những hạn chế của các phương pháp này.

Nội dung chương 1 sẽ bố cục theo các mục sau:

* 1. Bài toán xác định giới tính và ứng dụng của bài toán vào thực tiễn
  2. Các dạng dữ liệu lịch sử có thể dùng để dự đoán
  3. Các phương pháp xác định giới tính đã có
* 1.3.1. Phương pháp xác định giới tính sử dụng bài viết từ blog.
* 1.3.2. Phương pháp phân loại giới tính dựa trên hành vi và môi trường.
* 1.3.3. Phương pháp xác định giới tính sử dụng dữ liệu thông tin liên lạc di động hàng ngày.
* 1.3.4. Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy.
  1. Kết luận chương

**CHƯƠNG II: DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET SỬ DỤNG LỊCH SỬ TRUY CẬP**

Giới thiệu chương:

Trình bày tổng quan về kỹ thuật học máy, một số kỹ thuật học máy và tập trung vào kỹ thuật được sử dụng trong luận văn là kỹ thuật học máy SVM.

Dựa vào những đặc trưng việc truy cập thông tin của người dùng internet, đưa ra phương pháp dự đoán giới tính áp dụng kỹ thuật học máy và xếp hạng tỉ lệ độ chính xác nhằm tăng hiệu quả dự đoán so với các phương pháp đang tồn tại.

Nội dung chương 2 sẽ bố cục theo các mục sau:

* 1. Giới thiệu phương pháp học máy SVM
  2. Giới thiệu về dữ liệu sẽ sử dụng
  3. Các đặc trưng sẽ dùng trong phân lớp
     1. Dạng đặc trưng theo mốc thời gian
     2. Dạng đặc trưng dựa trên tần số hoạt động của người truy cập
     3. Dạng đặc trưng về danh mục và chủng loại sản phẩm
  4. Kết luận chương

**CHƯƠNG III: THỬ NGHIỆM VÀ ĐÁNH GIÁ**

**Giới thiệu chương:**

Sử dụng dữ liệu có sẵn PAKDD'15 được cung cấp bởi Công ty Cổ phần FPT (http://www.fpt.com.vn), thực hiện xây dựng bộ dữ liệu từ dữ liệu thực tế chưa chuẩn hóa hiện có PAKDD'15 cho một số lượng người dùng, sử dụng kỹ thuật học máy SVM ở chương 2 và một số công cụ để đưa ra tỉ lệ, độ chính xác của phương pháp dự đoán giới tính dựa trên lịch sử truy cập. Đánh giá kết quả so với các phương pháp dự đoán khác, và so sánh với cách làm việc hiện tại trong việc dự đoán giới tính.

Nội dụng chương 3 sẽ bố cục theo các mục sau:

* 1. Mô tả dữ liệu
  2. Các tiêu chuẩn đánh giá
  3. Phương pháp thực nghiệm
  4. Kết quả thực nghiệm
  5. So sánh với một số phương pháp khác
  6. Độ phức tạp và thời gian thực hiện phương pháp.
  7. Kết luận chương

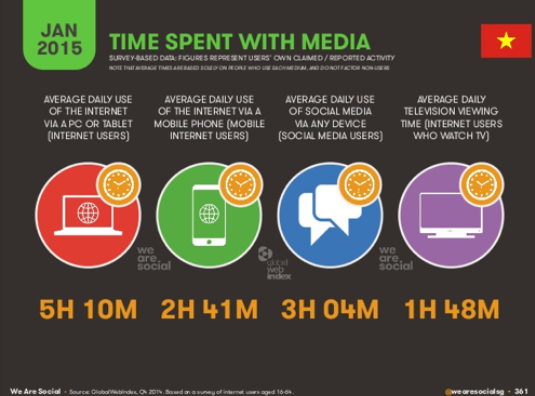
# Chương 1 – TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET

## Bài toán xác định giới tính và ứng dụng của bài toán vào thực tiễn

* + 1. ***Mở đầu***

Ngày nay, với sự phát triển không ngừng của khoa học công nghệ trên thế giới nói chung và ở Việt Nam nói riêng có những bước tiến vượt bậc. Cơ sở hạ tầng và các trang thiết bị tương đối hiện đại và không ngừng phát triển. Theo báo cáo tổng kết của Bộ TT&TT năm 2016, tỷ lệ người sử dụng Internet ở Việt Nam đạt 62,76% dân số, trong đó tỷ lệ hộ gia đình có truy cập Internet đạt 24,38%, tức là cứ 5 gia đình thì có một hộ sử dụng băng thông rộng cố định. Trong đó, theo thống kê của Cục Viễn thông (Bộ TT&TT) tháng 11/2016, tổng số thuê bao Internet băng rộng cố định đạt hơn 9 triệu thuê bao và số thuê bao băng rộng di động đạt hơn 12,6 triệu thuê bao.

Bên cạnh đó, theo thống kê của “wearesocial.net”, tháng 1-2015, người Việt Nam đang đứng thứ 4 trên thế giới về thời gian sử dụng Internet với 5,2 giờ mỗi ngày, chỉ sau Philippines đứng đầu là 6 giờ, tiếp đó là Thái Lan với 5,5 giờ, và Brazin là 5,4 giờ/ngày.



Hình 1.1 Thời gian sử dụng internet trung bình một ngày của người Việt Nam

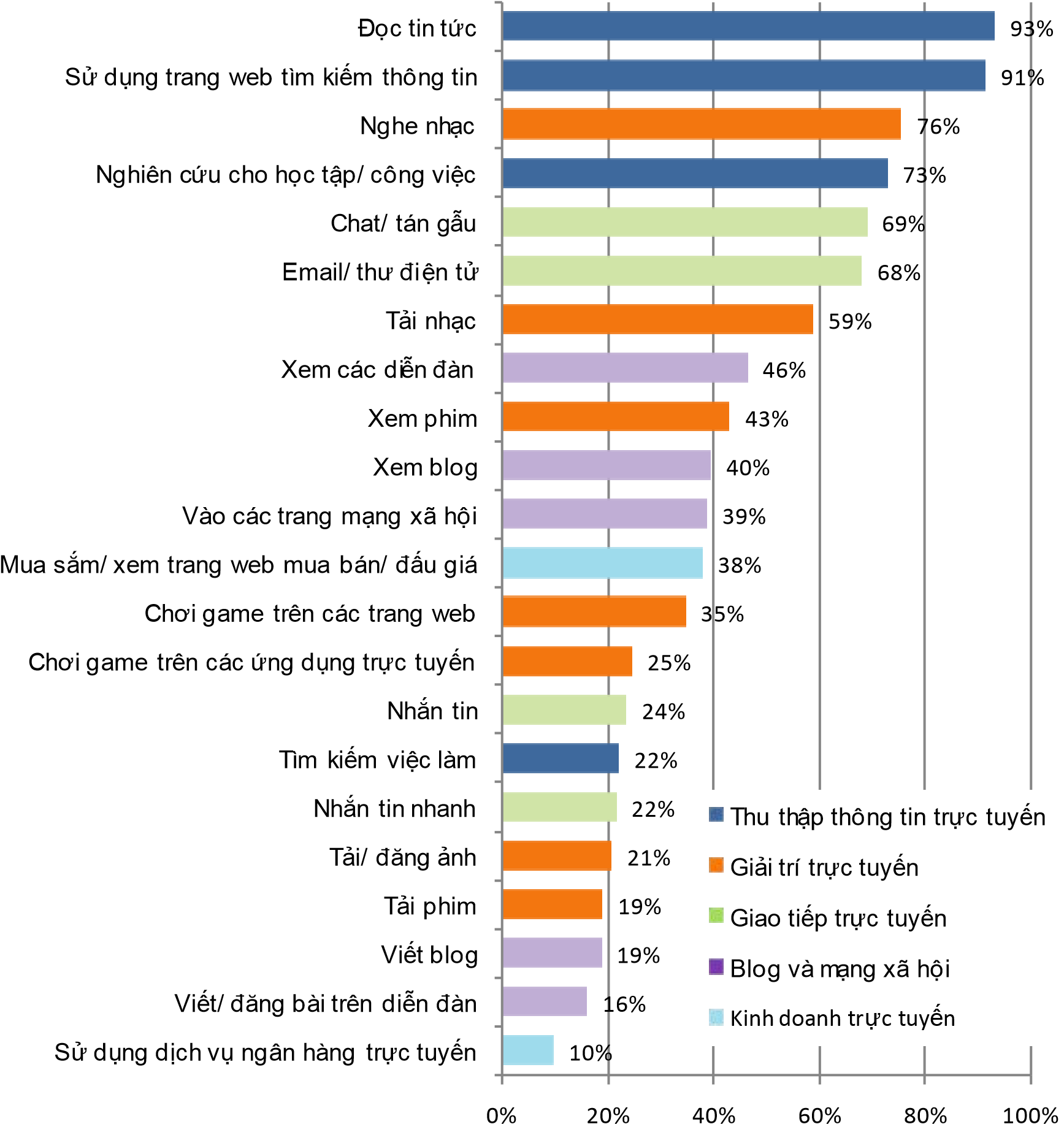
Chính vì sự phát triển không ngừng của công nghệ thông tin và mức độ phổ biến của internet ngày nay mà thông tin đến với người dùng vô cùng phong phú và liên tục. Người sử dụng internet hiện nay thường có thói quen truy cập và tìm kiếm đến những các vấn đề mình quan tâm. Hầu hết các thông tin được lưu vào như một phiên làm việc trên mạng. Các thông tin đó có thể là các bài báo, các tài liệu kinh doanh, sản phẩm, các thông tin kinh tế, thương mại điện tử, các thông tin cá nhân khác, ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin để phân loại các thông tin đó cho các mục đích khác nhau như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại...

Với thực tế đó, ta phải xác định và phân loại những thông tin hữu ích từ các nguồn dữ liệu phong phú từ các phiên làm việc, sử dụng internet của người dùng sao cho phù hợp với đối tượng cụ thể và hỗ trợ các công cụ tự động hoá trợ giúp trong việc phát hiện tri thức và khai thác thông tin.

* + 1. ***Bài toán xác định giới tính***

Nhìn chung, hoạt động thực hiện thường xuyên nhất trên Internet của người dùng là thu thập thông tin, như đọc tin tức hay sử dụng các trang web tìm kiếm. Hơn 90% số lượng người sử dụng Internet đã sử dụng những trang web tìm kiếm, khoảng một nửa trong số họ thậm chí sử dụng hàng ngày. Internet cũng được sử dụng để nghiên cứu cho việc học hay cho công việc bởi một nửa số người sử dụng Internet 1 lần 1 tuần hay thường xuyên hơn. Với các trang web và ứng dụng tương tác trực tuyến mới, người sử dụng không chỉ có cơ hội tìm được thông tin mà cũng đóng góp phần nội dung của riêng họ.

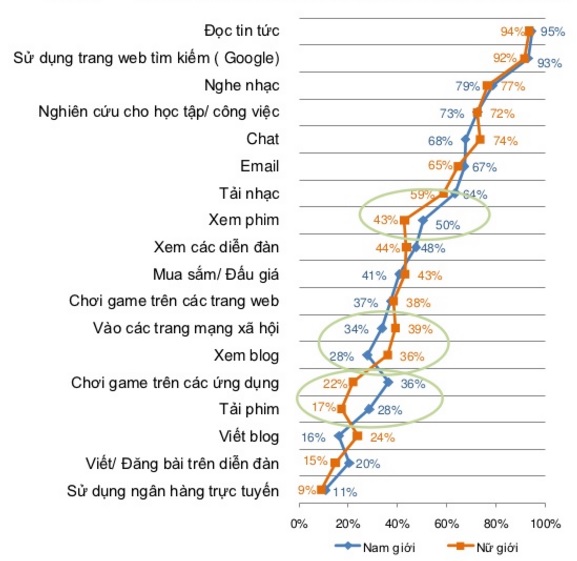
Thương mại điện tử hiện nay, số lượng truy cập đạt mức tăng trưởng đáng kể. Hầu hết các trang phổ biến là các trang web đấu giá và mua bán, nơi có 40% người sử dụng đã từng viếng thăm. Ngân hàng trực tuyến vẫn đang ở giai đoạn sơ khai. Mức độ sử dụng các trang web mua hàng trực tuyến và ngân hàng trực tuyến đã phát triển rất mạnh trong vòng vài năm trở lại đây.



Hình 1.2 Các hoạt động trực tuyến được người dùng mạng sử dụng

(Nguồn: Cimigo NetCitizens)

Việc sử dụng các hoạt động và truy cập internet có sự khác nhau giữa nam giới và nữ giới. Trung bình một ngày nam giới dành thời gian nhiều hơn cho Internet. Nam giới cũng có một số hoạt động trực tuyến giống với nữ giới. Tuy nhiên có những khác nhau cụ thể ví dự như nam giới có khuynh hướng truy cập những đặc trưng như tin tức thời sự, bóng đá, hay trò chơi và các mặt hàng dành cho nam giới. Trái lại nữ giới thường thích thú với các mục mua sắm, thương mại điện tử, chat và tham gia các trang mạng xã hội và blog.

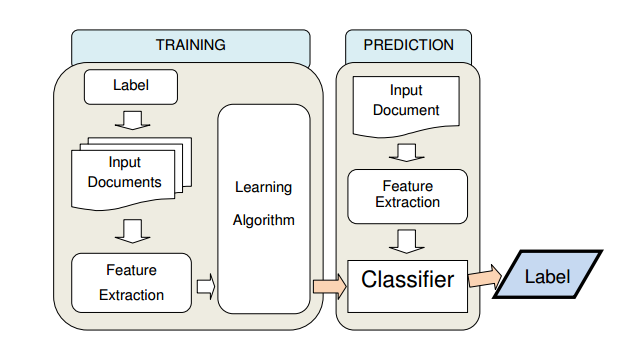


Hình 1.3 Các hoạt động trực tuyến được người dùng mạng theo giới tính

(Nguồn: Cimigo NetCitizens)

Dự đoán giới tính (hay Determination Gender hoặc Gender Prediction) là phương pháp phân loại và xác định các hoạt động được truy cập bởi giới tính Nam hoặc giới tính Nữ từ những hoạt động khác đã biết trước nhãn. Ví dụ một bài báo trong một trang web có thể được truy cập bởi giới tính nam hoặc giới tính nữ (như thể thao, giáo dục, pháp luật, công nghệ thông tin, mỹ phẩm, quần áo …). Việc phân loại có thể được tiến hành một cách thủ công: đọc nội dung của từng hoạt động và gán nó vào một nhãn nào đó. Tuy nhiên, đối với hệ thống gồm rất bản ghi thì phương pháp này sẽ tốn rất nhiều thời gian và công sức. Do vậy cần phải có phương pháp tự động để phân loại giới tính. Phương pháp này giúp cho việc xác định giới tính đạt độ chính xác cao và sử dụng cho các mục đích như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại.

Dưới đây là hình vẽ mô tả quy trình của bài toán xác định giới tính:



Hình 1.4 Quy trình phân loại xác định giới tính

Để tiến hành phân loại xác định giới tính nói chung, chúng ta sẽ thực hiện các bước sau đây:

* Bước 1: Xây dựng bộ dữ liệu huấn luyện dựa trên tập dữ liệu thu thập của người dùng đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lý và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng nội dung.
* Bước 2: Dữ liệu cần phân loại được xử lý, rút ra các đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và đưa ra kết quả.

Đặc điểm nổi bật của bài toán này là sự đa dạng của hoạt động và đặc trưng của nam giới và nữ giới. Các đặc trưng làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, nếu do con người thực hiện có thể dễ bị nhập nhằng. Ví dụ có hoạt động truy cập về xem thông tin mua sắm quần áo tại một trang web thương mại điện tử, hoạt động truy cập này vẫn có thể được truy cập bời nam giới hoặc nữ giới.

* + 1. ***Ứng dụng của bài toán vào thực tiễn***

Hiện nay trên thế giới đã có một số công trình nghiên cứu với các hướng tiếp cận khác nhau cho bài toán xác định giới tính, bao gồm các tập dữ liệu có đặc trưng thể hiện giới tính. Theo các kết quả trình bày trong các công trình đó thì những cách tiếp cận đều cho kết quả khả quan. Tuy nhiên khó có thể so sánh các kết quả ở trên với nhau vì tập dữ liệu thực nghiệm của mỗi phương pháp là khác nhau.

Hiện nay, với sự phát triển không ngừng của ngành công nghệ thông tin và truyền thông người dùng mạng hiện nay thường có thói quen cung cấp, trao đổi các thông tin một cách liên tục và thường xuyên. Hầu hết các thông tin đều là các hoạt động trực tuyến như tìm kiếm thông tin, chat, email, mua sắm trực tuyến ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin của người dùng internet để phân loại các thông tin đó sao cho phù hợp với giới tính như học tập, nghiên cứu, kinh doanh nhằm đưa ra các số liệu thống kê, kế hoạch quảng cáo giúp hệ thống tiếp cận cung cấp thông tin phù hợp, hữu ích cho từng đối tượng.

Trong những năm gần đây, phương pháp phân loại sử dụng Máy vector hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại 2 lớp và đa lớp cũng như trong nhiều ứng dụng khác (như phân loại văn bản theo chủ đề, phát hiện mặt người trong các ảnh, ước lượng hồi quy, dự đoán lỗi phần mềm...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể. Vì những lý do đó mà tôi đã chọn phương pháp này cho việc dự đoán giới tính của người dùng internet, cụ thể thuật toán và ứng dụng sẽ được trình bày trong các chương sau.

## Các dạng dữ liệu lịch sử có thể dự đoán

Có nhiều loại dữ liệu lịch sử có thể dự đoán. Ở giai đoạn đầu phân loại giới tính, hầu hết các nghiên cứu về lĩnh vực này tập trung vào việc nghiên cứu tác giả, đó là những nhiệm vụ xác định hoặc dự đoán các đặc điểm tác giả bằng cách phân tích các câu chuyện, tác phẩm, tiểu thuyết được tạo ra bởi tác giả nam hay tác giả nữ. Các phương pháp mà các nhà nghiên cứu sử dụng trong các nghiên cứu này chủ yếu dựa trên việc phân tích các phong cách viết, văn phong sử dụng các đặc trưng về ngữ pháp chẳng hạn như từ vựng, cú pháp, hoặc các đặc trưng dựa trên nội dung. Nghiên cứu đầu tiên trong lĩnh vực này bắt đầu vào thế kỷ 19 khi Mendenhall (1887) đã nghiên cứu các tác phẩm của Shakespeare.

Gần đây, do sự phát triển của Internet và các kênh truyền thông trực tuyến, các dạng dữ liệu lịch sử được thu thập là nội dung truyền thông:

* Email: Một dạng dữ liệu lịch sử, một phương tiện thông tin rất nhanh chứa đựng các văn bản đơn thuần và thường được dùng trong việc trao đổi thông tin. Chúng ta có thể dự đoán giới tính dựa trên địa chỉ email và văn bản có trong email.
* Blog: Là một tập san dữ liệu cá nhân trực tuyến. Nội dung và chủ đề của “blog” thì rất đa dạng, nhưng thông thường là những câu chuyện cá nhân, bản tin, danh sách các liên kết web, những bài tường thuật, phê bình một bộ phim hay tác phẩm văn học mới xuất bản và cuối cùng là những sự kiện xảy ra trong một nhóm người nào đó.
  1. **Các phương pháp xác định giới tính đã có**

**Bài toán**

Input: Cho tập văn bản miền ứng dụng D = {dj}, tập đặc trưng được chọn biểu diễn văn bản V = {wi}, ma trân trọng số W = (wi,j) .

Output: Tìm biểu diễn của các văn bản dj ∈D.

Một số mô hình: Mô hình Boolean, Mô hình không gian vector, Mô hình túi các từ (Mô hình xác suất)

* + 1. ***Phương pháp xác định giới tính sử dụng bài viết từ blog***

Một mô hình biểu diễn vector với hàm f cho ra giá trị rời rạc với duy nhất hai giá trị đúng và sai (true và false, hoặc 0 và 1) gọi là mô hình Boolean.

Hàm f tương ứng với từ khóa ti sẽ cho ra giá trị đúng nếu và chỉ nếu từ khóa ti xuất hiện trong văn bản đó.

Mô hình Boolean được xác định như sau:

Giả sử có một tập gồm m văn bản D = {d1, d2, d3, ..., dm}, T là một tập từ vựng gồm n từ khóa T = {t1, t2,......tn}. Gọi w = (wịj) là ma trận trọng số, trong đó wij là trọng số của từ khóa ti trong văn bản dj và được xác định như sau:

Ta xét 2 văn bản sau:

VB1: Life is not only life

VB2: To life is to fight

Áp dụng mô hình Boolean sau khi tiền xử lý văn bản ta có biểu diễn sau:

Bảng 1.1 Biểu diễn văn bản mô hình Boolean



* + 1. ***Phương pháp phân loại giới tính dựa trên hành vi và môi trường***

Cách biểu diễn văn bản thông dụng nhất là thông qua vector biểu diễn theo mô hình không gian vector (Vector Space Model). Đây là một cách biểu diễn tương đối đơn giản và hiệu quả.

Theo mô hình này, mỗi văn bản được biểu diễn thành một vector. Mỗi thành phần của vector là một từ khóa riêng biệt trong tập văn bản gốc và được gán một giá trị là hàm f chỉ mật độ xuất hiện của từ khóa trong văn bản.



Hình 1.0.2: Biểu diễn vector văn bản trong không gian 2 chiều

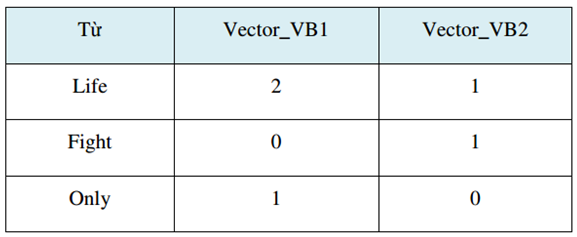
Giả sử ta có một văn bản và nó được biểu diễn bởi vector V(v1, v2,…,vn). Trong đó, vi là số lần xuất hiện của từ khóa thứ i trong văn bản. Ta xét 2 văn bản sau:

VB1: Life is not only life

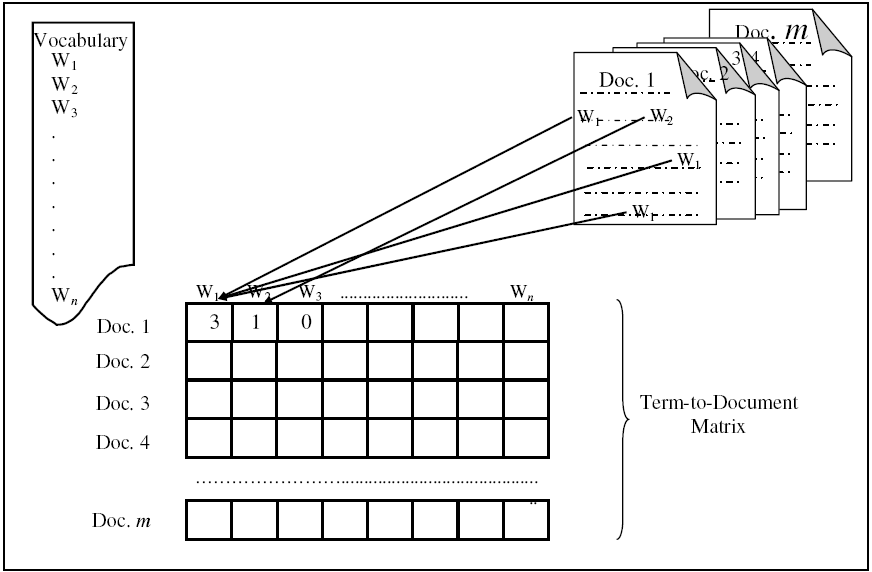
VB2: To life is to fight

Sau khi qua bước tiền xử lý văn bản, ta biểu diễn chúng như sau:

Bảng 2: Biểu diễn văn bản theo mô hình Vector



Trong các cơ sở dữ liệu văn bản, mô hình vector là mô hình biểu diễn văn bản được sử dụng phổ biến nhất hiện nay. Mối quan hệ giữa các trang văn bản được thực hiện thông qua việc tính toán trên các vector biểu diễn vì vậy được thi hành khá hiệu quả. Đặc biệt, nhiều công trình nghiên cứu về mối quan hệ "tương tự nhau" giữa các trang web (một trong những quan hệ điển hình nhất giữa các trang web) dựa trên mô hình biểu diễn vector.



Hình 1.0.3: Biểu diễn văn bản theo mô hình vector

* + 1. ***Phương pháp xác định giới tính sử dụng dữ liệu thông tin di động liên lạc hàng ngày***

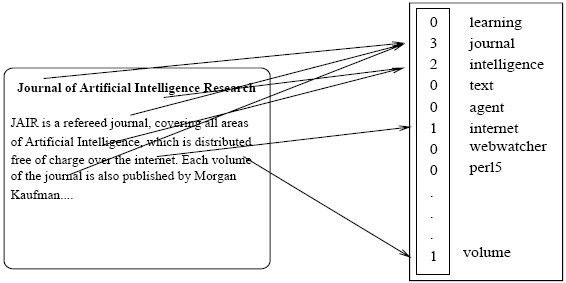
Mô hình xác suất là mô hình toán học làm việc với các biến ngẫu nhiên và phân bố xác xuất của nó. Theo thuật ngữ toán học, một mô hình xác suất có thể coi như một cặp (Y, P), trong đó Y là tập các quan sát (biến ngẫu nhiên) và P là tập các phân bố xác suất trên Y. Khi đó, sử dụng suy diễn xác suất sẽ cho ta kết luận về các phần tử của tập Y.

Văn bản trong mô hình xác suất được coi như một quan sát trong tập Y, trong đó:

* Các từ xuất hiện trong văn bản thể hiện nội dung văn bản
* Sự xuất hiện của các từ là độc lập lẫn nhau và độc lập ngữ cảnh
* Dạng đơn giản: chỉ liệt kê từ
* Dạng chi tiết: liệt kê từ và số lần xuất hiện

*Lưu ý*: Các giả thiết về tính độc lập không hoàn toàn đúng (độc lập lẫn nhau, độc lập ngữ cảnh) song mô hình thi hành hiệu quả trong nhiều trường hợp.

Khi đó văn bản sẽ gồm các từ mà nó chứa trong đó, chính vì vậy phương pháp này được gọi là biểu diễn túi - các - từ (bag-of-word)



Hình 1.0.4: Biểu diên văn bản theo mô hình túi các từ

* + 1. ***Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy***
  1. **Kết luận chương**

**Kết luận chương:**

Chương này tôi đã giới thiệu tổng quan về bài toán phân loại văn bản, một số phương pháp phân loại và các kỹ thuật liên quan đến biểu diễn văn bản. Bên cạnh đó, chương 1 còn đưa ra lý do tại sao lại chọn thuật toán SVM trong luận văn. Ngoài ra cần lưu ý đến 3 yếu tố quan trọng tác động đến kết quả phân loại văn bản:

1) Cần một tập dữ liệu huấn luyện chuẩn và đủ lớn để cho thuật toán học phân loại. Nếu chúng ta có được một tập dữ liệu chuẩn và đủ lớn thì quá trình huấn luyện sẽ tốt và khi đó chúng ta sẽ có kết qủa phân loại tốt sau khi đã được học.

2) Các phương pháp trên hầu hết đều sử dụng mô hình vector để biểu diễn văn bản, do đó phương pháp tách từ trong văn bản đóng vai trò quan trọng quá trình biểu diễn văn bản bằng vector. Yếu tố này rất quan trọng, vì có thể đối với một số ngôn ngữ như tiếng Anh thì thao tác tách từ trong văn bản đơn giản chỉ là dựa vào các khoảng trắng, tuy nhiên trong các ngôn ngữ đa âm tiết như tiếng Việt và một số ngôn ngữ khác thì sử dụng khoảng trắng khi tách từ là không chính xác, do đó phương pháp tách từ là một yếu tố quan trọng.

3) Thuật toán sử dụng để phân loại phải có thời gian xử lý hợp lý, thời gian này bao gồm: thời gian học, thời gian phân loại văn bản, ngoài ra thuật toán này phải có tính tăng cường (incremental function) nghĩa là không phân loại lại toàn bộ tập văn bản khi thêm một số văn bản mới vào tập dữ liệu mà chỉ phân loại các văn bản mới mà thôi, khi đó thuật toán phải có khả năng giảm độ nhiễu (noise) khi phân loại văn bản.

# Chương 2 – DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET SỬ DỤNG LỊCH SỬ TRUY CẬP



## Giới thiệu về phương pháp học máy SVM

* + 1. ***Giới thiệu về SVM***

Trong thời đại công nghệ thông tin hiện nay, sự phát triển của công nghệ kéo theo sự gia tăng rất lớn của lưu lượng thông tin lưu trữ và trao đổi. Do đó, yêu cầu về tổ chức lưu trữ và truy cập thông tin sao cho hiệu quả được đặt lên hàng đầu. Hướng giải quyết được đưa ra là tổ chức, tìm kiếm và phân loại thông tin một cách hiệu quả. Bản thân con người trong đời sống cũng tiếp nhận thế giới xung quanh thông qua sự phân loại và tổ chức ghi nhớ tri thức một cách hiệu quả. Phân loại thông qua các lớp và mô tả các lớp giúp cho tri thức được định dạng và lưu trữ trong đó. Support Vector Machines (SVM) là một phương pháp phân loại xuất phát từ lý thuyết học thống kê, dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (Structural Risk Minimisation). SVM sẽ cố gắng tìm cách phân loại dữ liệu sao cho có lỗi xảy ra trên tập kiểm tra là nhỏ nhất (Test Error Minimisation). Đây là một phương pháp mới trong lĩnh vực trí tuệ nhân tạo. Vào thời kỳ đầu khi SVM xuất hiện, khả năng tính toán của máy tính còn rất hạn chế, nên phương pháp SVM không được lưu tâm. Tuy nhiên, từ năm 1995 trở lại đây, các thuật toán sử dụng cho SVM phát triển rất nhanh, cùng với khả năng tính toán mạnh mẽ của máy tính, đã có được những ứng dụng rất to lớn.

1. Ý tưởng

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 phẳng f quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp “+” và lớp “−”. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác. Ý tưởng của nó là ánh xạ (tuyến tính hoặc phi tuyến) dữ liệu vào không gian các vector đặc trưng (space of feature vectors) mà ở đó một siêu phẳng tối ưu được tìm ra để tách dữ liệu thuộc hai lớp khác nhau.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất:



Hình 2.1 Mô tả phương pháp SVM

Đường tô đậm là siêu phẳng tốt nhất và các điểm được bao bởi hình chữ nhật là những điểm gần siêu phẳng nhất, chúng được gọi là các vector hỗ trợ (support vector). Các đường nét đứt mà các support vector nằm trên đó được gọi là lề (margin).

1. Cơ sở lý thuyết

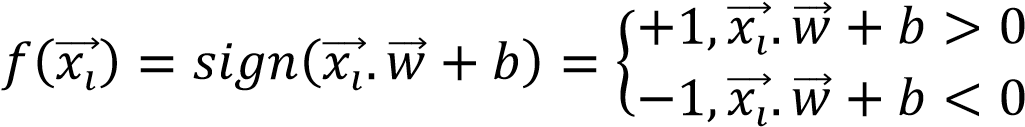
SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

Cho tập mẫu  với , thuộc vào hai lớp nhãn  là nhãn lớp tương ứng của các  (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vector  trong không gian:



Đặt:



Như vậy,  biểu diễn sự phân lớp của  vào hai lớp như đã nêu.

Ta nói  nếu  thuộc lớp I và  nếu  thuộc lớp II.

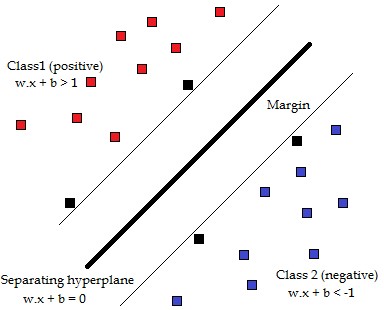
* + 1. ***Bài toán phân 2 lớp với SVM***

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới  thì cần phải xác định  được phân vào lớp  hay lớp .

Ta xét 3 trường hợp, mỗi trường hợp sẽ có 1 bài toán tối ưu, giải được bài toán tối ưu đó ta sẽ tìm được siêu phẳng cần tìm.

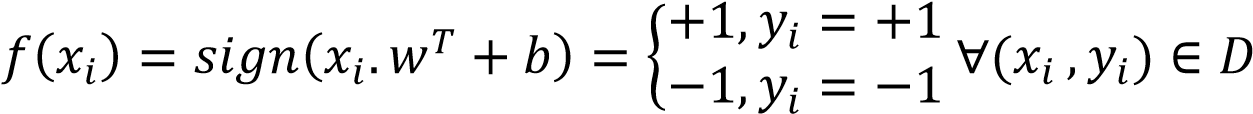
**Trường hợp 1:**

Tập D có thể phân chia tuyến tính được mà không có nhiễu (tất cả các điểm được gán nhãn +1 thuộc về phía dương của siêu phảng, tất cả các điểm được gán nhãn -1 thuộc về phía âm của siêu phẳng).

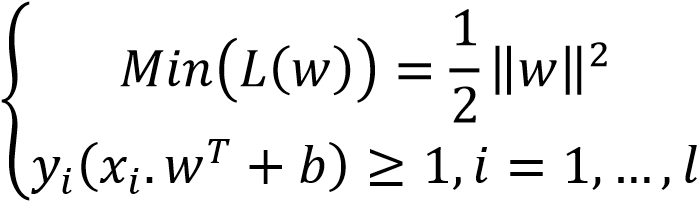


Hình 2.2 Tập dữ liệu được phân chia tuyến tính

Ta sẽ tìm siêu phẳng tách với 𝑤 ∈ ℝ𝑛 là vector trọng số, 𝑏 ∈ ℝ𝑛 là hệ số tự do, sao cho:

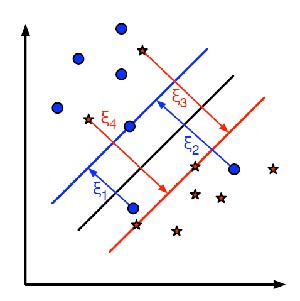


Lúc này ta cần giải bài toán tối ưu:



**Trường hợp 2:**

Tập dữ liệu D có thể phân chia tuyến tính được nhưng có nhiễu. Trong trường hợp này, hầu hết các điểm đều được phân chia đúng bởi siêu phẳng. Tuy nhiên có 1 số điểm bị nhiễu, nghĩa là: Điểm có nhãn dương nhưng lại thuộc phía âm của siêu phẳng, điểm có nhãn âm nhưng lại thuộc phía dương của siêu phẳng.

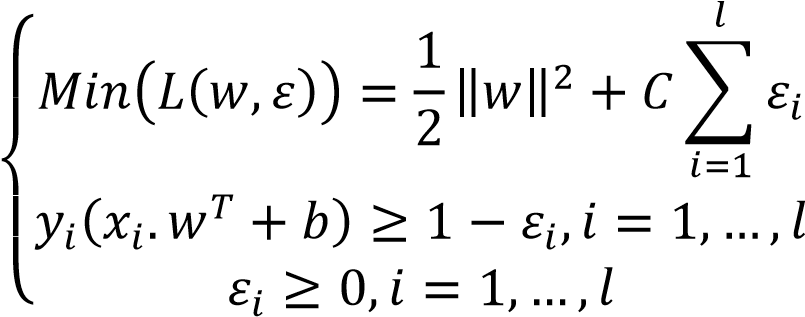


Hình 2.3 Tập dữ liệu được phân chia nhưng có nhiễu

Trong trường hợp này, ta sử dụng 1 biến mềm  sao cho: 



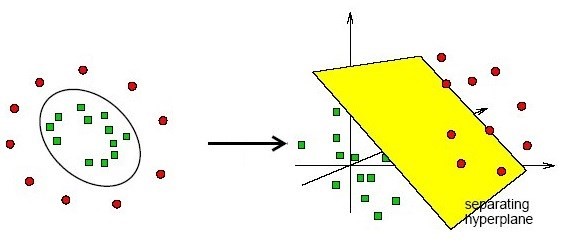
Bài toán tối ưu trở thành:



Trong đó C là tham số xác định trước, định nghĩa giá trị ràng buộc, C càng lớn thì mức độ vi phạm đối với những lỗi thực nghiệm (là lỗi xảy ra lúc huấn luyện, tính bằng thương số của số phần tử lỗi và tổng số phần tử huấn luyện) càng cao.

**Trường hợp 3:**

Tập dữ liệu D không thể phân chia tuyến tính được, ta sẽ ánh xạ các vector dữ liệu x từ không gian n chiều vào một không gian m chiều (m>n), sao cho trong không gian m chiều, D có thể phân chia tuyến tính được.

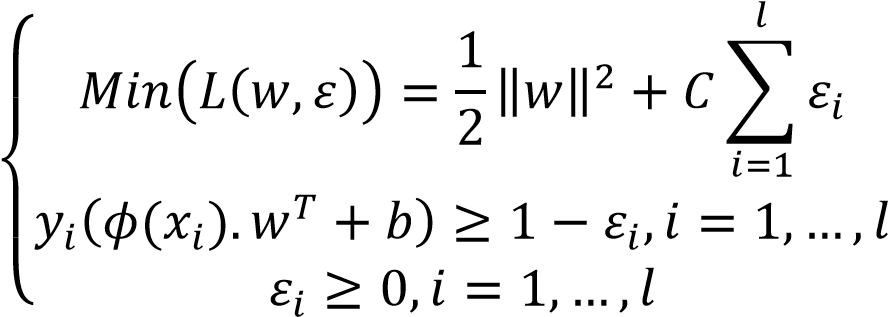


Hình 2.4 Tập dữ liệu không phân chia tuyến tính

Gọi  là một ánh xạ phi tuyến từ không gian  vào không gian .

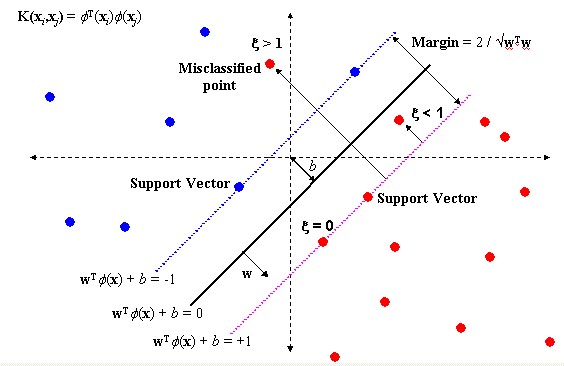


Bài toán tối ưu trở thành:



Ví dụ:

Để dễ hiểu hơn ta xét ví dụ mô tả hình học sau: Xét trong không gian 2 chiều (n=2), tập dữ liệu được cho bởi tập các điểm trên mặt phẳng.



Hình 2.5 Ví dụ biểu diễn tập dữ liệu trên không gian 2 chiều

Bây giờ ta tiến hành tìm siêu phẳng phân lớp dựa trên phương pháp SVM (1). Ta sẽ tìm 2 siêu phẳng song song (nét đứt trong hình …) sao cho khoảng cách giữa chúng là lớn nhất để có thể phân tách lớp này thành 2 phía (Ta gọi là 2 siêu phẳng phân tách). Siêu phẳng (1) nằm giữa 2 siêu phẳng trên (nét đậm trong hình).

Hình trên cho ta tập dữ liệu có thể phân tách tuyến tính. Bây giờ ta xét trường hợp tập dữ liệu không thể phân tách tuyến tính. Bây giờ ta sẽ xử lý bằng cách ánh xạ tập dữ liệu đã cho vào một không gian mới có số chiều lớn hơn không gian cũ (Gọi là không gian đặc trưng) mà trong không gian này tập dữ liệu có thể phân tách tuyến tính. Trong không gian đặc trưng ta sẽ tiếp tục tìm 2 siêu phẳng phân tách như trường hợp ban đầu.

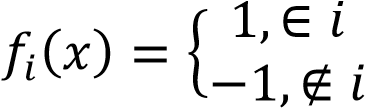
Các điểm nằm trên 2 siêu phẳng phân tách gọi là các vector hỗ trợ (Support vector). Các điểm này quyết định hàm phân tách dữ liệu. Từ đây, chúng ta có thể thấy phương pháp SVM không phụ thuộc vào các mẫu dữ liệu ban đầu, mà chỉ phụ thuộc vào các suport vector (quyết định 2 siêu phẳng phân tách). Cho dù các điểm khác bị xoá thì thuật toán vẫn cho ra các kết quả tương tự. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác do các điểm trong tập dữ liệu đều được dùng để tối ưu kết quả.

* + 1. ***Bài toán phân đa lớp với SVM***

Để phân đa lớp thì kỹ thuật SVM sẽ chia không gian dữ liệu thành 2 phần và tiếp tục với không gian đã được phân chia. Khi đó hàm quyết định phân dữ liệu vào lớp thứ i sẽ là:

𝑓𝑖(𝑥) = 𝑤𝑖𝑇𝑥 + 𝑏𝑖

Những phần tử x là support vector nếu thoả điều kiện:



Giả sử bài toán phân loại k lớp (𝑘 ≥ 2), ta sẽ tiến hành 𝑘(𝑘 − 1)/2 lần phân lớp nhị phân sử dụng phương pháp SVM. Mỗi lớp sẽ tiến hành phân tách với k-1 lớp còn lại để xác định k-1 hàm phân tách (chiến lược “một-đối-một” (one-against-one). Kỹ thuật phân đa lớp bằng phương pháp SVM hiện vẫn đang được tiếp tục nghiên cứu và phát triển.

* + 1. ***Các bước chính của phương pháp SVM***
* Tiền xử lý dữ liệu: Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thực thì ta cần phải tìm cách chuyển chúng về dạng số của SVM. Tránh các số quá lớn, thường nên co giãn dữ liệu để chuyển về đoạn [-1,1] hoặc [0,1].
* Chọn hàm hạt nhân: Cần chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.
* Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng.
* Sử dụng các tham số cho việc huấn luyện tập mẫu.
* Kiểm thử tập dữ liệu Test.
  + 1. ***Ưu điểm phương pháp SVM trong phân lớp dữ liệu***

Như đã biết, phân lớp dữ liệu là một tiến trình đưa các dữ liệu chưa biết nhãn vào các lớp dữ liệu đã biết nhãn tương ứng. Mỗi nhãn được xác định bởi một số tập dữ liệu mẫu của nhãn đó. Để thực hiện quá trình phân lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các bản ghi mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những bản ghi mới chưa biết nhãn.

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 dữ liệu phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các uớc 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…).

Trong công trình của mình năm 1999, Joachims đã so sánh SVM với Naïve Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003, Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính được đề cập trước đây của tập dữ liệu. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác.

Theo Xiaojin Zhu thì trong các công trình nghiên cứu của nhiều tác giả (chẳng hạn như Kiritchenko và Matwin vào năm 2001, Hwanjo Yu và Han vào năm 2003, Lewis vào năm 2004) đã chỉ ra rằng thuật toán SVM đem lại kết quả tốt nhất phân lớp văn bản.

* 1. **Giới thiệu về dữ liệu sử dụng**
     1. ***Giới thiệu thuật toán***

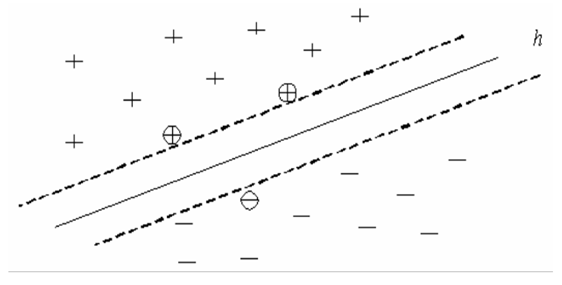
Máy vectơ hỗ trợ (SVM - viết tắt tên tiếng Anh support vector machine) là một khái niệm trong [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) và [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) cho một tập hợp các phương pháp [học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t) liên quan đến nhau để [phân loại](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_%28h%E1%BB%8Dc_m%C3%A1y%29&action=edit&redlink=1) và [phân tích hồi quy](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy).

SVM được định nghĩa như sau:

* Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác.
* Là thuật toán học giám sát (supervied learning) được sử dụng cho phân lớp dữ liệu.
* Là 1 phương pháp thử nghiệm, đưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu.
  + 1. ***Ý tưởng thuật toán***

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 phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp + và lớp -. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt, điều này được minh họa như sau:



**Hình 2.10: Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp + và – với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ (Support Vector được khoanh tròn)**

* + 1. ***Thuật toán***

Ban đầu, thuật toán SVM được thiết kế cho bài toán phân lớp nhị phân. Xét một ví dụ của bài toán phân lớp như hình vẽ; ở đó ta phải tìm một đường thẳng sao cho bên trái nó toàn là các điểm đỏ, bên phải nó toàn là các điểm xanh. Bài toán mà dùng đường thẳng để phân chia này được gọi là phân lớp tuyến tính (linear classification).



Hình 2.11: Phân lớp nhị phân

Hàm tuyến tính phân biệt hai lớp như sau:

(1)

Trong đó:

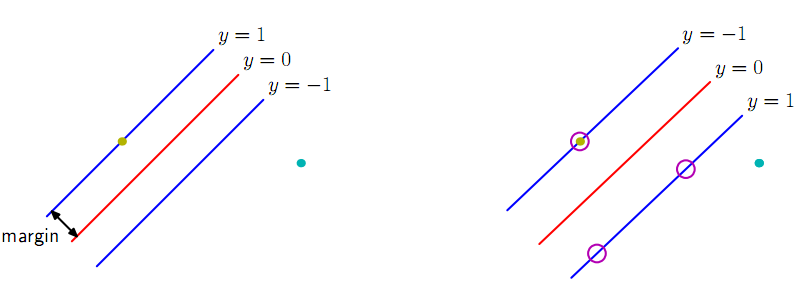
* là vector trọng số hay vector chuẩn của siêu phẳng phân cách, T là kí hiệu chuyển vị.
* là độ lệch
* là vector đặc trưng, làm hàm ánh xạ từ không gian đầu vào sang không gian đặc trưng.

Tập dữ liệu đầu vào gồm N mẫu input vector {x1, x2,...,xN}, với các giá trị nhãn tương ứng là {t1,…,tN} trong đó .

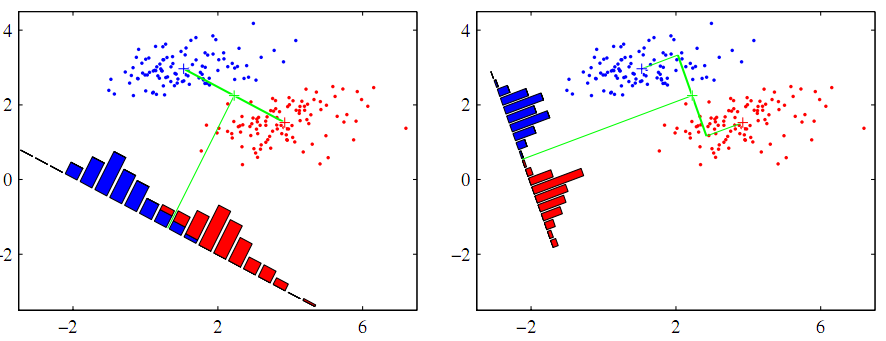
Lưu ý cách dùng từ ở đây: điểm dữ liệu, mẫu… đều được hiểu là input vector xi; nếu là không gian 2 chiều thì đường phân cách là đường thẳng, nhưng trong không gian đa chiều thì gọi đó là siêu phẳng.

Giả sử tập dữ liệu của ta có thể phân tách tuyến tính hoàn toàn (các mẫu đều được phân đúng lớp) trong không gian đặc trưng (feature space), do đó sẽ tồn tại giá trị tham số w và b theo (1) thỏa cho những điểm có nhãn và cho những điểm có , vì thế mà cho mọi điểm dữ liệu huấn luyện.

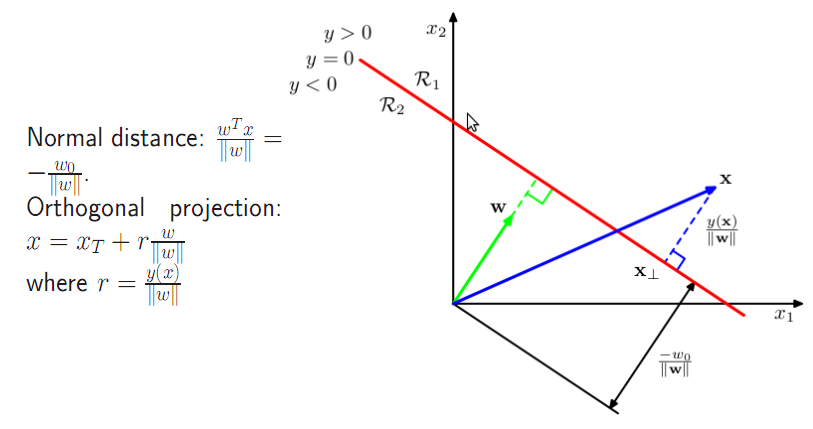
SVM tiếp cận giải quyết vấn đề này thông qua khái niệm gọi là lề, đường biên… (margin). Lề được chọn là khoảng cách nhỏ nhất từ đường phân cách đến mọi điểm dữ liệu hay là khoảng cách từ đường phân cách đến những điểm gần nhất.



Trong SVM, đường phân lớp tốt nhất chính là đường có khoảng cách margin lớn nhất (tức là sẽ tồn tại rất nhiều đường phân cách xoay theo các phương khác nhau, và ta chọn ra đường phân cách mà có khoảng cách margin là lớn nhất).



Ta có công thức tính khoảng cách từ điểm dữ liệu đến mặt phân cách như sau:



Do ta đang xét trong trường hợp các điểm dữ liệu đều được phân lớp đúng nên cho mọi n. Vì thế khoảng cách từ điểm xn đến mặt phân cách được viết lại như sau:

(2)

Lề là khoảng cách vuông góc đến điểm dữ liệu gần nhất xn từ tập dữ liệu, và chúng ta muốn tìm giá trị tối ưu của w và b bằng cách cực đại khoảng cách này. Vấn đề cần giải quyết sẽ được viết lại dưới dạng công thức sau:

(3)

Chúng ta có thể đem nhân tử ra ngoài bởi vì w không phụ thuộc n. Giải quyết vấn đề này một cách trực tiếp sẽ rất phức tạp, do đó ta sẽ chuyển nó về một vấn đề tương đương dễ giải quyết hơn. Ta sẽ scale và cho mọi điểm dữ liệu, từ đây khoảng cách lề trở thành 1, việc biến đổi này không làm thay đổi bản chất vấn đề.

(4)

Từ bây giờ, các điểm dữ liệu sẽ thỏa ràng buộc:

(5)

Vấn đề tối ưu yêu cầu ta cực đại được chuyển thành cực tiểu , ta viết lại công thức:

(6)

Việc nhân hệ số ½ sẽ giúp thuận lợi cho lấy đạo hàm về sau.

Lý thuyết Nhân tử Lagrange:

Vấn đề cực đại hàm f(x) thỏa điều kiện sẽ được viết lại dưới dạng tối ưu của hàm Lagrange như sau:

Trong đó x và λ phải thỏa điều kiện Karush-Kuhn-Tucker (KKT) như sau:

Nếu là cực tiểu hàm f(x) thì hàm Lagrange sẽ là:

Để giải quyết bài toán trên, ta viết lại theo hàm Lagrange như sau:

(7)

Trong đó là nhân tử Lagrange.

Lưu ý dấu (–) trong hàm Lagrange, bởi vì ta cực tiểu theo biến w và b, và là cực đại theo biến a.

Lấy đạo hàm L(w,b,a) theo w và b ta có:

(8)

(9)

Loại bỏ w và b ra khỏi L(w,b,a) bằng cách thế (8), (9) vào. Điều này sẽ dẫn ta đến vấn đề tối ưu:

(10)

Thỏa các ràng buộc:

(11)

(12)

Ở đây hàm nhân (kernel function) được định nghĩa là .

Vấn đề tạm thời gác lại ở đây, ta sẽ thảo luận kỹ thuật giải quyết (10) thỏa (11), (12) này sau.

Để phân lớp cho 1 điểm dữ liệu mới dùng mô hình đã huấn luyện, ta tính dấu của y(x) theo công thức (1), nhưng thế w trong (8) vào:

(13)

Thỏa các điều kiện KKT sau:

(14)

(15)

(16)

Vì thế với mọi điểm dữ liệu, hoặc là hoặc là . Những điểm dữ liệu mà có sẽ không xuất hiện trong (13) và do đó mà không đóng góp trong việc dự đoán điểm dữ liệu mới.

Những điểm dữ liệu còn lại được gọi là support vector, chúng thỏa , đó là những điểm nằm trên lề của siêu phẳng trong không gian đặc trưng.

Support vector chính là cái mà ta quan tâm trong quá trình huấn luyện của SVM. Việc phân lớp cho một điểm dữ liệu mới sẽ chỉ phụ thuộc vào các support vector.

Giả sử rằng ta đã giải quyết được vấn đề (10) và tìm được giá trị nhân tử a, bây giờ ta cần xác định tham số b dựa vào các support vector xn có . Thế (13) vào:

(17)

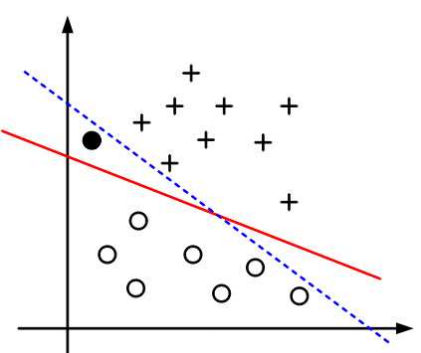
Trong đó S là tập các support vector. Mặc dù ta chỉ cần thế một điểm support vector xn vào là có thể tìm ra b, nhưng để đảm bảo tính ổn định của b ta sẽ tính b theo cách lấy giá trị trung bình dựa trên các support vector.

Đầu tiên ta nhân tn vào (17) (lưu ý , và giá trị b sẽ là:

(18)

Trong đó Ns là tổng số support vector.

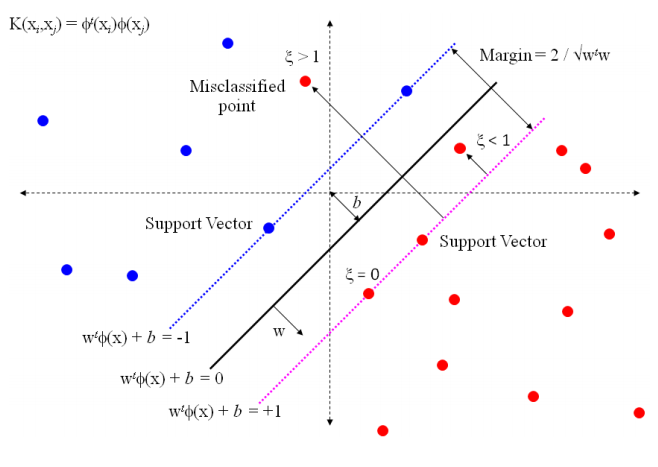
Ban đầu để dễ trình bày thuật toán ta đã giả sử là các điểm dữ liệu có thể phân tách hoàn toàn trong không gian đặc trưng . Nhưng việc phân tách hoàn toàn này có thể dẫn đến khả năng tổng quát hóa kém, vì thực tế một số mẫu trong quá trình thu thập dữ liệu có thể bị gán nhãn sai, nếu ta cố tình phân tách hoàn toàn sẽ làm cho mô hình dự đoán quá khớp.



Để chống lại sự quá khớp, chúng ta chấp nhận cho một vài điểm bị phân lớp sai.

Để làm điều này, ta dùng các biến slack variables cho mọi điểm dữ liệu.

* cho những điểm nằm trên lề hoặc phía trong của lề
* cho những điểm còn lại.
* Do đó những điểm nằm trên đường phân cách sẽ có
* Còn những điểm phân lớp sai sẽ có



Công thức (5) sẽ viết lại như sau:

(20)

Mục tiêu của ta bây giờ là cực đại khoảng cách lề, nhưng đồng thời cũng đảm bảo tính mềm mỏng cho những điểm bị phân lớp sai. Ta viết lại vấn đề cần cực tiểu:

(21)

Trong đó C > 0 đóng vai trò quyết định đặt tầm quan trọng vào biến hay là lề.

Bây giờ chúng ta cần cực tiểu (21) thỏa ràng buộc (20) và . Theo Lagrange ta viết lại:

(22)

Trong đó và là các nhân tử Lagrange.

Các điều kiện KKT cần thỏa là:

(23)

(24)

(25)

(26)

(27)

(28)

Với n = 1,…,N

Lấy đạo hàm (22) theo w, b và {}:

(29)

(30)

(31)

Thế (29), (30), (31) vào (22) ta được:

(32)

Từ (23), (26) và (31) ta có:

Vấn đề cần tối ưu giống hệt với trường hợp phân tách hoàn toàn, chỉ có điều kiện ràng buộc khác biệt như sau:

(33)

(34)

Thế (29) vào (1), ta sẽ thấy để dự đoán cho một điểm dữ liệu mới tương tự như (13).

Như trước đó, tập các điểm có không có đóng góp gì cho việc dự đoán điểm dữ liệu mới.

Những điểm còn lại tạo thành các support vector. Những điểm có và theo (25) thỏa:

(35)

Nếu theo (31) có , từ (28) suy ra và đó là những điểm nằm trên lề.

Những điểm có có thể là những điểm phân lớp đúng nằm giữa lề và đường phân cách nếu hoặc có thể là phân lớp sai nếu

Để xác định tham số b trong (1) ta sẽ dùng những support vector mà có vì thế :

(36)

Lần nữa, để đảm bảo tính ổn định của b ta tính theo trung bình:

(37)

Trong đó M là tập các điểm có

Để giải quyết (10) và (32) ta dùng thuật toán Sequential Minimal Optimization (SMO) do Platt đưa ra vào 1999. Thuật toán này sử dụng tập dữ liệu huấn luyện (còn gọi là tập làm việc) có kích thước nhỏ nhất bao gồm hai hệ số Lagrange. Bài toán quy hoạch toàn phương nhỏ nhất phải gồm hai hệ số Lagrange vì các hệ số Lagrange phải thỏa mãn ràng buộc đẳng thức (34). Phương pháp SMO cũng có một số heuristic cho việc chọn hai hệ số Lagrange để tối ưu hóa ở mỗi bước. Mặc dù có nhiều bài toán quy hoạch toàn phương con hơn so với các phương pháp khác, mỗi bài toán con này được giải rất nhanh dẫn đến bài toán quy hoạch toàn phương tổng thể cũng được giải một cách nhanh chóng.

## Phân loại đa lớp nội dung văn bản với SVM

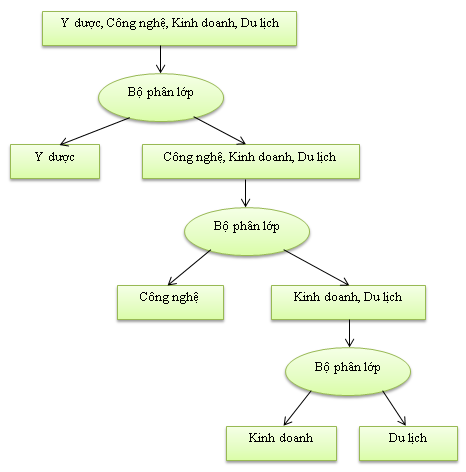
Bây giờ xét đến trường hợp phân nhiều lớp K > 2. Chúng ta có thể xây dựng việc phân K-class dựa trên việc kết hợp một số đường phân 2 lớp.

Sau đây ta tìm hiểu các chiến lược áp dụng trong bài toán phân lớp văn bản thuộc nhiều chủ đề khác nhau. Ý tưởng của bài toán phân lớp đa lớp là chuyển về bài toán phân lớp hai lớp bằng cách xây dựng nhiều bộ phân lớp hai lớp để giải quyết. Các chiến lược phân lớp đa lớp phổ biến này là: One-against-One (OAO), và One-against-Rest (OAR).

*Chiến lược One-against-Rest*

Trong chiến lược OAR, ta sẽ dùng K-1 bộ phân lớp nhị phân để xây dựng K-class. Bài toán phân lớp K lớp được chuyển thành K-1 bài toán phân lớp hai lớp. Trong đó bộ phân lớp hai lớp thứ I được xây dựng trên lớp thứ i và tất cả các lớp còn lại. Hàm quyết định thứ I dùng để phân lớp thứ i và những lớp còn lại có dạng:

Siêu phẳng yi(x) = 0 hình thành siêu phẳng phân chia tối ưu, các support vector thuộc lớp i thỏa yi(x) = 1 và các support vector thuộc lớp còn lại thỏa yi(x) = -1. Nếu vector dữ liệu x thỏa mãn điều kiện yi(x)>0 đối với duy nhất một I, x sẽ được phân vào lớp thứ i.



**Hình** 2.12**: Ví dụ phân lớp văn bản thuộc các chủ đề chiến lược OAR**

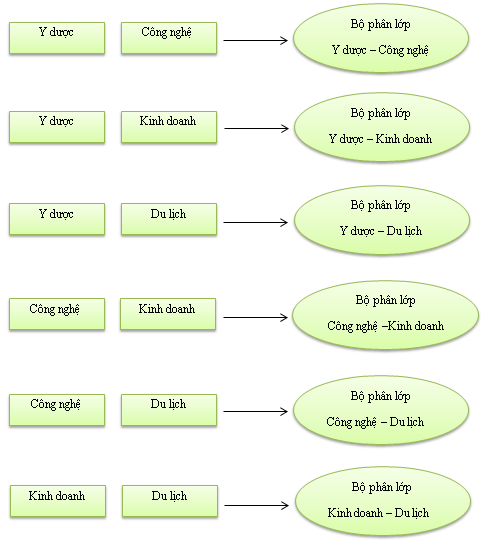
*Chiến lược One-against-One*

Chiến lược OAO, dùng K(K-1)/2 bộ phân lớp nhị phân được xây dựng bằng cách bắt cặp hai lớp một nên chiến lược này còn được gọi là pairwise và sử dụng theo phương pháp đa số kết lợp các bộ phận lớp này để xác định được kết quả phân lớp cuối cùng. Số lượng các bộ phân lớp không bao giờ vượt quá K(K-1)/2.

So với chiến lược OAR thì chiến lược này ngoài ưu điểm giảm bớt vùng không thể phân lớp mà còn làm tăng độ chính xác của việc phân lớp. Trong chiến lược OAR ta phải xây dựng một siêu phẳng để tách một lớp ra khỏi các lớp còn lại, việc này đòi hỏi sự phức tạp và có thể không chính xác. Tuy nhiên trong chiến lược OAO ta chỉ cần tách một lớp ra khỏi một lớp khác mà thôi.

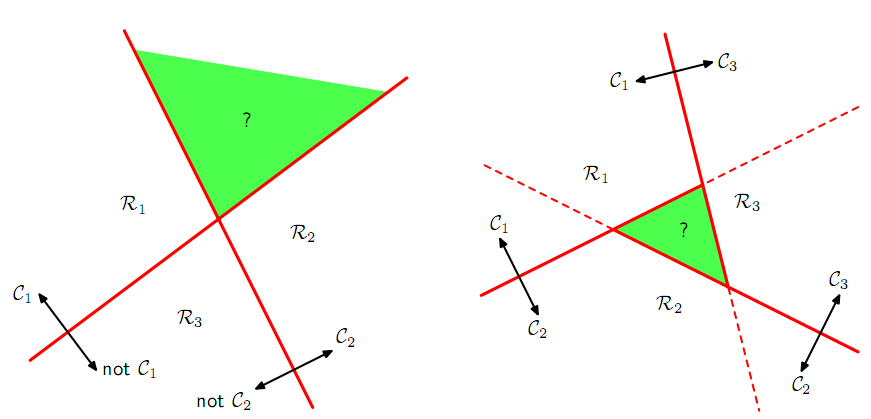
Chiến lược OAR chỉ cần K-1 bộ phân lớp cho K lớp. Trong khi đó chiến lược OAO lại cần đến K(K-1)/2 bộ phân lớp. Nhưng số mẫu huấn luyện cho từng bộ phân lớp trong OAO lại ít hơn và việc phân lớp cũng đơn giản hơn. Vì vậy chiến lược OAO có độ chính xác cao hơn nhưng chi phí để xây dựng lại tương đương với chiến lược OAR.

Hàm quyết định phân lớp của lớp i đối với lớp j trong chiến lược OAO là:



**Hình** 2.13**: Ví dụ phân lớp văn bản thuộc các chủ đề chiến lược OAO**

Tuy nhiên cả hai chiến lược trên đều dẫn đến vùng mập mờ trong phân lớp (như hình 2.14).



**Hình 2.14: Vùng mập mờ trong phân lớp**

Ta có thể tránh được vấn đề này bằng cách xây dựng K-Class dựa trên K hàm tuyến tính có dạng:

Và một điểm x được gán vào lớp Ck khi với mọi .

**Kết luận chương:**

Chương này đã giới thiệu chi tiết thuật toán Máy vector hỗ trợ SVM và trình bày cách thu thập, trích rút và xử lý một cách tự động các tin tức tuyển dụng. Sau đó, luận văn sẽ sử dụng SVM để thực hiện việc huấn luyện bộ phân loại cho bài toán phân loại thông tin tuyển dụng trên hệ thống website tuyển dụng.

# Chương 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ

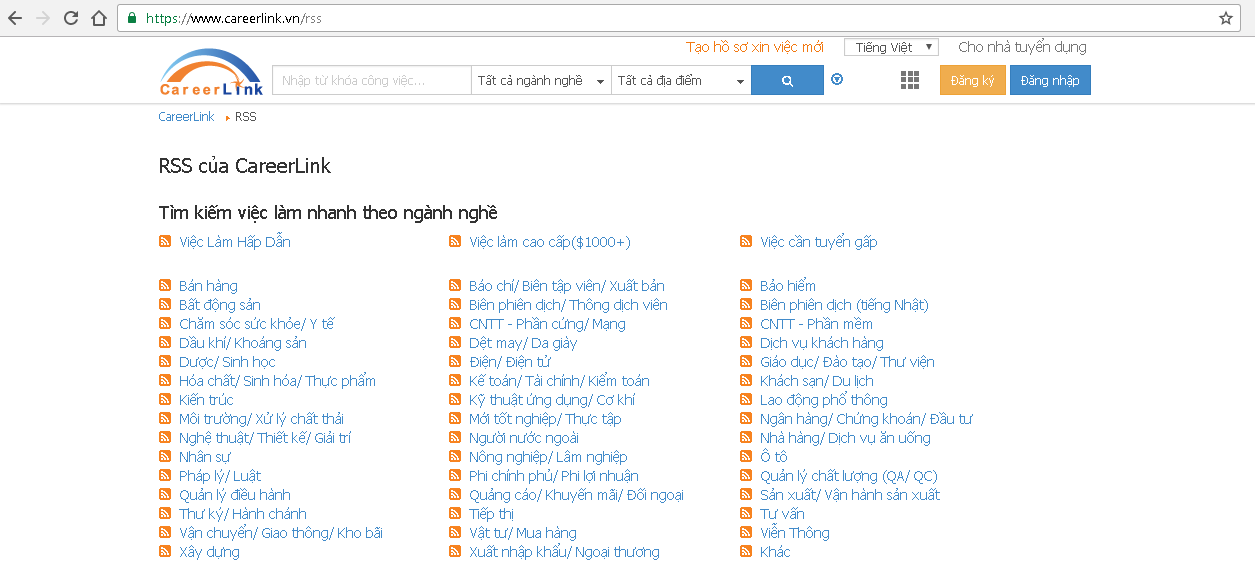


## Xây dựng bộ dữ liệu và tiền xử lý văn bản

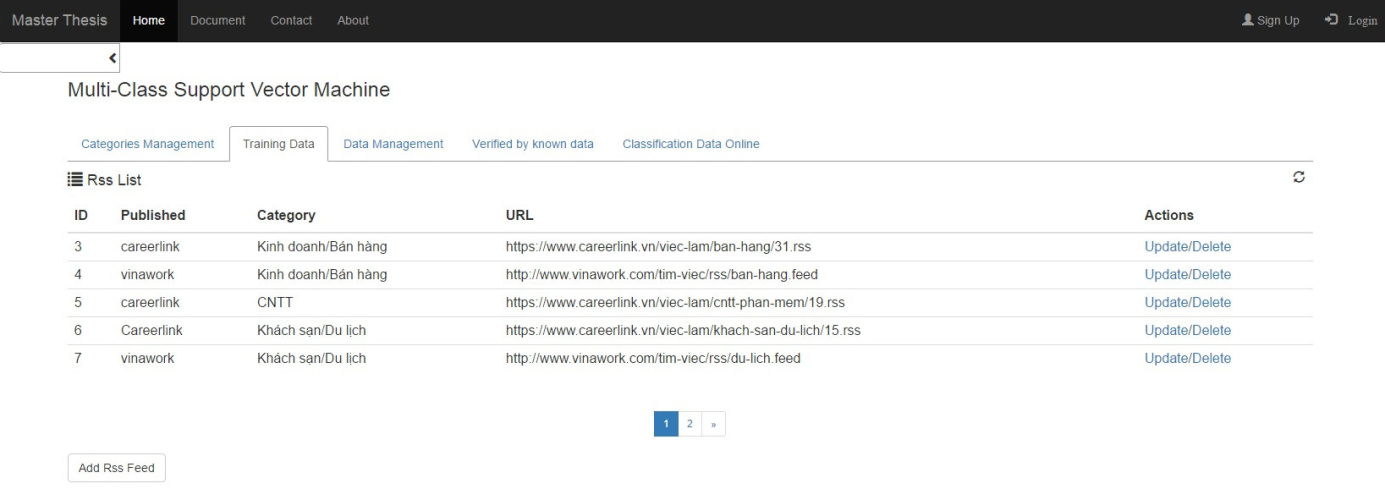
RSS (viết tắt từ Really Simple Syndication) là một tiêu chuẩn định dạng tài liệu dựa trên XML nhằm giúp người sử dụng dễ dàng cập nhật và tra cứu thông tin một cách nhanh chóng và thuận tiện nhất bằng cách tóm lược thông tin vào trong một đoạn dữ liệu ngắn gọn, hợp chuẩn.

Công nghệ của RSS cho phép người dùng Internet có thể đặt mua thông tin từ các websites có cung cấp khả năng RSS (RSS feeds); chúng thường là các site có nội dung thay đổi và được thêm vào thường xuyên. Để có thể dùng công nghệ này, các người quản trị site đó tạo ra hay quản lí một phần mềm chuyên dụng (như là một hệ thống quản lí nội dung - content management system - CMS) mà với định dạng XML máy có thể đọc được, có thể biểu diễn các bài tin mới thành một danh sách, với một hoặc hai dòng cho mỗi bài tin và một liên kết đến bài tin đầy đủ đó. Khác với việc mua nhiều ấn bản của các tờ báo hay tạp chí in giấy, hầu hết việc mua RSS là miễn phí.

Trong phạm vi luận văn, em đã sử dụng lấy tin tự động từ RSS để cập nhật các việc làm hằng ngày nhằm xây dựng bộ dữ liệu huấn luyện. Dữ liệu được lấy từ 2 trang web: <https://www.careerlink.vn/> và <http://www.vinawork.com/>



Hình 3.1: Thông tin RSS trên website <https://www.careerlink.vn/>



Hình 3.2: Danh sách các RSS được thu thập phục vụ giai đoạn huấn luyện



Hình 3.3: Các category được sử dụng

Các bài báo việc làm được thu thập được thông qua RSS từ tháng 8/2016. Các bài báo này được lấy về sau đó tiến hành tiền xử lý, cuối cùng được lưu dưới dạng file phục vụ cho huấn luyện và kiểm tra. Dữ liệu thu thập được gồm có 1409 bài báo, trong đó chọn 800 bài dùng cho huấn luyện và 609 bài dùng cho việc kiểm tra. Cụ thể số liệu như sau:

Bảng 3.1: Số lượng tập dữ liệu huấn luyện và kiểm tra

|  |  |  |
| --- | --- | --- |
| Chủ đề | Tập dữ liệu huấn luyện | Tập dữ liệu kiểm tra |
| Kinh doanh/Bán hàng | **200** | **233** |
| Công nghệ thông tin (CNTT) | **200** | **179** |
| Nhân sự | **200** | **117** |
| Khách sạn/Du lịch | **200** | **80** |

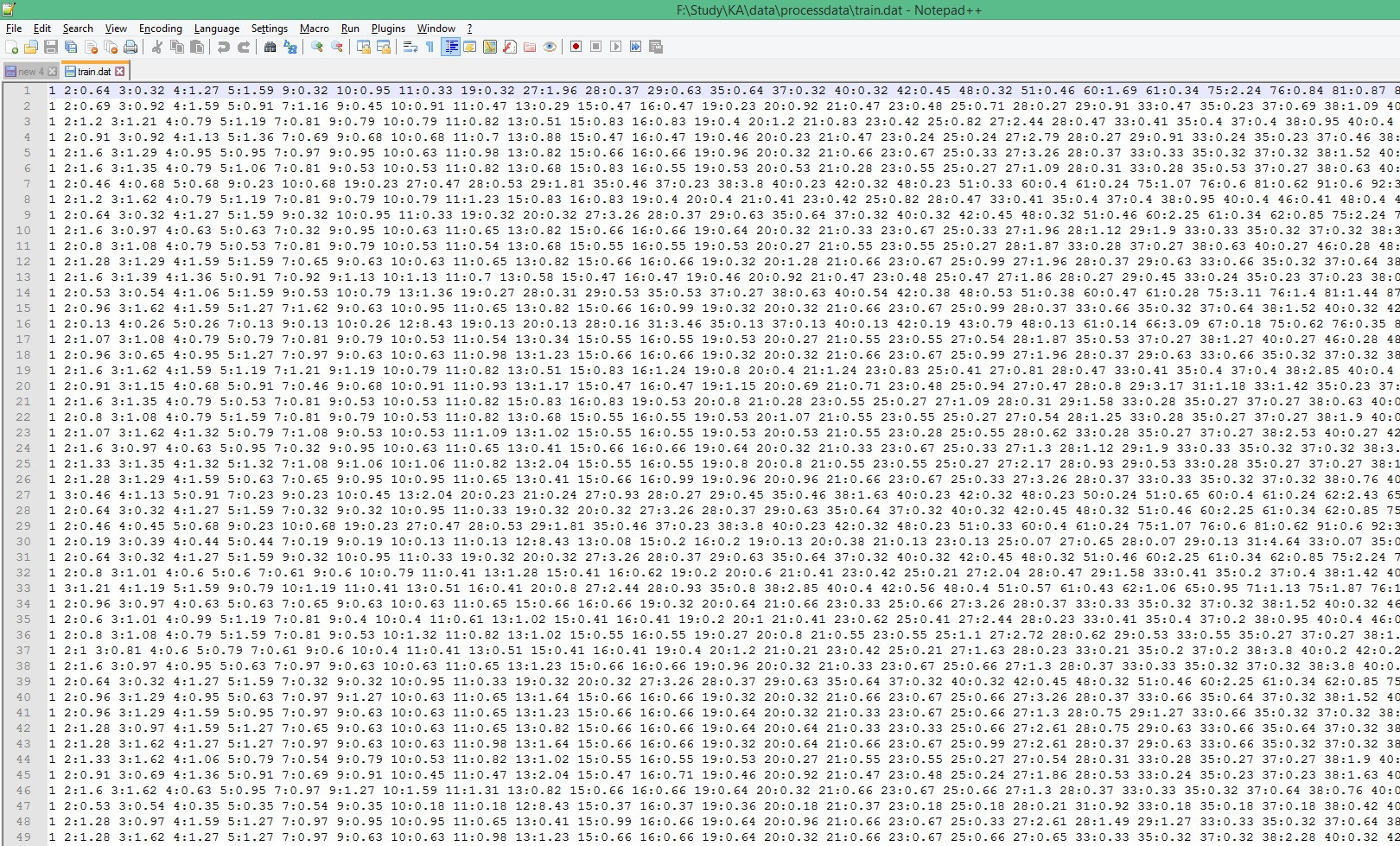
Sau các bước tiền xử lý, văn bản được biểu diễn dưới dạng:

<classi> <label1>:<value1> <label2>:<value2> ... <labeln>:<valuen>

Trong đó:

* classi là nhãn phân loại của mỗi chủ đề.
* labeli là chỉ số của từ đặc trưng thứ i trong không gian đặc trưng có xuất hiện ở phản hồi đó.
* valuei là trọng số của indexi được tính bởi công thức TF.IDF, nếu valuei = 0 thì không cần phải ghi đặc trưng đó.

Định dạng này tuân theo định dạng dữ liệu đầu vào của chương trình SVMMulticlass.



Hình 3.4: Biểu diễn văn bản dưới dạng vector

*Đánh giá độ chính xác của thuật toán:*

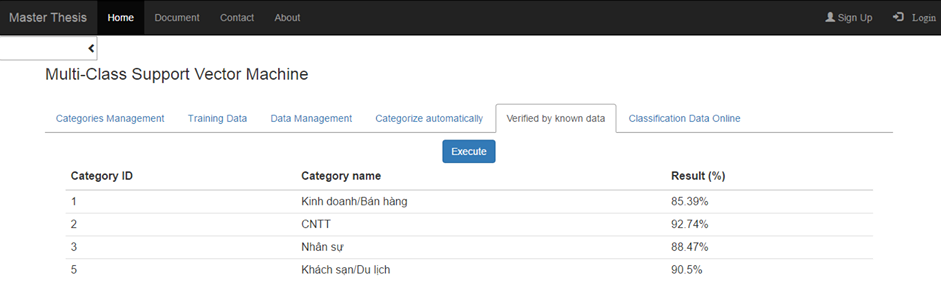
Chúng ta không thể khẳng định một phương pháp phân lớp văn bản cụ thể nào là chính xác hoàn toàn. Vì vậy việc đưa ra độ đo để đánh giá hiệu quả của thuật toán phân lớp giúp chúng ta có thể xác định được độ chính xác của thuật toán, từ đó áp dụng thuật toán đó vào việc phân lớp.

Độ chính xác có thể được tính theo các công thức:

Công thức đánh giá số 1:

Công thức đánh giá số 2:

Sau đây là kết quả thực nghiệm trên bộ test gồm 609 văn bản:



Hình 3.5: Kết quả độ chính xác của việc phân loại

* 1. **Giai đoạn phân lớp**
     1. ***URL phục vụ cho ứng dụng***

Mục đích của ứng dụng tìm kiếm thông tin việc làm đó là lấy những thông tin việc làm từ các website đưa vào hệ thống, sau đó sẽ tiến hành phân loại các việc làm đó theo chủ đề đã có. Danh sách các URL em sử dụng để phục vụ cho ứng dụng như sau:

[http://worklink.vn](http://worklink.vn/?gclid=CjwKEAiAjvrBBRDxm_nRusW3q1QSJAAzRI1tTwJW1a4LIhZQT60REsSbsTqh9dG44LdT6XBYtO2n7hoC5T_w_wcB)

<http://careerbuilder.vn/tim-viec-lam.html>

<https://www.timviecnhanh.com/>

<https://www.vietnamworks.com/tim-viec-lam>

<https://mywork.com.vn/>

* + 1. ***Triển khai ứng dụng***

Mục đích của ứng dụng là xây dựng một danh sách các việc làm được lấy từ các website được liệt kê trong phần 3.2.1 và tiến hành phân loại chúng vào 4 category có sẵn là Kinh doanh/Bán hàng, Công nghệ thông tin, Nhân sự và Khách sạn/Du lịch

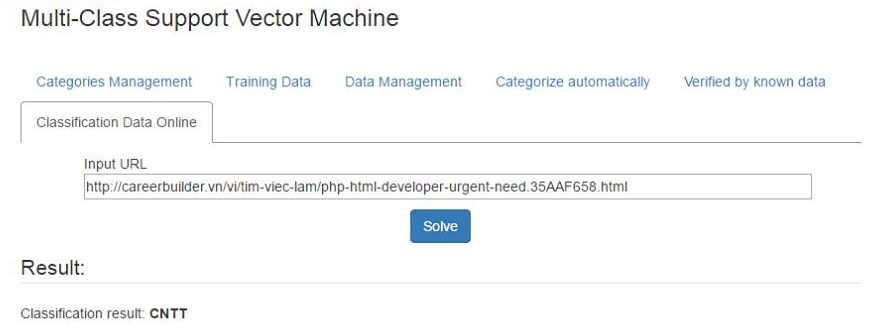
**Bước 1**: Input URL từ các danh sách các website việc làm không support RSS

**Bước 2**: Trích xuất văn bản từ các URL

**Bước 3**: Tiền xử lý văn bản

* Tách từ
* Loại bỏ từ dừng
* Làm sạch văn bản
* Vector hóa văn bản theo định dạng đầu vào của thuật toán SVM.

**Bước 4**: Thực hiện phân loại theo bốn chủ đề



Hình 3.6: Phân loại URL được input vào ứng dụng

**Kết luận chương**

Chương này đã trình bày cụ thể về việc xây dựng bộ dữ liệu huấn luyện và thử nghiệm, từ đó tính toán được độ chính xác của thuật toán; và đưa ra ứng dụng về việc phân loại URLs của các website việc làm vào nhãn tương ứng.

**KẾT LUẬN**

1. **Kết quả đạt được**

Luận văn tiến hành nghiên cứu giải quyết bài toán dự đoán giới tính người dùng Internet dựa trên lịch sử truy cập. Bài toán là nền tảng cho nhiều ứng dụng quan trọng thực tế như quảng cáo nhắm mục tiêu, các hệ thống cung cấp tiếp thị dịch vụ thương mại tới đúng người dùng, …

Những kết quả chính mà đồ án đạt được:

* Nghiên cứu và tìm hiểu về bài toán phân lớp, trình bày một số thuật toán phân lớp tiêu biểu
* Tìm hiểu một số đặc điểm của văn bản tiếng Việt phục vụ cho quá trình tiền xử lý
* Tìm hiểu và áp dụng các công cụ tiền xử lý dữ liệu đầu vào
* Nghiên cứu và tìm hiểu về thuật toán Support Vector Machine trên hai lớp và nhiều lớp
* Xây dựng chương trình lấy tin RSS hàng ngày
* Xây dựng chương trình huấn luyện và kiểm thử với bộ dữ liệu lấy từ chương trình lấy tin RSS.
* Xây dựng chương trình phân loại việc làm vào các category có sẵn từ các website không hỗ trợ RSS

1. **Hạn chế:**

* Một văn bản đầu vào sau khi phân loại chỉ thuộc về một loại nhãn.
* Hạn chế số lượng và chất lượng của kho dữ liệu tin tức ảnh hưởng đến chất lượng phân loại của hệ thống.
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào.

**3. Hướng phát triển**

* Xây dựng bộ dữ liệu lớn hoàn chỉnh, phong phú về các lĩnh vực việc làm và chia các lĩnh vực càng nhỏ càng tốt
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào
* Cải thiện hiệu xuất, tăng tốc độ xử lý dữ liệu
* Kiểm soát được thông tin lấy về theo ngày tháng năm mong muốn đảm bảo tính cập nhật của việc làm

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Do Viet Phuong and Tu Minh Phuong. “*Gender Prediction Using Browsing History”.* [KSE (1) 2013](http://dblp.uni-trier.de/db/conf/kse/kse2013-1.html#PhuongP13): 271-283.
2. Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z. (2007) “*Demographic prediction based onuser’s browsing behavior*”, Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada. [viewed 05.09.2016] Available from:  
   <http://wwwconference.org/www2007/papers/paper686.pdf>
3. Kabbur, S., Han, E.-H., Karypis, G. (2010) “*Content-based methods for predicting website demographic attributes*”, University of Minnesota Supercomputing Institute Research Report UMSI 2010/98 [viewed 06.09.2016]

* Available from:  
  <http://www.dtc.umn.edu/publications/reports/2010_01.pdf>

[4] Speltdoorn, S. (2010) “*Predicting demographic characteristics of web users using semisupervised classification techniques*” Master’s dissertation, Ghent University, Faculty of Economucs and Business Administration. [viewed 14.09.2016] Available from:  
<http://lib.ugent.be/fulltxt/RUG01/001/459/756/RUG01001459756_2011_0001_AC.pdf>

[5] Quanzeng You, Sumit Bhatia, Tong Sun, Jiebo Luo (2014) “*The eyes of the beholder: Gender prediction using images posted in Online Social Networks*”. Available from: <http://www.cs.rochester.edu/u/qyou/papers/gender_classification.pdf>

[6] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla (2014) “*Inferring User Demographics and Social Strategies in Mobile Social Networks*”. Available from: <http://www3.nd.edu/~ydong1/papers/KDD14-Dong-et-al-WhoAmI-demographic-prediction.pdf>

[7] Yan, X., Yan, L.: Gender classification of weblogs authors. In: Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, CA, March 27-29, pp. 228–230 (2006). Available from: <http://aaaipress.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-046.pdf>

[8] Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S. (2012). Demographic prediction based on users mobile behaviors. Mobile Data Challenge. Available from:<http://www.idiap.ch/project/mdc/publications/files/mdc-final241ying.pdf>

[9] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. Available from: <http://www.dongnguyen.nl/publications/nguyen-icwsm2013.pdf>

[10] Zhang, C., Zhang, P. (2010). Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA.

# PHỤ LỤC