Journal of Computer Science and Cybernetics, V.30, N.3 (2014), 203–215
DOI:10.15625/1813-9663/30/3/3253

ĐÁNH GIÁ MỘT SỐ KĨ THUẬT PHÁT HIỆN THƯ RÁC ỨNG DỤNG THUẬT TOÁN XẾP HẠNG NGƯỜI DÙNG TRONG MẠNG THƯ ĐIỆN TỬ TẠI TRƯỜNG ĐẠI HỌC HÀ NỘI

TRẦN QUANG ANH, VŨ MINH TUẤN, HÀ QUANG MINH

Khoa Công nghệ thông tin, Trường Đại học Hà Nội

anhtq@hanu.edu.vn; minhtuan_fit@hanu.edu.vn; minhhq_fit@hanu.edu.vn

Tóm tắt. Bài báo phân tích và kiểm nghiệm bốn phương pháp lọc thư rác dựa trên việc xếp hạng người dùng trong mạng thư điện tử: Phương pháp độ phân cụm, phương pháp độ phân cụm mở rộng, phương pháp sử dụng thuật toán PageRank và phương pháp sử dụng thuật toán PageRank có trọng số. Các thí nghiệm được thực hiện trên một số tập dữ liệu hoàn chỉnh của mạng thư điện tử Đại học Hà Nội. So sánh kết quả các thí nghiệm cho thấy, phương pháp sử dụng thuật toán PageRank và phương pháp độ phân cụm mở rộng mở rộng có kết quả tốt hơn các phương pháp còn lại. Tỷ lệ phát hiện thành công thư rác lên tới trên 99,5% trong khi tỷ lệ báo động nhầm thấp hơn 0,5%.

Từ khóa. phát hiện thư rác; mạng thư điện tử, phân cụm, thuật toán PageRank, xếp hạng người dùng.

Abstract. In this paper, four spam-filtering approaches based on user's ranking in the mail networks: Clustering, Extended Clustering Coefficient, PageRank Algorithm and Weighted PageRank Algorithm are analyzed. We also propose a couple of fully worked-out datasets from the email network of Hanoi University against which the experimental comparisons with the respect to the accuracy of email user ranking and spam filtering are conducted. The results indicate that PageRank algorithm and Extended Clustering Coefficient approaches are better than others. The rate of true detection is over 99.5% while the failed alarm remains below 0.5%.

Keywords. spam detection, email network, clustering, PageRank algorithm, user ranking.

1. MỞ ĐẦU

Trong những năm gần đây, ngăn chặn thư rác đã trở thành sứ mệnh toàn cầu trong lĩnh vực an ninh mạng. Các nhà nghiên cứu liên tục đề xuất các giải pháp lọc thư rác với nhiều phương pháp tiếp cận khác nhau [1] như: dựa trên tiêu đề, dựa trên nội dung, giao thức hay dựa trên tính xác thực của người gửi... Trong số rất nhiều cách thức đó, phương pháp sử dụng lý thuyết mạng phức hợp, cụ thể là mạng thư điện tử, đã từng bước khẳng định được tính ưu việt so với các phương pháp khác. Tuy nhiên, phương pháp này còn tương đối mới mẻ và chưa có những thí nghiệm chuyên sâu để đánh giá chính xác tiềm năng thực sự của nó. Chính vì thế, nhóm tác giả bài báo đã chọn bốn phương pháp lọc thư rác dựa việc xếp hạng người dùng trong mạng thư điện tử thực hiện các thí nghiệm nhằm: (1) đánh giá hiệu quả của cách tiếp cận dựa vào lý thuyết mạng phức hợp và (2) so sánh các phương pháp để tìm ra

đại diện hiệu quả nhất của cách tiếp cận này. Các phương pháp được thực hiện để triển khai thì nghiệm bao gồm: phương pháp độ phân cụm [2], phương pháp độ phân cụm mở rộng [1], phương pháp sử dụng thuật toán PageRank [3] và phương pháp sử dụng thuật toán PageRank có trọng số [4]. Kết quả của các thí nghiệm được so sánh để đánh giá tính chính xác của việc xếp hạng người dùng thư điện tử và ứng dụng kết quả đó để phát hiện phát hiện thư rác. Bên cạnh những đánh giá về các phương pháp, nhóm tác giả còn thực hiện một số tối ưu cho các thuật toán nhằm nâng cao hiệu quả thực thi và trình bày một số tập dữ liệu tương đối hoàn thiện phục vụ cho thí nghiệm.

Cấu trúc của bài báo được trình bày như sau: Phần II giới thiệu về bốn phương pháp phát hiện thư rác dựa trên lý thuyết mạng thư điện tử cùng vớinhững điểm mạnh, các giới hạn của các phương pháp này. Tiếp theo, phương pháp làm thí nghiệm và các tập dữ liệu mẫu được trình bày trong phần III. Phần IV phân tích và so sánh kết quả thu được từ các thì nghiệm; đồng thời, đưa ra các nhận xét, đánh giá về các phương pháp. Cuối cùng là phần Kết luận, tóm tắt lại vấn đề và đề cập đến hướng phát triển tiếp theo.

2. CÁC PHƯƠNG PHÁP LỌC THƯ RÁC DỰA TRÊN MẠNG THƯ ĐIỆN TỬ

2.1. Phương pháp độ phân cụm

P.O. Boykin và V. Roychowdhury [2] đã đề xuất một phương pháp phát hiện thư rác dựa trên độ phân cụm. Nhóm tác giả này thu thập thư điện tử từ hòm thư cá nhân để xây dựng một mạng thư điện tử mà trong đó, mỗi địa chỉ thư điện tử là một nút mạng, và liên kết giữa các nút mạng được coi là các cung. Theo mô hình này, việc trao đổi thông tin bằng thư điện tử giữa các nhóm người dùng được mô phỏng như một mạng xã hội (hay mạng thư điện tử). Dựa vào hai đặc tính chính của mạng thư điện tử (free-scale degree [5] và small-world degree [6]), độ phân cụm của nút i trong mạng thư điện tử được tính theo công thức sau:

$$C_i = \frac{2 * E_i}{k_i(k_i - 1)} \tag{1}$$

Trong đó, C_i là độ phân cụm; k_i số nút nối đến nút i; E_i là số cung giữa đỉnh liền kề với nút i. Nhóm tác giả phát hiện ra rằng độ phân cụm của nút i càng cao thì khả năng địa chỉ thư điện tử tương ứng với nút i là thư rác càng thấp. Đây là một phát hiện rất có ý nghĩa, tuy nhiên, việc tính độ phân cụm theo công thức () có một số hạn chế. Thứ nhất, công thức trên bỏ qua tất cả các đỉnh với với $k = 1$. Thứ hai, nghiêm trọng hơn, công thức này không phân biệt được các nút có chung giá trị $E_i = 0$ và khác giá trị k_i. Kết quả ghi nhận được qua thí nghiệm của nhóm tác giả, dựa trên thư điện tử từ hòm thư cá nhân, là 53% thư rác được phát hiện; 47% còn lại không đưa ra được kết quả.

2.2. Phương pháp độ phân cụm mở rộng

Để giải quyết những tồn tại của công thức tính độ phân cụm gốc (1), tác giả Bùi Ngọc Lan đã cải tiến công thức (1) của Boykin trong một nghiên cứu của mình [1] để đưa ra một công thức tính độ phân cụm mở rộng như sau:

$$C_i = \frac{2 * (E_i + 1)}{k_i(k_i - 1) + 1} \tag{2}$$

Tuy nhiên, để hướng tới mục đích tính toán độ tin cậy của người dùng, công thức (2) vẫn chưa thực sự thuyết phục. Thông thường, một người nhận được nhiều thư sẽ có mức độ đáng tin cậy cao hơn những người khác. Tuy nhiên, công thức (2) lại không phân biệt được trường hợp một người nhận được hàng loạt thư và một người gửi hàng loạt thư. Điều này làm cho việc tính toán sẽ có sự nhầm lẫn. Chính vì vậy, nhóm tác giả này đã đề xuất một công thức mới để tính độ phân nhóm:

$$C_i = \frac{2 * (E_i + 1)}{S_i(S_i - 1) + 1} + 0.2 * R_i \tag{3}$$

Trong đó, giá trị E_i vẫn giống như trông công thức (2.1). S_i là số nút mạng nhận được ít nhất một tin nhắn từ i; Ri là số nút mạng gửi đi ít nhất một tin nhắn tới i. Qua hai bước cải tiến, công thức (3) đã tính toán hiệu quả hơn giá trị độ phân cụm. Tuy nhiên, khi thực hiện thí nghiệm, nhóm tác giả này lại sử dụng tập dữ liệu rất giới hạn, thiếu nhiều thông tin về vai trò của người dùng và không bao gồm thư rác. Chính vì vậy mà tính chính xác của phương pháp này cần được xem xét kĩ hơn. Điều đó cho thấy vai trò quan trọng của một tập dữ liệu hoàn chỉnh phục vụ cho các thì nghiệm chuyên sâu sau này.

2.3. Phương pháp dựa trên thuật toán PageRank

Mạng WWW là một dạng của mạng phức hợp với các nút mạng là các trang web, cung là những đương liên kết từ trang này đến trang khác trong mạng. Năm 1998, Brin và Larry Page đã đề xuất thuật toán để xếp hạng các trang mang tên PageRank [3]. Ý tưởng của thuật toán là đánh giá một trang web được coi là quan trọng khi có nhiều trang web quan trọng liên kết đến nó, thể hiện qua công thức sau:

$$PR(A) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right) \tag{4}$$

Trong đó, gả sử trang web $T1$ đến Tn đều có liên kế tới trang web A; $C(A)$ là những liên kết từ trang web A ra ngoài. Chỉ số damping d được định nghĩa là xác suất người dùng chọn một liên kết trên trang mà họ đang xem. Theo như tính toán của tác giả, chỉ số damping d được đặt bằng 0.85.

Sự thành công của Google với thuật toán PageRank đã giúp thuật toán này trở nên nổi tiếng và được áp dụng trong nhiều ứng dụng xếp hạng khác nhau, trong đó có nghiên cứu của P.A. Chirita cùng các đồng nghiệp để xếp hạng thư điện tử nhằm phát hiện và ngăn chặn thư rác. Hệ thống xếp hạng thư điện tử của P.A. Chirita bao gồm máy chủ và máy trạm MailRank. Nguyên lý hoạt động của hệ thống là thu thập thông tin của người dùng và thực hiện những đánh giá dựa trên tập dữ liệu mẫu đó. Tuy rằng hệ thống có thể thu thập được thông tin của một số người dùng cụ thể nhưng không phải tất cả người dùng đều cài đặt máy trạm MailRank nên việc thu thập dữ liệu và đánh giá sẽ gặp khó khăn. Chính vì vậy, một lần nữa, ta lại thấy việc đánh giá thuật toán dựa trên một tập dữ liệu hoàn chỉnh từ máy chủ thư điện tử là rất cần thiết.

2.4. Phương pháp dựa trên thuật toán PageRank có trọng số

Năm 2004, hai tác giả Wenpu Xing và Ali Ghorbani đã đề xuất một phương pháp để cải tiến thuật toán xếp hạng trang PageRank, gọi là thuật toán PageRank có trọng số. Với

thuật toán cải tiến này, thay vì chia đều chỉ số xếp hạng của một trang web như thuật toán PageRank gốc (3), chỉ số này được chia cho các trang có số lượng liên kết tới (link-in) với trọng số khác nhau. Tác giả đưa ra hai giá trị $W^{in}_{(v,u)} = \frac{I_u}{\sum_{peR(v)} I_p}$ và $W^{out}_{(v,u)} = \frac{O_u}{\sum_{peR(v)} O_p}$. Trong đó, I_u và I_p lần lượt là số liên kết đến của trang web u và p; O_u và O_p là số liên kết ra đi của trang web u và p; $R(v)$ là tập hợp các trang web có liên kết từ trang web v. Công thức cải tiến được trình bày như sau:

$$PR(u) = (1-d) + d \sum_{veB(u)} PR(v) W^{int}_{(v,u)} W^{out}_{(v,u)} \tag{5}$$

Theo như những đánh giá của tác giả về dữ liệu web thì thuật toán PageRank có trọng số hoạt động hiệu quả hơn thuật toán PageRank gốc. Tuy nhiên, tác giả hoàn toàn không nhắc đến những đánh giá về ứng dụng của thuật toán này với tập dữ liệu thư rác. Chính vì thế, việc phân tích thuật toán PageRank có trọng số trên tập dữ liệu thư rác là rất có ý nghĩa.

3. THÍ NGHIỆM

3.1. Tập dữ liệu

a. Tập dữ liệu tacnghiep_1:

Đây là tập dữ liệu được thu thập từ hệ thống thư điện tử nội bộ của Trường Đại học Hà Nội trong thời gian từ 01/9/2008 đến 30/6/2009. Chúng tôi chọn thời gian này là vì đây là trọn một năm học tại Trường Đại học Hà Nội. Đồng thời, đây là giai đoạn giữa một nhiệm kỳ Hiệu trưởng, vì vậy ít có sự xáo trộn về các vị trí cán bộ chủ chốt trong Nhà trường. Tập này bao gồm 101 người dùng với 31 cán bộ chủ chốt (CBCC); số lượng thư điện tử là 14320. Tập được thu thập từ 01/9/2008 đến 30/6/2009 (tương đương quãng thời gian một năm học).Tập dữ liệu này hiện nay không có thư rác, vì vậy chưa thể sử dụng để nghiên cứu các thuật toán lọc thư rác mà chỉ dùng để tối ưu các tham số của các thuật toán và đánh giá tính chính xác của việc xếp hạng người dùng.

b. Tập dữ liệu tacnghiep_2:

Mặc dù tập dữ liệu tacnghiep_01 đã có thể sử dụng để nghiên cứu các thuật toán xếp hạng người dùng, tuy nhiên, trong tập dữ liệu tacnghiep_01 không chứa thư rác, vì vậy muốn có một tập dữ liệu có thể nghiên cứu các phương pháp lọc thư rác cần xây dựng một tập dữ liệu có cả thư rác [1] .

Trước hết, nhóm tác giả đã coi tập dữ liệu tacnghiep_01 là tập dữ liệu của các thư điện tử gửi nội bộ với nhau. Như vậy, vẫn cần những thư điện tử gửi từ trong ra và các thư điện tử gửi từ ngoài vào, trong đó có thư rác. Nhóm tác giả cần bổ sung thêm người gửi thư bình thường và người gửi thư rác từ bên ngoài vào. Đối với người gửi thư rác, nhóm thực hiện đề tài đã tạo ra 1000 địa chỉ gửi thư rác và thực hiện các cuộc gửi thư rác đến các địa chỉ ngẫu nhiên trong số những địa chỉ nội bộ trên. Quá trình thư rác được gửi đến mạng nội bộ theo thời gian tuân thủ quá trình phân bố Poisson. Số lượng thư rác được gửi đến trong khoảng

[1] Theo nhận định của GS. Commark (ĐH Toronto, trao đổi với nhóm tác giả bằng email), hiện tại không có một tập dữ liệu nào về thư rác mà vừa có thư đến, vừa có thư đi, vừa có thư rác trên mạng Internet.

thời gian từ 01/9/2008 đến 30/6/2009 là 32769, gấp khoảng 3 lần số lượng thư bình thường trong thời gian trên [2] . Trong quá trình tạo ra thư rác, nhóm tác giả đã tạo ra những thư gửi từ địa chỉ bên trong đến địa chỉ gửi thư rác với xác suất 0.001%.

Đối với thư bình thường gửi từ ngoài vào, nhóm tác giả sử dụng luôn những người dùng thật và thư thật mà không nằm trong nhóm người dùng của tập dữ liệu tacnghiep_01. Tập dữ liệu được thu thập từ 01/9/2008 đến 30/6/2009 (tương đương một năm học) bao gồm 101 người dùng nội bộ và 215 người dùng từ bên ngoài. Số lượng thư rác bổ sung vào tệp là 1000 thư; số thư điện tử nội bộ là 14320 thư; số lượng thư điện tử bình thường gửi ra ngoài và gửi từ ngoài vào trong là 7634 thư; số lượng thư rác gửi từ ngoài vào là 32769 thư.

3.2. Tối ưu tham số các thuật toán

a. Tối ưu hóa tham số PageRank và PageRank có trọng số

Trước khi có thể so sánh các phương pháp lọc thư rác dựa trên mạng phức hợp, nhóm tác giả đã thực hiện các thí nghiệm nhằm tối ưu hóa các tham số của thuật toán PageRank, trong đó bao gồm chỉ số damping và số chu kỳ tính toán điểm xếp hạng. Thuật toán PageRank sử dụng thuật toán điều chỉnh vòng lập (Iteration Algorithms) để tính các điểm xếp hạng của các nút mạng. Điểm xếp hạng của các nút mạng đầu tiên được thiết lập ở những giá trị nhất định, sau đó lần lượt được điều chỉnh dựa vào các phương trình thể hiện mối liên quan giữa mỗi điểm với các điểm còn lại. Như vậy số chu kỳ tính toán điểm xếp hạng ảnh hưởng đến khả năng hội tụ và thời gian tính toán của thuật toán. Ngoài ra, nhóm tác giả còn đặt một ngưỡng (threshold) để xác định có tồn tại một cung giữa hai người dùng hay không. Nếu số lượng thư gửi từ người dùng A đến người dùng B lớn hơn Threshold thì ta đặt một cung từ A đến B.

Nhóm tác giả đã sử dụng toàn bộ tập dữ liệu tacnghiep_01 để tối ưu hóa các tham số trên. Chúng tôi sử dụng 'time' trên linux để tính thời gian chạy.

Đầu tiên là thí nghiệm để tối ưu hóa chỉ số damping. Chúng tôi đặt giá trị ngưỡng Threshold = 3 và số chu kỳ tính toán bằng 2000 (theo kinh nghiệm). Chúng tôi thay đổi các giá trị của chỉ số damping từ 0 cho đến 1 và tính toán độ chính xác của thuật toán.

Độ chính xác của thuật toán có chiều hướng tăng lên khi ta tăng giá trị của chỉ số damping factor, trong khi tốc độ tính toán gần như không khác biệt nhau nhiều. Chỉ số damping thể hiện xác suất để một người dùng gửi thư cho một người trong nhóm những người bạn cũ (đã từng gửi thư). Theo kết quả trên, chúng tôi lựa chọn chỉ số damping = 0.9 cho các thí nghiệm tiếp theo.

Tiếp đến là thí nghiệm để tối ưu hóa chu kỳ tính toán của thuật toán PageRank. Chúng tôi đặt số chu kỳ tính toán bằng các giá trị từ 1 đến 5000, kết quả của độ chính xác của thuật toán và thời gian chạy của chương trình được ghi lại và đem so sánh.

Kết quả so sánh cho thấy thời gian tính toán tăng tuyến tính (tỷ lệ thuận) với số chu kỳ tính. Độ chính xác sau chu kỳ thứ 20 là không thay đổi. Chúng tôi chọn số chu kỳ tính toán bằng 100 cho các thí nghiệm về sau.

Các thí nghiệm tương tự được thực hiện với thuật toán PageRank có trọng số và thu được kết quả tương đồng.

[2] Theo kết quả phân tích số liệu thư điện tử HANU năm 2008, số lượng thư rác chiếm khoảng từ 70-80% tổng số thư.

b. Tối ưu hóa ngưỡng (threshold)

Mạng thư điện tử được xây dựng từ dữ liệu thư điện tử, trong đó người dùng là các nút mạng, còn giao dịch thư điện tử giữa những người dùng với nhau xác định các cung của mạng. Số lượng thư điện tử giữa hai người dùng lớn hơn số lượng (ngưỡng) nhất định sẽ xác định sự tồn tại của một cung giữa hai người dùng đó. Tất cả các phương pháp đều cần tối ưu hóa giá trị ngưỡng này. Tóm lại, threshold ở đây được hiểu là giá trị ngưỡng để hình thành một cung trong giữ hai người dùng trong mạng thư điện tử. Người đọc cần chú ý để phân biệt với một threshold khác được nhắc tới trong kết quả thí nghiệm, là ngưỡng của một thư điện tử bị coi là thư rác hay không.

Chúng tôi sử dụng phương pháp PageRank để tối ưu hóa giá trị ngưỡng. Với các giá trị ngưỡng tăng dần từ 1 đến 20, chúng tôi tính toán độ chính xác và thời gian chạy của chương trình và ghi lại kết quả để so sánh.

Đối chiều kết quả tại các ngưỡng, cho thấy tỷ lệ xếp hạng chính xác có xu hướng giảm khi ngưỡng tăng. Như vậy, việc đặt ngưỡng quá cao dẫn đến giảm lượng thông tin cần thiết dùng để xếp hạng người dùng. Tuy nhiên tốc độ xử lý tăng lên khi ngưỡng tăng lên. Lý do là khi ngưỡng tăng lên, lượng thông tin giảm đi, vì vậy tốc độ xử lý tăng lên. Cân nhắc giữa độ chính xác và tốc độ xử lý, chúng tôi chọn ngưỡng bằng 5 cho các thí nghiệm tiếp theo.

3.3. Thí nghiệm trên tập dữ liệu tacnghiep_1:

a. Dữ liệu trong thời gian 01 năm học:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/06/2009 (threshold để hình thành cung = 5). Độ chính xác và thời gian chạy của từng phương pháp được thể hiện ở Bảng 1.

Dữ liệu	Độ chính xác	Thời gian chạy
Độ phân cụm	0.5644	1.945
Độ phân cụm mở rộng	0.7624	1.993
PageRank	0.7624	0.635
PageRank có trọng số	0.7426	1.821

Bảng 1: Kết quả với dữ liệu 01 năm học của tacnghiep_01

Kết quả thí nghiệm cho thấy phương pháp độ phân cụm cho độ chính xác trong việc xếp hạng người dùng thấp nhất. Trên thực tế, độ chính xác bằng 0.5644 gần tương đương với độ chính xác khi ta sắp xếp ngẫu nhiên. Phương pháp độ phân cụm mở rộng, PageRank cho độ chính xác cao nhất. Phương pháp PageRank có trọng số có độ chính xác hơi thấp hơn phương pháp PageRank. Tốc độ xử lý của phương pháp PageRank là tốt nhất.

b. Dữ liệu trong thời gian 01 kỳ học:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 31/01/2009 (threshold để hình thành cung = 3). Độ chính xác và thời gian chạy của từng phương pháp được thể hiện ở Bảng 2.

Dữ liệu	Độ chính xác	Thời gian chạy
Độ phân cụm	0.5644	0.916
Độ phân cụm mở rộng	0.7030	0.943
PageRank	0.7228	0.574
PageRank có trọng số	0.7029	1.455

Bảng 2: Kết quả với dữ liệu 01 học kỳ của tacnghiep_01

Kết quả thu được trong Bảng 2 cũng gần giống với trong Bảng 1, tức là phương pháp độ phân cụm cho kết quả kém nhất. Phương pháp PageRank có trọng số cho kết quả kém hơn phương pháp PageRank. Phương pháp PageRank và phương pháp độ phân cụm mở rộng vẫn là hai phương pháp cho kết quả tốt nhất. Phương pháp PageRank có tốc độ xử lý nhanh nhất. Phương PageRank có trọng số có tốc độ xử lý chậm nhất.

c. Dữ liệu trong thời gian 03 tháng:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/11/2008 (threshold để hình thành cung = 1). Độ chính xác và thời gian chạy của từng phương pháp được thể hiện ở Bảng 3.

Dữ liệu	Độ chính xác	Thời gian chạy
Độ phân cụm	0.5644	0.619
Độ phân cụm mở rộng	0.7228	0.626
PageRank	0.7624	0.798
PageRank có trọng số	0.7030	3.195

Bảng 3: Kết quả với dữ liệu 03 tháng của tacnghiep_01

Đối với tập dữ liệu 3 tháng, phương pháp PageRank và phương pháp độ phân cụm mở rộng vẫn cho kết quả tốt nhất. Phương pháp PageRank có trọng số cho kết quả kém hơn, và

cuối cùng là phương pháp độ phân cụm mở rộng. Phương pháp PageRank có trọng số có tốc độ xử lý chậm nhất.

d. Dữ liệu trong thời gian 01 tháng:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/09/2008 (threshold để hình thành cung = 1). Đối với tập dữ liệu nhỏ, phương pháp độ phân cụm mở rộng cho kết quả tốt nhất với thời gian tính toán ít nhất. Phương pháp PageRank có độ chính xác thứ 2. Phương pháp độ phân cụm cho kết quả kém nhất.

e. Đánh giá:

Từ kết quả thí nghiệm trên, ta thấy phương pháp PageRank và phương pháp độ phân cụm mở rộng là hai phương pháp cho kết quả tốt nhất trong việc xếp hạng người dùng. Phương pháp PageRank có trọng số cho kết quả kém hơn và phương pháp độ phân cụm cho kết quả kém nhất.

Khi độ lớn của tập dữ liệu giảm dần từ 1 năm học đến 1 tháng, phương pháp PageRank cho độ chính xác giảm dần, trong khi tốc độ xử lý không thay đổi (có thể thấy rằng, tốc độ xử lý của phương pháp PageRank chủ yếu phụ thuộc vào số chu kỳ tính toán). Khi độ lớn của tập dữ liệu giảm dần, phương pháp độ phân cụm mở rộng cho tốc độ tính toán tăng dần, đồng thời độ chính xác cũng có chiều hướng tăng lên. Như vậy phương pháp PageRank thích hợp với những tập mẫu lớn, còn phương pháp độ phân cụm mở rộng thích hợp với những tập mẫu nhỏ.

3.4. Thí nghiệm trên tập dữ liệu tacnghiep_2:

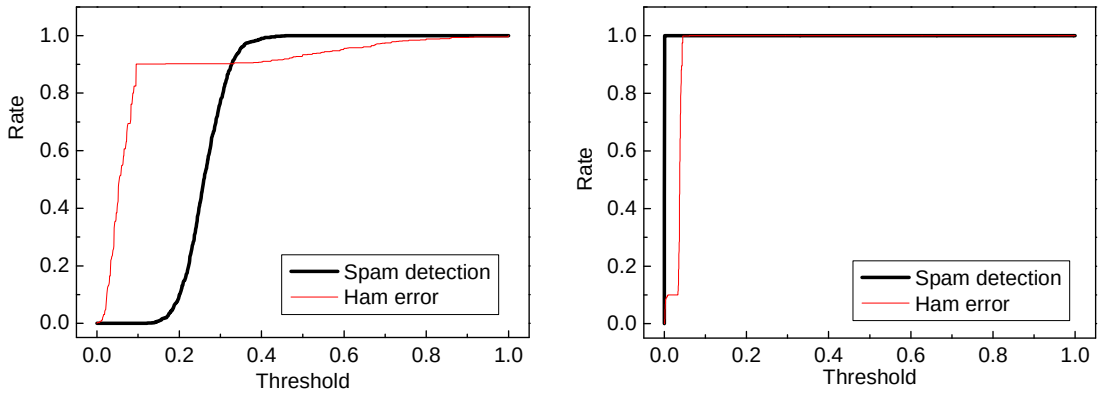
Đối với tập dữ liệu tacnghiep_2, sau khi đã được thêm vào một tỷ lệ thư rác nhất định, chúng tôi đánh giá hiệu quả của phương pháp dựa trên đồ thị ROC bao gồm 2 chỉ số: tỷ lệ lọc chính xác thư rác và tỷ lệ lọc nhầm thư thật.

a. Dữ liệu trong thời gian 01 năm học:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/06/2009 (threshold để hình thành cung = 5).

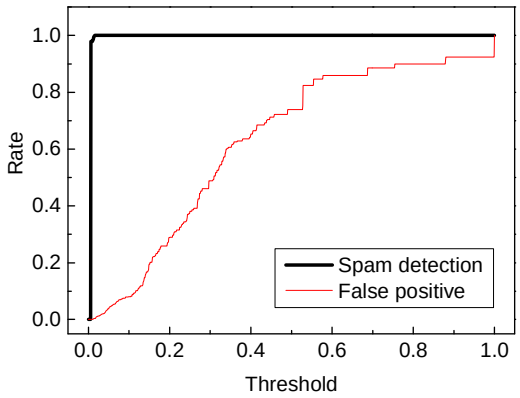
Trong Hình 1, ở mọi trường hợp ta đều có khi tăng giá trị ngưỡng, tỷ lệ lọc thành công thư rác tăng lên đồng thời tỷ lệ lọc nhầm thư thật cũng tăng theo. Phương pháp độ phân cụm cho kết quả kém (Hình 1-a) vì khi tỷ lệ lọc thành công thư rác đạt trên 90% thì tỷ lệ lọc nhầm thư thật cũng trên 80%. Các phương pháp còn lại đều cho kết quả khá tốt khi giá trị thereshold nhỏ, tỷ lệ lọc thành công thư rác là xấp xỉ 100%, trong khi tỷ lệ lọc nhầm thư thật là rất nhỏ. Chúng tôi lựa chọn các giá trị ngưỡng cho tỷ lệ lọc nhầm thư thật nhỏ hơn 0.5% và tỷ lệ lọc thành công thư rác lớn nhất.

Theo Bảng 4, phương pháp PageRank có tốc độ tính toán nhanh nhất. Phương pháp PageRank có trọng số có tốc độ tính toán chậm nhất. Về hiệu quả, phương pháp độ phân cụm mở rộng cho kết quả tốt nhất với tỷ lệ lọc thành thư rác là 100% trong khi tỷ lệ lọc nhầm thư thật là 0%.

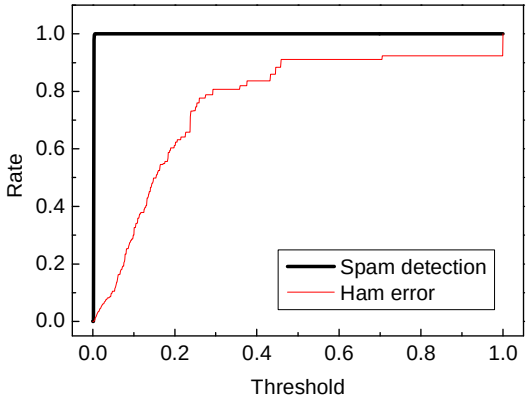


(a) Độ phân cụm

(b) Độ phân cụm mở rộng



(c) PageRank



(d) PageRank có trọng số

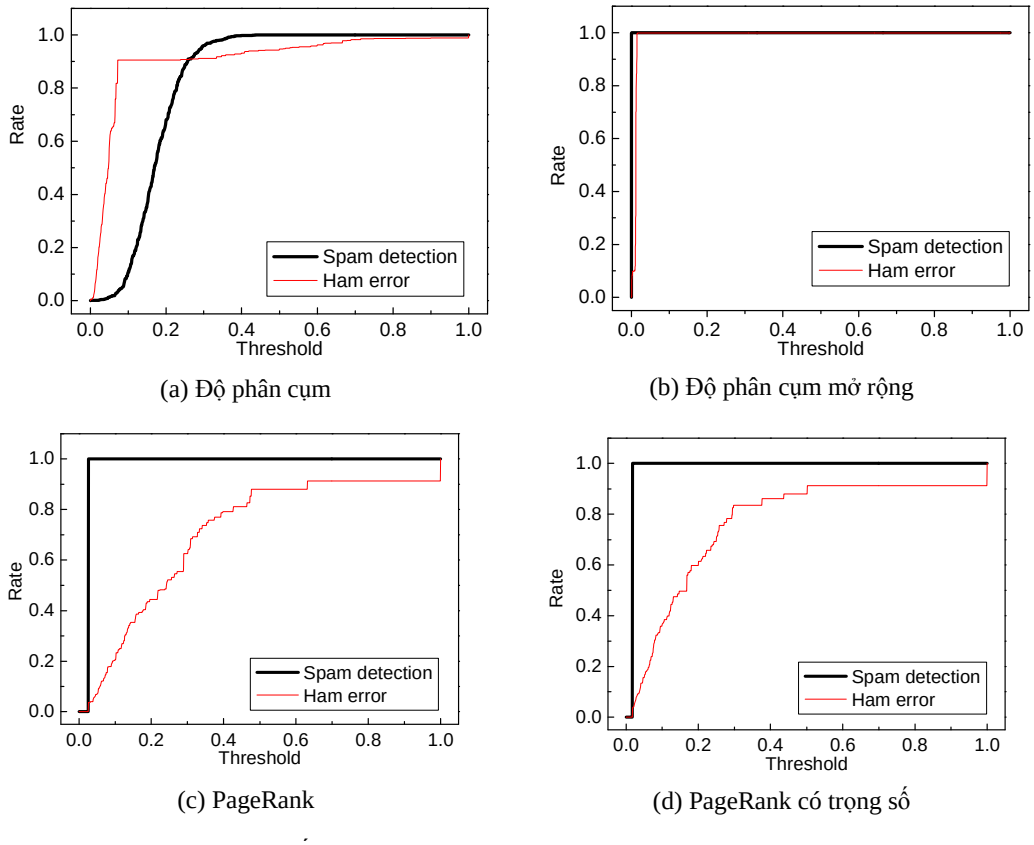
Hình 1: Kết quả so sánh với dữ liệu 01 năm của tacnghiep_02

Phương pháp	Ngưỡng (Spam hay Ham)	Tỷ lệ lọc thành công thư rác	Tỷ lệ lọc nhầm thư thật	Thời gian tính
Độ phân cụm	0.006	0	0.003	1m38.097s
Độ phân cụm mở rộng	0.005	1.0	0.0	1m37.366s
PageRank	0.016	1.0	0.0042	18.084s
PageRank có trọng số	0.004	0.9965	0.0060	3m58.702s

Bảng 4: Kết quả với dữ liệu 01 năm học của tacnghiep_02

b. Dữ liệu trong thời gian 01 kỳ học:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 31/01/2009 (threshold để hình thành cung = 3).



(a) Độ phân cụm (b) Độ phân cụm mở rộng

(c) PageRank (d) PageRank có trọng số

Hình 2: Kết quả so sánh với dữ liệu 01 học kỳ của tacnghiep_02

Phương pháp độ phân cụm vẫn cho kết quả kém (Hình 2-a) vì khi tỷ lệ lọc thành công thư rác đạt trên 90% thì tỷ lệ lọc nhầm thư thật cũng trên 80%. Các phương pháp còn lại đều cho kết quả khá tốt khi giá trị thereshold nhỏ, tỷ lệ lọc thành công thư rác là xấp xỉ 100%, trong khi tỷ lệ lọc nhầm thư thật là rất nhỏ. Chúng tôi lựa chọn các giá trị ngưỡng cho tỷ lệ lọc nhầm thư thật nhỏ hơn 0.5% và tỷ lệ lọc thành công thư rác lớn nhất. Về hiệu quả, phương pháp độ phân cụm mở rộng cho kết quả tốt nhất với tỷ lệ lọc thành thư rác là 100% trong khi tỷ lệ lọc nhầm thư thật chỉ có 0.13%. Phương pháp PageRank có tốc độ tính toán nhanh nhất.

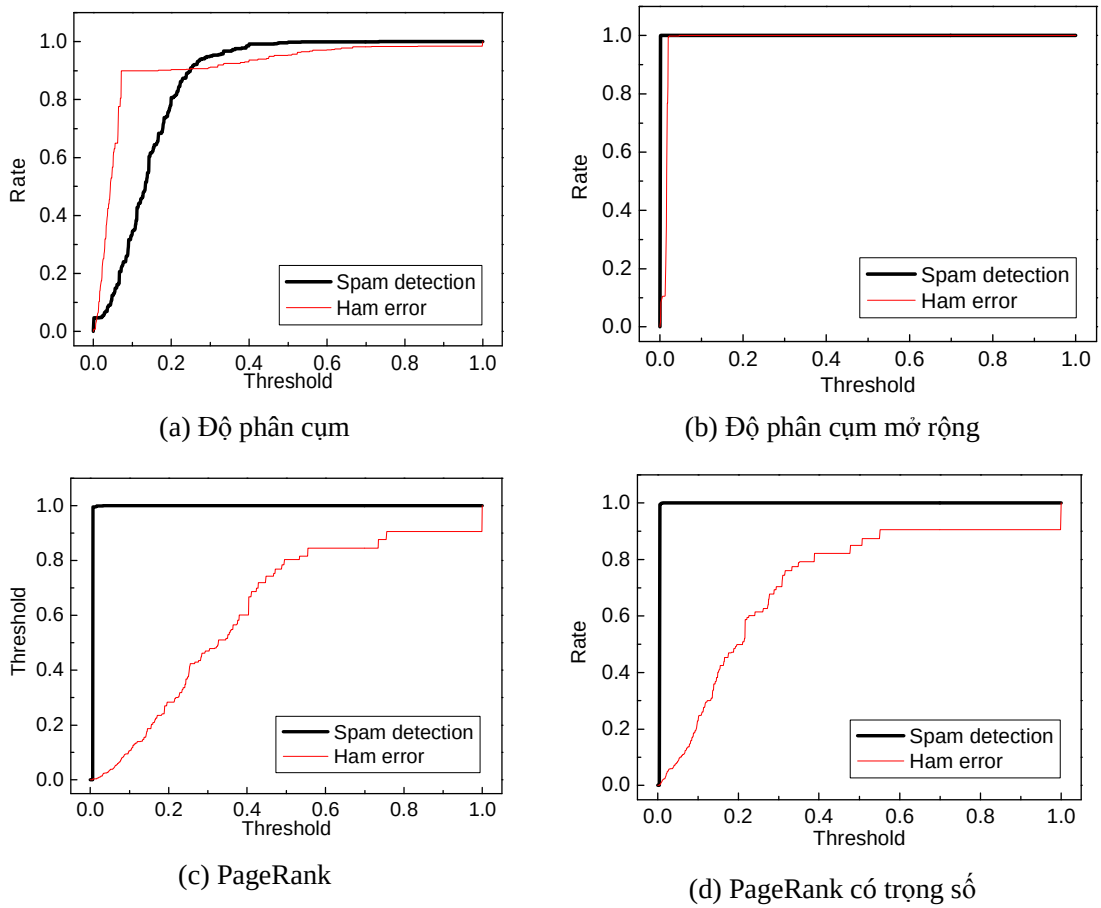
c. Dữ liệu trong thời gian 03 tháng:

Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/11/2008 (threshold để hình thành cung = 1).

Trong Hình 3, phương pháp độ phân cụm vẫn cho kết quả kém (Hình 3-a) vì khi tỷ lệ lọc thành công thư rác đạt trên 90% thì tỷ lệ lọc nhầm thư thật cũng trên 80%. Các phương pháp còn lại đều cho kết quả khá tốt khi giá trị thereshold nhỏ, tỷ lệ lọc thành công thư rác là xấp xỉ 100%, trong khi tỷ lệ lọc nhầm thư thật là rất nhỏ. Chúng tôi lựa chọn các giá trị ngưỡng cho tỷ lệ lọc nhầm thư thật nhỏ hơn 0.5% và tỷ lệ lọc thành công thư rác lớn nhất.

Về hiệu quả, phương pháp PageRank cho kết quả tốt nhất với tỷ lệ lọc thành thư rác là 99.8% trong khi tỷ lệ lọc nhầm thư thật chỉ có 0.35%. Phương pháp PageRank cũng có tốc

độ tính toán nhanh nhất.



(a) Độ phân cụm

(b) Độ phân cụm mở rộng

(c) PageRank

(d) PageRank có trọng số

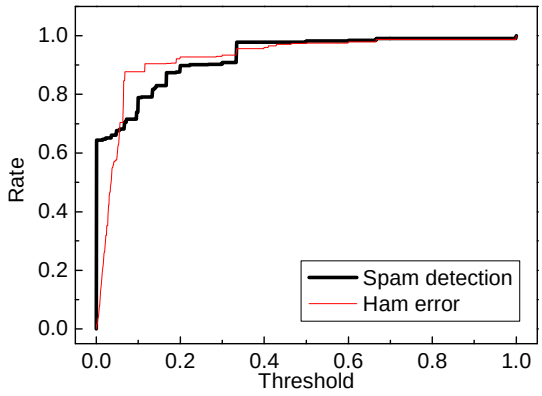
Hình 3: Kết quả so sánh với dữ liệu 03 tháng của tacnghiep_02

d. Dữ liệu trong thời gian 01 tháng:

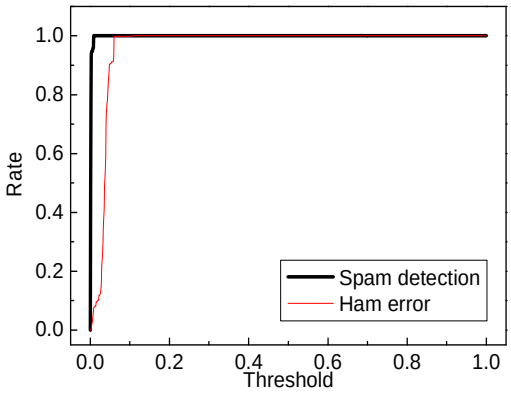
Chúng tôi sử dụng các tham số đã được tối ưu như bên trên. Thời gian dữ liệu từ 01/09/2008 – 30/09/2008 (threshold để hình thành cung = 1).

Phương pháp độ phân cụm vẫn cho kết quả kém (Hình 4-a), tuy nhiên hiệu quả lọc tốt hơn so với những thí nghiệm trước khi tỷ lệ lọc thành công thư rác đạt 64.4% thì tỷ lệ lọc nhầm thư thật là 1.68%. Các phương pháp còn lại đều cho kết quả khá tốt khi giá trị thereshold nhỏ, tỷ lệ lọc thành công thư rác là xấp xỉ 100%, trong khi tỷ lệ lọc nhầm thư thật là rất nhỏ. Chúng tôi lựa chọn các giá trị ngưỡng cho tỷ lệ lọc nhầm thư thật nhỏ hơn 0.5% và tỷ lệ lọc thành công thư rác lớn nhất.

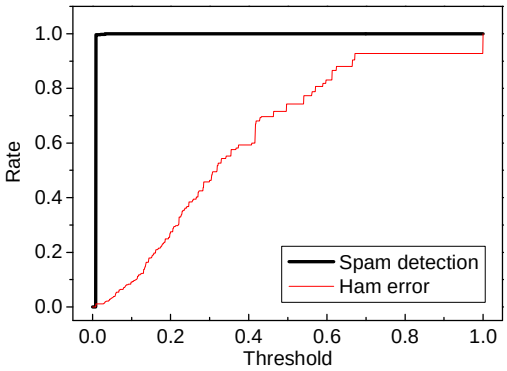
Về hiệu quả, phương pháp độ phân cụm mở rộng cho kết quả tốt nhất với tỷ lệ lọc thành thư rác là 93.82% trong khi tỷ lệ lọc nhầm thư thật chỉ có 0.5%. Phương pháp PageRank có tốc độ tính toán nhanh nhất.



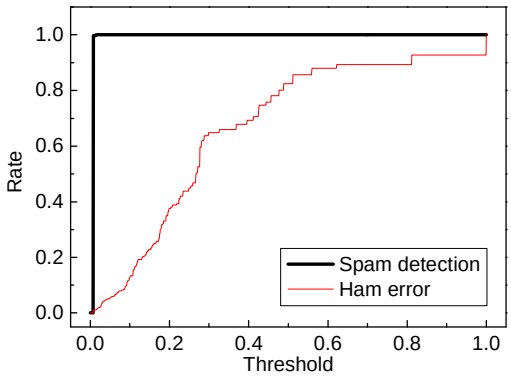
(a) Độ phân cụm



(b) Độ phân cụm mở rộng



(c) PageRank



(d) PageRank có trọng số

Hình 4: Kết quả so sánh với dữ liệu 01 tháng của tacnghiep_02

4. KẾT LUẬN

Phát hiện thư rác và xếp hạng người sử dụng thử điện tử dựa trên thuộc tính của mạng phức hợp (đại diện là mạng thử điện tử) là một phương pháp tiếp cận khá thuyết phục và có nhiều tiềm năng. Phương pháp này đã loại bỏ những hạn chế mà nhưng phương pháp mắc phải. Tuy nhiên, để đánh giá chính xác những điểm mạnh và tồn tại của phương pháp này, chúng ta cần đến một tập dữ liệu hoàn chỉnh với đầy đủ các yếu tố như các trao đổi nội bộ, thư đến và thư đi cũng như thư rác. Nhóm tác giả bài báo này đã đề xuất một tệp dữ liệu như vậy.

Với một tệp dữ liệu tương đối đầy đủ, một loạt các thí nghiệm đã được thực hiện để so sánh tính hiệu quả và thời gian thực thi của bốn phương pháp dựa trên thuộc tính của mạng thư điện tử để lọc thư rác và xếp hạng người dùng. Việc phân tích và so sánh kết quả của các thí nghiệm mang lại rất nhiều ý nghĩa.

TÀI LIỆU THAM KHẢO

[1] N. L. Bui, Q. A. Tran., Q. T. Ha, "User's authentic rating based on email networks," *The First International Conference on Mobile Computing, Communications and Applications (ICMOCCA 2006)*, pp 144-148

[2] P. O. Boykin and V. Roychowdhury, "Leveraging social networks to fight spam", *IEEE Computer*, vol. 38, no. (4), pp. :61-68, 2005; "Sorting e-mail friends from foes", *Nature News*, 19 Feb. 2004

[3] S. Brin, L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Proceedings of the 7th international conference on World Wide Web (WWW)*, Brisbane, Australia, pp. 107–117, 1998

[4] W. Xing, A. Ghorbani, "Weighted PageRank Algorithm", *Proceedings of the Second Annual Conference on Communication Networks and Services Research*, pp. 305 – 314, 2004

[5] H. Ebel, L-I. Mielsch and S. Bornholdt, "Scale-free topology of email networks", *Phys. Rev. E*, vol. 66, Article Id. 035103 (R), Sept., 2002

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

Ngày nhận bài 02 – 9 – 2013
Nhận lại sau sửa 30 – 7 – 2014