PHÂN LOẠI EMAIL SỬ DỤNG KỸ THUẬT XẾP HẠNG VÀ BAYES ĐƠN GIẢN

Phạm Bảo Thạch, Từ Minh Phương

Học viện Công nghệ Bưu chính viễn thông

Với số lượng thư điện tử ngày càng tăng, khả năng tự động phân loại thư thành các nhóm theo yêu cầu của người dùng là một trong những chức năng quan trọng của chương trình quản lý thư điện tử. Bài báo trình bày kết quả xây dựng và thử nghiệm hệ thống phân loại thư điện tử tự động có khả năng xử lý thư tiếng Việt. Quá trình phân loại thư được thực hiện bằng kỹ thuật phân loại văn bản tự động. Để đảm bảo hoạt động thuận tiện của hệ thống cả khi có ít và có nhiều dữ liệu huấn luyện, chúng tôi đã sử dụng kết hợp phương pháp xếp hạng (ranking) với phân loại Bayes đơn giản. Bài báo trình bày kết quả thử nghiệm bước đầu và phân tích ưu nhược điểm của hai phương pháp nói trên.

1. Đặt vấn đề

Sự tiện lợi và phổ dụng của thư điện tử dẫn tới tình trạng số lượng thư gửi và nhận không ngừng tăng lên. Việc nhận được lượng thư lớn hàng ngày khiến người dùng tốn nhiều thời gian cho xử lý, lưu trữ và tìm kiếm thư. Giải pháp hiệu quả cho vấn đề này là sắp xếp thư trong các thư mục (folder) khác nhau một cách hợp lý. Để giải phóng người dùng khỏi công đoạn tổ chức thư mục và sắp xếp thư bằng tay, một số phần mềm thư điện tử đã cung cấp chức năng sắp xếp thư tự động.

Việc phân chia thư vào các thư mục có thể thực hiện bằng hai cách. Cách thứ nhất yêu cầu người dùng xây dựng các quy tắc bằng dựa trên từ khoá và thực hiện phân loại thư theo những quy tắc này. Nhược điểm chính của cách phân loại này là người dùng khó xây dựng được những quy tắc tốt để phân loại. Trong bài báo này, chúng tôi sẽ đề cập tới cách thứ hai, đó là cách sử dụng kỹ thuật học máy và phân loại thư tự động dựa trên nội dung và một số đặc trưng của thư. Cách này được thực hiện như sau: trước tiên, người dùng xác định một số thư mục và sắp xếp một số thư vào các thư mục này. Nội dung các thư đã được xác định thư mục sẽ được sử dụng để huấn luyện bộ phân loại. Khi có thư mới, bộ phân loại đã được huấn luyện sẽ tự động phân loại và xếp thư vào thư mục tương ứng dựa trên nội dung thư. Dĩ nhiên, phương pháp này chỉ có thể cho kết quả tốt nếu các thư trong cùng thư mục có đặc điểm chung về mặt ngữ nghĩa hoặc nội dung.

Phương pháp phân loại thư tự động theo nội dung là một trường hợp riêng của phân loại văn bản tự động [11] và đã được nghiên cứu nhiều cho bài toán lọc thư rác [1, 9]. Tuy

nhiên, tự động phân loại thư theo yêu cầu của người dùng là bài toán khó hơn và ít được nghiên cứu hơn. Một số kết quả tiêu biểu về những nghiên cứu này được trình bày trong [2, 6, 12].

Trong các nghiên cứu nói trên, phương pháp chung được sử dụng là biểu diễn phần nội dung văn bản của thư (có thể bổ sung một số thành phần khác như người gửi, người nhận, tiêu đề thư) dưới dạng vectơ các đặc trưng và sử dụng biểu diễn vectơ làm đầu vào cho bộ phân loại. Thuật toán phân loại thường sử dụng là phân loại Bayes đơn giản (Naïve Bayes), Support Vector Machines (SVM), Entropy cực đại, cây quyết định, random forest. Đây là những thuật toán đã được sử dụng cho phân loại văn bản tự động và đều có chung một đặc điểm là quá trình huấn luyện được thực hiện theo dạng "mẻ" (batch). Trong giai đoạn huấn luyện, thuật toán phân loại được cung cấp toàn bộ dữ liệu huấn luyện và mô hình phân loại được xây dựng dựa trên toàn bộ dữ liệu đó. Khi có thêm dữ liệu huấn luyện mới, bộ phân loại sẽ được huấn luyện lại trên toàn bộ dữ liệu - cả cũ và mới.

Do đặc điểm bài toán phân loại thư điện tử, việc huấn luyện theo mẻ có những nhược điểm nhất định khi sử dụng cho bài toán này. Thông thường, ở giai đoạn đầu, lượng dữ liệu huấn luyện còn ít (mới có ít thư được người dùng phân loại), độ chính xác phân loại chưa cao. Khi dữ liệu dần được bổ sung hoặc khi có thư bị phân loại sai, bộ phân loại được huấn luyện được huấn luyện và cập nhật mô hình lại. Việc huấn luyện theo mẻ đòi hỏi tính toán trên toàn bộ dữ liệu cũ và mới, do vậy tốn thời gian. Khác với huấn luyện theo mẻ, một phương pháp huấn luyện khác được gọi là huấn luyện trực tuyến (online). Mỗi khi có một ví dụ huấn luyện mới, phương pháp này tiến hành cập nhật tham số của mô hình theo ví dụ đó mà không cần tính toán lại với những ví dụ cũ. Đây là một giải pháp tốt cho ứng dụng trong đó dữ liệu được cung cấp dần dần như bài toán đang xét.

Trong bài báo này, chúng tôi trình bày kết quả xây dựng chương trình phân loại thư điện tử sử dụng phân loại Bayes đơn giản và một phương pháp phân loại trực tuyến dựa trên việc xếp hạng các nhãn phân loại [4] đồng thời so sánh độ chính xác và ưu nhược điểm của hai phương pháp. Theo chúng tôi được biết, đây là nghiên cứu đầu tiên sử dụng huấn luyện trực tuyến cho phân loại thư. Một kết quả quan trọng khác việc xử lý thư tiếng Việt và kết quả thử nghiệm cho các giải pháp. Chương trình phân loại thư của chúng tôi được xây dựng dưới dạng một add-in có thể tích hợp với chương trình thư điện tử thông dụng Outlook của Microsoft.

2. Phân loại thư theo nội dung

Phân loại thư theo nội dung là trường hợp riêng của phân loại văn bản tự động và có thể phát biểu dưới dạng bài toán học máy có giám sát như sau. Cho dữ liệu huấn luyện gồm m mẫu D = { (d1, y1) ,..., (dm, ym) } trong đó di là nội dung thư thứ i và yi là nhãn

phân loại của thư đó. Nhãn phân loại yi có thể nhận một trong k giá trị của tập nhãn phân loại C = { c1, ..., ck}. Trong trường hợp đang xét, mỗi giá trị ci tương ứng với một thư mục mà thư được xếp vào, k là số lượng thư mục như vậy. Yêu cầu của bài toán là sử dụng dữ liệu huấn luyện để xây dựng hàm phân loại cho phép gán cho thư mới một trong các nhãn phân loại y ∈ C.

Để giải quyết bài toán phân loại thư phát biểu như trên cần giải quyết hai vấn đề:

- Thứ nhất, cần biểu diễn nội dung thư dưới dạng phù hợp với thuật toán phân loại.
- Thứ hai, lựa chọn phương pháp phân loại phù hợp và tiến hành huấn luyện bộ phân loại.

Dưới đây, chúng tôi sẽ trình bày giải pháp cho hai vấn đề trên.

2.1. Biểu diễn nội dung văn bản

Phương pháp thường dùng để biểu diễn nội dung văn bản là sử dụng phương pháp "túi từ" ("bag-of-words"). Phần nội dung d của mỗi thư được biểu diễn bởi một vectơ $\vec{x}$ = (x1, x2, ..., xn), trong đó x1, x2, ..., xn là giá trị của đặc trưng X1, X2,..., Xn. Mỗi đặc trưng có thể là một từ hoặc một cụm từ. Ở đây, n là số lượng đặc trưng được xác định từ toàn bộ tập dữ liệu huấn luyện, tức là số lượng từ/cụm từ khác nhau trong tập dữ liệu huấn luyện. Cách xác định n cũng như việc quyết định từ hay cụm từ nào được coi là đặc trưng sẽ được đề cập trong một phần sau của bài báo.

Có nhiều cách tính xi (i = 1,..., n) (tổng quan về các cách tính giá trị đặc trưng khác nhau có thể xem trong [11]). Trong nghiên cứu này chúng tôi sử dụng hai cách tính xi: 1) xi là số lần xuất hiện đặc trưng Xi trong văn bản đang xét và 2) xi = 1 nếu Xi xuất hiện trong văn bản và xi = 0 nếu ngược lại. Trong trường hợp thứ hai, đặc trưng được gọi là đặc trưng nhị phân vì có thể có một trong hai giá trị.

Trong nghiên cứu đang thực hiện, nội dung thư được bổ sung thêm nội dung các trường "From", "To", và "Subject" của thư. Các trường này được coi như những đoạn văn bản thông thường khác.

2.2. Phân loại Bayes đơn giản

Để xác định nhãn phân loại cho thư, bộ phân loại Bayes tính các xác suất điều kiện

P (y = ci | X1 = x1,..., Xn = xn) , i = 1,..., k

Hay viết đơn giản là P (ci | $\vec{x}$) , i = 1,..., k

tức là xác suất một thư với nội dung (x1, x2, ..., xn) nhận nhãn phân loại ci ∈ C. Sử dụng công thức Bayes, xác suất trên được tính như sau

$$P(y = c_i \mid X_1 = x_1,..., X_n = x_n) = \frac{P(X_1 = x_1,..., X_n = x_n \mid y = c_i) \cdot P(y = c_i)}{P(X_1 = x_1,..., X_n = x_n)} \quad (1)$$

Thư sẽ được gán nhãn co tương ứng với giá trị xác suất điều kiện tính theo (1) lớn nhất

$$c_o = \arg\max_{c_i \in C} P(y = c_i \mid X_1 = x_1,..., X_n = x_n) \quad (2)$$

Trong công thức (1), giá trị mẫu số không phụ thuộc vào ci, do vậy công thức (2) có thể viết thành

$$c_o = \arg\max_{c_i \in C} P(X_1 = x_1,..., X_n = x_n \mid y = c_i) \cdot P(y = c_i) \quad (3)$$

Trên thực tế, có thể giá trị P (co | $\vec{x}$) không chênh lệch nhiều so với các giá trị P (ci | $\vec{x}$), i ≠ o, tức là xác suất thư thuộc một loại cụ thể là không rõ ràng. Để giải quyết vấn đề này, thư chỉ được gán nhãn co nếu

$$\frac{P(c_o \mid \vec{x})}{P(c_s \mid \vec{x})} > T \quad (4)$$

trong đó cs là phân loại có xác suất lớn thứ nhì sau co và T là giá trị ngưỡng thể hiện sự chắc chắn của quyết định. Trong trường hợp không vượt qua được ngưỡng này, thư sẽ được coi là có phân loại không rõ ràng và không được gán nhãn phân loại. Chương trình do chúng tôi xây dựng sử dụng T = 100, tuy nhiên người dùng có thể thay đổi giá trị này.

Để xác định co ta cần tính các xác suất trong vế phải của (3). Xác suất P(y = ci) trên tập dữ liệu huấn luyện có thể tính dễ dàng bằng cách đếm tần suất xuất hiện của thư có nhãn ci. Việc xác định P ($\vec{x}$|ci) phức tạp hơn nhiều do phải tính tất cả các tổ hợp giá trị của vectơ $\vec{x}$ và đòi hỏi lượng dữ liệu huấn luyện lớn tương ứng. Để khắc phục vấn đề này, phương pháp Bayes đơn giản sử dụng một số giả thiết về tính độc lập xác suất của các đặc trưng nếu đã biết nhãn phân loại. Có một số cách tính giá trị khác nhau P ($\vec{x}$| ci) tương ứng với các phiên bản khác nhau của phương pháp phân loại văn bản sử dụng Bayes đơn giản. Trong nghiên cứu này, chúng tôi sẽ sử dụng phiên bản Bayes đơn giản với mô hình đa thức (multinomial naïve Bayes) do mô hình này cho kết quả tốt trong bài toán tương tự [9].

Mô hình đa thức coi nội dung thư d sinh ra bằng cách lấy ngẫu nhiên có lặp |d| đặc trưng từ tập đặc trưng chung F với xác suất P(fi| c) cho mỗi fi (|d| là số lượng đặc trưng trong thư d). Sử dụng thêm giả thiết là |d| không phụ thuộc vào nhãn c cho phép tính P ($\vec{x}$| ci) theo xác suất đa thức như sau:

$$P(X_1 = x_1,..., X_n = x_n \mid y = c) = P(|d|).|d|!.\prod_{i=1}^{n} \frac{P(f_i \mid y = c)^{x_i}}{x_i!} \quad (5)$$

Xác suất P(fi | y = c) được tính từ dữ liệu huấn luyện theo công thức

$$P(f_i \mid y = c) = \frac{N_{c,f_i} + 1}{N_c + n} \quad (6)$$

Trong đó Nc là số lượng thư với nhãn phân loại c và Nc,fi là tổng số lần đặc trưng fi xuất hiện trong các thư có nhãn c.

Thay (5) vào (3) ta được công thức cho phép xác định co.

Cũng như trong [9], chúng tôi sử dụng các đặc trưng nhị phân (chỉ nhận giá trị 1 hoặc 0) cho mô hình đa thức do cách biểu diễn này cho kết quả tốt hơn.

2.3. Phân loại bằng cách xếp hạng

Với một thư d và tập nhãn phân loại C = {c1, ..., ck}, có thể thực hiện phân loại bằng cách xếp hạng các nhãn phân loại theo mức độ phù hợp của nhãn đối với d. Phân loại xếp hạng cao nhất sẽ được gán cho thư. Chương trình phân loại thư điện tử của chúng tôi sử dụng một thuật toán phân loại dựa trên xếp hạng có tên là Multi-class Multilabel Perceptron (MMP) [4]. Đây là thuật toán dựa trên nguyên lý huấn luyện perceptron và được đề xuất cho bài toán phân loại văn bản trong đó mỗi văn bản có thể được phân thành nhiều loại khác nhau. Ở đây, chúng tôi sử dụng phiên bản cho trường hợp văn bản chỉ có thể nhận một nhãn phân loại duy nhất.

Tương tự như với phân loại Bayes, văn bản d được biểu diễn bằng vectơ $\vec{x}$ = (x1, x2, ..., xn). Thuật toán xếp hạng sử dụng một tập gồm k mẫu $\vec{w}_1,...,\vec{w}_k$, mỗi mẫu là một vectơ gồm n phần tử và tương ứng với một nhãn phân loại. Cần nhắc lại rằng, k là số lượng phân loại của bài toán. Các mẫu được xếp hạng dựa trên độ tương tự của mẫu đó với vectơ $\vec{x}$ của văn bản. Độ tương tự giữa mẫu $\vec{w}_i$ với $\vec{x}$ được tính bằng tích vô hướng $\vec{w}_i.\vec{x}$ giữa hai vectơ. Nhãn phân loại tương ứng với mẫu được xếp hạng cao nhất sẽ được gán cho văn bản.

Nhiệm vụ của giai đoạn huấn luyện là xác định giá trị các vectơ $\vec{w}_1,...,\vec{w}_k$ từ dữ liệu huấn luyện. Thuật toán huấn luyện MMP là thuật toán huấn luyện trực tuyến: thuật toán nhận một ví dụ huấn luyện, xác định nhãn cho ví dụ đó bằng cách chọn nhãn xếp hạng cao nhất, nếu nhãn tính được không trùng với nhãn của ví dụ thì cập nhật các mẫu $\vec{w}_i$.

Nguyên lý cập nhật tham số tương đối đơn giản: với mỗi ví dụ phân loại sai, tính lại vectơ $\vec{w}_i$ sao cho vectơ tương ứng với nhãn phân loại đúng di chuyển về phía $\vec{x}$ (và do vậy tăng giá trị của tích vô hướng) trong khi các vectơ khác di chuyển ra xa $\vec{x}$. Khoảng cách di chuyển các vectơ được tính như sau. Gọi ej là số lượng nhãn phân loại bị xếp loại sai so với nhãn cj. Nếu cj là phân loại thực thì ej là số nhãn phân loại được xếp loại cao hơn hoặc bằng cj. Nếu cj không phải nhãn thực nhưng lại được xếp hạng cao nhất thì ej = 1 (nhãn thực bị xếp hạng sai), ngược lại ej = 0. Mỗi vectơ $\vec{w}_j$ khi đó được cập nhật một lượng bằng $e_j / \sum_j e_j$. Thuật toán huấn luyện MMP được thể hiện trên hình 1.

1. Khởi tạo $\vec{w}_1 = ... = \vec{w}_k = 0$
2. Lặp với i = 1, ..., m
   a. Chọn ví dụ huấn luyện $(\vec{x}_i, y_i)$
   b. Tính nhãn phân loại z_i cho $\vec{x}_i$
   c. Nếu $z_i \neq y_i$ thì thực hiện:
      i. với $c_j = y_i$ tính e_j = số lượng nhãn phân loại xếp trên hoặc bằng c_j
      ii. với $c_j \neq y_i$ gán $e_j = 1$ nếu xếp hạng trên y_i, $e_j = 0$ nếu ngược lại
      iii. với $c_j = y_i$, cập nhật

$$\vec{w}_j \leftarrow \vec{w}_j + \frac{e_j}{\sum_j e_j} \vec{x}_i$$

      iv. Với $c_j \neq y_i$, cập nhật

$$\vec{w}_j \leftarrow \vec{w}_j - \frac{e_j}{\sum_j e_j} \vec{x}_i$$

3. Trả về $\vec{w}_1, ..., \vec{w}_k$

***Hình 1* - Thuật toán huấn luyện MMP**

3. XÂY DỰNG BỘ PHÂN LOẠI EMAIL TIẾNG VIỆT

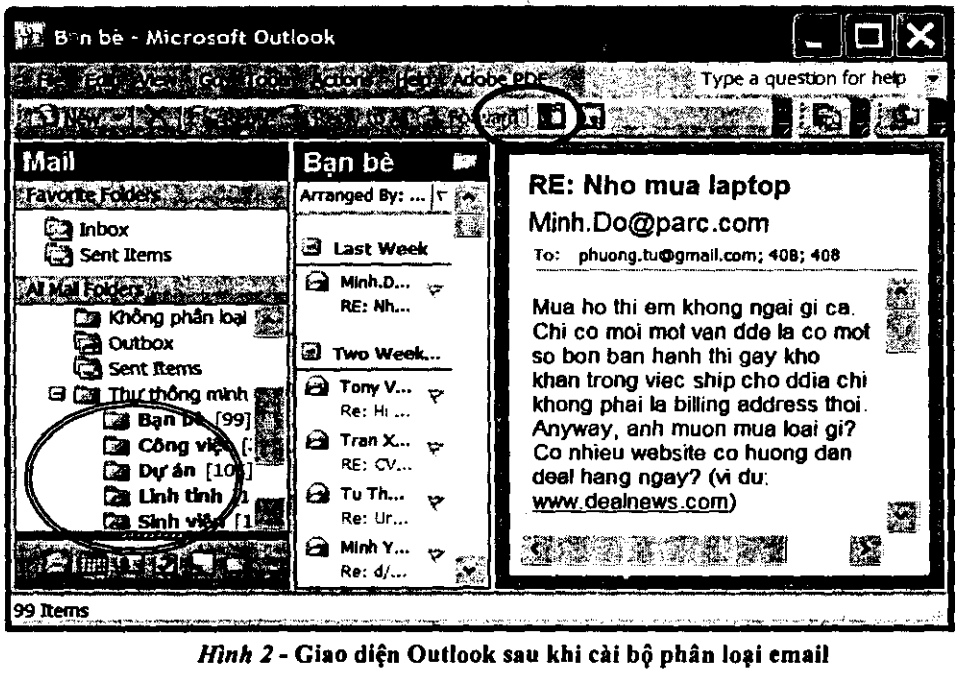
Trong phần này, chúng tôi mô tả chương trình phân loại thư điện tử do chúng tôi xây dựng sử dụng hai thuật toán phân loại mô tả ở trên. Ngoài các mô tả chung về chương trình, chúng tôi cũng trình bày giải pháp biểu diễn thư và lựa chọn đặc trưng phù hợp với thư viết bằng tiếng Việt.

3.1. Mô tả chung

Bộ phân loại thư được xây dựng dưới dạng một add-in có thể tích hợp vào chương trình thư điện tử thông dụng Outlook của Microsoft. Việc xây dựng add-in cho Outlook tương đối thuận lợi (ví dụ nếu so sánh với Outlook Express) do Microsoft có cung cấp giao diện lập trình API cho ứng dụng này.

Sau khi cài đặt bộ phân loại, trên thanh công cụ của Outlook sẽ được bổ sung hai nút cho phép người dùng kích hoạt chế độ huấn luyện và đặt tham số cho bộ phân loại thư (các nút trong hình ôvan có viền dầy trên hình 2).

Để sử dụng bộ phân loại, người dùng cần tạo ra các thư mục con trong thư mục "Thư thông minh" theo nhu cầu riêng của mình. Số lượng thư mục con có thể tạo ra là không hạn chế. Trong hình ôvan với đường viền đôi trên hình 2 là ví dụ các thư mục "Bạn bè", "Công việc", "Dự án".v.v.



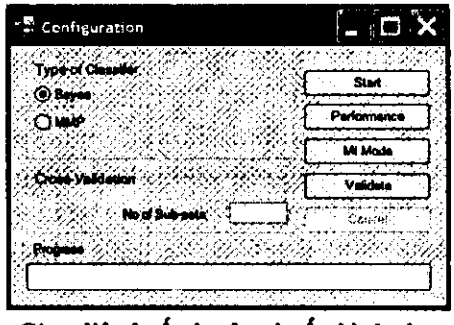
***Hình 2* - Giao diện Outlook sau khi cài bộ phân loại email**

Mỗi khi có một thư mới xuất hiện, bộ lọc sẽ sử dụng thuật toán Bayes hoặc MMP để phân loại thư. Trong trường hợp có thể phân loại chắc chắn (theo biểu thức (4) đối với Bayes hoặc xếp hàng cao hơn hẳn đối với MMP), thư sẽ được đặt vào thư mục tương ứng. Nếu phân loại là không chắc chắn, thư sẽ không được phân loại mà đặt vào thư mục quy định trước có tên "Không phân loại được". Thử nghiệm cho thấy, ở giai đoạn đầu khi chưa có hoặc ít dữ liệu huấn luyện, bộ phân loại sẽ chuyển hầu hết thư vào "Không phân loại

được". Trong trường hợp này, người dùng phải chuyển thư từ "Không phân loại được" vào thư mục cần thiết. Khi phát hiện thư bị phân loại sai, người dùng cũng cần chuyển thư về đúng thư mục bằng tay.

Sau khi chuyển thư về thư mục đúng, người dùng cần kích hoạt bộ phân loại để huấn luyện lại. Dữ liệu huấn luyện khi đó là những thư đã được đặt đúng trong thư mục con của mình. Thông thường, đối với Bayes, việc huấn luyện lại nên thực hiện sau khi đã có sự cập nhật đáng kể trong dữ liệu huấn luyện, ngược lại, có thể kích hoạt MMP trong trường hợp dữ liệu được cập nhật tối thiểu, được gọi là huấn luyện tăng dần.

Để thay đổi các tuỳ chọn của bộ phân loại, người dùng có thể kích vào nút bên phải trong hai nút đã nhắc tới ở hình 2. Giao diện tuỳ chọn được thể hiện trên hình 3. Trên giao diện này có chế độ "Validate" cho phép người dùng tự kiểm tra độ chính xác phân loại.



Hình 3 - **Giao diện huấn luyện và cấu hình chương trình**

3.2. Lựa chọn đặc trưng

Trong hầu hết các nghiên cứu lọc thư rác tiếng Anh, đặc trưng được sử dụng là những từ riêng lẻ (word). Do đặc điểm của tiếng Anh nên việc xác định từ trong câu rất đơn giản, mỗi từ được phân cách với từ khác bằng dấu cách hoặc các dấu trắng khác.

Đối với tiếng Việt, từ có thể bao gồm nhiều tiếng, ví dụ từ "đặc trưng" bao gồm hai tiếng "đặc" và "trưng". Trong khi có thể tách từng tiếng một cách dễ dàng thì việc xác định từ hoàn toàn không đơn giản. Ngoài ra, do từ bao gồm nhiều tiếng, việc sử dụng tần suất các đặc trưng như trong phân loại Bayes có thể bị ảnh hưởng và cho kết quả khác với thư tiếng Anh.

Theo kết quả thực nghiệm trong [9] đối với trường hợp văn bản bao gồm cả tiếng Việt và tiếng Anh, đặc trưng nên được xác định dưới dạng các k-gram với k = 1,2. Tức là nội dung thư sẽ được tách thành các tiếng, mỗi tiếng riêng lẻ được coi là một đặc trưng, kết hợp của hai tiếng gần nhau cũng được coi là một đặc trưng. Phương pháp này rất đơn giản so với một số kỹ thuật tách từ tiếng Việt khác, trong khi vẫn cho kết quả phân loại tương đối tốt, và do vậy được lựa chọn sử dụng trong bộ phân loại thư của chúng tôi. Trong khi

tách từ, chúng tôi không phân biệt tiếng Anh và tiếng Việt, như vậy tập đặc trưng sẽ bao gồm cả đặc trưng tiếng Việt và tiếng Anh (nếu có).

Sau khi tách được các đặc trưng dưới dạng k-gram như trên, vấn đề tiếp theo là quyết định số lượng đặc trưng sẽ sử dụng. Nếu thống kê trong toàn bộ tập dữ liệu huấn luyện thì số lượng đặc trưng có thể lên tới vài chục nghìn. Rất nhiều đặc trưng - từ không liên quan tới phân loại của thư và cần loại bỏ.

Chúng tôi sử dụng hai phương pháp chọn đặc trưng. Phương pháp thứ nhất loại bỏ những đặc trưng xuất hiện trong quá ít thư hoặc xuất hiện trong quá nhiều thư. Nếu đặc trưng xuất hiện trong quá ít thư thì đó là những đặc trưng xuất hiện tình cờ và không phụ thuộc vào nhãn phân loại. Trong thực nghiệm, những đặc trưng xuất hiện trong ít hơn 3 thư sẽ bị loại. Ngược lại, nếu đặc trưng xuất hiện trong hầu hết các thư thì đó là những đặc trưng phổ biến trong bất kỳ thư nào và do vậy cũng không chứa thông tin về phân loại của thư. Trong các thực nghiệm, chúng tôi sẽ loại bỏ những đặc trưng xuất hiện thường xuyên nhất.

Phương pháp thứ hai sử dụng độ đo thông tin tương hỗ (mutual information – MI) để lựa chọn đặc trưng. MI là độ đo mức độ liên quan về thông tin giữa hai biến ngẫu nhiên, tức là khi biết giá trị biến này thì ta có thể biết được gì về giá trị biến kia. Trong trường hợp lọc thư, hai biến ngẫu nhiên là giá trị đặc trưng và nhãn phân loại. MI được tính như sau

$$MI(X,c_i) = \sum_{x\in\{0,1\}, c_i\in C} P(X=x, y=c_i)\log\frac{P(X=x, y=c_i)}{P(X=x)P(y=c_i)} \qquad (7)$$

Các xác suất P(X,ci), P(X) và P(ci) được tính bằng tần suất xuất hiện của các sự kiện tương ứng trên dữ liệu huấn luyện. Sau khi đã tính MI cho tất cả các đặc trưng k-gram, n đặc trưng có MI cao nhất sẽ được lựa chọn.

Cũng theo kết quả được trình bày trong [9] chúng tôi sử dụng hai phương pháp lựa chọn đặc trưng vừa trình bày để chọn ra tối đa 3000 đặc trưng có MI cao nhất. Lúc đầu, khi dữ liệu huấn luyện chưa nhiều, số lượng đặc trưng sẽ ít hơn 3000, những phần tử còn lại của mẫu $\vec{w}_i$ sẽ được lấp đầy bằng giá trị 0.

4. Thử nghiệm

Trong phần này, chúng tôi trình bày kết quả thử nghiệm về độ chính xác phân loại thư với số lượng dữ liệu huấn luyện khác nhau. Nói chung, độ chính xác phân loại thư phụ thuộc rất nhiều vào việc xác định thư mục (phân loại) của người dùng. Nếu các thư mục liên quan nhiều tới nội dung thư và không bị chồng chéo nhau thì độ chính xác sẽ cao hơn trong trường hợp ngược lại. Do vậy, kết quả thử nghiệm trình bày dưới đây chủ yếu để so

sánh hai phương pháp phân loại chứ không mang tính tổng quát cho mọi đối tượng người dùng và cách phân chia thư mục khác nhau.

4.1. Dữ liệu thử nghiệm

Do không có các bộ dữ liệu thư tiếng Việt chuẩn dùng cho thử nghiệm nên chúng tôi tự xây dựng hai bộ dữ liệu để dùng trong các thử nghiệm của mình. Bộ dữ liệu thứ nhất bao gồm 1536 thư là thư của một trong hai tác giả nhận được từ giữa năm 2006 đến nay. Các thư này bao gồm các thư tiếng Anh, thư tiếng Việt có dấu và không dấu. Bộ dữ liệu thứ hai khoảng 500 thư được thu thập từ hộp thư của một số đồng nghiệp và bạn bè.

Đối với các thư bình thường nhận được, trong trường hợp thư nhận từ cùng một nguồn qua nhiều phiên gửi và reply thì đối với những thư gửi sau sẽ được xoá phần đã gửi từ trước, chỉ giữ lại nội dung thư nhận được cuối cùng. Đối với những thư bao gồm cả văn bản và hình ảnh, chỉ có phần văn bản được sử dụng, phần hình ảnh bị bỏ qua không xem xét.

Trong quá trình tiền xử lý, tất cả các thẻ HTML, XML, các đoạn script đều bị loại bỏ. Những thông tin này, xét về mặt nào đó cũng có thể sử dụng trong quá trình phân loại. Tuy nhiên, do phương pháp chính được sử dụng ở đây là phân loại theo nội dung nên ta không xét đến những thông tin này.

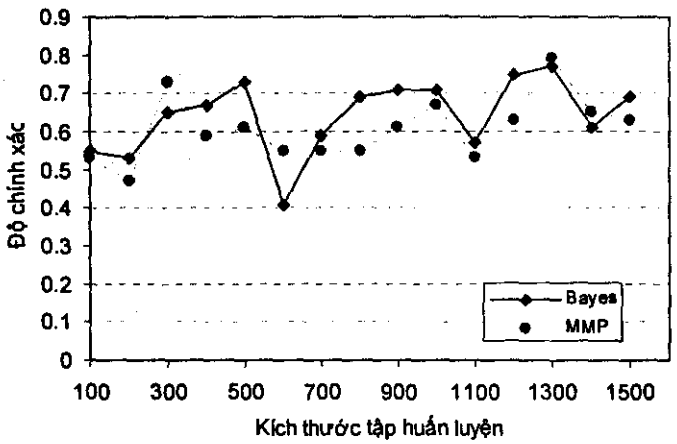
4.2. Phương pháp thử nghiệm và kết quả

Việc thử nghiệm phân loại thư cần tính đến yếu tố tuần tự thời gian của thư, cụ thể, thư nhận trước được dùng huấn luyện để phân loại thư sau. Do đó, cách thường dùng thử nghiệm là sắp xếp thư theo thứ tự thời gian, sau thực hiện tiếp như sau: sử dụng N thư đầu tiên để huấn luyện và tính độ chính xác phân loại trên N thư kiểm tra tiếp theo, sau đó sử dụng 2N thư đầu để huấn luyện và tính độ chính xác trên N thư tiếp theo. Tiếp tục như vậy cho đến khi sử dụng hết bộ dữ liệu. Trong các thử nghiệm đã tiến hành, giá trị N được chọn bằng 100 cho bộ dữ liệu thứ nhất và 40 cho bộ dữ liệu thứ hai. Độ chính xác phân loại được tính bằng tỷ lệ số thư được phân loại đúng vào thư mục của mình cho từng tập huấn luyện/kiểm tra với kích thước phần huấn luyện tăng dần như trên.

Chúng tôi phân chia thư trong hai tập dữ liệu thành các thư mục sau: "Các dự án", "Công việc", "Bạn bè", "Sinh viên", "Linh tinh". Cần lưu ý rằng, bộ dữ liệu thứ hai không có đủ cả năm thư mục trên do phần "Sinh viên" không có thư nào. Ngoài ra, trên thực tế có thể chia nhỏ nhiều thư mục nhưng chúng tôi chỉ dùng năm thư mục cho đơn giản.

Trong quá trình thực nghiệm, chúng tôi sử dụng ngưỡng T = 1 cho công thức (4), tức là bất cứ thư nào cũng đều được phân loại. Nếu nhiều thư mục có xếp hạng hoặc xác suất điều kiện cùng cao nhất và bằng nhau thì một thư mục sẽ được chọn ngẫu nhiên.

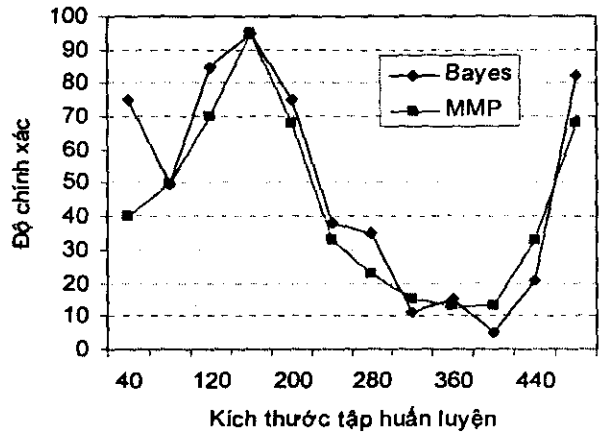
Độ chính xác phân loại cho từng tập huấn luyện/kiểm tra với kích thước tập huấn luyện tăng dần cho bộ dữ liệu thứ nhất và thứ hai được thể hiện tương ứng trên hình 4 và 5. Trong hai đồ thị này, trục x là kích thước phần huấn luyện và y là độ chính xác phân loại tính bằng. Đường liền nét thể hiện độ chính xác của phương pháp Bayes và đường không liền nét là độ chính xác của MMP.



***Hình 4* - Kết quả phân loại trên bộ dữ liệu thứ nhất**

Kết quả thử nghiệm cho thấy, độ chính xác phân loại phụ thuộc vào đặc điểm và thứ tự xuất hiện thư. Tại những thời điểm có thư thuộc loại mới xuất hiện trong khi tập huấn luyện chưa có nhiều thư loại này, độ chính xác sẽ bị ảnh hưởng rõ rệt. Do vậy, việc huấn luyện lại sau khi có các thư mới là cần thiết và do vậy cần có chế độ huấn luyện tăng dần. Đối với chế độ huấn luyện tăng dần, ưu điểm về thời gian rõ ràng thuộc về MMP do không đòi hỏi tính toán lại trên toàn bộ thư trước đó. Thậm chí, việc huấn luyện lại MMP có thể thực hiện rất nhanh trên từng thư mới, trong khi điều này là không thực tế với Bayes, đặc biệt khi lượng thư cũ lớn.

Nhìn chung, độ chính xác phân loại của hai phương pháp không chênh lệch nhau đáng kể. Nếu lưu ý rằng, phân loại Bayes đơn giản là phương pháp đại diện và phổ biến nhất cho bài toán phân loại văn bản nói chung và email nói riêng thì có thể kết luận phương pháp MMP đại diện cho thuật toán huấn luyện trực tuyến có thể sử dụng tương đương cho bài toán này trong khi có ưu thế hơn về việc huấn luyện tăng dần.



Hình 5 - **Kết quả phân loại trên bộ dữ liệu thứ hai**

Kết luận

Bài báo đã trình bày kết quả xây dựng chương trình phân loại thư điện tử sử dụng các thuật toán xếp hạng và phân loại Bayes đơn giản. Với việc sử dụng kỹ thuật phù hợp để biểu diễn nội dung thư, chương trình có khả năng phân loại cả thư viết bằng tiếng Anh và thư viết bằng tiếng Việt có dấu cũng như không dấu. Một kết quả quan trọng của bài báo là việc thử nghiệm sử dụng một thuật toán xếp hạng trực tuyến cho bài toán phân loại thư. Kết quả thử nghiệm trên hai bộ dữ liệu thực cho thấy, thuật toán xếp hạng cho hiệu quả phân loại tương đương với phân loại Bayes đơn giản trong khi rất phù hợp với chế độ huấn luyện tăng dần và do vậy có thể là một lựa chọn tốt cho bài toán phân loại thư điện tử.

Do việc gán nhãn cho thư trong giai đoạn đầu đòi hỏi nhiều thời gian từ phía người dùng, việc tận dụng những thư đã được phân loại cộng với thư chưa được phân loại cho quá trình huấn luyện bằng các phương pháp học máy nửa giám sát (semi-supervised learning) là một giải pháp cần xem xét cho bài toán này. Đây cũng là hướng phát triển tiếp theo cho nghiên cứu phân loại thư của chúng tôi.

Tài liệu tham khảo

[1] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Spyropoulos, C. D. 2000. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-

Spam Filtering with Personal E-mail Messages. In Proc. of the 23 rd Annual International ACM SIGR Conference on Research and Development in Information Retrieval, pp. 160-167, Athens, Greece, 2000.

[2] R. Bekkerman, A. McCallum and G. Huang. Automatic Categorization of Email into Folders: Benchmark Experiments on {E}nron and {SRI} Corpora. UMass CIIR Technical Report IR-418, 2004.

[3] K. Crammer and Y. Singer. A New Family of Online Algorithms for Category Ranking. Proceedings of SIGIR*02, 2002.

[4] K. Crammer and Y. Singer. A Family of Additive Online Algorithms for Category Ranking. Journal of Machine Learning Research 3, 2003.

[5] E. Crawford, I. Koprinska, J. Patrick, A multi-learner approach to e-mail classification, in: Proc. 7th Australasian Document Computing Symposium (ADCS), 2002.

[6] Irena Koprinska, Josiah Poon, James Clark and Jason Chan. Learning to classify e-mail. Information siences. 177(10) 2007.

[7] K. Mock. An Experimental Framework for Email Categorization and Management. SIGIR 01.

[8] Koprinska, F. Trieu, J. Poon, J. Clark, E-mail classification by decision forests, in: Proc. 8th Australasian. Document Computing Symposium (ADCS), 2003.

[9] Nguyễn Duy Phương, Phạm Văn Cường, Từ Minh Phương. Một số giải pháp lọc thư rác tiếng Việt. Kỷ yếu hội thảo quốc gia Một số vấn đề chọn lọc về CNTT. Đà lạt 2006.

[10] Schneider K., A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering, Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003

[11] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (2002) 1–47.

[12] Richard B. Segal and Jeffrey O. Kephart. MailCat: An Intelligent Assistant for Organizing E-Mail. Proceedings of the Third International Conference on Autonomous Agents. 1999.

[13] Y. Yang, J. Pedersen, A comparative study on feature selection in text categorization, in: Proc. 4th International Conference on Machine Learning, 1997.

[14] Popfile: http://popfile.sourceforge.net/