**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 8](#_Toc482278250)

[1.1 Tính cấp thiết của đề tài 8](#_Toc482278251)

[1.2 Tổng quan về vấn đề nghiên cứu 8](#_Toc482278252)

[1.3 Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu 9](#_Toc482278253)

[1.4 Cấu trúc luận văn 11](#_Toc482278254)

[Chương 1 – TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET 14](#_Toc482278255)

[1.1. Bài toán xác định giới tính và ứng dụng của bài toán vào thực tiễn 14](#_Toc482278256)

[***1.1.1.*** ***Mở đầu*** 14](#_Toc482278257)

[***1.1.2.*** ***Bài toán xác định giới tính*** 15](#_Toc482278258)

[***1.1.3.*** ***Ứng dụng của bài toán vào thực tiễn*** 18](#_Toc482278259)

[1.2. Các dạng dữ liệu lịch sử có thể dự đoán 19](#_Toc482278260)

[1.3. Các phương pháp xác định giới tính đã có 19](#_Toc482278261)

[***1.3.1.*** ***Phương pháp xác định giới tính sử dụng bài viết từ blog*** 20](#_Toc482278262)

[***1.3.2.*** ***Phương pháp phân loại giới tính dựa trên hành vi và môi trường*** 20](#_Toc482278263)

[***1.3.3.*** ***Phương pháp xác định giới tính sử dụng dữ liệu thông tin di động liên lạc hàng ngày*** 20](#_Toc482278264)

[***1.3.4.*** ***Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy*** 20](#_Toc482278265)

[1.4. Kết luận chương 20](#_Toc482278266)

[Chương 2 – DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET SỬ DỤNG LỊCH SỬ TRUY CẬP 21](#_Toc482278267)

[2.1. Giới thiệu về phương pháp học máy SVM 21](#_Toc482278270)

[***2.1.1.*** ***Giới thiệu về SVM*** 21](#_Toc482278271)

[***2.1.2.*** ***Bài toán phân 2 lớp với SVM*** 22](#_Toc482278272)

[***2.1.3.*** ***Bài toán phân đa lớp với SVM*** 26](#_Toc482278273)

[***2.1.4.*** ***Các bước chính của phương pháp SVM*** 27](#_Toc482278274)

[***2.1.5.*** ***Ưu điểm phương pháp SVM trong phân lớp dữ liệu*** 27](#_Toc482278275)

[2.2. Giới thiệu về dữ liệu sử dụng 28](#_Toc482278276)

[***2.2.1.*** ***Giới thiệu thuật toán*** 28](#_Toc482278277)

[***2.2.2.*** ***Ý tưởng thuật toán*** 28](#_Toc482278278)

[***2.2.3.*** ***Thuật toán*** 28](#_Toc482278279)

[1.5 Phân loại đa lớp nội dung văn bản với SVM 28](#_Toc482278280)

[Chương 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ 33](#_Toc482278281)

[3.1. Xây dựng bộ dữ liệu và tiền xử lý văn bản 33](#_Toc482278285)

[3.2. Giai đoạn phân lớp 36](#_Toc482278286)

[***3.2.1.*** ***URL phục vụ cho ứng dụng*** 36](#_Toc482278287)

[***3.2.2.*** ***Triển khai ứng dụng*** 36](#_Toc482278288)

[KẾT LUẬN 38](#_Toc482278289)

[DANH MỤC TÀI LIỆU THAM KHẢO 40](#_Toc482278290)

[PHỤ LỤC 42](#_Toc482278291)

**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

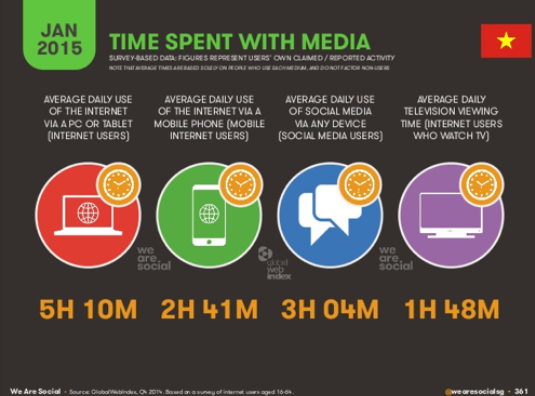
# Chương 1 – TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET

## 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

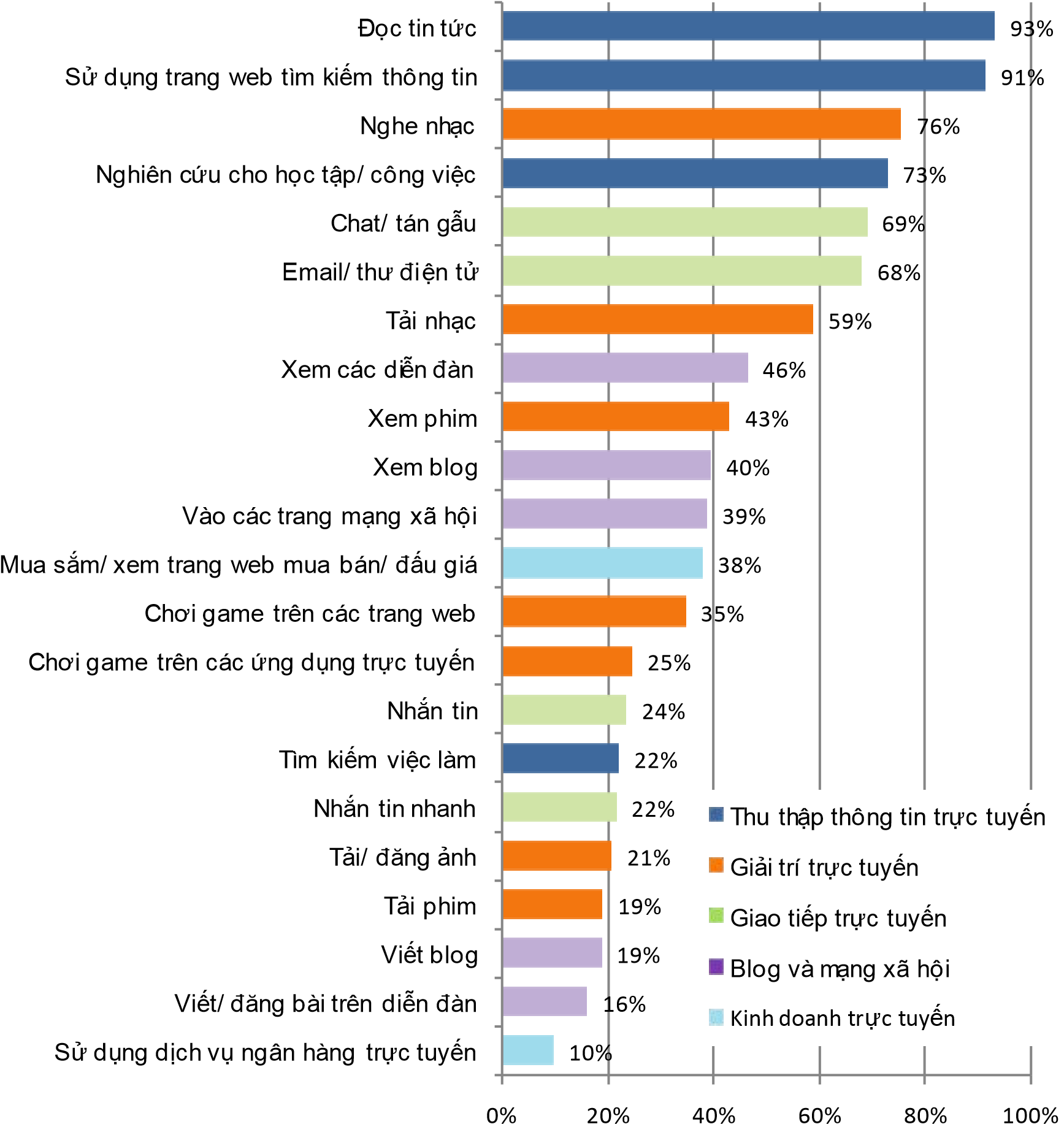
Chính vì sự phát triển không ngừng của công nghệ thông tin và mức độ phổ biến của internet ngày nay mà thông tin đến với người dùng vô cùng phong phú và liên tục. Người sử dụng internet hiện nay thường có thói quen truy cập và tìm kiếm đến những các vấn đề mình quan tâm. Hầu hết các thông tin được lưu 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.

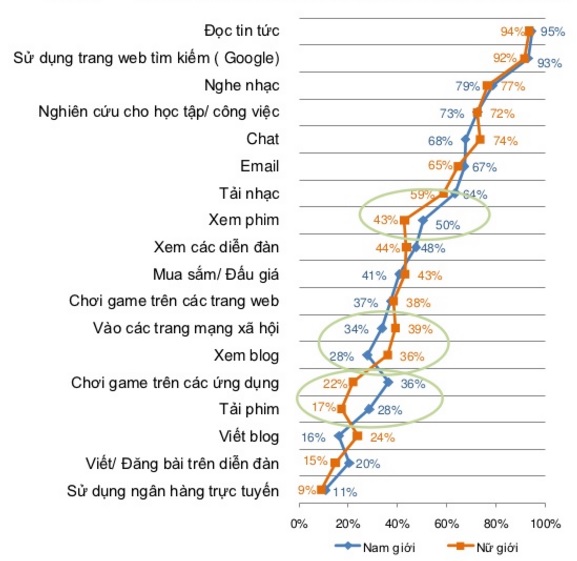
Thương mại điện tử hiện nay, số lượng truy cập đạt mức tăng trưởng đáng kể. Hầu hết các trang phổ biến là các trang web đấu giá và mua bán, nơi có 40% người sử dụng đã từng viếng thăm. Ngân hàng trực tuyến vẫn đang ở giai đoạn sơ khai. Mức độ sử dụng các trang web mua hàng trực tuyến và ngân hàng trực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

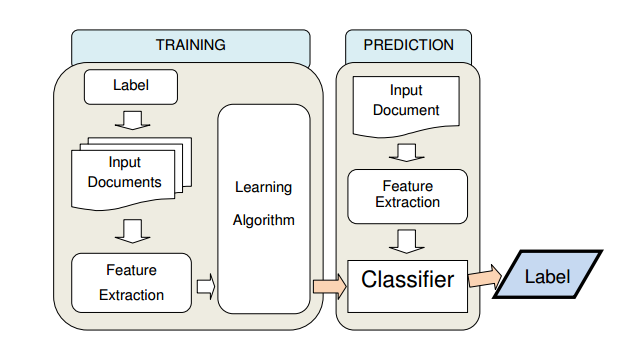


Hình 1.3 Các hoạt động trực tuyến được người dùng mạng theo giới tính

(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***

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, công nghệ ngày càng phát triển, đặc biệt với sự ra đời của các trang mạng xã hội, thương mại điện tử nên lượng thông tin lớn, phi cấu trúc, phức tạp, thậm chí là các thông tin rác cũng rất nhiều. Cần thiết phải có những nghiên cứu để xác định được thông tin gì là cần thiết và thông tin nào là dư thừa. Các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin đều đi tìm câu trả lời cho câu hỏi đó. 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 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 ví dụ như:

* 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 bài viết 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.
* Twitter: Là một mạng xã hội và các thông điệp trên twitter của người dùng được sử dụng như một văn phong, hành vi để xác định xem thông điệp này được viết bởi giới tính nào.
  1. **Các phương pháp xác định giới tính đã có**

Trên thế giới, một số công trình đi trước đã nghiên cứu các phương pháp dựa trên phân tích văn bản như De Vel et al. (2001) đã sử dụng 221 đặc trưng để xác định tác giả của email. Argamon và Koppel et al. (2003) đã nghiên cứu sự khác biệt trong phong cách viết của nam và nữ trong 604 tài liệu của National Corpus của Anh. Schler et al. (2007) khám phá việc sử dụng các đặc trưng và dựa trên nội dung để dự đoán giới tính và độ tuổi của các blogger trên bộ dữ liệu với hơn 71,000 bài viết blog từ blogger.com. Mô hình này đã đạt được kết quả 80% cho dự đoán giới tính và 76% đối với các dự đoán tuổi. Đồng Nguyễn et al. (2013) đã tiến hành một nghiên cứu để dự đoán giới tính và độ tuổi của các thông điệp twitter và diễn đàn bài viết bằng cách sử dụng phương pháp hồi quy với độ chính xác khoảng 80%.

* + 1. ***Phương pháp xác định giới tính sử dụng bài viết từ blog***
    2. ***Phương pháp phân loại giới tính dựa trên hành vi và môi trường***
    3. ***Phương pháp xác định giới tính sử dụng dữ liệu thông tin di động liên lạc hàng ngày***
    4. ***Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy***

1. Giới thiệu

Xác định giới tính sử dụng dữ liệu từ các thông điệp Twitter là phương pháp phân loại cho từng bình luận theo đặc trưng dựa trên nội dung bình luận. Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, ta tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa, biểu diễn văn bản theo mô hình túi từ (được trình bày bởi Dong Nguyen năm 2013).

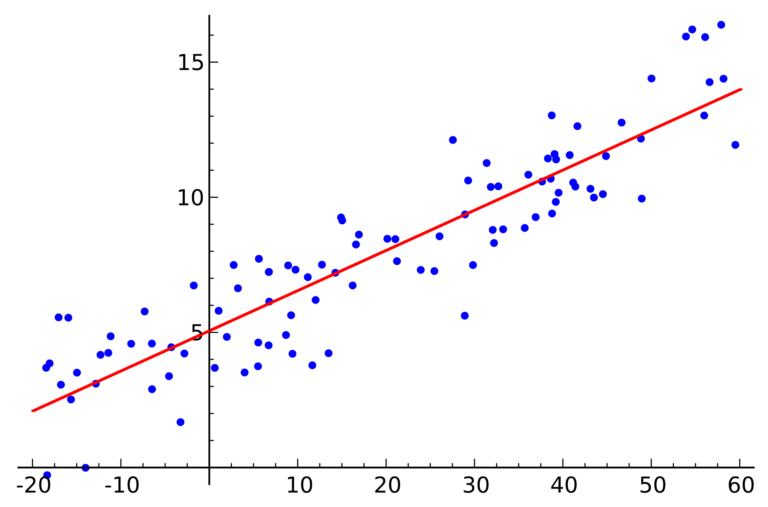
1. Ý tưởng

Đọc nội dung twitter của ai đó, người ta thường nhìn thấy được giới tính của người dùng. Ví dụ, Bạn có thể biết giới tính người dùng phía sau twitter sau đây?

AS LONG AS YOU LOVE ME <3

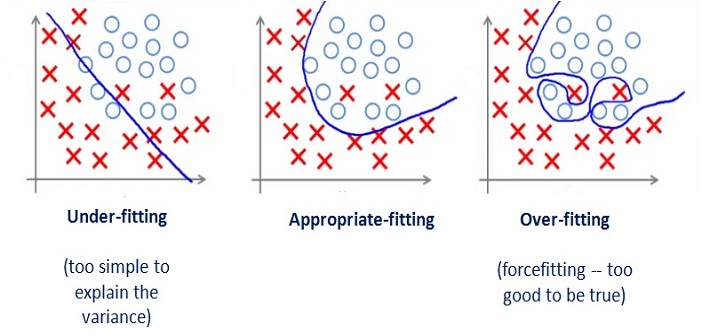
Hồi Quy (regression) là một phương pháp học có giám sát (supervised learning) trong Máy Học. Mục tiêu chính là tìm ra mối quan hệ giữa các đặc trưng của một vấn đề nào đó. Cụ thể hơn, từ một tập dữ liệu cho trước, ta xây dựng một mô hình (phương trình, đồ thị, …) khớp nhất với tập dữ liệu, thể hiện được xu hướng biến thiên và mối quan hệ giữa các đặc trưng. Khi có một mẫu dữ liệu mới vào, dựa vào mô hình, chúng ta có thể dự đoán giá trị của mẫu dữ liệu đó. Lấy ví dụ như chúng ta cần dự đoán **giới tính của một twitter**dựa vào**nội dung** và đặc trưng viết của twitter đó. Như vậy chúng ta cần tìm mối quan hệ giữa **giới tính** phụ thuộc vào **nội dung** và **điểm giữa kỳ**. Dựa vào tập dữ liệu (giả sử thu thập nội dung, đặc trưng viết và các ký tự đặc biệt của 100 người dùng twitter), ta xây dựng một phương trình *y*=*θ*​0​​+*θ*​1​​*x*​1​​+*θ*​2​​*x*​2 trong đó y là giới tính phụ thuộc x1 (nội dung) và x2 (đặc trưng viết). Khi có thêm một mẫu dữ liệu của một người dùng mới, chỉ cần áp vào phương trình như vậy ta sẽ dự đoán được giới tính của người đó.

Ta thấy phương trình y =*θ*​0​​+*θ*​1​​*x*​1​​+*θ*​2​​*x*​2​​ là phương trình của mặt phẳng trong không gian 3 chiều. Những mô hình tương tự như phương trình đường thẳng, phương trình mặt phẳng chính là những mô hình tuyến tính. Hồi quy tuyến tính (linear regression) là một mô hình đơn giản trong bài toán hồi quy, trong đó chúng ta dùng đường thẳng, mặt phẳng, hay phương trình tuyến tính nói chung để dự đoán xu hướng của dữ liệu. Giải bài toán hồi quy tuyến tính chính là đi tìm các tham số *θ*​0​​, *θ*​1​​,...để xác định phương trình tuyến tính.



Hình 1.5 Ví dụ về hồi quy tuyến tính

Một trong những vấn đề gặp phải trong khi chạy mô hình Hồi Quy Tuyến Tính chính là hiện tượng quá khớp (overfitting). Overfitting là vấn đề xảy ra khi mô hình ta tạo ra cố gắng quá mức để khớp với các mẫu trong tập huấn luyện. Mô hình tuy rằng khớp với các mẫu huấn luyện nhưng lại không thể hiện được xu hướng của dữ liệu dẫn đến việc mô hình chỉ đúng với các giá trị trong tập huấn luyện và sai hoàn toàn với các giá trị test.



Hình 1.6 Quá trình khớp

Vấn đề quá khớp thường xảy ra khi bộ dữ liệu twitter của ta có nhiều đặc trưng nhưng lại có ít mẫu dữ liệu. Ví dụ như chúng ta muốn tạo ra một mô hình có dạng đường thẳng tức là cần hai đặc trưng x1, x2 (đặc trưng tọa độ trong mặt phẳng) nhưng lại chỉ có một mẫu dữ liệu được biểu diễn thành một điểm. Để xác định đường thẳng cần ít nhất hai điểm và nếu chỉ có một điểm thì có vô số mô hình phù hợp với mẫu dữ liệu nhưng trong đó chỉ có một mô hình là thật sự đúng với thực tế.

* 1. **Kết luận chương**

**Kết luận chương:**

Chương này đã giới thiệu tổng quan về bài toán xác định giới tính, ứng dụng của bài toán vào thực tiễn và một số phương pháp xác định giới tính và dữ liệu lịch sử liên quan đến việc phân loại giới tính nam hay giới tính nữ. Bên cạnh đó, chương 1 còn đưa ra lý do và thực trạng các hoạt động của người dùng internet trong luận văn. Ngoài ra cần lưu ý đến yếu tố quan trọng tác động đến kết quả phân loại giới tính đó là:

Phải có một tập dữ liệu lịch sử để 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.

# 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

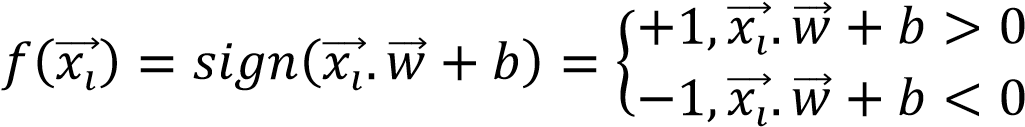
SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

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.

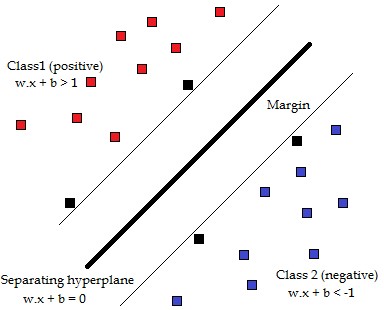
* + 1. ***Bài toán phân 2 lớp với SVM***

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.

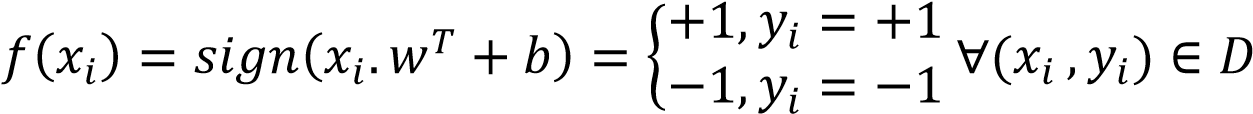
**Trường hợp 1:**

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).

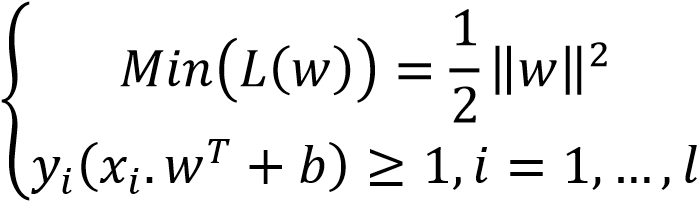


Hình 2.2 Tập dữ liệu được phân chia tuyến tính

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

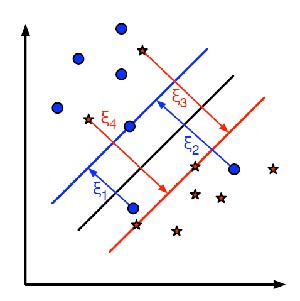


Lúc này ta cần giải bài toán tối ưu:



**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.

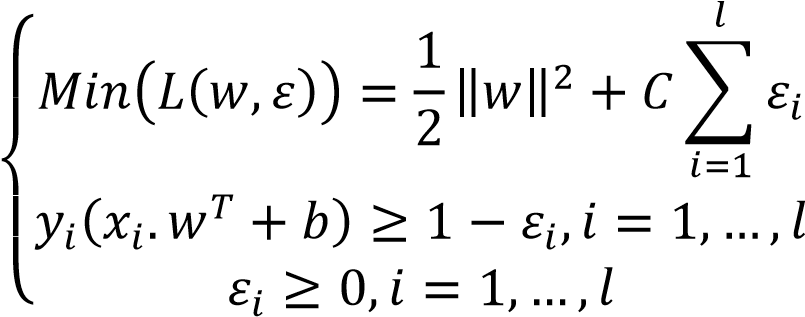


Hình 2.3 Tập dữ liệu được phân chia nhưng có nhiễu

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



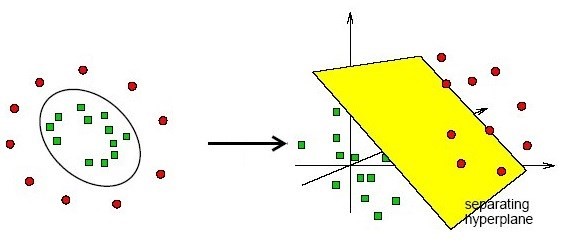
Bài toán tối ưu trở thành:



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.

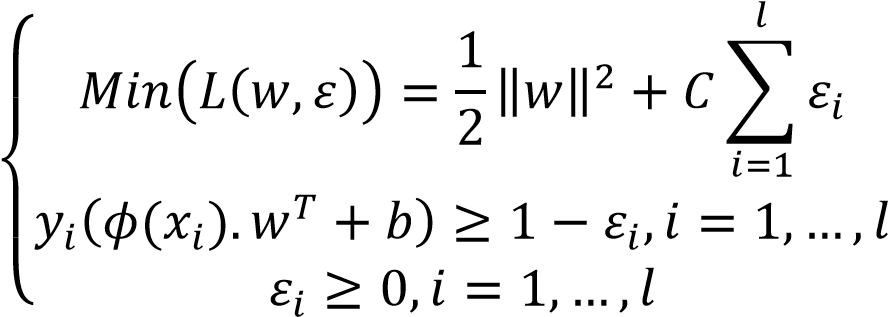


Hình 2.4 Tập dữ liệu không phân chia tuyến tính

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

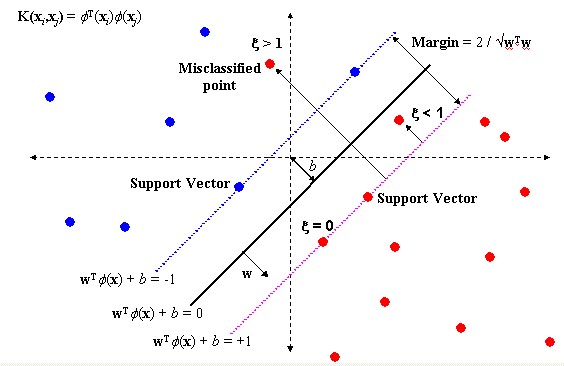


Bài toán tối ưu trở thành:



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.

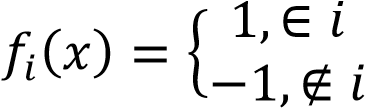
Các điểm nằm trên 2 siêu phẳng phân tách gọi là các vector hỗ trợ (Support vector). Các điểm này quyết định hàm phân tách dữ liệu. Từ đây, chúng ta có thể thấy phương pháp SVM không phụ thuộc vào các mẫu dữ liệu ban đầu, mà chỉ phụ thuộc vào các suport vector (quyết định 2 siêu phẳng phân tách). Cho 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**

Dữ liệu được sử dụng trong luận văn để dự đoán giới tính là tập dữ liệu có sẵn PAKDD'15 được cung cấp bởi Tập đoàn Công ty Cổ phần FPT cung cấp ([www.fpt.com.vn](http://www.fpt.com.vn)).  Nó được lấy từ mô phỏng các hoạt động xem sản phẩm, với người dùng có giới tính nhất định. Dữ liệu này theo sát sự phân bố thực tế trong mối quan hệ này. Nhiệm vụ trong luận văn này là tái tạo lại thông tin về giới tính của người dùng từ các bản ghi xem sản phẩm. Mục tiêu của luận văn là đưa ra được độ chính xác để phục vụ cho các ứng dụng thực tế trong thương mại điện tử phân tích số liệu lớn.

**Định dạng dữ liệu :** Các dữ liệu đã cho được chia thành một bộ huấn luyện và bộ thử nghiệm riêng biệt [**trainingData.csv**](https://knowledgepit.fedcsis.org/mod/folder/view.php?id=838) và [**testData.csv**](https://knowledgepit.fedcsis.org/mod/folder/view.php?id=838). Mỗi tệp này chứa 15.000 bản ghi tương ứng với nhật ký xem sản phẩm. Một phiên đăng nhập bao gồm**bốn cột**, được phân cách bằng**dấu phẩy**. Cột đầu tiên là ID phiên hoạt động. Cột thứ hai và thứ ba tương ứng với thời gian bắt đầu phiên và thời gian kết thúc phiên, tương ứng. Cột cuối cùng chứa**danh sách ID sản phẩm** đã được xem trong phiên, (thứ tự xem được bảo toàn). ID sản phẩm liên tiếp được tách bằng **dấu chấm phẩy**. Ngoài ra còn có các file nhãn **Label.**

Vì phân phối ID sản phẩm duy nhất trong dữ liệu rất thưa thớt nên các ID này chứa thông tin bổ sung về phân cấp loại sản phẩm. Mỗi ID sản phẩm có thể được phân tách thành **bốn ID** khác nhau được tách ra bằng **dấu gạch chéo**. Các ID bắt đầu bằng chữ 'A' là các loại phổ biến nhất và những từ bắt đầu bằng 'D' tương ứng với từng sản phẩm. Các ID bắt đầu bằng 'B' và 'C' được liên kết với các danh mục phụ và tiểu mục con, tương ứng.

Một bản ghi duy nhất trong các tập tin dữ liệu huấn luyện bao gồm bốn loại thông tin:

• Session ID

• Start time (bao gồm cả ngày)

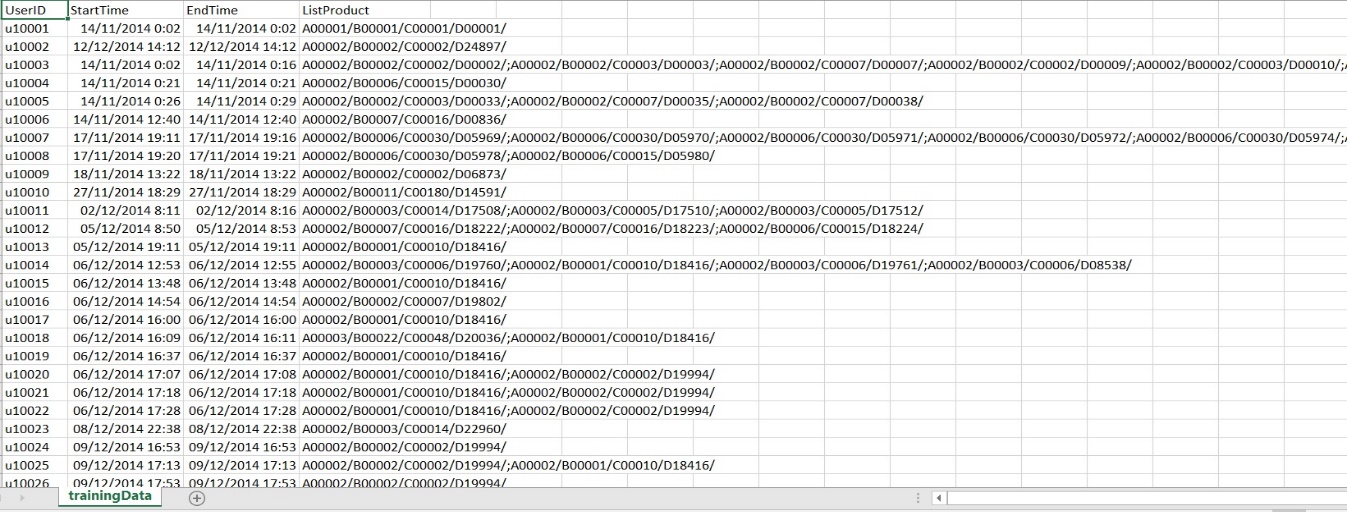
• End time (bao gồm cả ngày)

• Danh sách các ID sản phẩm

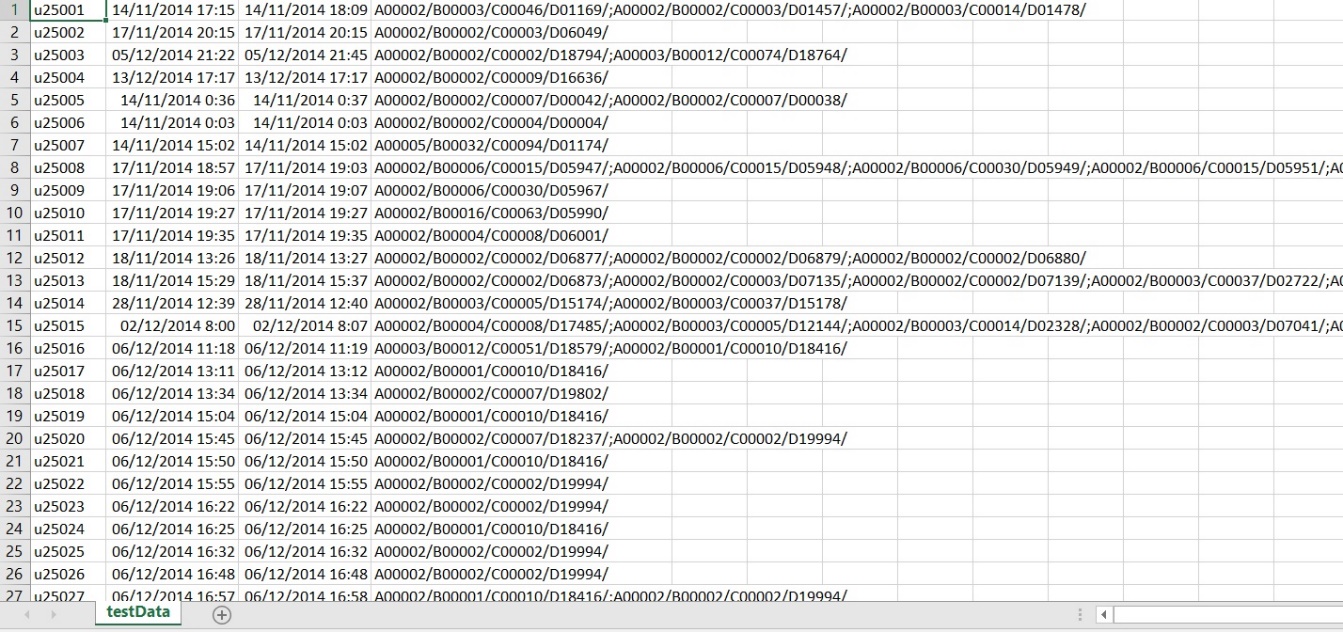
Một ví dụ về một bản ghi như sau:

u10008, 2014/11/17 19:20:06, 2014/11/17 19:21:54,

A00001 / B00001 / C00001 / D00001 /; A00001 / B00002 / C00002 / D00002

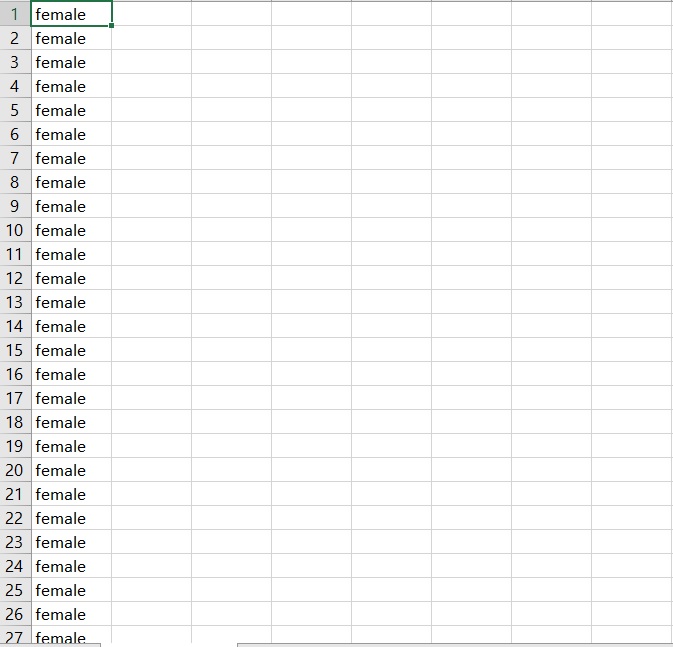


Hình 2.6 Bộ huấn luyện – TranningData



Hình 2.7 Bộ thử nghiệm –TestData

Về cơ bản, tập huấn luyện TrainningData và tập thử nghiệm TestData có cùng định dạng dữ liệu giống nhau để phù hợp với công việc dự đoán.



Hình 2.8 Các nhãn trong tập dữ liệu

Với tập dữ liệu ở trên được phân loại thành 2 nhãn male và female (nam và nữ), từ 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 luận văn của em đã chọn phương pháp này cho việc dự đoán giới tính của người dùng internet, cụ thể kiểm nghiệm và đánh giá sẽ được trình bày trong chương sau.

* 1. **Các dạng đặc trưng sẽ dùng trong phân lớp**

Trong luận văn này, em đã phát triển một hệ thống có thể lấy dữ liệu từ các bản ghi xem sản phẩm của người dùng với giới tính đã được biết đến, các đặc trưng yếu tố và nhãn lớp để tạo ra một tập dữ liệu huấn luyện. Từ đó sẽ xây dựng mô hình từ các tập dữ liệu huấn luyện và sử dụng một phương pháp phân loại để có thể dự đoán giới tính của người dùng chưa biết dựa trên các hoạt động xem sản phẩm của người đã biết. Các tập tin dữ liệu huấn luyện chứa các bản ghi tương ứng với các bản ghi đang xem sản phẩm. Một bản ghi duy nhất chứa các thông tin về sản phẩm xem dữ liệu của người dùng, chẳng hạn như thời gian đầu phiên, thời gian kết thúc phiên giao dịch, danh mục sản phẩm, chủng loại ID. Các nhãn lớp cho mỗi mẫu huấn luyện là nam và nữ. Trong các phần tiếp theo, em mô tả các đặc trưng và kỹ thuật mà đã được sử dụng để dự đoán chi tiết.

* + 1. ***Dạng đặc trưng theo mốc thời gian***

Thời gian trong ngày, ngày trong tuần, ngày lễ, xem thời gian, số lượng sản phẩm được xem trong một phiên, vv là những yếu tố có thể được sử dụng để dự đoán giới tính của một khách hàng.

* + 1. ***Dạng đặc trưng dựa trên tần số hoạt động của người truy cập***
    2. ***Dạng đặc trưng về danh mục và chủng loại sản phẩm***
  1. **Kết luận chương**

**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

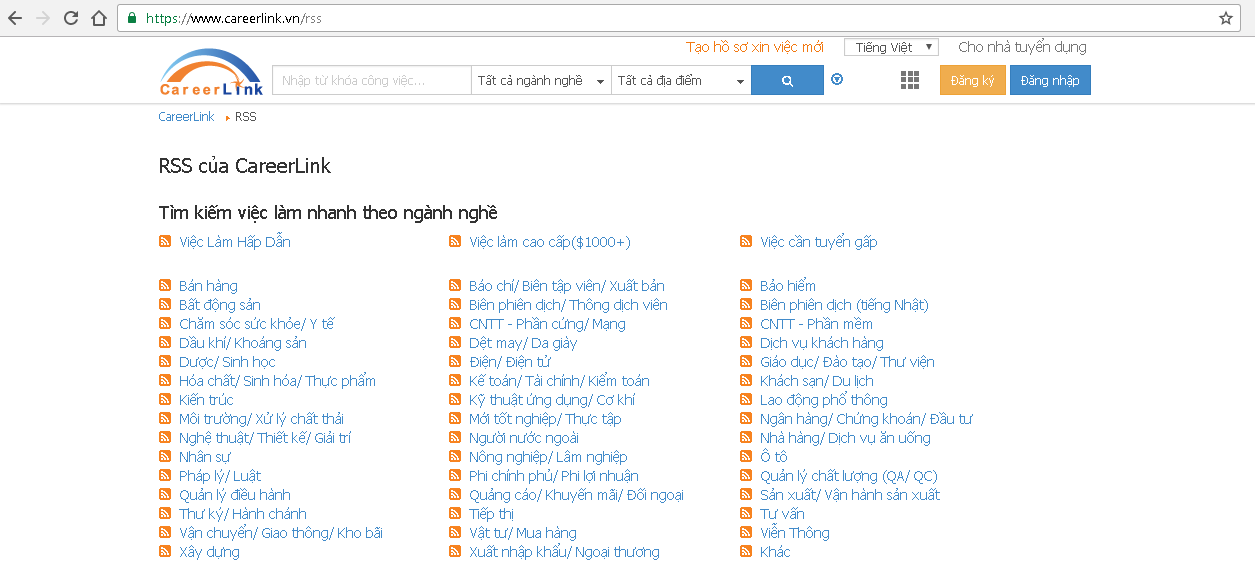


## Xây dựng bộ dữ liệu và tiền xử lý văn bản

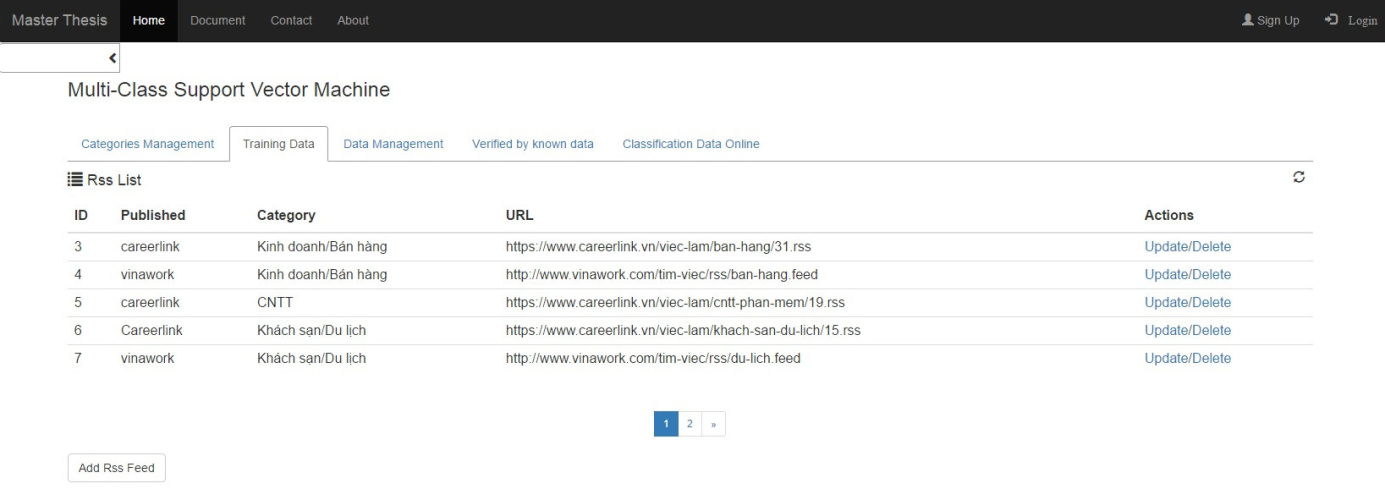
RSS (viết tắt từ Really Simple Syndication) là một tiêu chuẩn định dạng tài liệu dựa trên XML nhằm giúp người sử dụng dễ dàng cập nhật và tra cứu thông tin một cách nhanh chóng và thuận tiện nhất bằng cách tóm lược thông tin vào trong một đoạn dữ liệu ngắn gọn, hợp chuẩn.

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** |

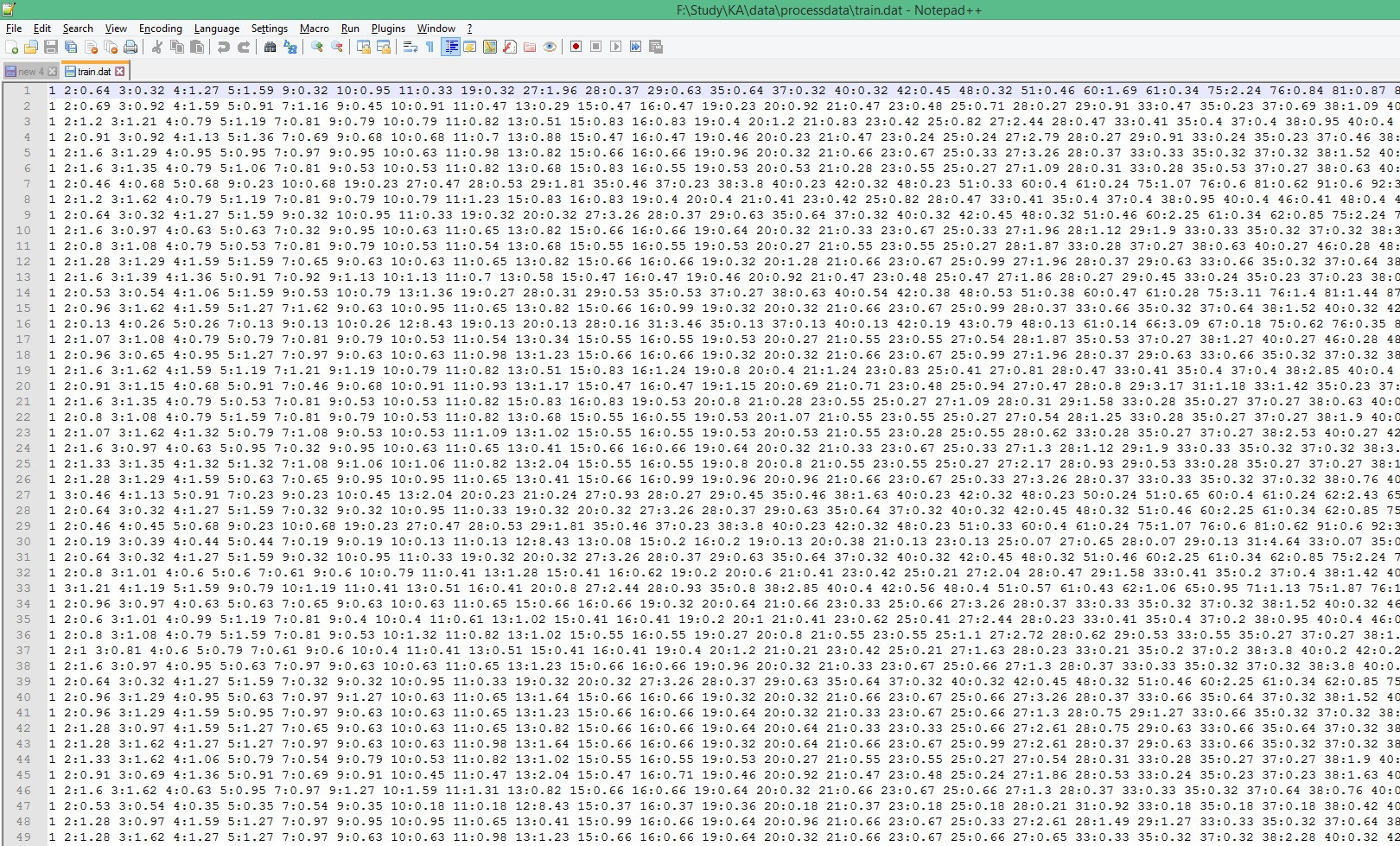
Sau các bước tiền xử lý, văn bản được biểu diễn dưới dạng:

<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:*

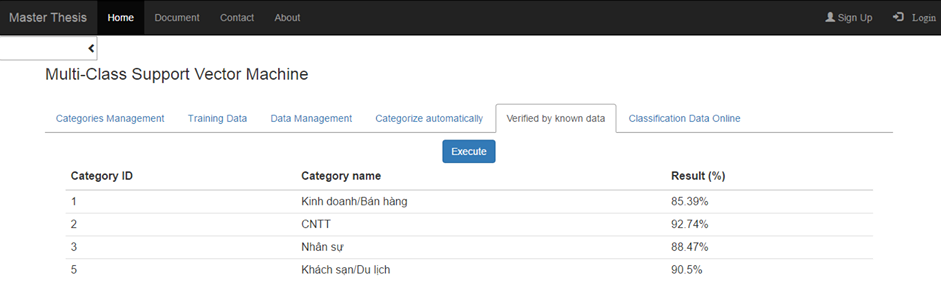
Chúng ta không thể khẳng định một phương pháp phân lớp văn bản cụ thể nào là chính xác hoàn toàn. Vì vậy việc đưa ra độ đo để đánh giá hiệu quả của thuật toán phân lớp giúp chúng ta có thể xác định được độ chính xác của thuật toán, từ đó áp dụng thuật toán đó vào việc phân lớp.

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

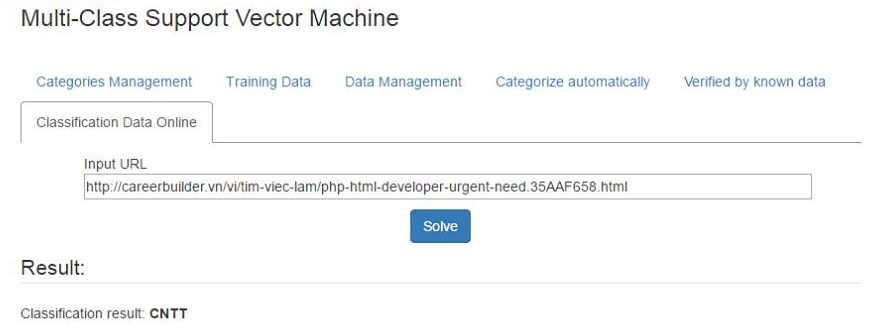
**Bước 1**: Input URL từ các danh sách các website việc làm không support RSS

**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