Tóm tắt nội dung khóa luận

Khóa luận trình bày một số nội dung cơ bản nhất về thư rác (khái niệm, tác hại, các hình thức phát tán thư rác...), tập trung định hướng tới các phương pháp lọc thư rác, đặc biệt là phương pháp lọc dựa trên nội dung.

Trong các phương pháp lọc theo nội dung, khóa luận quan tâm mô tả, phân tích hệ thống hệ thống *Email Classification Using Examples* (ECUE), một phương pháp lọc spam dựa trên nội dung do Delany và Cunningham đề xuất năm 2004 [4]. Khóa luận mô tả kiến trúc của CBR và kiến trúc hệ thống ECUE. Hệ thống ECUE có khả năng giải quyết được vấn đề concept drift, hệ thống được xây dựng dựa trên phương pháp Case-Based Reasoning (CBR) [1] với việc coi các email là các case, tập các case đã được phân lớp spam, non-spam được sử dụng làm tập dữ liệu huấn luyện gọi là case-base. Để giải quyết vấn đề concept drift ECUE có hai thành phần chính là: Case-base Editing và case-base update policy [5]. Phần cuối cùng của khóa luận trình bày về kết quả thực nghiệm tiến hành trên hệ thống lọc thư rác sử dụng thuật toán Bayes theo chương trình Spambayes.

Mở đầu

Một trong những dịch vụ mà Internet mang lại đó là dịch vụ thư điện tử, đó là phương tiện giao tiếp rất đơn giản, tiện lợi, rẻ và hiệu quả giữa mọi người trong cộng đồng sử dụng dịch vụ Internet. Tuy nhiên chính vì những lợi ích của dịch vụ thư điện tử mang lại mà số lượng thư trao đổi trên Internet ngày càng tăng, và một số không nhỏ trong số đó là thư rác (spam). Thư rác thường được gửi với số lượng rất lớn, không được người dùng mong đợi, thường với mục đích quảng cáo, đính kèm virus, gây phiền toái khó chịu cho người dùng, làm giảm tốc độ truyền internet và tốc độ xử lý của email server, gây thiệt hại rất lớn về kinh tế.

Đã có rất nhều phương pháp đưa ra để giảm số lượng thư rác. Như việc đưa ra các luật lệ để hạn chế việc gửi thư rác, đưa ra các phương pháp kĩ thuật lọc thư rác như: lọc dựa trên địa chỉ IP (whitelist, balacklist), lọc dựa trên danh tính người gửi, lọc dựa trên chuỗi hỏi đáp, phương pháp lọc dựa trên mạng xã hội, và phương pháp lọc nội dung…Mỗi phương pháp đều có ưu nhược điểm riêng, không có phương pháp nào là hoàn hảo vì vậy để có bộ lọc thư rác tốt cần phải kết hợp các phương pháp với nhau. Trong các phương pháp lọc thư rác phương pháp lọc dựa trên nội dung hiện đang được quan tâm nhiều, và được đánh giá là có triển vọng đưa ra kết quả cao. Phương pháp lọc nội dung dựa trên việc phân tích nội dung của email để phân biệt spam email và nonspam email.

Tuy đã có nhiều biện pháp ngăn chặn thư rác nhưng số lượng thư rác vẫn càng ngày càng nhiều, tác hại gây ra càng lớn, cấu trúc nội dung của thư càng ngày càng thay đổi tinh vi hơn để vượt qua các bộ lọc vì vậy cần có một hệ thống lọc có khả năng giải quyết được vấn đề thư rác ngày càng tăng, nội dung, cấu trúc của thư ngày càng phức tạp tinh vi hơn (concept drift).

Đã có nhiều hệ thống học máy lọc thư rác sử dụng các thuật toán Naïve bayes, phân lớp dựa trên thống kê (Lewis and Ringuette 1994, Lewis 1998), Support Vector Machines (Joachims 1998, Dumais et al. 1998) các phương pháp này đều cho kết quả lọc khá tốt[17]. Tuy nhiên các mô hình này chưa giải quyết được vấn đề concept drift . Một mô hình mới đã được Delany(2006) đề xuất, dựa trên hệ thống học máy sử dụng phương

pháp Case-Based Reasoning (CBR)(Riesbeck and Shank 1989)[17] có khả năng giải quyết được concept drift. Phương pháp CBR, sử dụng các vấn đề trước đây đã được giải quyết để đưa ra giải pháp cho vấn đề mới. Các vấn đề đã được giải quyết được lưu vào tập dữ liệu dùng để huấn luyện gọi là case-base. Các case được biểu diễn dưới dạng véc tơ n chiều, mỗi thành phần là một token đã được trích chọn từ việc phân tích cú pháp, phân tích từ tố của tài liệu (email). Các vector cũng chứa thêm một thành phần nữa chỉ lớp mà tài liệu đó được phân (nonspam, spam).

Trong việc ứng dụng CBR để lọc thư rác có hai vấn đề chính là: làm thế nào để quản lý được tập dữ liệu huấn luyện(case-base), chứa một số lượng lớn email của người dùng. Thứ hai là làm thế nào để điều khiển được vấn đề concept drift. Để quản lý được dữ liệu huấn luyện CBR áp dụng các luật để điều chỉnh case-base(case-base Editing), nhằm đưa ra tập case-base chứa các case có khả năng dự đoán cao nhất cho việc phân lớp case mới. Để giải quyết được concept drift CBR thực hiện việc lựa chọn lại các đặc trưng và case mới tốt nhất cho việc xác định lớp cho case mới.

Trong khóa luận này tôi xin trình bày hướng tiệp cận của Email Classification Using Example (ECUE)(Delany, Cunningham, 2004), phương pháp học máy lọc thư rác dựa trên CBR. Trong ECUE có hai phần chính cần quan tâm là: Công nghệ sử dụng cho Case-base Editing là Competence Based Editing(CBE)(Smyth và McKenna 1998); và Case-base update policity. CBE có hai chức năng chính là loại bỏ case nhiễu và case dư thừa, việc loại bỏ case nhiễu áp dụng thuật toán Blame Based Noise Reduction (BBNR), việc loại bỏ case dư thừa áp dụng thuật toán Conservative Redundancy Reduction (CRR)(Riesbeck and Shank 1989) [17]. Case-base update policy thực hiện việc đưa các case đã được phân lớp là spam, nonspam vào case-base để đưa dự đoán lớp cho case tiếp theo, trong trường hợp cho case học lại, case-base update policy thực hiện lựa chọn lại các đặc trưng để tìm ra đặc trưng có ích trong việc dự đoán lớp cho case mới.

Chương 1

THƯ RÁC VÀ CÁC PHƯƠNG PHÁP LỌC THƯ RÁC

Một trong những dịch vụ mà Internet mang lại đó là dịch vụ thư điện tử, đó là phương tiện giao tiếp rất đơn giản, tiện lợi, rẻ và hiệu quả giữa mọi người trong cộng đồng sử dụng dịch vụ Internet. Tuy nhiên chính vì những lợi ích của dịch vụ thư điện tử mang lại mà số lượng thư trao đổi trên Internet ngày càng tăng, và đa số trong số những thư đó là thư rác (spam). Thư rác thường được gửi với số lượng rất lớn, không được người dùng mong đợi, thường với mục đích quảng cáo, đính kèm virus, gây phiền toái khó chịu cho người dùng, làm giảm tốc độ truyền internet và tốc độ xử lý của email server, gây thiệt hại rất lớn về kinh tế. Chương này sẽ khái quát các vấn đề về khái niệm thư rác, ảnh hưởng của thư rác trong cuộc sống của chúng ta và các phương pháp ngăn chặn thư rác.

1.1 Một số khái niệm cơ bản

1.1.1 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). Sau đây sẽ đưa ra một định nghĩa thông dụng nhất về thư rác và giải thích các đặc điểm của nó để phân biệt thư rác với thư thông thường [18,19]:

Thư rác (spam mail) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận.

Một bức thư nếu gửi không theo yêu cầu có thể đó là thư làm quen hoặc thư được gửi lần đầu tiên, còn nếu thư được gửi hàng loạt thì nó có thể là thư gửi cho khách hàng của các công ty, các nhà cung cấp dịch vụ. Vì thế một bức thư bị coi là rác khi nó không được yêu cầu, và được gửi hàng loạt.

Tuy nhiên yếu tố quan trọng nhất để phân biệt thư rác với thư thông thường là nội dung thư. Khi một người nhận được thư rác, người đó không thể xác định được thư đó được gửi hàng loạt hay không nhưng có thể xác định được đó là thư rác sau khi đọc nội dung thư. Đặc điểm này chính là cơ sở cho giải pháp phân loại thư rác bằng cách phân tích nội dung thư.

1.1.2 Phân loại thư rác

Có rất nhiều cách phân loại thư rác[18] .

- ***Dựa trên kiểu phát tán thư rác***: Tính tới thời điểm hiện tại, thư rác có thể bị gửi thông qua thư điện tử, nhóm thảo luận (newsgroups), điện thoại di động (Short Message Service - SMS) và các dịch vụ gửi tin nhắn trên mạng (như Yahoo Messenger, Windows Messenger...)
- ***Dựa vào quan hệ với người gửi thư rác***: bao gồm người lạ mặt, bạn bè, người quen và các dịch vụ quyên góp giúp đỡ…
- ***Dựa vào nội dung của thư rác*****:** các kiểu nội dung phổ biến như thư về thương mại, thư về chính trị, thư về công nghệ, chuỗi thư (chain e-mail) và các loại khác (như thư phát tán virus...).
- ***Dựa trên động lực của người gửi*****:** Thông thường, thư rác được gửi đi cho những mục đích quảng bá thông tin. Ngoài ra, còn có một số loại thư rác được gửi tới một người nhận xác định nào đó nhằm mục đích phá vỡ và gây cản trở công việc của người nhận hay mạng của nhà cung cấp dịch vụ thư điện tử (ESP) được gọi là "bom thư". Thư rác còn được cố ý gửi đi nhằm thông báo tin sai lệch, làm xáo trộn công việc và cuộc sống của người nhận.

Sự phân loại thư rác rất quan trọng không chỉ trong lĩnh vực tạo những bộ lọc thư rác có hiệu quả cao mà còn giúp cho việc ban hành các bộ luật chống thư rác phù hợp.

1.1.3 Tác hại thư rác

Theo thống kê thư rác hiện chiếm hơn một nửa số e-mail truyền trên Internet và chính thư rác là nguồn lây lan virus nhanh nhất. Thiệt hại do chúng gây ra rất lớn đối với sự phát triển internet nói chung và người sử dụng thư điện tử nói riêng.

Theo thống kê toàn cầu của hãng nghiên cứu Ferris Research ở San Francisco [18], thư rác gây thiệt hại 50 tỷ USD trong năm 2005. Chỉ tính riêng ở Mỹ, thiệt hại do thư rác gây ra đối với các doanh nghiệp ước tính khoảng 17 tỷ USD/năm.

Thư rác chiếm khoảng 80% lưu lượng thư điện tử thế giới trong quý 1/2006, đó là kết luận của nhóm hợp tác chống thư rác gồm các công ty AOL, Bell Canada, Cigular Wireless, EarthLink, France Telecom, Microsoft, Verizon, và Yahoo. Microsoft và AOL cho biết hai hãng này trung bình mỗi ngày chặn gần 5 tỷ thư rác. Ước tính, cứ 9 trong 10 email sử dụng dịch vụ MSN Hotmail của Microsoft là thư rác[18].

Tại Việt Nam, tình hình thư rác cũng đang rất phức tạp. Công ty Điện toán và Truyền số liệu (VDC) - ISP lớn nhất Việt Nam - cho biết, thư rác hiện nay chiếm phần lớn lưu lượng email qua hệ thống máy chủ thư của ISP này.

Các thư phàn nàn gửi đến ISP nếu không giải quyết, các khách hàng của ISP đó có thể bị liệt vào danh sách đen, không gửi được email ra địa chỉ nước ngoài. Một số ISP cho biết, cuối năm ngoái, khách hàng của nhiều ISP ở Việt Nam thường xuyên bị tê liệt do bị liệt vào danh sách đen. Mỗi lần thoát ra khỏi danh sách này ISP phải mất khoảng 40 USD. Tại trang web Spamhaus.org (tổ chức theo dõi các nguồn gửi thư rác), có lần vnn.vn đã có trong danh sách top 10 ISP cung cấp nhiều rác nhất.

Không chỉ gây thiệt hại về tiền bạc, thư rác còn làm giảm hiệu quả làm việc, gây stress, tiêu tốn thời gian của nhân viên... Những điều này cũng đồng nghĩa với việc, năng suất lao động giảm, ảnh hưởng tới tình hình kinh doanh và doanh thu của công ty.

Một số lời khuyên cho người dùng thư điện tử:

- Yêu cầu và đòi hỏi nhà chức trách phải đưa ra những luật lệ nghiêm cấm thư rác và có hình phạt đích đáng cho kẻ cố tình gửi thư rác.
- Mỗi người dùng nên tạo nhiều địa chỉ email, với mục đích khác nhau nên dùng địa chỉ email khác nhau.
- Hạn chế việc đăng kí các dịch vụ vô ích: nên tìm hiểu kĩ thông tin về dịch vụ trước khi cung cấp địa chỉ email của mình.
- Kích hoạt các dịch vụ chống thư rác của ISP.
- Cài đặt một số chương trình xử lý thư trong máy tính cá nhân để xóa thư rác ngay khi chuyển về máy.

- Bảo vệ mật khẩu của mình: chọn mật khẩu lạ, khó đoán chứa chữ cái, xen lẫn chữ số và chữ hoa xen lẫn chữ thường.
- Thường xuyên ghi dự phòng dữ liệu quan trọng. Đồng thời cảnh giác với những thư từ người quen biết nhưng không được báo trước, bởi có thể chúng được gửi đi mà người gửi không biết.

Số lượng Spam vẫn luôn luôn tăng và ngày càng tinh vi hơn, người ta nhận định rằng việc chống Spam sẽ luôn luôn phải thực hiện, tùy vào ý thức của cư dân Internet và sức mạnh của công nghệ mà việc Spam chỉ được hạn chế phần nào.

1.2 Các phương pháp lọc thư rác

1.2.1 Lọc thư rác thông qua việc đưa ra luật lệ nhằm hạn chế, ngăn chặn việc gửi thư rác

Khi tình trạng thư rác ngày càng tăng trên đường truyền internet gây ra nhiều phiền toái và thiệt hại lớn trên thế giới rất nhiều các quốc gia đã đưa ra các luật để ngăn chặn thư rác. Dưới đây là một số nội dung cơ bản liên quan tới giải pháp ngăn chặn thông qua luật lệ pháp lý được đưa ra trên báo điện tử của bộ viễn thông .

Mỹ là một những nước đầu tiên trên thế giới cố gắng ban hành các văn bản pháp luật để giải quyết vấn đề thư điện tử rác tràn ngập. Từ tháng 7 năm 1997, bang Nevada đã dẫn đầu trong việc ban hành các quy phạm pháp luật quy định về hành vi phục vụ và sử dụng thư tín điện tử. Tính đến tháng 3 năm 2003, đã có 26 bang ban hành quy phạm pháp luật quy định về dịch vụ và hành vi sử dụng thư tín điện tử. Đến tháng 11 năm 2003, con số này lên đến 36. Về phía chính quyền liên bang, từ những năm 1990, cả Thượng nghị viện và Hạ nghị viện đều quan tâm đến sự lan rộng của thư tín điện tử quấy rối và thư rác, và đã đưa ra nhiều dự án luật như "Luật bảo vệ hộp thư không bị quấy rối" (1999), "Luật Bảo vệ người sử dụng thư điện tử", "Luật Khống chế thư điện tử không được phép" (2000), "Luật Khống chế thư rác truyền qua đường điện thoại vô tuyến" (2000) , "Luật Chống thư rác" (2001).

Mười năm gần đây, Liên minh Châu Âu cũng đã ban hành một số chỉ lệnh, đưa ra các quy phạm và chỉ dẫn đối với các vấn đề thương mại điện tử, thông tin điện tử, bảo hộ dữ liệu.

Trong các chỉ lệnh nói trên, có không ít các qui định có liên quan mật thiết, thậm chí là trực tiếp với phục vụ và sử dụng thư điện tử như "Chỉ lệnh Bảo vệ dữ liệu cá nhân ở Châu Âu", "Chỉ lệnh về thông tin điện tử và bảo mật dữ liệu" ... Ngày 12 tháng 7 năm 2002, Nghị Viện Liên minh Châu Âu đã thông qua "Chỉ lệnh Bảo mật riêng tư và Thông tin điện tử trong Liên minh Châu Âu". Chỉ lệnh quy định: Từ 31 tháng 10 năm 2003, trong phạm vi Liên minh Châu Âu, nếu chưa được người nhận đồng ý trước, không được gửi thư điện tử thương mại hay nhằm mục đích tuyên truyền cho cá nhân. Tiếp theo sau

khi Liên minh Châu Âu đưa ra các qui định về phục vụ và sử dụng thư điện tử, các nước thành viên Liên minh Châu Âu, như Italia, Anh, Đan Mạch, Tây Ban Nha ... đều đã ban hành quy phạm pháp luật trong nước quy định hành vi cung cấp và sử dụng thư điện tử, ngăn chặn sự tràn ngập của thư rác.

Tại Việt Nam vấn đề thư rác bắt đầu nhận được sự quan tâm từ phía các cơ quan có trách nhiệm. Bộ Thương mại đang soạn thảo Thông tư quản lý hoạt động quảng cáo thương mại trên các phương tiện điện tử. Trên trang báo điện tử của bộ viễn thông, Bà Lại Việt Anh, Trưởng Phòng chính sách, Vụ Thương mại điện tử, Bộ Thương mại, nhận xét: mục tiêu của Thông tư này trước mắt tập trung quản lý ba hình thức quảng cáo đang bức xúc: thư điện tử, tin nhắn điện thoại di động và quảng cáo trên trang thông tin điện tử.

1.2.2 Lọc thư rác dựa trên địa chỉ IP

Phương pháp lọc thư rác thông qua địa chỉ IP là phương pháp đơn giản và được sử dụng sớm nhất trong công cuộc chống thư rác. Dựa vào địa chỉ IP của người gửi để xác định thư đó bị ngăn chặn hoặc cho qua. Có hai cách để thực hiện việc lọc thư: một là duy trì một danh sách các địa chỉ IP bị chặn (còn gọi là danh sách đen blacklist); thứ hai là sử dụng một danh sách các địa chỉ IP cho phép qua (danh sách trắng whitelist).

Danh sách đen (Blacklist)

Người ta lập ra một danh sách các địa chỉ gửi thư rác. Các nhà cung cấp dịch vụ thư điện tử (ISP) sẽ dựa trên danh sách này để loại bỏ những thư nằm trong danh sách này. Danh sách này thường xuyên được cập nhật và được chia sẻ giữa các nhà cung cấp dịch vụ. Một số danh sách đen điển hình được lập ra như: SpamCop Blocking List và Composite Block List.

Ưu điểm của phương pháp này là các ISP sẽ ngăn chặn được khá nhiều địa chỉ gửi thư rác. Mặc dù danh sách đen này luôn được cập nhật nhưng với sự thay đổi liên tục địa chỉ, sự giả mạo địa chỉ hoặc lợi dụng một mail server hợp pháp để gửi thư rác đã làm số lượng thư rác gửi đi vẫn ngày càng tăng cao. Do đó phương pháp này chỉ ngăn chặn được một nửa số thư rác gửi đi và sẽ mất rất nhiều thư hợp pháp nếu ngăn chặn nhầm.

Danh sách trắng (Whitelist)

Danh sách các địa chỉ tin cậy (Safe Sender List), danh sách này có thể do một nhà cung cấp dịch vụ nào đó cung cấp. Những địa chỉ thuộc danh sách sẽ được cho qua bộ lọc. Người dùng phải đăng ký với nhà cung cấp danh sách để được nằm trong danh sách.

Ưu điểm: số lượng địa chỉ trong danh sách trắng sẽ ít hơn trong danh sách đen vì thế sẽ dễ cập nhật hơn danh sách đen và giải quyết được tình trạng chặn nhầm thư.

Tuy nhiên cả hai phương pháp trên đều có nhược điểm là khó cập nhật, nhất là khi ai đó thay đổi địa chỉ IP. Ngoài ra người gửi cũng có thể lợi dụng server mail có trong danh sách trắng để gửi thư rác, khi đó rất khó kiểm soát.

1.2.3 Lọc dựa trên chuỗi hỏi/đáp (Challenge/Response filters)

Đặc trưng của phương pháp này là khả năng tự động gửi thư hồi đáp cho người gửi để yêu cầu một số hành động chắc chắn về việc gửi thư của họ. Chương trình kiểm tra này được đặt tên là "Turing Test" sau một vài kiểm tra được nghĩ ra bởi nhà toán học người anh tên là Alan Turing.

Trong một vài năm gần đây xuất hiện của một vài dịch vụ Internet tự động xử lý hàm Challenge/Response này cho người dùng, chương trình yêu cầu người gửi thư phải vào website của họ và trả lời một số câu hỏi để chắc chắn về e-mail mà người này đã gửi.Việc này chỉ được yêu cầu trong lần gửi thư đầu tiên.

Đối với một số người dùng có lượng thư trao đổi thấp, hệ thống đơn lẻ này có thể chấp nhận được như một phương pháp hoàn hảo để loại trừ hoàn toàn thư rác từ hòm thư của họ.

1.2.4 Phương pháp lọc dựa trên mạng xã hội.

Các nghiên cứu gần đây đã bắt đầu khai thác thông tin từ *mạng xã hội* cho việc xác định thư rác bằng cách xây dựng một đồ thị (các đỉnh là địa chỉ email, cung được thêm vào giữa 2 node A và B nếu giữa A và B có sự trao đổi thư qua lại). Người ta đã sử dụng một số tính chất đặc trưng của mạng xã hội để xây dựng một công cụ lọc thư rác [18].

Đầu tiên, người ta phân đồ thị thành các thành phần con rồi tính độ phân cụm cho từng thành phần này. Mỗi thành phần con là một đồ thị mạng xã hội của một node, bao gồm tất cả các node xung quanh là "node hàng xóm" (các node có cung liên kết với node này) và những cung liên kết giữa các node hàng xóm này với nhau. Nếu thành phần nào có độ phân cụm thấp thì node tương ứng với thành phần đó là một địa chỉ gửi thư rác. Trong thành phần mạng xã hội của những node gửi thư rác, những node hàng xóm của nó thường là những node rất ngẫu nhiên, không có mối quan hệ (không có sự trao đổi email qua lại với nhau) nên độ phân cụm của mạng xã hội của những node này rất thấp. Ngược lại, mạng xã hội ứng với những người dùng bình thường có độ phân cụm cao hơn.

Dựa vào độ phân cụm, người ta tạo được danh sách đen (Blacklist) gồm địa chỉ email tương ứng với những node có độ phân cụm rất thấp, danh sách trắng (Whitelist) ứng với node có độ phân cụm cao, số node còn lại sẽ được đưa vào danh sách cần xem xét (Greylist). Phương pháp này có thể phân loại được 53% tổng số email một cách chính xác là ham hay spam. Nhược điểm của phương pháp là những spammer có thể xây dựng mạng xã hội của chính họ nên khó có thể phát hiện ra.

1.2.5 Phương pháp định danh người gửi

Giả mạo thư điện tử - là việc giả mạo địa chỉ thư điện tử của công ty hoặc của người khác để khiến người sử dụng tin tưởng và mở thư - đang là một trong những thử thách lớn nhất mà cộng đồng sử dụng Internet và các kỹ thuật viên chống thư rác hiện đang phải đối mặt. Nếu không có sự thẩm định quyền, xác nhận và khả năng truy tìm danh tính của người gửi, các hãng cung cấp dịch vụ thư điện tử không bao giờ có thể biết chắc một bức thư là hợp pháp hay bị giả mạo. Do đó việc xác nhận danh tính của người gửi là rất cần thiết. Phương pháp được đề xuất đó là phương pháp Domainkeys, đây là phương pháp hiện đang rất được quan tâm chú ý nghiên cứu phát triển.

Domainkeys là một phương thức mã hóa định danh, được đề xuất bởi Yahoo vào tháng 5 năm 2004. Domainkeys không những chỉ cho phép xác định domain của người gửi mà còn cho phép kiểm tra tính toàn vẹn của chính nội dung của email. Domainkeys sử dụng mã hóa khóa công cộng RSA để xác minh tính toàn vẹn của người gửi email tại mức domain. Domainkeys được thực hiện và sử dụng bởi cả yahoo! Mail và Google mail.

Nội dung cơ bản của Domainkeys được trình bày như sau. Mỗi domain phải sinh ra một cặp khóa bí mật và khóa công khai. Khóa công khai được công bố trong bản ghi vùng DNS. Khóa bí mật được giữ lại tại dịch vụ MTA gửi thư.

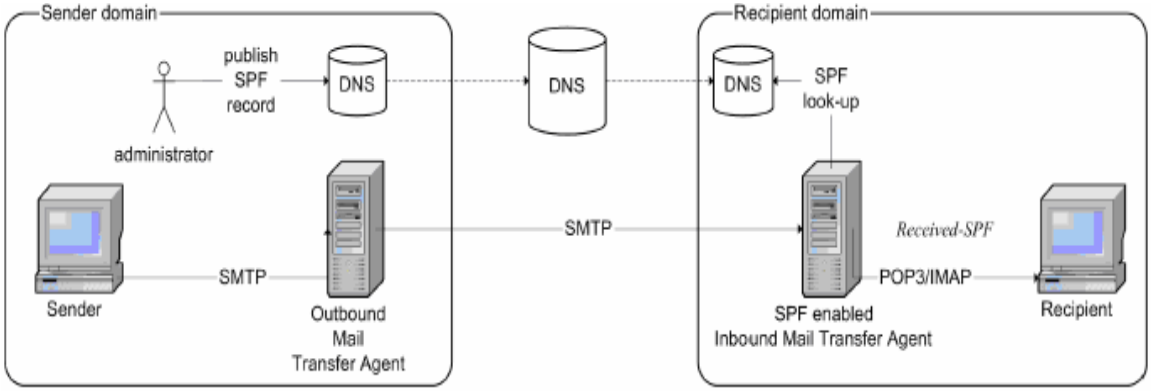
Sau khi email đã được gửi đi, dịch vụ gửi thư MTA ký số vào nội dung của email bằng khóa bí mật. Chữ ký được thêm vào trường Domainkey_signature.

Ví dụ:

DomainKey-Signature: a=rsa-sha1 s=brisbane;

d=example.net;c=simple; q=dns; b=dzdVyOfAKCd…ZHRNiYzR;

Hình vẽ dưới đây (hình1) mô tả hệ thống gửi và nhận thư, chỉ ra vị trí sử dụng domainkeys.



Hình 1.1 Khung ID người gửi được thi hành trên MTA [6]

Domainkeys yêu cầu cả bên gửi Mail Transfer Agent(MTA) và bên nhận MTA thực hiện domainkey. Việc xác minh của Domainkeys_signature có thể cũng được thực hiện tại Domainkeys_enabled của Mail User Agent (MUA).

Khi server nhận được tên của domain từ mail gốc (string-domainkey) thì bộ selector thực hiện tra cứu DNS. Dữ liệu trả về chứa khóa công khai của domain đó. Người nhận có thể giải mã giá trị băm chứa trong trường tiêu đề và đồng thời tính lại giá trị băm cho phần thân của mail nhận được. sau đó so sánh hai giá trị này nếu giống nhau chứng tỏ mail được gửi là thật, đảm bảo tin cậy nếu không là mail không đáng tin.

Ưu điểm:

- xác định nguồn gốc domain của email một cách rõ ràng, sẽ hiệu quả hơn nếu kết hợp với sử dụng danh sách đen và danh sách trắng. Giúp dễ dàng phát hiện ra sự tấn công phising.
- Loại bỏ những email giả mạo tại phần mềm email người dùng cuối (mail user agents) hoặc bởi ISP's mail transfer agents.
- Theo dõi việc lạm dụng domain của những cá nhân một cách dễ dàng hơn.

Khả năng tương thích:

Domainkeys tương thích với cấu trúc hiện tại của email. Trong trường hợp đặc biệt, đối với hệ thống email mà không có sự hỗ trợ của domainkeys thì nó là trong suốt.

Nhược điểm

Domainkeys là một công nghệ xác định danh tính, nó không tham gia trực tiếp trong việc lọc spam. Ví dụ: Domainkeys cho người nhận thư biết mẩu tin đó từ example.net, nhưng không thể cho biết liệu mail từ example đó có phải là spam hay không. Chỉ chữ ký không khẳng định thư đó có được mong muốn hay không, và các Spammer cũng có thể ký mail, cũng có thể giả mạo chữ ký…

Ngoài ra còn có một số phương pháp khác như:

- ***SPF classic*** : được IETF đề xuất đầu tiên vào tháng 7 năm 2003. SPF sử dụng return_path hay SMPT "MAIL FROM" để xác nhận danh tính của người gửi.

Nhà quản trị domain sẽ phát hành một bản ghi SPF dịnh dạng là file txt trong Domain Name System. Bản ghi SPF chỉ rõ những host đã được định danh gửi mail.

Sau khi nhận một emai, dịch vụ nhận thư MTA sẽ kiểm tra bản ghi SPF, nếu người gửi với đặc tính "Mail From" thỏa mãn sẽ được phép gửi mail .Trong trường hợp người gửi không được phép gửi thư, MTA sẽ đánh dấu email đó hoặc là đẩy mail đó ra và thông báo lỗi SMPT 550. Trong trường hợp đánh dấu, email đươc sử lý tiếp bởi một bộ lọc dựa trên các luật. SPF được thực hiện ngay trên dịch vụ nhận MTA.

- ***Sender ID Framework (SIDF)***: SIDF là kỹ thuật định danh IP được chuẩn IETF đề xuất , nó kết hợp với SPF và Microsoft CallID (MIC04). Rất nhiều nhà sản xuất phần mềm có hỗ trợ SID Framework.
- ***Identified Internet Mai (IMM):*** Cũng giống như Domainkeys, IMM là phương thức mã hóa danh tính (authentiaction) . Nó sử dụng mã hóa khóa công cộng RSA. IMM được phát triển bởi Cisco Systems và IETF đưa ra tháng 7 năm 2004. Ý tưởng Domainkeys và IMM là tương tự nhau, nhưng có một vài điểm khác.
- ***FairUCE*** : fair use of Unsolicited Commercial Email, được phát triển bởi IBM. FairUCE là kỹ thuật dựa trên xác định tính đúng đắn của IP. IBM không cố gắng đạt tới hệ thống FairUCE hoàn hảo, nhưng là một cơ cấu đơn giản hiệu quả để xác định tính đúng đắn.

Tất cả những kỹ thuật nêu ra ở trên nhằm cải tiến vấn đề an toàn cho giao thức SMTP. Kỹ thuật nổi bật là Domainkeys và Identified Internet Mail. IIM hiện tại chỉ được đưa ra với phiên bản alpha. Domainkeys đã được đưa vào sử dụng, nhưng chỉ được thực hiện bởi 2 nhà sản xuất. Vì thế tỉ lệ chấp nhận của những đề xuất này là rất thấp. Tuy nhiên một chuẩn mới Domainkeys Identified Mail, sự kết hợp của hai kỹ thuật Domainkeys và IIM đang được phát triển làm thay khả năng chấp nhận của chúng được tăng lên

1.2.6 Phương pháp lọc nội dung

Phương pháp lọc nội dung để phân loại thư rác đã và đang được quan tâm, nghiên cứu và ứng dụng nhiều nhất. Phương pháp này dựa vào nội dung và chủ đề bức thư để phân biệt thư rác và thư hợp lệ. Phương pháp này có ưu điểm đó là chúng ta có thể dễ dàng thay đổi bộ lọc để nó có thể lọc các loại thư rác cho phù hợp. Nhược điểm của phương pháp này là: do biết được cách thức lọc nội dung nên các spammer luôn luôn thay đổi hình thức nội dung của thư rác.

Phần dưới đây trình bày những nét cơ bản nhất về các phương pháp lọc nội dung thông dụng [18,19].

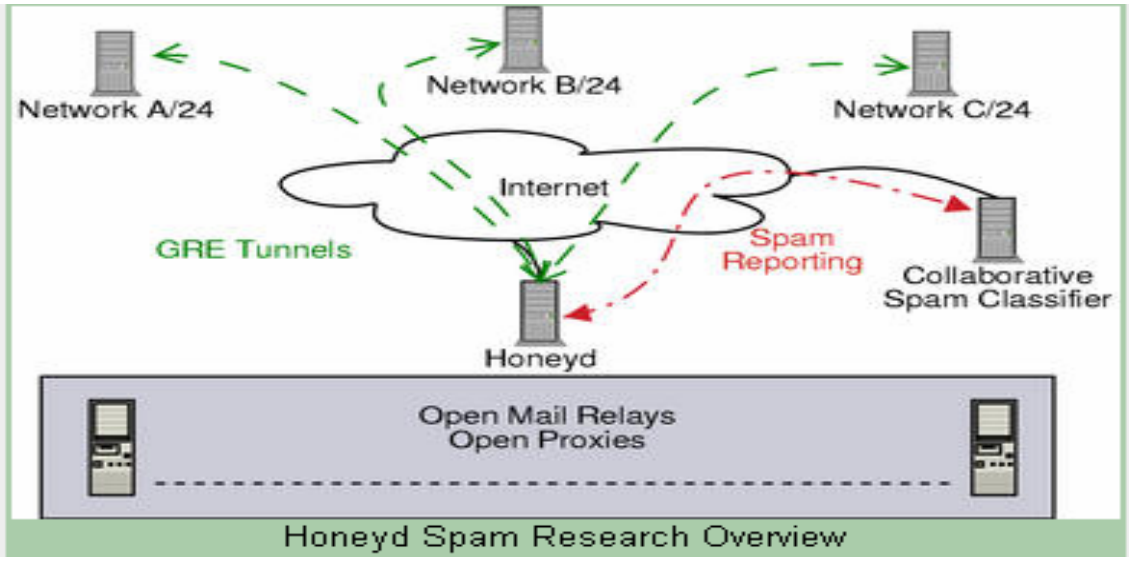
<u>Lọc dựa trên các dấu hiệu nhận biết</u>

Trước tiên, tạo ra các địa chỉ email để bẫy thư rác, gọi là honeypots, phương pháp này được nghiên cứu phát triển nhiều vào năm 2003. Honeypots chứa các địa chỉ sao cho không bao giờ thư bình thường có thể gửi đến. Do đó thư gửi đến bẫy địa chỉ này ta có thể coi đó là thư rác.

Sau đó hệ thống so sánh thư mới đến với thư đã được bẫy. Sự so sánh dựa trên dấu hiệu nhận biết, nếu chúng có dấu hiệu giống nhau thì có thể kết luận thư mới đến là thư rác.

Ưu điểm của phương pháp này là đơn giản, nhanh và không lọc nhầm thư thường thành thư rác. Tuy nhiên spammer có thể dễ dàng vượt qua hệ thống bằng cách sinh ngẫu nhiên các mẩu thư rác sau đó gộp lại nhằm làm cho dấu hiệu của các bức thư rác khác nhau. Bởi vậy tỉ lệ lọc thư rác của hệ thống luôn nhỏ hơn 70%. Do không lọc thư thường thành thư rác nên phương pháp này được triển khai trên server.

Một hệ thống lọc thư rác dựa trên honeypots hoạt động rất hiệu quả đó là eTrap. Hệ thống eTrap sử dụng honeypots để thu thập thông tin về spam. Những thông tin về spam được lưu trữ trong cơ sở dữ liệu chia sẻ chung. Hệ thống eTrap lọc thư rác dựa trên những thông tin về spam này.



Hinh 1.2 : Mô tả tổng quan quá trình hoạt động của honeyd : Trước tiên honeyd bẫy các địa chỉ gửi thư rác, sau đó toàn bộ thông tin về thư rác thu được sẽ được gửi tới Collaborative Spam Classifier để tổng hợp thông tin. Dựa vào những thông tin đó bộ phân loại thư rác sẽ phân tichsm để phân loại thư rác.

Lọc thư rác thông qua bỏ phiếu trên danh sách trắng, đen.

Hệ thống tìm xem các từ trong danh sách đen/trắng có nằm trong thư mới đến không và đếm số lần xuất hiện của chúng. Nếu số lượng từ thuộc danh sách trắng nhiều hơn rất nhiều số từ thuộc danh sách đen thì bức thư đó là hợp pháp và ngược lại sẽ là thư rác.

Đặc trưng của bộ lọc thông qua bỏ phiếu trên danh sách đen/trắng:

- Không có biến đổi dữ liệu ban đầu.
- Biểu thức chính quy để tách từ ra khỏi thư là: [[:graph:]]+
- Việc chọn đặc trưng đơn giản chỉ là các từ đơn

- Cơ sở dữ liệu về đặc trưng chỉ được nạp khi các từ nằm trong danh sách đen hoặc trắng. Nếu nằm trong danh sách đen thì đặt là -1, trong danh sách trắng là +1, các trường hợp còn lại đặt là 0.
- Luật tổ hợp là : "Điểm mới = Điểm cũ + trọng số đặc trưng"
- Ngưỡng lọc cuối cùng là : Nếu Điểm mới > 0 là thư hợp pháp, nếu < 0 là thư rác.

Như vậy bộ lọc thực hiện chấm điểm các từ trong danh sách đen và các từ trong danh sách trắng bằng nhau. Một số cải biên của phương pháp này là đánh trọng số cho các từ trong danh sách đen cao hơn trong danh sách trắng hoặc ngược lại.

Lọc thư rác dựa vào phương pháp heuristic.

Cách thức hoạt động của phương pháp này là dựa trên việc xác định những từ đặc trưng thuộc về thư rác, từ đặc trưng thuộc về thư hợp pháp, sau đó phát hiện những đặc trưng đó trong thư mới nhận để đưa ra kết luận thư đó là thư rác hay thư hợp lệ.

Người ta đánh trọng số cho các đặc trưng trên bằng tay hoặc bằng thuật toán và lập một ngưỡng để phân loại thư. Nếu bức thư có trọng số lớn hơn ngưỡng quy định sẽ bị coi là thư rác.

Các chương trình lọc thư rác sử dụng phương pháp này có hiệu suất khác nhau. Vì mỗi chương trình sử dụng một luật lọc khác nhau.

Một số chương trình lọc theo phương pháp này như hệ thống chấm điểm cho email sử dụng phương pháp hueristic của mail server Mdaemon, SpamAssassin hay SpamGuard của Yahoo.

Phương pháp này có ưu điểm là dễ cài đặt và hiệu suất chặn thư rác khá cao khi xây dựng được hệ thống luật tốt. Nhược điểm chính của phương pháp này là tỉ lệ chặn nhầm thư hợp pháp cũng khá lớn 0.5%. Phương pháp này không linh hoạt do các luật được xây dựng luôn chậm hơn sao với sự biến đổi của từ ngữ trong thư rác.

Phương pháp này thường được áp dụng cho các bộ lọc thư ở server.

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.

Tỉ lệ chặn thư rác của bộ lọc sử dụng phương pháp này rất cao, khoảng 99%. Chương trình SpamProbe có thể đạt tới tỉ lệ lọc thư rác tới 99.9%. Các phương pháp học máy và xác suất thống kê cho phép phân loại cả những thư rác chưa từng xuất hiện trước đó. Phương pháp này còn có tỉ lệ chặn thư hợp pháp rất thấp, thấp hơn nhiều so với phương pháp heuristic.

Nhược điểm của phương pháp này là phải có một tập hợp các thư để huấn luyện. Hiệu suất của bộ lọc sẽ phụ thuộc nhiều vào tập huấn luyện này. Tập dữ liệu càng lớn càng chứa nhiều dạng khác nhau thì kết quả phân loại về sau sẽ càng chính xác.

Hiện nay phương pháp lọc thư rác theo học máy và xác suất thống kê là một phương pháp có triển vọng với nhiều ứng dụng thương mại như Hotmail, Google, Yahoo.

Để có một bộ lọc hoàn hảo dường như không thể thực hiện được, một bộ lọc tốt nhất là bộ lọc kết hợp nhiều bộ lọc. Việc Spam ngày càng được thực hiện tinh vi hơn đòi hỏi các bộ lọc phải có khả năng biến đổi theo sự thay đổi của Spam, sự thay đổi về số lượng, về nội dung và cấu trúc của các thư spam. Vì vậy yêu cầu đặt ra phải có một bộ lọc có khả năng cập nhật để có thể thay đổi, chống lại những thư spam có cấu trúc nội dung mới, bộ lọc học máy lọc dựa trên nội dung Email Classification Using Example(ECUE) đã được chứng minh là có khả năng thực hiện được điều đó. Trong khuôn khổ khóa luận này em xin trình bày hệ thống lọc thư rác ECUE mới do Delany đề xuất và đã xây dựng thử nghiệm thành công.

Lọc thư rác dựa trên thuật toán bayes [8,15]:

Coi mỗi email được biểu diễn bởi một vectơ thuộc tính đặc trưng

$\vec{x} = (x_1, x_2,\ldots,x_n)$. với $(x_1, x_2,\ldots,x_n)$ là các giá trị thuộc tính $X_1, X_2, \ldots, X_n$ tương ứng trong không gian đặc trưng (space model) . Ta sử dụng giá trị nhị phân 0 và 1 để mô tả email đó có đặc điểm X_i hay không, giả xử nếu email đó có đặc điểm X_i thì ta đặt thuộc tính $X_i = 1$, còn nếu email đó không có đặc điểm X_i thì ta có thuộc tính $X_i = 0$.

Từ thuyết xác suất của Bayes và xác suất đầy đủ chúng ta có công thức tính xác suất mail với vectơ $\vec{x} = (x_1, x_2,\ldots,x_n)$ thuộc vào lớp c như sau:

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c)P(\vec{X} = \vec{x} \mid C = c)}{\sum_{k\in\{Spam,legitimate\}} P(C = k)P(\vec{X} = \vec{x} \mid C = k)} \quad (1)$$

Để đơn giản khi tính $P(\vec{X} \mid C)$ ta phải giả sử $X_1, X_2,\ldots,X_n$ là độc lập. Khi đó biểu thức (1) tương đương với biểu thức sau:

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c)\prod_{i=1}^{n} P(X_i = x_i \mid C = c)}{\sum_{k \in \{Spam, legitimate\}} P(C = k)\prod_{i=1}^{n} P(X_i = x_i \mid C = k)}$$

Giá trị được sử dụng rất rộng rãi để đánh hạng cho thuộc tính là giá trị tương hỗ MI(mutual information), ta lấy những thuộc tính có giá trị MI lớn nhất. Ta có thể tính giá trị tương hỗ MI(Mutual information) mà mỗi đại diện của X thuộc về loại C như sau :

$$MI = \sum_{x \in \{0,1\}, c \in \{spam, legitimate\}} P(X = x, C = c) \log \frac{P(X = x, C = c)}{P(X = x)P(X = c)}$$

Một email được coi là spam nếu:

$$\frac{P(C = spam \mid \vec{X} = \vec{x})}{P(C = legitimate \mid \vec{X} = \vec{x})} > \lambda \quad (2)$$

Giả sử các thuộc tính X_i là độc lập khi đó ta có:

P(C=spam| $\vec{X} = \vec{x}$) = 1 - P(C=legitimate| $\vec{X} = \vec{x}$)

Khi đó (2) tương đương với :

P(C=spam| $\vec{X} = \vec{x}$) > t, với t = $\frac{\lambda}{1+\lambda}$, $\lambda = \frac{t}{1-t}$.

Thuật toán bayes đã được áp dụng vào chương trình lọc thư rác spambayes, và cho kết quả lọc khá hiệu quả.

Chương 2

CASE-BASED REASONING

Đã có nhiều hệ thống học máy lọc thư rác sử dụng các thuật toán Naïve bayes, phân lớp dựa trên thống kê (Lewis and Ringuette 1994, Lewis 1998), Support Vector Machines (Joachims 1998, Dumais et al. 1998) các phương pháp này đều cho kết quả lọc khá tốt[4]. Tuy nhiên các mô hình này chưa giải quyết được vấn đề concept drift . Một mô hình mới đã được Delany(2006) đề xuất, dựa trên hệ thống học máy sử dụng phương pháp Case-Based Reasoning (CBR)(Riesbeck and Shank 1989)[17] có khả năng giải quyết được concept drift. Phương pháp CBR, sử dụng các vấn đề trước đây đã được giải quyết để đưa ra giải pháp cho vấn đề mới. Các vấn đề đã được giải quyết được lưu vào tập dữ liệu dùng để huấn luyện gọi là case-base. Các case được biểu diễn dưới dạng véc tơ n chiều, mỗi thành phần là một token đã được trích chọn từ việc phân tích cú pháp, phân tích từ tố của tài liệu (email). Các vector cũng chứa thêm một thành phần nữa chỉ lớp mà tài liệu đó được phân (nonspam, spam). Trình bày về cấu trúc Case-based Reasoning(CBR), chu trình thực hiện của CBR Retrieve, Reuse, Revise, Retain; sự biểu diễn case; việc trích chọn các đặc trưng, biểu diễn đặc trưng; Và đưa ra ưu điểm của CBR trong việc giải quyết vấn đề concept; ứng dụng CBR trong lĩnh vực phân lớp Textual CBR.

2.1 Case-based Reasoning.

Case-Base Reasoning(CBR) (Smyth và McKenna 1998) là phương pháp kĩ thuật giải quyết vấn đề, thực hiện giải quyết các vấn đề mới bằng việc sử dụng lại những giải pháp đã có của những vấn đề trước. Những vấn đề trước đây được mã hóa gọi là các case, mỗi case chứa những thuộc tính đặc trưng của vấn đề đó và giải pháp cho nó. Một tập các case được gọi là case-base, là kiến thức nền tảng đã qua trải nghiệm, case-base được sử dụng cho quá trình đưa giải pháp cho vấn đề mới[17].

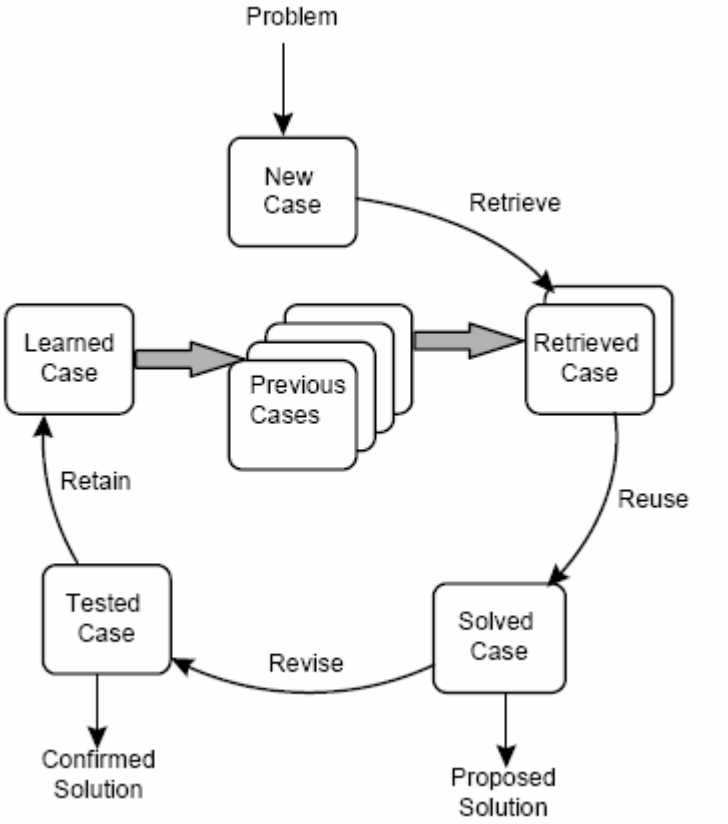
CBR thực hiện theo một chu trình gồm các tiến trình sau (theo Aamodt and Plaza 1994) (được mô tả trong hình 2.1):

1. Lấy từ casebase những case tương đồn với case mới (case cần được đưa ra giải pháp).

2. Sử dụng lại những case trên để đưa ra giải pháp cho case mới.

3. Kiểm tra lại giải pháp cho case mới, nếu cần.

4. Giữ lại giải pháp đã được giải quyết đó để giải quyết những vấn đề mới tiếp theo.



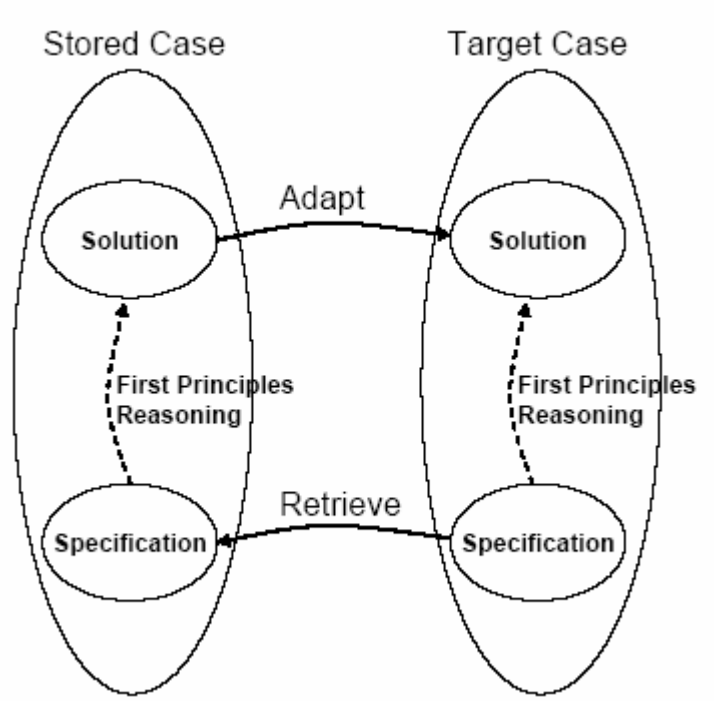
Hình 2.1 Biểu diễn chu trình thực hiện Case-based Reasoning.[17]

Quy trình thực hiện như sau:

Khi có một vấn đề mới cần phải giải quyết, vấn đề đó sẽ được biểu diễn dưới dạng case. Case mới này sẽ được so sánh với các case trong case-base, những case có độ tương đồng cao nhất với case mới sẽ được trích ra từ case-base. Tập hợp case được trích ra đó sẽ được phân tích để đưa ra giải pháp cho case mới. Giải pháp đưa ra cho case mới có thể sẽ được kiểm tra lại, nếu giải pháp đó chưa được thỏa đáng thì thực hiện tinh toán lại để đưa ra giải pháp thỏa đáng hơn. Giải pháp cho vấn đề mới sẽ được lưu lại vào tập hợp các vấn đề đã có giải pháp.

2.1.1 Biểu diễn Case

"Một case là mảnh kiến thức biểu diễn sự trải nghiệm" (theo Watson và Marir năm 1994). Case biểu diễn kiến thức cụ thể ở mức sẵn dùng, một case gồm đặc tả của một vấn đề và giải pháp cho vấn đề đó và có thể có thêm kết luận logic của vấn đề đó (outcome). Các case đó được lưu lại và được sử dụng để giải quyết case mới. Hình 2.2 biểu diễn tiến trình hoạt động của CBR, target case là case cần được đưa ra giải pháp, stored case là tập các case đã có giải pháp. Các case chứa giải pháp (solution) và đặc trưng của case (specification).



Hình 2.2: Tiến trình của CBR (Cunningham, 1994)[17]

Thông thường mô tả của một case chứa một tập các đặc trưng. Những đặc trưng này được xác định qua một quá trình kiểm tra kiến thức: hệ chuyên gia phỏng vấn trong lĩnh vực mà nó liên quan đến, việc đưa ra những yêu cầu và việc sử dụng các phương pháp kĩ thuật tập hợp dữ liệu. Ví dụ như một vấn đề về một chương trình quản lý quỹ tín dụng. Một khách hàng tiếp cận với ngân hàng và yêu cầu vay tiền. Người quản lý ngân hàng sẽ quyết định có nên cho vay hay không như thế nào? Vấn đề này được thực hiện bằng cách sử dụng hệ thống các tri thức hay hệ thống dựa trên các luật (còn gọi là hệ chuyên gia). Trong trường hợp cho ứng dụng này case biểu diễn một sự trải nghiệm, nó nên biểu diễn những đặc trưng của ứng dụng đẻ xác định nên hay không nên cho khách hàng vay tiền. Trong case sẽ phải chứa số lượng tiền mà khách hàng muốn vay, thời hạn trả tiền, giới tính của khách hàng, tình trạng hôn nhân, tuổi, tình trạng và những chi tiết

mô tả việc làm như tiền lương, vị trí đảm trách…mục đich vay tiền làm gì, và có thể thêm một vài đặc trưng khác nữa.

Feature	Value
Amount required	2,500
Type of Purchase	Personal
Repayment Period (months)	6
Gender	Male
Age	30
Married	No
Employed	Yes
Weekly wage	€260
Years in Employment	1.5
Recommendation	Accept
Outcome	Good

Bảng 2.1: Biểu diễn các case, người vay tiền ngân hàng.[17]

2.1.2 Case Retrieval

Quá trình lấy các case gồm việc tìm kiếm các case có độ tương đồng cao nhất với case hiện tại, những case này có tiềm năng cao cho việc dự đoán cho case mới. Kolodner(1992) khẳng định việc tìm các case phù hợp chính là phần quan trọng nhất của case-base reasoning[17].

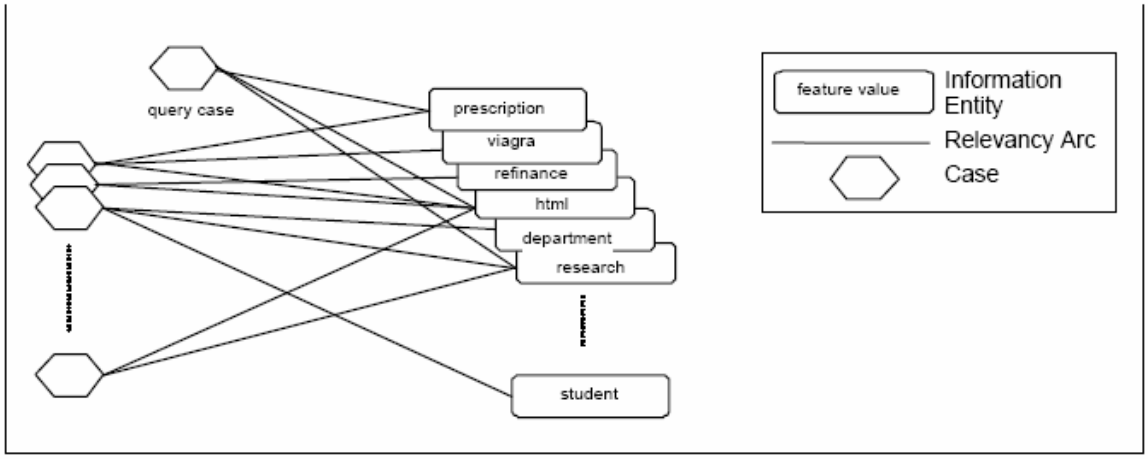
Trong CBR có hai phương thức chính để lấy các case có độ tương đồng cao với case mới từ case-base, đó là sử dụng thuật toán cây quyết định và thuật toán k-Nearnest Neighbour(k-NN). Thuật toán cây quyết định (Wess et al .1994) thực hiện phân tích các đặc trưng để tìm ra đặc trưng nào là tốt nhất cho việc so sánh các case với nhau. Các đặc trưng tốt đó được sắp xếp vào vào một cấu trúc cây, đặc trưng tốt nhất được đặt ở đỉnh của cây. Sau đó các case sẽ được tổ chức lưu trữ trong bộ nhớ theo cấu trúc của cây quyết định, thuật toán retrieval sẽ tìm kiếm trên cây quyết định những node case mới. Khi các case được sắp xếp theo cấu trúc có thứ tự thì thời gian thực hiện retrieval tăng theo hàm logarithm của số case. Tuy nhiên phương pháp này đòi hỏi một số lượng đáng kể các case để có thể nhận ra các đặc trưng và xác định cấu trúc có cấp bậc thích hợp. Sự phân tích

này đòi hỏi rất nhiều thời gian và luôn phải thực hiện khi có một case mới được thêm vào case-base.

Thuật toán k-NN thực hiện so sánh các case trong casebase với case mới và tính toán ma trận tương đồng của case đó với case mới. Ma trận tương đồng được tính dựa trên mức độ gần (‘close’) của những đặc trưng giữa case được lựa chọn và case mới. Mỗi đặc trưng được so sánh và tính điểm được dựa vào mức độ khác nhau giữa hai đặc trưng, các đặc trưng càng gần nhau thì điểm số sẽ càng cao. Điểm số của các đặc trưng cũng phụ thuộc vào mối quan hệ của chúng với giải pháp đưa ra. Ta sẽ chọn ra k case có giá trị tương đồng cao nhất, và dựa vào k cây đó để đưa ra giải pháp cho case mới.

Hạn chế của phương pháp k-NN là thời gian tìm kiếm sẽ tăng tuyến tính theo số lượng case có trong case base. Lenz et al. (1998a) đưa ra đề xuất cải tiến mới là thuật toán Case Retrieval Nets(CRN), thuật toán tính toán độ tương đồng. CRN là một kiểu cấu trúc bộ nhớ giúp cho việc lấy các case về được thực hiện một cách mềm dẻo, hiệu quả. Nó dựa theo ý tưởng mạng neural và sự kết hợp các mô hình bộ nhớ . CRN gồm các thành phần sau:

- Mỗi case được lưu trữ dưới dạng các node.
- Thông tin chứa trong các node Information Entity Nodes (IEs) là các cặp feature-value của case.
- Relevance Arcs liên kết các node case (IEs)với nhau. Chúng được đánh trọng số thể hiện độ quan trọng của IE.
- Similarity Arcs kết nối với các IE cùng tham chiếu đến một số các thuộc tính, và được đánh trọng số dựa vào độ tương đồng giữa các IE nối với nhau.



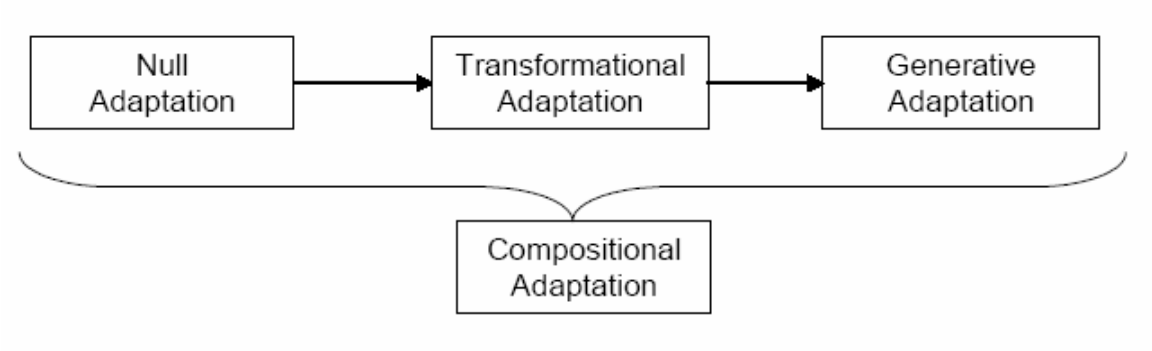
Hình 2.3: Mô hình CRR[17]

Hoạt động của CRR:

Các case mới được kích hoạt bằng cách kết nối nó vào mạng qua một tập các Relevance Arc và sự kích hoạt này sẽ được lan rộng khắp mạng. Mỗi một case node có điểm số kích hoạt tương ứng với độ tương đồng của nó với case mới. Những case node có điểm số kích hoạt cao sẽ là những case có độ tương đồng cao với case mới. CRNs khai thác được lợi ích của sự dư thừa đặc trưng và chúng có thể bỏ qua các giá trị đặc trưng bị lỗi hoặc vắng mặt.

2.1.3 Reuse

Trong trường hợp mà ở đó các case đã được lấy về giống hệt case mới, khi đó giải pháp của các case lấy về sẽ áp dụng cho case mới. Trong các trường hợp khác, giải pháp cũ cần được thay thế để tương ứng với case mới, tiến trình thực hiện sự thay thế được gọi là adaptation. Một vài kĩ thuật adaptation được sử dụng trong CBR được mô tả theo hình 2.4



Hình 2.4: Quy trình Adaptation(Wilke and Bergmann 1998, Wilke et al. 1998)[17]

- Đơn giản nhất là null adaptation: Không cần adaption, giải pháp đưa ra được áp dụng trực tiếp cho case mới, trường hợp này thường gặp trong bài toán phân lớp.

- Transformational adaptation: Sử dụng một tập các luật để điều chỉnh những giải pháp đã thu được trên cơ sở sự khác nhau giữa những đặc trưng của case mới và case lấy về.

- Mô hình Generative: Phức tạp hơn và yêu cầu một bộ giải quyết vấn đề để có thể tích hợp được vào hệ thống CBR., bộ giải quyết vấn đề này được sử dụng để sinh những phần nhỏ của giải pháp.

- Compositional Adaptation: Đưa ra giải pháp hoàn chỉnh cho case mới bằng việc kết hợp các thành phần của giải pháp vừa được hiệu chỉnh của các case lấy về.

2.1.4 Revision và Retension

Hai tiến trình cuối cùng trong chu trình của CBR được thực hiện đồng thời, cả hai tạo nên quá trình học của CBR. Khi giải pháp cho case mới được đưa ra từ tiến trình Reuse không thỏa mãn thì sẽ tiến hành cho case mới học lại. Giải pháp đưa ra đó phải được phân tích lại để có được giải pháp đúng hơn, khi giải pháp đó thỏa mãn rồi sẽ được lưu lại, và case mới được thêm vào case-base.

Tiến trình Revision gồm hai bước: Định giá giải pháp đưa ra và chuẩn đoán sửa chữa lỗi nếu cần. Bước định giá gồm việc xét xem giải pháp đưa ra dựa trên đánh giá mức độ tốt case-base . Sự đánh giá có thể dựa trên thông tin có trên thực tế, kết hợp với việc hỏi các chuyên gia hoặc kiểm tra giải pháp đó trên môi trường thực tế. Sự đánh giá cũng có thể dựa trên kết quả kết quả của mô hình mô phỏng áp dụng giải pháp đó. Case sau khi được học lại sẽ có thể được đưa vào case base.

Tiến trình Retension thực hiện việc đưa thêm các case đã được học lại vào case-base . Giải pháp mới tốt sẽ được thêm vào bộ nhớ case để thuận tiện cho việc giải quyết các vấn đề tương tự tiếp theo và cả những giải pháp lỗi cũng được đưa vào nhằm tránh lặp lại những lỗi tương tự.

Việc tăng số lượng các case trong case base có thể sẽ làm cho hệ thống bao phủ nhiều vấn đề hơn, giải quyết vấn đề mới tốt hơn. Nhưng không phải vì vậy mà ta sẽ tăng số lượng đó một cách bừa bãi, nó có thể làm giảm hiệu quả việc sử dụng case-base.(Smyth and Cunningham 1996). Vì vậy việc điều chỉnh case-base cho thích hợp là rất cần thiết, tôi sẽ mô tả chi tiết trong phần 3.3 dưới đây. Điểm quan trọng trong việc điều chỉnh case-base (case-base editing) là giảm bớt kích cỡ của case-base trong khi đó phải duy trì được hiệu suất thực hiện.

2.1.5 Những ưu điểm của CBR

Case-based reasoning có nhiều ưu điểm hơn so với những phương pháp kĩ thuật học máy khác. CBR là một phương pháp học máy do Aha phát triển năm 1997. Phương pháp này có ưu điểm đó là những mẫu huấn luyện mới có thể được thêm vào một cách rất dễ dàng. Sự hạn chế của việc học này là tất cả các mẫu được sử dụng cần phải lưu trữ và hệ thống có thể truy cập được mỗi khi có yêu cầu. Để thực hiện các tiến trình xử lý khi có yêu cầu đòi hỏi tính toán nhiều. Tuy nhiên tốc độ phát triển của phần cứng máy tính rất nhanh do đó sự hạn chế này dễ dàng khắc phục được.

2.1.6 Ứng dụng phương pháp CBR vào việc phân lớp văn bản (Textual CBR)

Một lĩnh vực trong nghiên cứu CBR đó là lọc spam Textual Case-Based Reasoning(TCBR)(Lenz et 1998). TCBR là một lĩnh vực thuộc CBR với các case là tài liệu dạng text. Có rất nhiều lĩnh vực ứng dụng TCBR như: trợ giúp trên bàn giấy (hepl desks) (Lenz 1998, Lenz 1998), hỗ trợ khách hàng (customer support) (Gupta and Aha

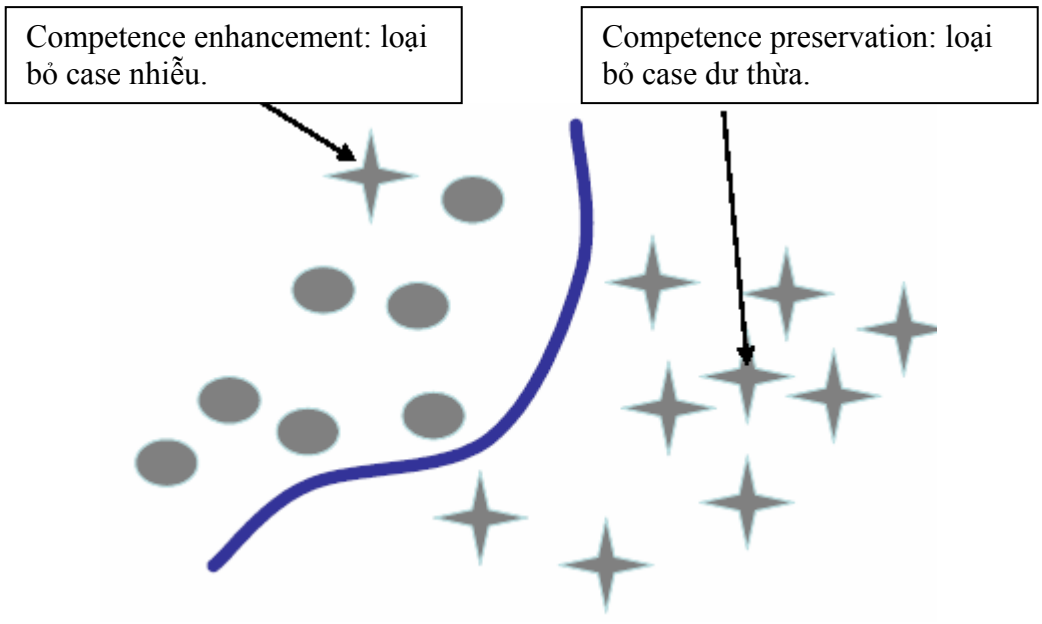
2004), người dạy học thông minh (intelligent tutoring) (Ashley and Aleven 1991) and luật sư (law) (Bruninghaus and Ashley 2001, 2003).

TCBR dựa trên ưu điểm của việc trích chọn thông tin (IR- Information Retrieval) (Baeza-Yates and Ribeiro-Neto 1999). IR lấy một tập các mục (đó là các từ thông thường) nằm trong tập tài liệu thu được và dựa trên thống kê để đánh chỉ số cho mục đó, ví dụ như: xác suất hay tần suất xuất hiện của từ đó trong tài liệu. Tần suất xuất hiện của mục đó trong tài liệu cũng được sử dụng để xác định độ quan trọng của nó và độ quan trọng này được sử dụng để tính toán độ tương đồng giữa các tài liệu với nhau. Ý tưởng này đã được áp dụng trong TCBR. Trong TCBR, những case phải được trích ra từ những tài liệu dạng text và sự biểu diễn những tài liệu này chính là khóa để tính toán độ tương đồng giữa các case. Sự biểu diễn của case trong TCBR có thể phức tạp hơn trong IR. Chúng có thể bao gồm cả công nghệ xử lý ngôn ngữ tự nhiên như Part-of-Speech tagging và thông tin có cấu trúc dưới dạng cặp thuộc tính – giá trị (Lenz 1998).

2.2 Case-base Editing

Một lĩnh vực nghiên cứu về CBR gần đây được nghiên cứu nhiều đó là hiệu chỉnh case-base (case-base editing), giảm bớt số lượng case trong case-base cho phù hợp. Công nghệ case-base editing được quan tâm đặc biệt trong lĩnh vực lọc spam, mỗi mail được coi là một case, mỗi một cá nhân có thể nhận được một số lượng rất lớn email, và địa chỉ của các email đó cần được kết hợp vào kiến thức cơ sở (case-base) để phân lớp những email mới đến. Trong khóa luận này tôi sẽ trình bày chi tiết về phương pháp edit case-base.

Phương pháp Case-base Editing đã được Brighton và Mellish (2002) chia ra làm hai nhiệm vụ chính là Competence preservation và competence enhancement. Competence preservation thực hiện việc giảm bớt sự dư thừa, những case không có đóng góp gì vào việc phân lớp cho case mới. Competence enhancement thực hiện loại bỏ những case nhiễu ra khỏi dữ liệu huấn luyện.



Hình 2.4 mình họa cả hai trường hợp này, các case cùng một lớp có hình sao, các case thuộc lớp khác có hình tròn.[17]

Có hai chiến lược được thực hiện trong Edit case-base: incremental: thêm các case ở tập dữ liệu huấn luyện vào edited set rỗng, và decremental: giảm bớt tập dữ liệu huấn luyện bằng cách loại bỏ một số case.

Một phương pháp Compentence perservation do Hart (1968) đề suất sớm nhất đó là Condensed Nearest Neighbour (CNN). CNN là một phương pháp thực hiện incremental, thêm vào tập edited set (được khởi tạo là tập rỗng) case bất kì từ tập dữ liệu huấn luyện mà những case này không thể được phân lớp đúng bởi case trong edited set. Ritter năm 1975 đưa ra cải tiến dựa trên CNN đó là phương pháp Selective Nearest Neighbour (SNN) áp dụng them các luật cho case trong tập dữ liệu huấn luyện. Năm 1972 Gates giới thiệu phương pháp decremental: Đầu tiên tập edited set bằng với tập dữ liệu huấn luyện sau đó sẽ loại bỏ các case từ tập edited set, sự loại bỏ case đó phải thỏa mãn các case còn lại vẫn được phân lớp đúng.

Thuật toán Edited Nearest Neighbour (ENN) của Wilson (năm 1972), thực hiện chiến lược decremental – loại bỏ các case (case không phù hợp với k hàng xóm gần nhất của nó) ra khỏi tập dữ liệu huấn luyện. Các case này bị coi là nhiễu, các case này không nằm cùng lớp với một cụm case cùng lớp. Tomek (năm 1976) đã cải tiến thuật toán ENN thành repeated ENN (RENN). RENN thực hiện lặp đi lặp lại thuật toán ENN cho đến khi không thể loại trừ được case nào ra khỏi tập dữ liệu huấn luyện thì dừng lại.

Hướng nghiên cứu gần đây cho case-base editing là xây dựng mô hình competence của tập dữ liệu huấn luyện, sử dụng các thuộc tính competence (khả năng) để xác định case sẽ được đưa vào tập edited set. Việc đánh giá và sử dụng case competence được

Smyth và Keane (năm 1995) đưa ra đầu tiên và được phát triển bởi Zu và Yang (năm 1997). Smyth và Keane (năm 1995) đưa ra hai thuộc tính competence quan trọng đó là reachability và coverage. Tập reachability của case c gồm tất cả các case mà c có thể được phân lớp đúng dựa vào các case đó. Tập coverage của case c gồm tất cả các case mà c đóng góp vào việc phân lớp đúng cho những case đó. Trong chương 3 sẽ trình bày một phương pháp mới để edit Case-base do Delany đề xuất, phương pháp BBNR dựa trên phương pháp của Smyth và Keane.

Chương 3

EMAIL CLASSIFICATION USING EXAMPLE

Chương này mô tả thiết kế của hệ thống lọc spam dựa trên case-based là Email Classification Using Examples(ECUE). Đầu tiên sẽ mô tả thiết kế việc sử dụng case-based trong ECUE, mô tả việc trích chọn, lựa chọn các đặc trưng và sự biểu diễn các đặc trưng của case trong case-base và việc lấy case như thế nào, và công nghệ case-base editing(Delany 2005)[17].

3.1 Mô hình thiết kế Case-base áp dụng trong hệ thống ECUE

Phần này sẽ trình bày thiết kế của case-base áp dụng trong hệ thống ECUE, chỉ ra những đặc trưng của case. Mô tả việc trích chọn những đặc trưng từ email messagess như thế nào, đặc trưng nào sẽ được trích chọn, đặc trưng đó được biểu diễn trong case-base như thế nào. Mô tả tiến trình lựa chọn các đặc trưng, chọn những thuộc tính này để dự đoán thư đó là spam hay là thư hợp lệ. Mô tả việc lấy các case từ case-base để đưa vào phân lớp như thế nào, và mô tả công nghệ case-editing..

3.1.1 Trích chọn đặc trưng

Để có thể nhận dạng các đặc trưng từ tập dữ liệu huấn luyện email, mỗi một email được phân tích từ loại và từ tố. Những phần đính kèm email sẽ được loại bỏ trước khi phân tích cú pháp, mã html trong email vẫn được đưa vào bộ phân tích từ tố. Tập dữ được sử dụng trong suốt quá trình đánh giá đó là tập dữ liệu của cá nhân, ví dụ như các email trong tập dữ liệu được gửi tới một người nhận. Do đó những thông tin chứa trong trường header của email là rất hữu ích, bao gồm Subject, To và From cũng sẽ được đưa vào bộ phân tích từ tố. Theo nhiều nghiên cứu đã đưa ra kết luận những thông tin trong trường header của email có tầm quan trọng tương đương với nội dung của email.

Ba loại đặc trưng được xác nhận đó là:

- Đặc trưng từ (ví dụ: các chuỗi kí tự được phân cách nhau bởi kí tự trắng hoặc được phân cách nhau bởi thẻ đánh dấu bắt đầu và thẻ đánh dấu kết thúc trong mã HTML).

- Đặc trưng kí tự đơn.

- Đặc trưng có tính chất cấu trúc, chữ hoa, chữ thường, dấu chấm câu và kí tự phân cách.

3.1.2 Biểu diễn đặc trưng

Trong lĩnh vực lọc spam, mỗi một ví dụ học là một case được biểu diễn dưới dạng một vector các giá trị thuộc tính e_j= (f_{1j}, f_{2j} , . . . f_{nj}, s). Trong phân lớp văn bản những đặc trưng của từ vựng thường được biểu diễn dưới hai dạng[17]:

(a) mã nhị phân ví dụ như: nếu đặc trưng f_{ij} thuộc vào email e_i thì f_{ij}=1, ngược lại bằng 0.

(b) biểu diễn dưới dạng số, trong đó f_{ij} là số lần xuất hiện của đặc trưng đó trong email.

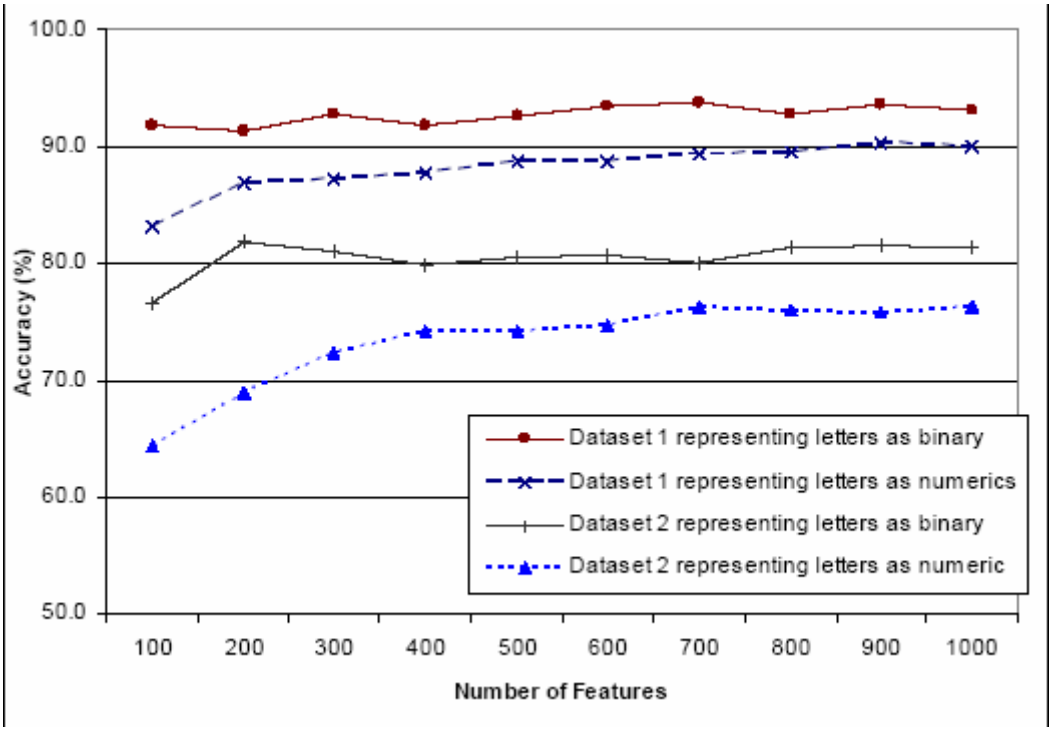
Thuộc tính s biểu diễn cho lớp email đó là spam hay là nonspam.

Thường giá trị của f_{ij} cho f_i trong email e_j được tính dựa vào tần suất xuất hiện của đặc trưng đó trong email. Công thức tính như sau:

$$f_{ij} = \frac{freq_{ij}}{\max_k freq_{kj}}$$

$freq_{ij}$ là số lần xuất hiện của f_i trong email e_j. Công thức trên được tính cho cả đặc trưng từ và đặc trưng chữ cái và đặc trưng thống kê.

Trong phương pháp biểu diễn dưới dạng nhị phân. Đối với các đặc trưng từ, sử dụng luật tồn tại để xác định: nếu từ đó xuất hiện trong email thì giá trị của đặc trưng f_{ij}=1 và ngược lại f_{ij}=0. Tuy nhiên với đặc trưng chữ cái thì không thể sử dụng luật tồn tại được vì hầu như các chữ cái đều xuất hiện trong email. Với đặc trưng chữ cái chúng ta sử dụng giá trị Information Gain (Quinlan năm 1997) của đặc trưng đó để từ đó kết luận giá trị f_{ij} của nó bằng 1 hay bằng 0. Hình 3.1 dưới đây biểu diễn độ chính xác khi sử dụng biểu diễn kí tự dưới dạng binary của hai tập dữ liệu và dưới dạng numeric, ta thấy khi biểu diễn kí tự dưới dạng binary cho độ chính xác cao hơn.



Hình 3.1 : Biểu diễn sự so sánh độ chính xác thu được khi biểu diễn dưới dạng binary và dạng số[17].

3.1.3 Lựa chọn các đặc trưng

Việc phân tích thành từ tố của hàng nghìn email sẽ dẫn đến một số lượng khổng lồ các đặc trưng, vì vậy việc lựa chọn các đặc trưng để làm giảm kích cỡ không gian các đặc trưng là rất cần thiết. Yang và Pedersen (1997) đưa ra đề xuất sử dụng phương pháp đánh giá độ Information Gain (IG) (Quinlan 1997) của đặc trưng để lựa chọn đặc trưng tốt nhất. Information Gain của một đặc trưng là độ đo lượng thông tin mà đặc trưng đó đóng góp vàp tập dữ liệu huấn luyện. Công thức tính IG của đặc trưng A trong tập dữ liệu huấn luyện T như sau[17]:

$$IG(T,A) = Entropy(T) - \sum_{v \in values(A)} \frac{|T_v|}{|T|} Entropy(T_v)$$

T_v là tập con của tập T

Entropy là độ đo xác định trong một tập dữ liệu có bao nhiêu tạp chất. công thức tính như sau[4]:

$$Entropy(T) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

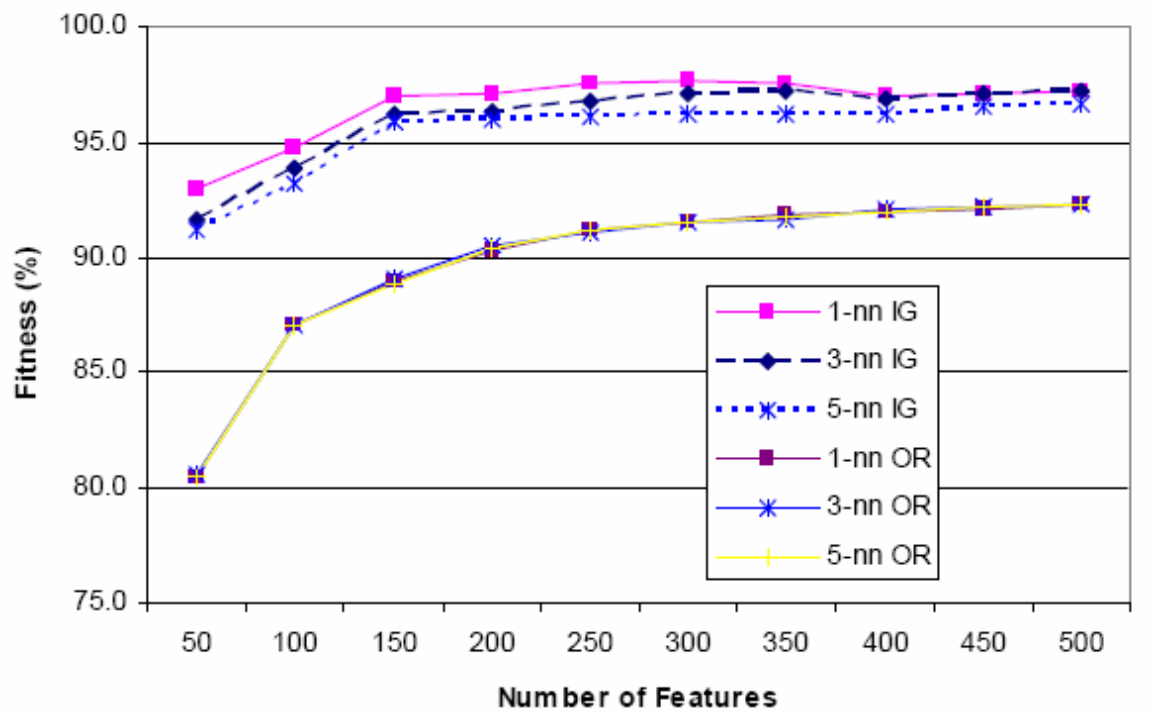
c là số lớp trong tập dữ liệu huấn luyện (trong lĩnh vực lọc spam có 2 lớp là lớp spam và nonspam).

Trong công nghệ lựa chọn đặc trưng Cunningham cũng đưa ra một phương pháp mới đó là sử dụng Odds Ratio (OR) (Mladenic 1998). OR là phương pháp lựa chọn đặc trưng trong bài toán phân lớp nhị phân, sử dụng tỉ lệ chênh lệch (odd) của các đặc trưng xuất hiện trong một lớp với sự xuất hiện của đặc trưng đó trong một lớp khác. Công thức tính OR như sau:

$$OR(f_i, c_j) = \frac{P(f_i|c_j)}{P(f_i|\bar{c_j})}$$

Với $P(f_i|c_j)$ là xác suất xuất hiện đặc trưng f_i trong lớp c_j

Hình 4.2 sẽ biểu diễn sự chính xác của việc lựa chọn đặc trưng khi sử dụng IG và OR. Rõ ràng ta thấy sử dụng IG cho độ chính xác cao hơn OR.



Hình 3.2: So sánh sử dụng IG và OR. Với tập dữ liệu gồm 1000 emails, 500 spam và 500 nonspam, chỉ sử dụng đặc trưng từ[17].

3.1.4 Phân lớp dựa trên thuật toán k-Nearest Neighbour(k-NN).

Bộ phân lớp dựa trên thuật toán k-Nearest Neighbour (k-NN) sẽ phân tích bộ case có độ tương đồng lớn với case mới để phân lớp cho case mới. Độ tương đồng Sim giữa case mới e_t và case e_c trong case-base được tính theo công thức sau[17]:

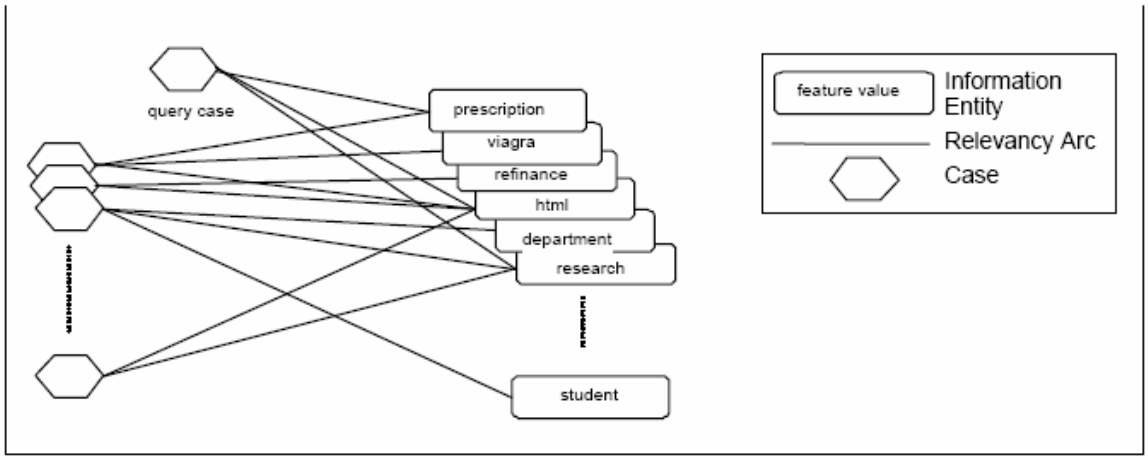
$$Sim = \sum_{i=1}^{n} |f_{it} - f_{ic}|$$

f_{it}: là tần số xuất hiện của đặc trưng thứ i trong case e_t

Khi chọn được những case có độ tương đồng cao nhất với case mới, sử dụng thuật toán bình chọn để xác định lớp gán cho case mới.

3.1.5 Case Retrieval:

Theo thuật toán k-NN chuẩn tính độ tương đồng cho từng case trong case-base với case mới. Cách tính này không hiệu quả, những case spam chứa rất nhiều đặc trưng không thể nhận biết, những đặc trưng được biểu diễn dưới dạng nhị phân vì do đó có một cách tiếp cận mới là Case Retrieval Nets (CRNs)(Lenz et al. 1998). Khi những đặc trưng được biểu diễn dưới dạng nhị phân, IEs chỉ gồm những đặc trưng có giá trị true không cần thiết chứa độ tương đồng.



Hình 3.3 Mô tả một ví dụ áp dụng CRN để lọc spam. Quá trình thực hiện CRN có một vài nét tương tự như Concept Network Graph (CNG)) (Ceglowski et al. 2003)[16]

3.2 Case-Base Maintenance

Chiến lược quản lý case-base trong ECUE gồm có hai phần chính, quản lý kích thước của dữ liệu huấn luyện và thứ hai là việc kế thừa những email gồm cả spam và

nonspam. Phần chính trong quản lý tập dữ liệu huấn luyện là thực hiện edit case-base, xóa bỏ những mẫu nhiễu, loại bỏ những case dư thừa trong case-base. Các nhà nghiên cứu Smyth và Keane năm 1995, McKenna và Smyth năm2000, Wilson và Martinez năm 1997, Brighton và Mellish năm 2002 đã có những nghiên cứu đáng kể về vấn đề edit case-base. Trong ECUE công nghệ edit case-base được sử dụng là Competence Based Editting (Delany và Cunningham năm 2004), sử dụng thuộc tính competence của case để xác định ra case nhiễu và case dư thừa, loại bỏ case đó ra khỏi case-base.

CBE xác định competence của case-base bằng cách xác định những case có đóng góp vào việc phân lớp chính xác cho case mới, và cả những case làm cho việc phân lớp đó bị sai. Những thuộc tính competence của mỗi case được sử dụng trong hai giai đoạn xử lý để tìm ra case cần loại bỏ: thứ nhất incremental và decremental.

Nhiệm thứ hai trong việc duy trì case base là cập nhật case-base với những mẫu email mới đã được phân loại là spam, nonspam. Việc cập nhật case-base được thực hiện ở hai mức, mức đơn giản nhất chỉ là việc đưa các case mới đã được phân lớp vào case-base, mức cao hơn là khi việc phân lớp case mới chưa được thỏa đáng, hệ thống sẽ cho case mới học lại và việc cập nhật case-base sẽ thực hiện lựa chọn lại các đặc trưng có độ dự đoán lớp cho case mới nhất.

3.3 Competence Based Editing

Case-base Editing sử dụng phương pháp Competence Based Editing (CBE) để xác định case không hữu ích trong việc dự đoán phân lớp cho case mới. CBE có hai chức năng chính là loại bỏ case nhiễu và case dư thừa, việc loại bỏ case nhiễu áp dụng thuật toán Blame Based Noise Reduction (BBNR), việc loại bỏ case dư thừa áp dụng thuật toán Conservative Redundancy Reduction (CRR)(Riesbeck and Shank 1989) [16]. Case-base update policy thực hiện việc đưa các case đã được phân lớp là spam, nonspam vào case-base để đưa dự đoán lớp cho case tiếp theo, trong trường hợp cho case học lại, case-base update policy thực hiện lựa chọn lại các đặc trưng để tìm ra đặc trưng có ích trong việc dự đoán lớp cho case mới.

3.3.1 Thuật toán Blame Based Noise Reduction

Mô hình Case-Base Competence ban đầu được Smyth và McKenna đề suất có hai tập: tập reachability và tập coverage. Tập reachability của case t là tập gồm case trong case-base giúp phân lớp đúng cho case t. Tập coverage của case t gồm case mới t mà được phân lớp đúng. Ta có thể biểu diễn hai tập đó như sau:[16]

Reachability Set($t \in C$) = $\{c \in C : Classifies(t, c)\}$

Coverage Set($t \in C$) = $\{c \in C : Classifies(c, t)\}$

Classifies(a,b) nghĩa là case b góp phần vào việc phân lớp đúng cho case mới a, case mới a được phân lớp đúng và case b được coi là hàng xóm cùng lớp gần nhất của case a.

Phát triển mô hình Case-Base Competence, Delany đã mở rộng mô hình với việc thêm các thuộc tính mới; tập Liability của case t là tập các case mà làm phân lớp case t bị sai, tập Liability có thể được biểu diễn như sau:

Liability Set($t \in C$) = {$c \in C$: Misclassifies(c, t)}

Với Misclassifies(a,b) nghĩa là case b gây ra việc phân lớp sai cho case mới a, khi case mới a bị phân lớp sai thì case b sẽ coi như là hàng xóm khác lớp của a.

Thuật toán BBNR: Thuật toán giảm thiểu nhiễu của Wilson 1972. Những case nhiễu là những case trong tập dữ liệu case huấn luyện nhưng nó bị gán nhãn sai (phân lớp sai). Theo phương pháp của Wilson thì loại bỏ những case bị phân lớp sai, sẽ bị gán nhãn sai case đó bị coi là nhiễu. Theo hướng tiếp cận BBNR, những case gây ra phân lớp sai sẽ được chú ý hơn là những case bị phân lớp sai. Trong bộ luật áp dụng để giảm bớt nhiễu chúng ta cố gắng loại bỏ những case bị gán nhãn sai, và loại bỏ những case không hữu ích - case gây ra việc phân lớp sai: ví dụ case là email, một email thực tế là spam nhưng nó lại có nhiều đặc trưng giống như là một thư hợp lệ.

Theo phương pháp BBNR sẽ xem xét tất cả các case trong case-base gây ra việc phân lớp sai. Đối với mỗi một case c sẽ có một liability chứa ít nhất là một phần tử, nếu những case trong tập coverage của c vẫn được phân lớp đúng nếu không có c thì c sẽ bị loại bỏ. Thuật toán BBNR được mô tả như sau[16]:

```
T = Training Set
/* Build case−base competence model */
for (each c in T)
    CSet(c) = Coverage Set of c
    LSet(c) = Liability Set of c
endfor
/* remove noisy cases */
TSet = T sorted in descending order of LSet(c) size
and
    ascending order of CSet(c) size
c = first case in TSet
while (|LSet(c)| > 0)
    TSet = TSet − {c}
    misClassifiedFlag = false
    for (each x in CSet(c))
        if (x cannot be correctly classified by TSet)
            misClassifiedFlag = true
            break
        endif
    endfor
    if (misClassifiedFlag == true)
        TSet = TSet + {c}
    endif
    c = next case in TSet
endwhile
return TSet
```

3.3.2 Conservative Redundancy Reduction

Thuật toán loại bỏ case nhiễu dựa trên việc xác định những case nằm trên đường biên giữa hai lớp. Tập coverage lớn gồm những case nằm trong cụm các case cùng lớp, tập coverage nhỏ gồm những case mà một số hàng xóm của nó thuộc cùng lớp. những case thuộc biên giữa 2 lớp sẽ thuộc vào tập coverage nhỏ, những case này sẽ được thêm vào edited set đầu tiên.
Thuật toán sử dụng cho việc loại bỏ các case dư thừa được biểu diễn như sau:[16]

```
T = Training Set
/* Build case−base competence model */
for (each c in T)
    CSet(c) = Coverage Set of c
endfor
/* remove redundant cases from case−base */
ESet = {}, /* Edited Set */
TSet = T sorted in ascending order of CSet(c) size
c = first case in TSet
while TSet <> {}
    ESet = ESet + {c}
    TSet = TSet − CSet{c}
    c = next case in TSet
endwhile
return ESet
```

3.4 Mô hình thiết kế ECUE online

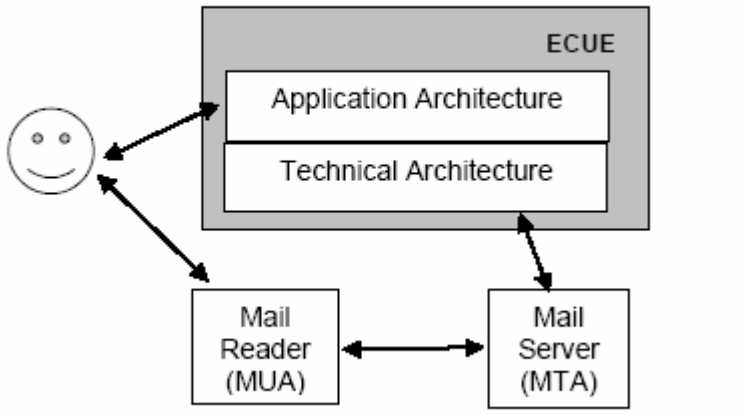
Phần này sẽ mô tả về thiết kế của hệ thống ứng dụng online ECUE (Delany)[17], những công nghệ sử dụng cho phép hệ thống tích hợp với việc nhận thư của từng cá nhân và thực hiện các chức năng học, lọc spam.

3.4.1 Cấu trúc của hệ thống

Cấu trúc của ứng dụng ECUE lọc thư rác được minh họa trên hình 4.4, có hai phần chính; cấu trúc liên quan đến kĩ thuật và cấu trúc liên quan đến ứng dụng. Cấu trúc liên quan đến công nghệ là bộ khung thực hiện các chức năng lọc, nó chịu trách nhiệm tích hợp với mailbox của người dùng để thực hiện các công việc sau:

(1) Lấy thư mới đến và thực hiện lọc thư đó

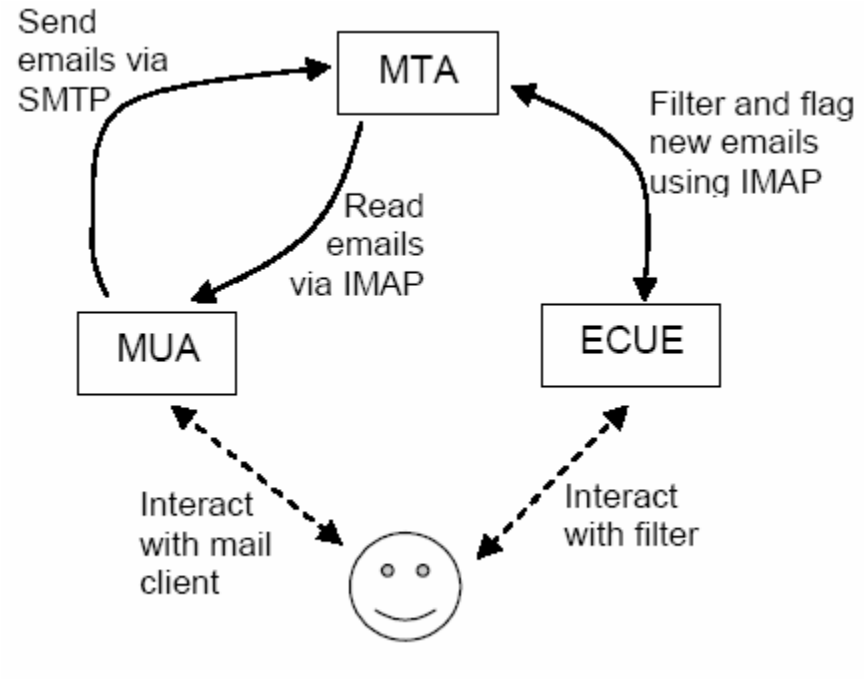
(2) Khi người dùng nhận được độ đo False Positive(FP) hoặc là False Negative (FN) của email thì ứng dụng lọc spam đưa những email này vào tiến trình học.



Hình 3.4 Kiến trúc hệ thống ECUE[17].

Application architecture hỗ trợ những chức năng lọc thực sự. Nó tích hợp với technical architecture thông báo khi email mới cần được lọc hoặc khi bộ lọc gặp lỗi và quá trình học lại được tiếp tục. Yêu cầu chính đòi hỏi hệ thống lọc phải tích hợp được với hệ thống mail user agent hoặc hệ thống mail reader. Điều này cho phép người dùng vẫn tiếp tục sử dụng phần mềm đọc mail mà không gây ảnh hưởng gì đến hệ thống lọc. Cấu trúc của hệ thống lọc cũng được thiết kế hỗ trợ cho giao thức Internet Message Acces Protocol (IMAP)(Hughes 1998). Giao thức IMAP là một trong hai giao thức mail (giao thức IMAP và POP3) có thể nhận email. Ưu điểm của IMAP so với POP3 đó là IMAP hỗ trợ việc lưu trữ các email nhận được trên server trung tâm, do đó có thể thực hiện nhiều truy cập cùng một lúc từ các vị trí khác nhau. Bằng việc sử dụng IMAP để truy cập vào mailbox, các email có thể được lọc và gán cờ trên server và điều này cho phép người dùng bất kì một trình đọc thư nào có hỗ trợ IMAP trên máy khách để truy cập và đọc thư của họ. Có rất nhiều ứng dụng đọc thư có hỗ trợ giao thức IMAP phổ biến như: MS Outlook, Mozilla, Netscape và Thunderbird.

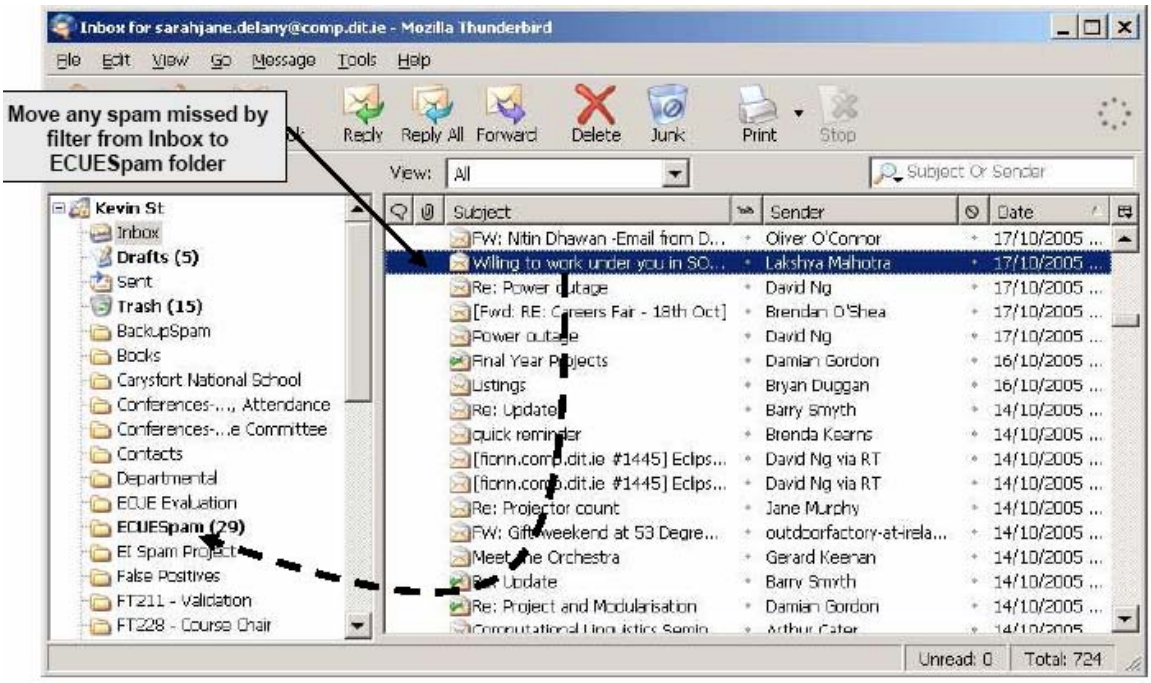
Hình 3.5 mình họa hệ thống lọc spam thực hiện như thế nào với hệ thống đọc thư. Cả mail reader và hệ thống lọc spam đều thăm dò qua MTA hoặc mail server theo định kì để kiểm tra xem có thư mới hay không.



Hình 3.5 Sơ đồ minh họa sự tích hợp giữa hệ thống lọc ECUE và mail client[17]

3.4.2 Tương tác với người dùng

Phải có sự tương tác giữa người dùng và hệ thống lọc, vì hai nguyên nhân chính sau: Thứ nhất là bộ lọc phải cho phép người dùng biết những email đã bị phân loại thành sapm, thứ hai là ngừoi dùng phải được phép cảnh báo bộ lọc là email đã bị phân lớp sai. Hệ thống lọc đặt những email là spam vào thư mục spam cho người dùng tạo, còn những thư không phải là spam sẽ được đưa vào Inbox. Nếu người dùng tìm thấy thư bị phân lớp sai họ có thể chỉ ra cho hệ thống bằng cách di chuyển những thư đó từ thư mục đó sang thư mục mà lẽ ra nó ở đó. Thư mục mail cũng được sử dụng để làm dữ liệu huấn luyện ban đầu cho hệ thống. người dùng xác nhận tập email dùng để huấn luyện

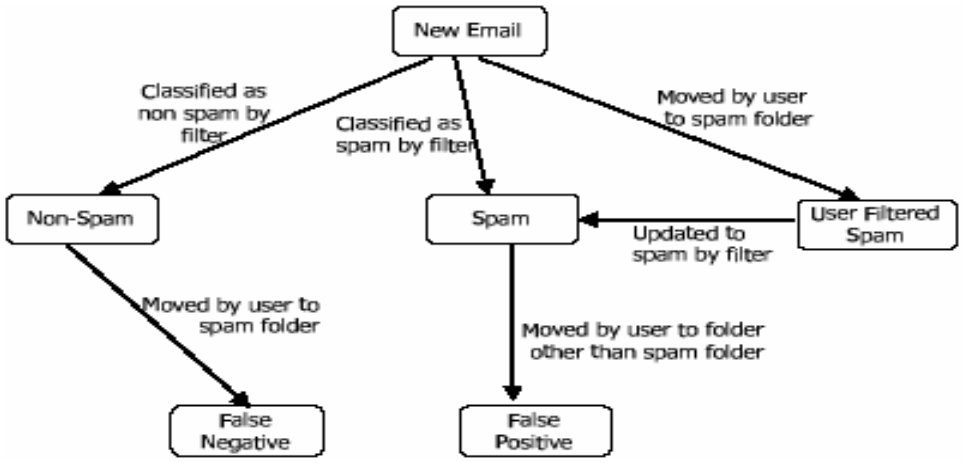


Hình 3.6: Người dùng tương tác với hệ thống ECUE[17]

3.4.3 Theo dõi Emails

Để theo dõi những thư đến và thư đã được lọc ứng dụng ECUE gắn thêm một trường vào header của email. Khi một email đã được lọc, một trường header được thêm vào email đó để chỉ ra email đó là spam hay là nonspam. Nếu người dùng tìm thấy một email trong Inbox của họ mà email đó đã được phân vào lớp spam họ có thể di chuyển thư đó đến thư mục spam. Do đó nếu email đó có trường header xác định là nonspam(do hệ thống lọc) thì email này là FN. Tương tự nếu email có trường header là spam được người dùng di chuyển đến Inbox thì thư đó là FP.

Trong trường hợp người dùng có thể truy cập vào thư mới đến trước khi bộ lọc thực hiện lọc thư đó, nếu người dùng xác nhận thư đó là spam và di chuyển nó đến thư mục spam thì hệ thống lọc sẽ coi thư đó là thư spam do người dùng lọc và hệ thống sẽ cập nhật thư đó là thư spam (thêm giá trị xác định là spam vào trường header của thư đó). Trong trường hợp khác, khi người dùng coi một thư là thư nonspam và di chuyển nó đến thư mục khác không phải là thư mục spam, khi đó trong thời gian tiếp theo bộ lọc sẽ truy cập vào thư mục đó và thư đó sẽ được lọc.



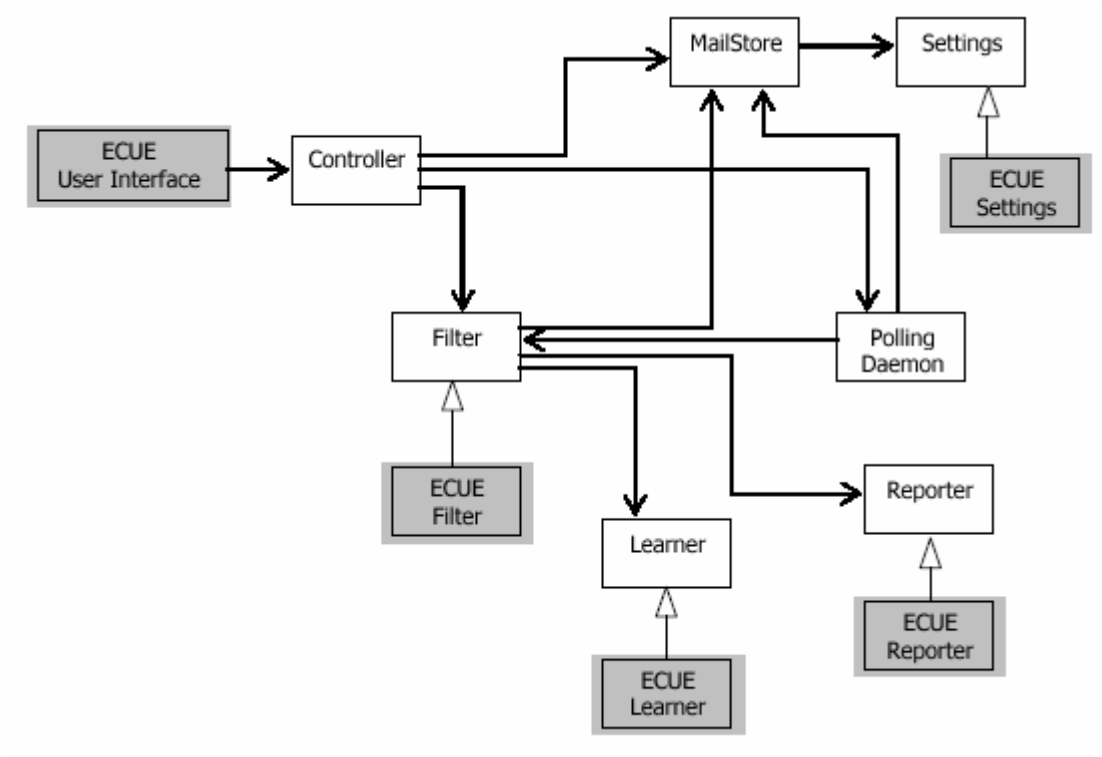
Hình 3.7 Mô tả sơ đồ các trạng thái di chuyển có thể xảy ra đối với một email[17]

3.5 Mô hình thiết kế ở mức cao

3.5.1 Mô hình thiết kế tầng Technical Architecture

Technical architecture gồm hai lớp chính là Daemon và Filter. Lớp Filter làm nhiệm vụ lọc email và xác định thư đó là thư spam hay nonspam và quản lý những thư FP và FN do người dùng xác nhận. Lớp Daemon quản lý giao tiếp giữa mailbox của người dùng trên MTA và bộ lọc. Daemon và Filter được thực hiện theo những thread riêng biệt. Cấu trúc của Daemon và Filter có thể được thay đổi mà không ảnh hưởng đến ứng dụng lọc sử dụng chúng. Lớp Daemon thực hiện thăm dò mailbox của người dùng theo định kì, kiểm tra xem số lượng thư trong các folder có sự thay đổi nào không. Nếu số lượng email có sự thay đổi, có thể là có thư mới đến hoặc do người dùng di chuyển thư giữa các folder với nhau, khi đó daemon sẽ thông báo cho bộ lọc, và bộ lọc sẽ biết được folder nào cần phải lọc.

Khi nhận được thông báo từ Daemon thì lớp Filter sẽ được thực thi, nó được kích hoạt ở cấp thư mục. Lớp Filter kiểm tra trường header của email trong thư mục xem email đó là thư mới đến hay là thư FP hoặc FN. Nếu đó là thư mới đến, thư đó sẽ được lọc và được phân loại là spam hay nonspam, khi đó giá trị tương ứng của header sẽ được thêm vào email. Nếu email đó là FP hoặc FN thì một bản báo cáo được ghi lại (do bộ Reporter thực hiện) và lớp Learner được kích hoạt và thực hiện việc học.



Hình 3.8: Sơ đồ các lớp của ECUE[17]

Lớp MailStore cung cấp các phương thức để kết nối với mailbox và truy cập vào các thư mục trong mailbox. Lớp Settings thiết lập cấu hình cần thiết cho hệ thống gồm những chi tiết cần thiết để người dùng có thể truy cập vào mailbox, như: host, username và password và tên những folder do người dùng tạo lập. Những tham số cấu hình này được thiết lập trong file cấu hình, nó được truy cập và tải khi ứng dụng hoạt động.

Lớp Controller là lớp điều khiển chính của ứng dụng. Nó thực hiện chức năng điều khiển khởi động hoặc ngừng Filter và kiểm soát Daemon. Nó thực hiện các thread riêng biệt và độc lập với Filter và Daemon. Giao diện của lớp Learner và Reporter chỉ định việc học và báo cáo khi hệ thống lọc yêu cầu

3.5.2 Mô hình thiết kế tần Application Architecture

Tầng ứng dụng (Application) cung cấp các chức năng lọc của case-base reasoning. Tầng ứng dụng liên quan đến những phần sau:

1. thiết lập và lưu giữ case-base, những email mẫu huấn luyện.

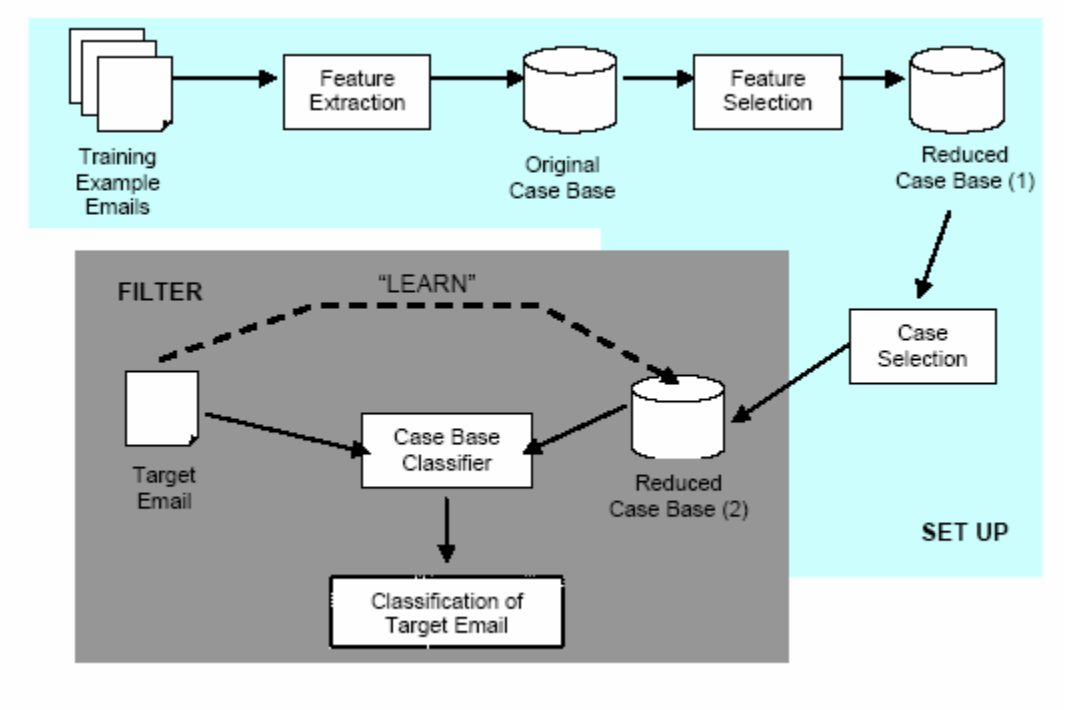
2. tiến trình phân lớp sử dụng để xác định lớp cho email mới

3. tiến trình cập nhật để hệ thống học những mẫu email mới.

Hình 3.8 mô tả cấu trúc của các chức năng ở tầng ứng dụng. Nó gồm hai tiến trình chính:

Tiến trình SetUp làm nhiệm vụ tạo case-base từ những email của người dùng(gồm cả spam và nonspam)

Tiến trình Filter thực hiện lọc những email mới và cập nhật lại case-base khi có bất kì một email bị phân lớp nỗi.



Hình 3.9 : Cấu trúc của tầng Application[17]

Setting up a Case-base

Hệ thống sử dụng những mẫu email trước, gồm cả spam và nonspam làm tập dữ liệu huấn luyện. Đầu tiên những email đó sẽ được qua bước Feature Extraction (trích chọn các thuộc tính), thực hiện phân tích cú pháp, phân tích từ tố những email huấn luyện đó thu được các đặc trưng. Có ba loại đặc trưng được trích chọn đó là: đặc trưng từ (word features), đặc trưng chữ cái(letter features), và đặc trưng statistical(statistical features). Output của tiến trình trích chọn các thuộc tính này là một case-base được khởi tạo với các cặp giá trị feature-value cho mỗi email trong mẫu huấn luyện.

Tiến trình trích chọn các đặc trưng sẽ đưa ra một số lượng lớn các đặc trưng cho mỗi email huấn luyện. Thêm vào đó, sự biểu diễn mỗi email bị thưa thớt, chỉ một số lượng nhỏ số feature được thiết lập với giá trị lớn hơn 0. công việc của Feature Selection là xác định những đặc trưng(được trích ra nhờ bộ Feature Extraction) có khả năng dự báo tốt nhất một email là spam hay nonspam. Phương pháp được sử dụng để lựa chọn các đặc trưng này là Information Gain. Output của tiến trình Feature Selection là giảm bớt số đặc

trưng trong mỗi email huấn luyện, giảm tập các tập chứa cặp giá trị feature-value để trong tập đó chỉ chứa những đặc trưng có khả năng dự báo cao nhất. Trong hệ thống ECUE ta có thể cấu hình để xác định số lượng đặc trưng cần thiết. Ở đây các đặc trưng được thể hiện dưới dạng nhị phân.

Nhiệm vụ của Case Selection là áp dụng phương pháp Competence-Base Editting, sử dụng các thuộc tính competence của các mẫu trong case-base để loại bỏ các case nhiễu và case thừa trong case-base. Output của tiến trình này là làm nhỏ case-base.

Hệ thống sử dụng nhiều tập dữ liệu huấn luyện khác nhau phụ thuộc vào case-base cần phải được xây dựng. Nếu hệ thống được thực thi lần đầu tiên, tiến trình SetUp sẽ sử dụng tập dữ liệu huấn luyện được được người dùng đưa vào thư mục huấn luyện trên mailbox. Tổng số email được sử dụng để huấn luyện được cấu hình.

Filtering and Learning

Bộ phân lớp một email mới được thực hiện dựa trên thuật toán k-Nearest Neighbour đã được trình bày ở phần 4.1.4. Giá trị của k được thiết lập ở file cấu hình. Tiến trình phân lớp sử dụng bỏ phiếu đồng nhất để giúp bộ phân lớp gặp lỗi phân lớp FP, tức là đòi hỏi tất cả k hàng xóm được xác định bởi thuật toán k-NN phân vào lớp spam trước khi case mới có thể bị phân lớp là spam.

Khi người dùng xác nhận rằng email đó bị hệ thống phân lớp sai, tiến trình học sẽ được thực hiện. Có hai cấp học được thiết lập trong hệ thống:

(i) Đưa case mới sau khi được phân lớp vào case-base

(ii) Thực hiện học lại, trích chọn lại các đặc trưng để phân lớp đúng cho case mới.

Hệ thống cũng cung cấp feedback (hồi âm) đến người đọc qua lớp ECUE Reporter để cung cấp thống kê cho người đọc về sự thực hiện lọc của hệ thống và những thống kê đó cũng được sử dụng vào mục đích định giá.

Whitelisting

Để giảm và loại trừ lỗi false positive, hệ thống lưu ý đến danh dách trắng, được thảo luận trong chương 2, phần 2.3.2, nó hoạt động ở hai mức sau:

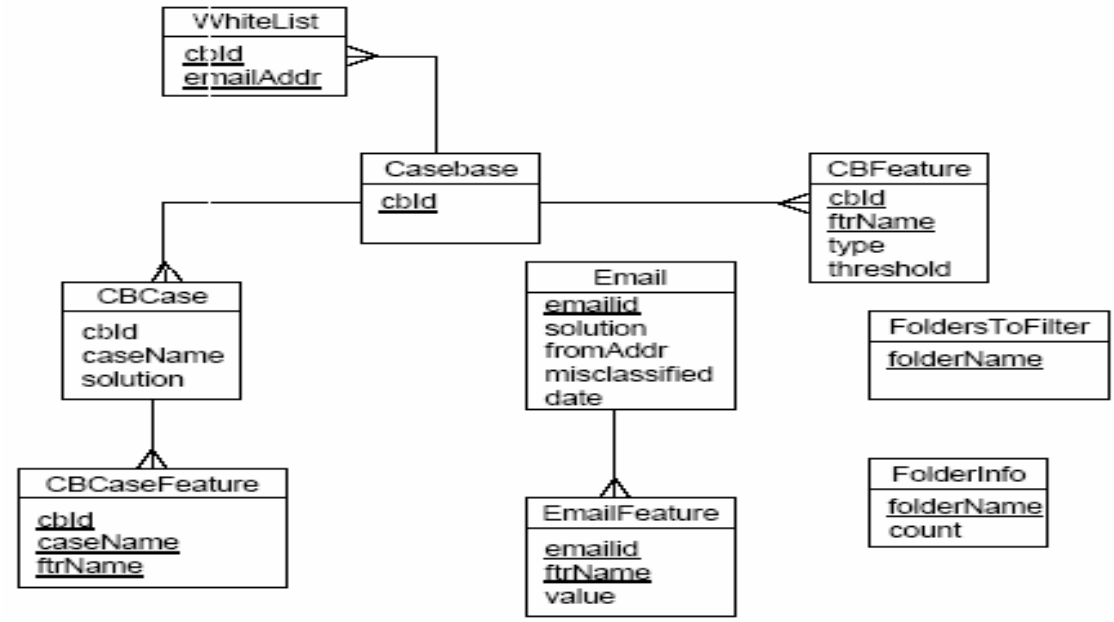
1. Người dùng có thể định nghĩa các miền được phép trong file cấu hình. Bất kì email nào đến từ những vùng này được coi là hợp lệ.

2. Người gửi những email hợp lệ được lưu lại trong danh sách, những email đó sẽ có thêm đặc trưng để xác định người gửi đó nằm trong danh sách trắng hay không. Đặc trưng này được sử dụng trong tiến trình thu thu thập case-base (case-base retrieval process), xác định những case hàng xóm của case mới.

Database

Hệ thống sử dụng cơ sở dữ liệu MySql để lưu trữ cả case-base và những email của người dùng được định dạng dưới form chứa cặp giá trị feature-value. Cấu trúc của dữ liệu được mô tả trên hình 4.9.

CBFeature lưu chi tiết những đặc trưng được lựa chọn cho một case-base. CBCase lưu chi tiết của case trong case-base, còn CBCaseFeature lưu chi tiết những đặc trưng trong một case. Email và EmailFeature lưu những chi tiết của các email (dưới định dạng chứa cặp giá trị feature-value, để thuận tiện trong việc xây dựng lại case-base) để bộ lọc thực hiện phân lớp. FolderInfo và FoldersToFilter giữ những thông tin về trạng thái của mailbox của người dùng giữa những lần bộ lọc thực thi, hai thực thể này chỉ có tác dụng làm cho việc thực thi được thuận lợi hơn, nó cho phép ứng dụng xác định thư mục cần phải lọc khi hệ thống khởi động lại. Thực thể WhiteLisst giữ các thông tin về danh sách trắng.



Hình 3.10: cơ sở dữ liệu ECUE[17]

3.6 Đánh giá kết quả lọc của hệ thống ECUE

Phần này đưa ra đánh giá của Delany về mức độ lọc chính xác của hệ thống ECUE, đồng thời đánh giá hiệu quả của thuật toán BBRN sử dụng để edit Case-base. Sự đánh giá dựa trên so sánh các tham số tỉ lệ Error, FP và FN.

3.6.1 Kết quả so sánh về mức độ lọc chính xác của hệ thống ECUE khi sử dụng thuật toán BBRN và thuật toán RENN(Delany, 2006)[17]

4 tập dữ liệu(500 spam, 500 nonspam), email được biểu diễn dạng nhị phân:

- Tập dữ liệu 1.1, 1.2: email nhận được của một người dùng trong tháng 2/2003

- Tập dữ liệu 2.1, 2.2: email nhận được từ tháng 2-12/2003

Sử dụng thuật toán IG để trích chọn feature (k=3): 700 feature được trích chọn.

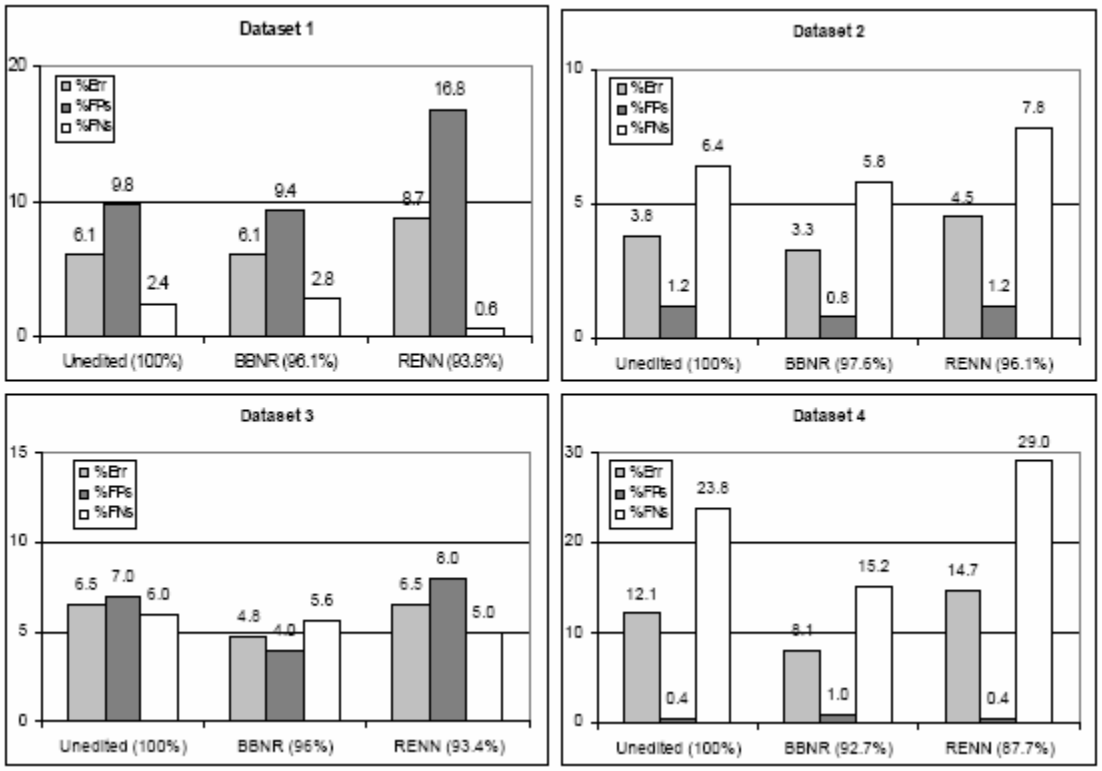
Mỗi tập dữ liệu được chia thành 20 mục. 1 mục làm dữ liệu test, 19 mục làm dữ liệu huấn luyện.

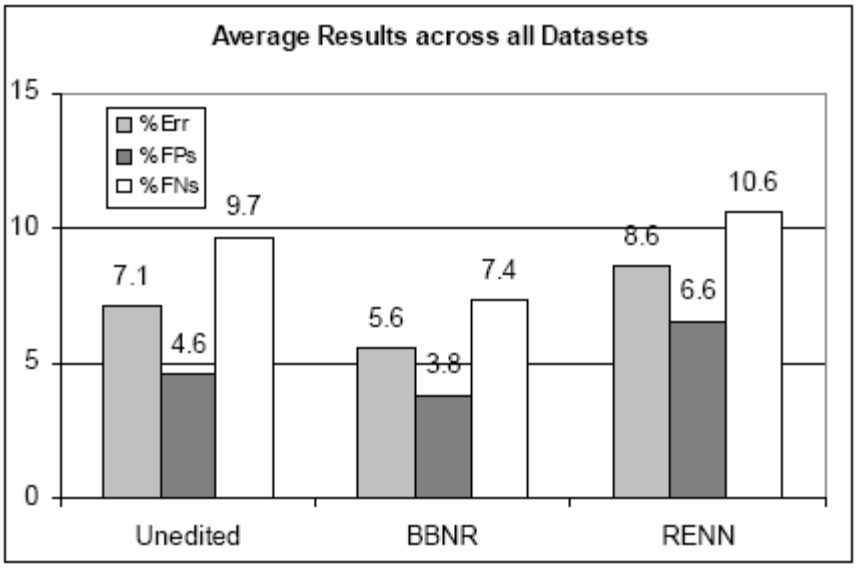
Thực hiện đánh giá dựa trên 3 tham số:

- Error rate: phần trăm email bị ECUE phân lớp sai.

- FN: số email thực sự là spam nhưng bị ECUE phân loại sai thành non-spam

- FP: số email thực sự là non-spam nhưng bị ECUE phân loại sai thành spam.





Hình 3.x: Kết quả so sánh khi sử dụng thuật toán BBNR và RENN để loại bỏ case nhiễu trong case-base, thực hiện trên 4 tập dữ liệu huấn luyện (case-base), và kết quả trung bình cho cả 4 tập dữ liệu.[17]

Dựa vào đồ thị ta thấy rõ:

- Khi sử dụng thuật toán BBNR để loại bỏ case nhiễu trong case-base kết quả thu được rất tốt, tỉ lệ error thấp hơn so với sử dụng thuật toán RENN.

- Và sử dụng thuật toán BBNR rõ ràng cho kết quả tốt hơn nhiều, có tỉ lệ lỗi thấp hơn khi ta không thực hiện điều chỉnh case-base (unedited), loại bỏ case nhiễu

3.6.2 Kết quả đánh giá hoạt động của hệ thống ECUE online.

Tiến hành cài ECUE trên máy PCs[17]

- Có gate way spam filter (SpamAssasin)

- Một số người dùng tắt chức năng SpamAssasin.

Thực hiện đánh giá trên 4 người dùng, dựa vào tham số ER, FN, FP.

Bảng 4.1 chứa các thông số sau:

(i) số ngày hệ thống ECUE thực hiện lọc thư trên máy PC của người dùng

(ii) số lượng thư spam và nonspam được lọc qua thời gian

(iii) Thông tin về tập dữ liệu huấn luyện được sử dụng: gồm số lượng email được dùng huấn luyện và tỉ lệ phần trăm số thư được gán nhãn là spam (có nhãn %spam).

(iv) số lần trung bình update case-base trong 1 ngày (có nhãn (#days))

(v) thông tin về số email mới(spam hoặc nonspam) được thêm vào case-base: gồm tổng số email mới được thêm, và kích thước của case-base (có nhãn Final size)

(vi) tỉ lệ FN: số thư spam mail mà ECUE đã phân lớp sai, những thư spam này đã bị phân lớp sai thành thư hợp lệ (có nhãn %FNs)

(vii) tỉ lệ FP: số thư hợp lệ bị hệ thống ECUE lọc thành thư spam (có nhãn FPs%)

(viii) tỉ lệ error: tổng số thư mà hệ thống ECUE phân lớp sai (có nhãn Error%)

User		1	2	3	4
Filter Period	Start date	18-11-04	9-3-05	20-4-04	7-9-05
	End date	15-07-05	15-07-05	16-10-05	1-11-05
Emails Filtered	#spam	3689	4081	742	75
	#legit	1161	469	1480	917
Casebase	Initial size	308	299	201	308
	%spam	56%	54%	79%	60%
	#ftr reselects	3	3	1	2
% Error	No update	32.8	21.8	17.3	8.1
	With update	6.1	4.7	12.1	4.3
% FPs	No update	0.3	0.0	1.3	7.3
	With update	0.7	0.2	1.1	0.4
% FNs	No update	43.2	24.4	49.3	17.3
	With update	7.8	5.2	34.0	52

Bảng 4.1: kết quả đánh giá ECUE cho 4 user[17]

Từ bảng 3.x ta thấy:

ECUE thực hiện rất tốt việc nhận ra các mẫu spam mới, giảm đáng kể độ FN của các user trừ user 4, và user 3 độ FN ko giảm nhiều như các user khác vì:

- Thư của user 4 và user 3 được lọc qua gate way spam filter trước khi qua bộ lọc ECUE.

- Nhưng kết quả ECUE xác định đúng 66% thư spam của user 3.

- User 1 và user 2 nhận được số thư spam nhiều hơn so với nonspam nên độ FP chỉ giảm nhẹ.

- Trung bình 90% số lượng thư của 4 người dùng được phân lớp đúng.

- User 1 và user 2 email được phân lớp đúng tới 93.9% và 95,3%.

- Average error: 6.8%

Chương 4

THỰC NGHIỆM

Trong phần thực nghiệm này tôi tiến hành thực nghiệm sử dụng chương trình phần mềm nguồn mở *SpamBayes anti-spam*, một dự án được tiến hành của http://sourceforge.net/project/. Chương trình SpamBayes anti-spam thực hiện lọc thư rác dựa trên nội dung áp dụng thuật toán Bayes. Chương trình xây dựng trên ngôn ngữ Perl, có khả năng tích hợp được với hệ thống Mail reader. Trong phần thực nghiệm này tôi tiến hành biên dịch, cài đặt chương trình Bayes tích hợp với hệ thống Microsoft Office Outlook 2003 để lọc thư.

Thực hiện cài đặt:

Download:

- Phython installer: http://www.python.org/download/
- Pywin32 extensions: https://sourceforge.net/projects/pywin32/
- SpamBayes source: spambayes-1.1a3 : http://sourceforge.net/project/showfiles.php?group_id=61702

Chạy Phython installer, pywin32 installer. Click vào addin.py trong thư mục outlook2000 của thư mục spambayes-1.1a3.

Dữ liệu:

Tập dữ liệu huấn luyện gồm 27 ham email và 27 spam email được trích ra từ Dữ liệu corpus: 20030228_hard_ham.tar (gồm 500 ham email) và 20021010_spam.tar(gồm 250 spam email) (http://spamassassin.apache.org/publiccorpus/)

Tập dữ liệu kiểm tra gồm: 16 thư ham và 16 thư spam trích ra từ 20030228_hard_ham.tar + 1518 thư (từ hòm thư: hienhst@yahoo.com và anhtuan_it@yahoo.com – chứa 953 spam và 565 ham), như vậy tổng cộng dữ liệu lọc

gồm 1540 thư, có 1021 thư spam (66,3%) và 519 thư ham (33,7%). Thư spam chủ yếu có nội dung quảng cáo, chứa nhiều link liên kết.

Thực nghiệm

Lần 1: Khi chưa huấn luyện:

Kết quả: 10 thư (trong đó 2 thư spam) đều được xác định là ham.

Lần 2: Khi tập huấn luyện chỉ chứa thư spam:

Kết quả: 10 thư (trong đó 2 thư spam) đều được xác định là spam

Lần 3: Khi tập huấn luyện chỉ chứa thư ham:

Kết quả: 10 thư (trong đó 2 thư spam) đều được xác định là ham

Lần 4: Khi thực hiện huấn luyện với tập dữ liệu huấn luyện đã nêu ở trên, số thư được lọc: 1540 thư

Kết quả: Thư ham: 1239 (80%)

Thư spam: 28 (1,8%)

Thư unsure: 273 (17,7%)

Lần 5: cho hệ thống huấn luyện 16 thư thực chất là spam nhưng hệ thống coi là thư ham.

Kết quả: Thư ham: 1445 (52,4%)

Thư Spam: 1005 (36,5%)

Thư unsure: 306 (11,1%)

Lần 6: Tiến hành huấn luyện 11 thư trong hợp lệ nhưng bị hệ thống lọc là spam:

Kết quả: Thư ham: 2043 (44,6%)

Thư spam: 2063 (45,0%)

Thư unsure: 476 (10,4%)

Lần 7: Thực hiện huấn luyện 19 thư là spam nhưng bị bộ lọc phân vào lớp unsure

Kết quả: Thư ham: 2652 (34,1%)

Thư spam: 4504 (57,9%)

Thư unsure: 618 (7,9%)

Lần 8: Thực hiện huấn luyện 20 thư là ham nhưng bị bộ lọc phân vào lớp unsure

Kết quả: Thư ham: 2996 (32,1%)

Thư spam: 5693 (61,0%)

Thư unsure: 642 (6,9%)

Kết quả sau lần 8:

Số lượng thư trong unsure: 17 thư

Thư spam: 1028 thư, trong đó 940 thư lọc đúng là thư spam.

Thư ham: 495 thư, trong đó 421 thư được lọc đúng là ham.

Đánh giá:

Từ kết quả thu được các lần thực nghiệm ta nhận thấy rõ ràng hệ thống Spambayes có khả năng học rất tốt, sau khi được học thư spam và thư ham hệ thống lọc chính xác hơn. Từ kết quả cuối cùng của lần 8 ta có:

Số thư ham bị hệ thống lọc thành thư spam là: 1028 – 940 = 88 thư

=> FP = 100*88/1028 = 8.56%

Số thư spam được hệ thống lọc thành thư ham là: 495 – 421 = 74 thư

=> FN = 100* 74/495 = 14,9%

Chúng ta tiến hành tính toán độ hồi tưởng, độ chính xác và độ đo F1 đối với kết quả trên đây.

Kết luận

Hiện nay thư rác ngày càng phát triển gây thiệt hại lớn về kinh tế cũng như gây nhiều phiền toái cho người dùng. Số lượng thư rác ngày càng tăng, nội dung cấu trúc của chúng càng thay đổi vì vậy cần có một hệ thống học máy lọc thư để có thể cập nhật, loại bỏ được những mẫu thư mới. Hệ thống học máy lọc thư rác dựa trên nội dung sử dụng phương pháp CBR – hệ thống ECUE đã được xây dựng và đáp ứng được điều đó. Khóa luận đã đạt được một số kết quả như sau:

- Khái quát một số nội dung cơ bản về thư rác, các phương pháp lọc thư rác.

- Trình bày chi tiết về hai phương pháp lọc thư rác theo nội dung theo thuật toán Bayes, trong đó tập trung tới giải pháp của Delany. Đã trình bày về cấu trúc CBR và hệ thống lọc thư rác ECUE.

- Đã tiến hành khai thác chương trình nguồn mở *SpamBayes anti-spam*, cho chạy thực nghiệm và phân tích sơ bộ kết quả.

Để xây dựng được hệ thống ECUE hoàn chỉnh cần nhiều người cùng tham gia. Bước đầu em đã tìm hiểu về cấu trúc cũng như phương pháp để xây dựng hệ thống ECUE, trong tương lai, em hy vọng với sự giúp đỡ của các thày cô và các bạn chúng ta có thể xây dựng được hệ thống học máy lọc thư rác dựa trên nội dung trên cơ sở các nội dung tương tự như hệ thống ECUE.

Tài liệu tham khảo

[1] Aha, D. W.: 1997, Editorial, Artificial Intelligence Review, Special Issue on Lazy Learning

[2] Aha, D. W., Kibler, D. and Albert, M. K.: 1991, Instance-based learning algorithms, Machine Learning

[3] Deborah Fallows (2003). Spam: How it is hurting email and degrading life on the internet. *Technical report, Pew Internet and American Life Project*, Oct 2003

[4] Ion Androutsopoulos, John Koutsias V.Chandrinos and Contstantine D.Spyropoulos. "An Experimental Comparision of Naïve Bayes and keyword-based anti-spam Filtering with persional email message"

[5] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos (). Learning to filter spam email: a comparison of a naïve bayes and a memory-based approach.

[6] J.W.L.Boelen, S.P.Ekkebus (2005). Dealing with spam in the near future Overview of sender authentication techniques. *University of Twente*, Nertherland.

[7] Joachims T. (1998). Text Categorization with Support Vetor Machines: Learning with Many Relevant Feature, *Proceeding of ECML-98, 10th European Conference on Machine Learning*, 1998.

[8] Johan Hovold (). Naïve Bayes Spam filtering using Word-Position-Based attributes. *Department of Computer Science Lund University*.

[9] Kasun De Zoysa, Lakmal Warusawithana (). An innovative Method to Prevent Spam. Department of Communication and Media Technologies, University of Colombo School of Computing, 35, Reid Avenue, Colombo 7, Sri Lanka.

[10] M. Perone (2004). An overview of spam blocking techniques. *Technical report, Barracuda Networks*, 2004

[11] Mehran Sahami, Susan Dumais, David Heckerman and Eric Horvitz (1998). A Bayesian Approach to Filtering Junk E-Mail. *Proceedings of AAAI-98 Workshop on Learning for Text Categorization.*

[12] Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz (). "A bayesian approach to filtering junk email (mehran sahami, susan dumais, david heckerman, eric horvitz)".

[13] Newman, M. E. J. and Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A* 263, 341–346.

[14] S. J. Delany and P. Cunningham, 'An analysis of case-based editing in a spam filtering system', in *7th European Conference on Case-Based Reasoning (ECCBR*

2004), eds., P. Funk and P. Gonz´alez-Calero, volume 3155 of *LNAI*, pp. 128–141. Springer, (2004).

[15] OReilly.SpamAssassin.Jul.2004.eBook-DDU. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472

[16] Delany SJ, P Cunningham & B Smyth (2006) **ECUE: A Spam Filter that Uses Machine Learning to track Concept Drift**, In: *Proc of the 17th Eur. Conf. on Artificial Intelligence (PAIS stream),* p627-631.

[17] Delany SJ (2006) **Using Case-Based Reasoning for Spam Filtering**, PhD Thesis, March 2006]

[18] Bùi Ngọc Lan (2006). Lọc thư rác dựa trên tính chất của mạng xã hội. Khóa luận tốt nghiệp đại học. Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[19] Từ Minh Phương, Phạm Văn Cường, Nguyễn Duy Phương, Hoàng Trọng Huy (2006). Báo cáo đề tài "Nghiên cứu xây dựng hệ thống lọc thư rác có khả năng lọc thư rác tiếng Anh và tiếng Việt". Học viện Bưu chính Viễn thông, 2006.