HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN VĂN CẢNH

NGHIÊN CỨU PHƯƠNG PHÁP ĐÁNH GIÁ MỨC ĐỘ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - NĂM 2019

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **TS. ĐỖ XUÂN CHỢ**

Phản biện 1: **TS. PHÙNG XUÂN ỔN**

Phản biện 2: **TS. HOÀNG XUÂN DẬU**

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông
Vào lúc: 9 giờ 40 ngày 11 tháng 1 năm 2020

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. Lý do chọn đề tài

Thư điện tử là một ứng dụng được sử dụng rộng rãi trên toàn cầu. Thư điện tử giúp rút ngắn thời gian, khoảng cách giữa việc gửi và nhận thư, tiết kiệm chi phí cho quá trình gửi thư. Do hàng ngày người dùng nhận được rất nhiều thư điện tử khác nhau nên sẽ khó khăn trong việc xác định và nhận dạng những thư điện tử nào quan trọng cần đọc và trả lời sớm. Công cụ hỗ trợ phân loại mức độ ưu tiên cho thư điện tử là cần thiết. Từ những lý do trên, học viên với sự giúp đỡ của TS. Đỗ Xuân Chợ lựa chọn đề tài: "**Nghiên cứu phương pháp đánh giá mức độ ưu tiên trong thư điện tử**". Luận văn bao gồm 3 chương:

Chương 1**: Tổng quan về hệ thống thư điện tử**

Chương này trình bày tổng quan về hệ thống thư điện tử, một số công cụ mã nguồn mở để xây dựng hệ thống thư điện tử.

Chương 2: Đánh giá mức độ ưu tiên của thư điện tử Luận văn sẽ trình bày một số công nghệ lọc thư rác hỗ trợ phân loại mức độ ưu tiên của thư điện tử.Tiếp đó là phương pháp nhằm đánh giá, phân loại mức độ ưu tiên cho thư điện tử.

Chương 3: Cài đặt và thử nghiệm

Tiến hành thử nghiệm phương pháp đánh giá độ ưu tiên của thư điện tử ở chương hai.

CHƯƠNG 1 - TỔNG QUAN VỀ THƯ ĐIỆN TỬ

Nội dung chương 1 đề cập đến khái niệm hệ thống thư điện tử bao gồm: định nghĩa, thành phần, chức năng, kiến trúc, vai trò và tầm quan trọng và sự cần thiết của việc phân loại độ ưu tiên của thư điện tử.

1.1 Khái niệm thư điện tử

Thư điện tử còn gọi tắt là E-Mail, là một dịch vụ được triển khai trên các mạng máy tính cho phép người dùng có thể trao đổi thư từ với nhau. Nó là một thông điệp gửi từ máy tính này đến máy tính khác trên mạng máy tính và mang nội dung cần thiết từ người gửi đến người nhận. Thư điện tử truyền gửi được nội dung chữ và các nội dung đa phương tiện như hình ảnh, âm thanh, video…

1.2 Lịch sử phát triển

Năm 1971 Ray Tomlinson thực hiện gửi thành công một thông báo thư tín điện tử đầu tiên trong mạng RPANET. Tomlinson đã sửa đổi hệ thống xử lý thông báo để người sử dụng có thể gửi các thông báo cho các đối tượng nhận không chỉ trong một hệ thống mà trên các hệ thống ARPANET khác Sau đó nhiều công trình nghiên cứu khác đã được tiến hành và thư tín điện tử đã nhanh chóng trở thành một ứng dụng được sử dụng nhiều nhất trên ARPANET trước đây và Internet ngày nay

1.3 Thành phần cấu trúc hệ thống thư điện tử

Hệ thống Mail Server là một hệ thống tổng thể bao gồm nhiều thành phần hoạt động tương tác với nhau. Hầu hết hệ thống thư điện tử bao gồm ba thành phần cơ bản là MUA, MTA và MDA.

1.3.1 MTA(Mail Transfer Agent)

Khi các bức thư được gửi đến từ MUA, MTA có nhiệm vụ nhận diện người gửi và người nhận từ thông tin đóng gói trong phần header của thư và điền các thông tin cần thiết vào header. Sau đó MTA chuyển thư cho MDA để chuyển đến hộp thư ngay tại MTA, hoặc chuyển cho Remote-MTA.

1.3.2 MDA (Mail Delivery Agent)

Là một chương trình được MTA sử dụng để đẩy thư vào hộp thư của người dùng. Ngoài ra MDA còn có khả năng lọc thư, định hướng thư... Thường là MTA được tích hợp với một MDA hoặc một vài MDA.

1.3.3 MUA (Mail User Agent)

MUA là chương trình quản lý thư đầu cuối cho phép người dùng có thể đọc, viết và lấy thư về từ MTA. Đằng sau những công việc vận chuyển thì chức năng chính của MUA là cung cấp giao diện cho người dùng tương tác với thư, gồm có:

- Soạn thảo, gửi thư.
- Hiển thị thư, gồm cả các tệp đính kèm.
- Gửi trả hay chuyển tiếp thư.
- Gắn các tệp vào các thư gửi đi (Text, HTML, MIME v.v…).
- Thay đổi các tham số(ví dụ như server được sử dụng, kiểu hiển thị thư, kiểu mã hoá thư v.v…).
- Thao tác trên các thư mục thư địa phương và ở đầu xa.
- Cung cấp số địa chỉ thư (danh bạ địa chỉ).
- Lọc thư.

1.4 Các giải pháp thư điện tử mã nguồn mở

Hiện nay trên thế giới đã xuất hiện rất nhiều sản phẩm xây dựng một hệ thống Mail Server. Trong thế giới mã nguồn mở

hiện nay, đã có rất nhiều hệ thống truyền tải thư điện tử MTA (Mail Transfer Agent) được phát triển. Nổi tiếng và phổ biến trong số đó gồm có: Zimbra, Sendmail, Qmail, Postfix, Exim, Courier. Mỗi MTA đều có những ưu điểm và nhược điểm riêng.[9]

1.4.1 Zimbra

Zimbra, hệ thống thư điện tử thế hệ mới, được xây dựng bởi cộng đồng phầm mềm tự do nguồn mở và công ty VMWare, đáp ứng các nhu cầu về trao đổi thư tín điện tử và hỗ trợ làm việc. Hệ thống thư điện tử Zimbra đó là công nghệ trên mã nguồn mở cho phép người dùng tiết kiệm được tối đa chi phí mà vẫn đảm bảo được nguyên tắc tôn trọng bản quyền.

1.4.2 Sendmail

Sendmail (http://www.sendmail.org) là MTA đơn giản và lâu đời nhất trên các dòng Unix thời xưa. Ngày nay, Sendmail đa được thương mại hóa bên cạnh sản phẩn miễn phí và vẫn được tiếp tục duy trì, phát triển. Tuy nhiên, vì được thiết kế theo cấu trúc khối và ảnh hưởng từ cấu trúc cũ, nên Sendmail chưa đạt được tính năng ổn định và bảo mật của một MTA như mong muốn.

1.4.3 Qmail

Qmail được viết bởi Bernstein, là một MTA dành cho hệ điều hành tựa Unix, bao gồm Linux, FreeBSD, Sun Solaris. Qmail ra đời như một tất yếu thay thế cho Sendmail và các yếu điểm của nó. Do Qmail được thiết kế module hóa và tối ưu hóa các tính năng ngay từ đầu, nên nó có tốc độ thực thi rất nhanh và ổn định.

1.4.4 Postfix

Weitse Venema, tác giả của các phần mềm miễn phí nổi tiếng như TCP Wrappers, SATAN và Logdaemon, ông không hài lòng khi sử dụng các MTA hiện có (bao gồm cả Qmail), vì

vậy, ông đã viết ra Postfix (http://www.postfix.org). Postfix là một MTA mới, có khả năng thực thi cao, thừa kế cấu trúc thiết kế tốt từ Qmail, trong khi đó vẫn giữ được tính tương thích tối đa với Sendmail.[14]

1.4.5 Exim

Philip Hazel đa phát triển Exim (http://www.exim.org) tại trường đại học Cambridge. Nó được thiết kế theo xu hướng nhỏ và đơn giản nhưng vẫn đảm bảo các tính năng. Tuy nhiên, Exim vẫn được thiết kế theo cấu trúc khối, và hai yếu tố quan trọng là bảo mật và khả năng thực thi lại không được coi trọng. [14]

1.5 Kiến trúc hệ thống thư điện tử mã nguồn mở Zimbra

Kiến trúc hệ thống thư điện tử nguồn mở Zimbra bao gồm những lõi sau [6]:

- Các mã nguồn mở tích hợp trong Zimbra: Linux®, Apache Tomcat, Postfix, MySQL®, OpenLDAP®.
- Giao thức chuẩn được sử dụng là: SMTP, LMTP, SOAP, XML, IMAP, POP.
- Công nghệ được sử dụng để thiết kế là: Java, JavaScript thin client, DHTML.
- Trình duyệt dựa trên giao diện giao diện khách hàng, giao diện này cho phép người dùng dễ dàng truy cập vào tất cả các chức năng của Zimbra Collaboration Suite (ZCS).

1.6 Triển khai Zimbra MTA

1.6.1 Tiếp nhận và gửi thư thông qua Zimbra MTA

1.7 Những tiện ích và vai trò của thư điện tử trong cuộc sống ngày nay

Thư điện tử là một ứng dụng được sử dụng rộng rãi trên toàn cầu. Thư điện tử giúp rút ngắn thời gian, khoảng cách giữa việc gửi và nhận thư, tiết kiệm chi phí cho quá trình gửi thư.

Việc viết thư điện tử cũng nhanh chóng tiện lợi, truyền tải đầy đủ thông điệp mà người dùng muốn gửi đi bao gồm hình ảnh, âm thanh, nội dung văn bản … với dung lượng lớn theo dạng nhập trực tiếp vào khung soạn thảo hoặc đính kèm.

Có hơn 3,9 tỷ người dùng email trên toàn thế giới. Năm nay, số lượng người dùng email đạt mốc 3,9 tỷ, điều đó có nghĩa là hơn 50% dân số thế giới hiện đang sử dụng email. Năm 2020, số lượng người dùng email sẽ tăng lên 4 tỷ. Theo số liệu thống kê tiếp thị qua email gần đây, tốc độ tăng trưởng người dùng dự đoán trong bốn năm tới là 3%, tức là khoảng 100 triệu người dùng mỗi năm. Vì vậy, vào năm 2023, số lượng người dùng email trên toàn thế giới sẽ xấp xỉ 4,3 tỷ, có khoảng 5,59 tỷ tài khoản email đang hoạt động.

1.8 Kết luận chương

Qua những thống kê trên, hàng ngày mỗi người dùng thường nhận được rất nhiều thư điện tử khác nhau nên sẽ khó khăn trong việc xác định và nhận dạng những thư điện tử nào quan trọng cần đọc và trả lời sớm, những thư nào có thể chỉ để theo dõi. Vì vậy ta phải dùng đến khái niệm "Mức độ ưu tiên" với thư điện tử. Theo định nghĩa tiếng Anh "Mức độ ưu tiên" - "Priority" được sử dụng để so sánh hai vật hoặc hai điều kiện, khi mà một vật/điều kiện phải quan tâm nhiều hơn những vật/điều kiện khác và phải được giải quyết trước khi chuyển sang (những) vật/điều kiện tiếp theo. Công cụ hỗ trợ nhận dạng và phân loại mức độ ưu tiên cho thư điện tử là cần thiết. Chương tiếp theo của luận văn xin được trình bày phương pháp đánh giá độ ưu tiên cho thư điện tử.

CHƯƠNG 2 – ĐÁNH GIÁ MỨC ĐỘ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

Chương 2 sẽ trình bày phương pháp nhằm đánh giá, phân loại mức độ ưu tiên cho thư điện tử.

2.1 Một số công nghệ phân loại thư rác hỗ trợ phân loại mức độ ưu tiên của thư điện tử

2.1.1 Định nghĩa thư rác

Hiện nay vẫn chưa có một định nghĩa hoàn chỉnh, chặt chẽ về thư rác. Có quan điểm coi thư rác là những thư quảng cáo không được yêu cầu (Unsolicited Commercial Email-UCE), có quan điểm rộng hơn cho rằng thư rác bao gồm thư quảng cáo, thư quấy rối, và những thư có nội dung không lành mạnh (Unsolicited Bulk Emai -UBE).

Nội dung thông dụng nhất về định nghĩa thư rác: Thư rác (spam mail) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận.

2.1.2 Các phương pháp loc thư rác

2.1.2.1 Phương pháp dùng danh sách trắng đen

2.1.2.2 Phương pháp lọc theo từ khóa

2.1.2.3 Phương pháp lọc dựa trên mạng xã hội

2.1.2.4 Phương pháp lọc thư rác dùng chuỗi hỏi đáp

2.1.2.5 Lọc thư rác dựa trên xác suất thống kê và học máy

Đầu tiên sẽ phân loại các bức thư thành thư rác và thư hợp lệ. Một thuật toán được áp dụng để trích chọn và đánh trọng số cho các đặc trưng của thư rác theo một cách nào đó (thường là áp dụng công thức xác suất). Sau khi trích chọn đặc trưng, hai tập thư rác và thư hợp lệ sẽ được sử dụng để huấn luyện một bộ phân loại tự động. Quá trình huấn luyện dựa trên một phương pháp học máy.

2.1.2.6 Phương pháp lọc SpamAssassin

2.2 Tổng quan về học máy

2.2.1 Khái niệm cơ bản

Sự phát triển nhanh chóng của các kỹ thuật khai phá dữ liệu đã đưa Học máy thành một lĩnh vực riêng biệt của Khoa học máy tính. Học máy là một lĩnh vực của Trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể [11].

Quy trình chung của một tiến trình học máy gồm 5 bước sau:

- Nhập dữ liệu.
- Xử lý dữ liệu. Tại bước này, dữ sẽ được chuyển đổi, làm sạch và chuẩn hóa để phù hợp với thuật toán. Sau đó, dữ liệu được chia ra thành hai tập – ‘tập huấn luyện’ và ‘tập thử nghiệm’.
- Huấn luyện mô hình.
- Thử nghiệm mô hình.
- Triển khai mô hình.

2.2.2 Trích chọn đặc trưng

Trong các ví dụ đã đưa ra ở trên, cần phải trích xuất các thuộc tính từ dữ liệu đầu vào để đưa vào thuật toán. Ví dụ, với trường hợp tính giá nhà, dữ liệu có thể được biểu diễn dưới dạng ma trận đa chiều, với mỗi cột là một thuộc tính và mỗi dòng là giá trị của thuộc tính đó. Trong trường hợp hình ảnh, dữ liệu có thể được biểu diễn dưới dạng giá trị RGB của mỗi pixel. Các thuộc tính này được gọi là đặc trưng, và ma trận là vector đặc trưng. Quá trình trích xuất dữ liệu từ tệp tin được

gọi là trích xuất đặc trưng. Mục đích của quá trình này là thu được một tập dữ liệu chi tiết và không dư thừa.

2.2.3 Phân loại học máy

2.2.3.1 Học có giám sát và học không giám sát

Đối với **học có giám sát**, việc học được dựa trên các dữ liệu được dán nhãn. Trong trường hợp này, chúng ta sẽ dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước.

Học có giám sát được chia nhỏ thành hai loại chính:

- **Phân lớp (Classification)**. Dựa vào tập dữ liệu đã được dán nhãn, với mỗi nhãn định nghĩa một lớp, dự đoán xem một dữ liệu mới chưa biết thuộc vào lớp nào. Số lớp thường nhỏ và hữu hạn.
- **Hồi quy (Regression)**. Nhãn không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ về dự đoán mức giá của một ngôi nhà thuộc loại này.

Ngược lại với học có giám sát, trong **học không giám sát**, dữ liệu không được dán nhãn. Ở đây, mục tiêu là tìm một số mẫu trong tập dữ liệu chưa được phân loại, thay vì dự đoán một số giá trị. Một bài toán quen thuộc của học không giám sát là phân cụm (clustering). **Phân cụm** là việc tìm kiếm điểm chung giữa các dữ liệu trong tập dữ liệu và chia chúng thành các cụm tương ứng dựa vào điểm chung này. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng.

2.2.3.2 Một số kỹ thuật học máy

K-Nearest Neighbors

K-Nearest Neighbors (KNN) là một trong những thuật toán đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong số các thuật toán của học máy. KNN là một thuật toán phi tham số, tức là nó không đưa ra bất kỳ dự đoán nào về cấu trúc của dữ liệu. Khi huấn luyện, thuật toán này không học một

điều gì từ dữ liệu huấn luyện (đây cũng là lý do thuật toán này được xếp vào loại lazy learning).

KNN có thể áp dụng được vào cả hai loại của bài toán học có giám sát là Phân lớp và Hồi quy. Trong cả hai bài toán, kết quả dự đoán của một điểm dữ liệu mới được suy ra trực tiếp từ k điểm dữ liệu gần nhất trong tập dữ liệu huấn luyện. Đối với bài toán phân lớp, kết quả đầu ra sẽ là lớp mà dữ liệu thuộc về, dựa trên việc bình chọn (majority vote) của k điểm gần nhất.

Có nhiều phương pháp đo khoảng cách giữa các điểm để tìm ra điểm gần nhất. Các phương pháp phổ biến nhất bao gồm khoảng cách Hamming, khoảng cách Manhattan, khoảng cách Minkowski:

Khoảng cách Hamming: $d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$ (2.1)

Khoảng cách Manhattan:

$$d_1(p,q) = ||p-q||_1 = \sum_{i=1}^{n} | p_i - q_i | \quad (2.2)$$

Khoảng cách Minkowski = $(\sum_{i=1}^{n} | x_i - y_i|^p)^{1/p}$ (2.3)

Phương pháp phổ biến nhất đối với các biến liên tục là khoảng cách Euclidean, được định nghĩa bởi công thức (2.4) dưới đây:

$d_{Euclidean} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$; p và q là các điểm trong không gian n (2.4)

Thuật toán Random Forest

Random Forest dựa trên tính ngẫu nhiên (random) và được tạo nên từ nhiều cây quyết định (forest – "rừng").

Thuật toán có thể được mô tả như sau :

- Các cây được xây dựng dựa trên 2/3 dữ liệu của tập dữ liệu huấn luyện (62.3%). Dữ liệu được lựa chọn ngẫu nhiên.
- Một số biến dự đoán được chọn ngẫu nhiên từ tổng số các biến dự đoán. Sau đó, cách phân chia tốt nhất của các biến được lựa chọn sẽ được dùng để phân chia nút. Theo mặc định, số lượng biến được chọn sẽ là căn bậc hai của tổng số các thuộc tính dùng để dự đoán và không đổi đối với các cây.
- Tỉ lệ dự đoán sai được tính toán dựa vào phần dữ liệu còn lại (dữ liệu out-of- bag).
- Mỗi cây huấn luyện sẽ đưa ra một kết quả phân loại, được gọi là "bỏ phiếu". Lớp nhận được nhiều "phiếu" nhất sẽ được chọn là kết quả cuối cùng. [1]

Thuật toán Logistic Regression

Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu ra *rời rạc* (*discrete target variable*) *y* ứng với một véc-tơ đầu vào **x**. Việc này tương đương với chuyện phân loại các đầu vào **x** vào các nhóm *y* tương ứng. Ví dụ, xem một bức ảnh có chứa một con mèo hay không. Thì ở đây ta coi đầu ra $y = 1$ nếu bước ảnh có một con mèo và $y = 0$ nếu bức ảnh không có con mèo nào. Đầu vào **x** ở đây sẽ là các pixel một bức ảnh đầu vào. Sử dụng phương pháp thống kê ta có thể coi rằng khả năng đầu vào x nằm trong nhóm y_0 là xác xuất nhóm y_0 khi biết x: p(y_0|x). Ta có hàm **sigmoid** (logistic sigmoid function).[8]

$$p(y_0|\mathbf{x}) = \frac{1}{1+\exp(-a)} = \sigma(a) \qquad (2.6)$$

Vận dụng thuyết phân phối chuẩn, ta có thể chỉ ra rằng: $a = \boldsymbol{w}^T\mathbf{x} + \boldsymbol{w}_0$

Đặt $\boldsymbol{x}_0$=[1,…..,1] ta có thể viết gọn : $a = \boldsymbol{w}^T\mathbf{x}$

Thay vào công thức (2.6) bên trên ta có : $p(y_0|\mathbf{x})=\frac{1}{1+\exp(-a)}=\sigma(\boldsymbol{w}^{\boldsymbol{T}}\mathbf{x})$ Trong đó x là thuộc tính đầu vào còn w là trọng số tương ứng.

Ta phải tối ưu hàm mất mát

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}log\sigma^{(i)} + (1-y^{(i)})log(1-\sigma^{(i)})\right)$$

Theo phương pháp Gradient Descent ta cập nhật tham số sau mỗi vòng lặp [12]:

$$\mathbf{w} = \mathbf{w} - \eta\frac{1}{m}\mathbf{X}^{\intercal}(\sigma - \mathbf{y})$$

2.2.4 Thuật toán khai phá dữ liệu văn bản

Thuật toán TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) là một kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản.

TF: Term Frequency(Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản(tổng số từ trong một văn bản).

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

Trong đó:

tf(t, d): tần suất xuất hiện của từ t trong văn bản d

f(t, d): Số lần xuất hiện của từ t trong văn bản d

max({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d.

IDF: Inverse Document Frequency(Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ . Khi tính toán TF , tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như "is", "of" và "that" thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Trong đó: idf(t, D): giá trị idf của từ t trong tập văn bản

|D|: Tổng số văn bản trong tập D

|{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

Việc sử dụng logarit nhằm giúp giá trị tf-idf của một từ nhỏ hơn, do chúng ta có công thức tính tf-idf của một từ trong 1 văn bản là tích của tf và idf của từ đó. Cụ thể, chúng ta có công thức tính TF-IDF hoàn chỉnh như sau:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

2.3 Phương pháp phân loại độ ưu tiên của thư điện tử

2.3.1 Các thành phần của một thư điện tử

Các thành phần của một thư điện tử thông thường bao gồm người gửi, người nhận, thời gian, tiêu đề, phần nội dung, các tệp tin đính kèm.

2.3.2 Lựa chọn đặc trưng để xét độ ưu tiên

Thư điện tử là một phương tiện dựa trên sự trao đổi qua lại. Mọi người gửi và nhận thư theo thời gian. Thư điện tử là phương tiện dựa trên giao dịch, các đặc trưng xã hội sẽ là tối

quan trọng trong việc đánh giá tầm quan trọng của thư [3]. Đặc trưng đáng xem xét là địa chỉ người gửi, người nhận, tần suất phản hồi giữa họ. Đặc trưng quan trọng mà ta chú ý là thời gian nhận được email. Tiếp theo xem xét email đó có đang ở trong một luồng email nào đó không. Những email cùng luồng thường cùng chủ đề, và có thể là để trả lời lại một thư khác. Ví dụ như ở Gmail thì nó được đánh dấu là "RE". Ta trích xuất đặc trưng từ nội dung của thư bằng các kỹ thuật khai thác văn bản. Cụ thể, nếu có các thuật ngữ phổ biến trong các chủ đề và nội dung email mà người dùng nhận được, thì các email trong tương lai có chứa các thuật ngữ này ở trong chủ đề và nội dung có thể quan trọng hơn thuật ngữ không xuất hiện. [2]

2.3.3 Cách tính trọng số dựa vào các đặc trưng

Đặc trưng tần suất thư gửi đến: Đếm số lần xuất hiện của mỗi địa chỉ email trong số email dùng để trainning. Với số lần xuất hiện của một địa chỉ email là x_i. Trọng số thứ nhất: $w_1 = \log_{10} x_i$

Đặc trưng tần suất thư phản hồi: Lọc các email là email phản hồi. Gọi số lần xuất hiện một địa chỉ email trong số các email phản hồi là x_j. Trọng số thứ hai: $w_2 = \log_{10} x_j$

Đặc trưng tỉ lệ số lượng thư trên thời gian của luồng email: Lọc các thread thư, Loại các thread không có reply, tính tổng thời gian của thread đó. Với thread i. Gọi tổng thời gian của thread là t với t tính bằng giây, số lượng thư qua lại của luồng thư i là n. Trọng số thứ ba : $w_3 = \log_{10} \frac{n}{t}$

Sử dụng phương pháp TF-IDF, tính được độ quan trọng của các từ trong nội dung của các email trong tập mẫu. Với m là số lượng từ của nội dung thư, x_j là độ quan trọng của từng từ.

Đặc trưng độ quan trọng của nội dung thư: Trọng số thứ tư là : $w_4 = \log_{10} \sum_{m}^{i=1} x_j$

Đặc trưng độ quan trọng của tiêu đề: Với n là số lượng từ của tiêu đề của mỗi thư, x_i là độ quan trọng của từng từ. Trọng số thứ năm là $w_5 = \log_{10} \sum_{n}^{i=1} x_i$. [2]

2.3 Kết luận chương

Chương 2 đưa ra một số phương pháp lọc thư rác để hỗ trợ cho việc phân loại độ ưu tiên của thư điện tử như dùng danh sách trắng, đen; lọc theo từ khóa, lọc dựa vào mạng xã hội, lọc dùng chuỗi hỏi đáp, lọc dựa trên học máy, phương pháp dùng lọc SpamAssassin. Bên cạnh đó là cái nhìn tổng quan nhất về khái niệm học máy và giới thiệu một số thuật toán được sử dụng trong luận văn, bao gồm KNN, Logistic Regression, Random Forest. Phương pháp phân loại độ ưu tiên của thư điện tử được giới thiệu ở cuối chương. Dữ liệu đầu vào đã được xử lý là các thư điện tử tiếng Việt gồm 4 trường thông tin: người gửi, thời gian, tiêu đề, nội dung. Chương 2 cũng đưa ra cách lựa chọn đặc trưng. Các đặc trưng được xét đến là đặc trưng về xã hội: người gửi, đặc trưng về thời gian: thời gian nhận thư, đặc trưng về nội dung : tiêu đề thư, nội dung thư. Từ cách đặc trưng được lựa chọn, chương 2 trình bày cách tính năm trọng số để đưa vào mô hình học máy. Kết quả mong muốn đạt được là khi có một thư mới ta sẽ phân nó vào hai nhóm: Quan trọng và Không quan trọng. Chương 3 luận văn sẽ trình bày về thực nghiệm và các kết quả đạt được.

CHƯƠNG 3 - CÀI ĐẶT VÀ THỬ NGHIỆM

Chương 3 sẽ tiến hành áp dụng phương pháp phân loại đã giới thiệu ở chương 2 vào tập dữ liệu mẫu. Sau đó, đưa ra kết quả thu được và kết luận.

3.1 Thu thập và tiền xử lý dữ liệu

3.1.1 Thu thập dữ liệu

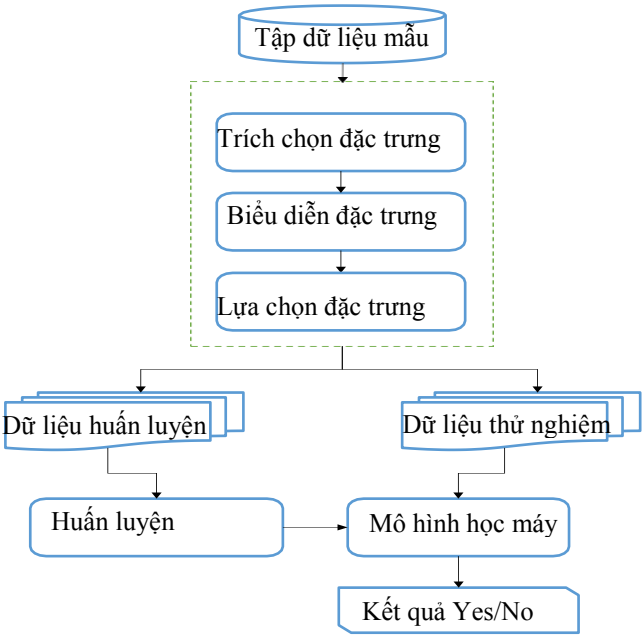
Trong phần chương 3, bộ dữ liệu được sử dụng là bộ dữ liệu thu thập trên mạng internet. Sử dụng Goolge takeout để lấy flie Mbox dữ liệu mail của tên miền @fpt.edu.vn Bộ dữ liệu thực nghiệm gồm 30 user:

Tổng số mail	Số mail quan trọng	Số mail không quan trọng
61733	20054	41679

3.1.2 Tiền xử lý dữ liệu

Với mỗi email có tối đa 12 trường dữ liệu Các email đều được lấy từ tên miền @fpt.edu.vn. Với mỗi email được lấy ra với 4 trường dữ liệu {subject',’from',’date',’body’} lọc bỏ các email có loại ngôn ngữ khác chỉ để lại các thư là tiếng Việt. Các email được lưu trong tệp định dạng mbox chuyển về định dang csv.

3.2 Thực nghiệm đánh giá



Hình 3.3 : Mô hình quá trình phân loại thư điện tử

Quá trình thực hiện bao gồm 2 giai đoạn:

- Giai đoạn huấn luyện : Đầu vào của giai đoạn này là các dữ liệu đã được tiền xử lý để đưa ra các vector đặc trưng. Trong bước huấn luyện, dữ liệu sẽ được phân loại theo nhãn phân loại tương ứng, sau đó sử dụng các thuật toán học máy để đưa ra bộ phân loại tương ứng phục vụ cho giai đoạn phát hiện.
- Giai đoạn phát hiện: Dữ liệu trong giai đoạn này được xử lý tương tự như dữ liệu trong giai đoạn huấn luyện. Đầu vào của giai đoạn phát hiện là dữ liệu đã được tiền xử lý và model (bộ phân loại – kết quả của giai đoạn huấn luyện).

Áp dụng các tính các trọng số ở chương 2 ta sẽ có điểm số cụ thể cho mỗi e-mail và được tính bằng hàm log của tích các đặc trưng. Môi trường thử nghiệm: Hệ điều hành window 10, Ngôn ngữ python.

3.3 Kết quả chạy thực nghiệm

Bảng 3.1: Kết quả chạy thử nghiệm

User	Model								
	Random Forest			KNN			Logistic Regression		
	AUC	F1	Recall	AUC	F1	Recall	AUC	F1	Recall
chiennthe14174	0.912	0.892	0.896	0.835	0.862	0.876	0.795	0.84	0.885
dangnhha14019	0.713	0.666	0.670	0.667	0.625	0.632	0.551	0.431	0.571
datntse04909	0.953	0.915	0.916	0.846	0.878	0.885	0.715	0.812	0.853
ducnmhe13066	0.676	0.617	0.618	0.673	0.637	0.637	0.598	0.551	0.563
ducnmse05559	0.834	0.745	0.745	0.675	0.631	0.631	0.495	0.5	0.515
hiepphse04711	0.838	0.767	0.768	0.683	0.643	0.646	0.67	0.635	0.646
hieudtse04712	0.882	0.841	0.843	0.800	0.794	0.802	0.796	0.782	0.797
linhnptsb02246	0.832	0.795	0.802	0.666	0.689	0.705	0.658	0.568	0.694
phucnhse04534	0.849	0.772	0.772	0.722	0.664	0.664	0.65	0.609	0.613
quangnvse0583	0.884	0.795	0.795	0.762	0.702	0.703	0.692	0.644	0.645
quynhthse0464	0.869	0.777	0.776	0.758	0.691	0.692	0.708	0.657	0.659
sanglqse04676	0.949	0.894	0.895	0.862	0.841	0.846	0.778	0.782	0.807
toannbsb02527	0.843	0.775	0.776	0.720	0.673	0.675	0.645	0.606	0.638
tuanntse04733	0.925	0.877	0.879	0.809	0.806	0.815	0.695	0.702	0.767
tuanthsb01889	0.808	0.762	0.769	0.684	0.677	0.689	0.631	0.585	0.682
tungptse04569	0.901	0.819	0.819	0.788	0.718	0.719	0.564	0.431	0.528
tungtmse05324	0.847	0.803	0.809	0.724	0.736	0.754	0.714	0.666	0.753

Bảng 3.2 Độ hiệu quả trung bình của từng thuật toán

Model								
Random Forest			KNN			Logistic Regression		
AUC	F1	Recall	AUC	F1	Recall	AUC	F1	Recall
0.854	0.795	0.797	0.746	0.722	0.728	0.668	0.635	0.683

Accuracy: tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử. **Recall:** là tỷ lệ số điểm true positive trong tổng số những điểm thực sự là positive (TP+FN). Giá trị recall cao đồng nghĩa với việc TPR (true positive Rate) cao, tức là tỷ lệ bỏ sót các điểm thực sự positive là thấp.**F1-score**: là harmonic mean của precision và recall. F1 càng cao, bộ phân loại càng tốt. Từ kết quả trên ta thấy được với thuật toán Random Forest các chỉ số là tốt nhất với các chỉ số : AUC : 0.854,F1 : 0.795,Recall : 0.797 Cho kết quả phân loại tốt nhất trong ba thuật toán.

3.3 Kết luận chương 3

Chương 3 trình bày về quá trình thử nghiệm trên môi trường windown 10, ngôn ngữ lập trình python. Các dữ liệu được thu thập bằng công cụ google **takeout.** Các dữ liệu là các thư điện tử có tối đa tới 12 trường dữ liệu, bao gồm các thư có các loại ngôn ngữ khác nhau dưới định dạng Mbox. Dữ liệu được tiền xử lý còn lại các thư Tiếng Việt bao gồm trường thông tin chính dưới dạng file csv. Chương 3 cũng nêu mô hình thực hiện phân loại độ ưu tiên của thư điện tử, kết quả thực nghiệm của quá trình phân loại. Dựa vào kết quả thực nghiệm ta thấy được với thuật toán Random Forest các chỉ số là tốt nhất với các chỉ số : AUC : 0.854,F1 : 0.795,Recall : 0.797 Cho kết quả phân loại tốt nhất trong ba thuật toán.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết quả đạt được

- Trình bày sự phổ biến vai trò của thư điện tử trong cuộc sống hiện đại.

- Trình bày kết quả nghiên cứu về thư điện tử: định nghĩa, lịch sử phát triển thư điện tử, các thành phần cấu trúc của hệ thống thư điện tử.

- Các giải pháp hệ thống thư điện tử mã nguồn mở. Chi tiết cài đặt kiến trúc hệ thống, các thành phần của mã nguồn mở Zimba.

- Trình bày các phương pháp hỗ trợ đánh giá giá mức độ ưu tiên thư điện tử.

- Trình bày cơ sở lý thuyết, phương pháp đánh giá mức độ ưu tiên của thư điện tử

- Tiến hành thực nghiệm, đánh giá kết quả. Quá trình thực nghiệm học viên xử lý dữ liệu là các email thu thập được trên internet. Sử dụng thuật toán, phương pháp được trình bày ở chương 2 để tính toán các trọng số từ các đặc trưng của thư, đưa vào các thuật toán học máy để thực hiện phân lớp. Thuật toán hiệu quả nhất khi thực nghiệm là thuật toán Random Forest.

2. Hướng phát triển của luận văn

Một số hướng phát triển tiếp theo của luận văn:

- Nghiên cứu các công nghệ mới được ứng dụng trong phân loại và đánh giá mức độ ưu tiên của thư điện tử.
- Nghiên cứu cái tiến phương pháp đánh giá mức độ ưu tiên của thư điện tử bằng tiếng Việt