BỘ KHOA HỌC VÀ CÔNG NGHỆ BỘ CÔNG AN

CHƯƠNG TRÌNH KHCN CẤP NHÀ NƯỚC KC.01/06-10

BÁO CÁO TỔNG HỢP

KẾT QUẢ KHOA HỌC CÔNG NGHỆ ĐỀ TÀI

NGHIÊN CỨU, PHÁT TRIỂN HỆ THỐNG LỌC NỘI DUNG HỖ TRỢ QUẢN LÝ VÀ ĐẢM BẢO AN TOÀN – AN NINH THÔNG TIN TRÊN MẠNG INTERNET

MÃ SỐ ĐỀ TÀI: KC.01.02/06-10

Cơ quan chủ trì đề tài: Cục Công nghệ tin học nghiệp vụ,
Tổng cục Kỹ thuật - Bộ Công An

Chủ nhiệm đề tài: Thiếu tướng, TS. Nguyễn Viết Thế

8195

Hà Nội - 2009

BỘ KHOA HỌC VÀ CÔNG NGHỆ BỘ CÔNG AN

CHƯƠNG TRÌNH KHCN CẤP NHÀ NƯỚC KC.01/06-10

BÁO CÁO TỔNG HỢP

KẾT QUẢ KHOA HỌC CÔNG NGHỆ ĐỀ TÀI

NGHIÊN CỨU, PHÁT TRIỂN HỆ THỐNG LỌC NỘI DUNG HỖ TRỢ QUẢN LÝ VÀ ĐẢM BẢO AN TOÀN – AN NINH THÔNG TIN TRÊN MẠNG INTERNET

MÃ SỐ ĐỀ TÀI: KC.01.02/06-10

Chủ nhiệm đề tài/dự án: Cơ quan chủ trì đề tài/dự án:

(ký tên) *(ký tên và đóng dấu)*

Thiếu tướng, **TS. Nguyễn Viết Thế** **Đại tá Nguyễn Văn Thủy**

Ban chủ nhiệm chương trình Bộ Khoa học và Công nghệ

(ký tên) *(ký tên và đóng dấu khi gửi lưu trữ)*

Hà Nội - 2009

TỔNG CỤC KỸ THUẬT
CỤC CÔNG NGHỆ TIN HỌC NGHIỆP VỤ

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày tháng năm 200...

BÁO CÁO THỐNG KÊ
KẾT QUẢ THỰC HIỆN ĐỀ TÀI

I. THÔNG TIN CHUNG

1. Tên đề tài: Nghiên cứu, phát triển hệ thống lọc nội dung hỗ trợ quản lý và đảm bảo an toàn – an ninh thông tin trên mạng Internet.

Mã số đề tài: KC.01.02/06-10

Thuộc: Chương trình khoa học và công nghệ trọng điểm cấp Nhà nước giai đoạn 2006-1010 "Nghiên cứu, phát triển và ứng dụng công nghệ thông tin và truyền thông ", mã số KC.01/06-10

2. Chủ nhiệm đề tài/dự án:

Họ và tên: Nguyễn Viết Thế

Ngày, tháng, năm sinh: 1951 Nam/ Nữ: Nam

Học hàm, học vị: Tiến sỹ

Chức danh khoa học:

Chức vụ: Cục Trưởng

Điện thoại: Tổ chức: 06947801; Nhà riêng: 06942624;

Mobile: 0913239801

Fax: 04.7537.7997; E-mail: the_nv52@yahoo.com

Tên tổ chức đang công tác: Cục Công nghệ Tin học nghiệp vụ - Tổng cục Kỹ Thuật - Bộ Công an

Địa chỉ tổ chức: 80 Trần Quốc Hoàn, Cầu Giấy, Hà Nội

Địa chỉ nhà riêng: Số 10-A12 Đầm Trấu, Phường Bạch Đằng, Quận Hai Bà Trưng, Hà Nội

3. Tổ chức chủ trì đề tài/dự án:

Tên tổ chức chủ trì đề tài: Cục Công nghệ Tin học nghiệp vụ - Tổng cục Kỹ Thuật - Bộ Công an

Điện thoại: 069.47801

Fax: .....

E-mail: .....

Website: www.e15.bca

Địa chỉ: 80 Trần Quốc Hoàn, Cầu Giấy, Hà Nội

Họ và tên thủ trưởng tổ chức: Nguyễn Viết Thế

Số tài khoản: .....

Ngân hàng: .....

Tên cơ quan chủ quản đề tài: Tổng cục Kỹ Thuật - Bộ Công an

II. TÌNH HÌNH THỰC HIỆN

1. Thời gian thực hiện đề tài/dự án:

- Theo Hợp đồng đã ký kết: từ tháng 04/2007 đến tháng 4/ 2009
- Thực tế thực hiện: từ tháng 04/2007 đến tháng 10/2009
- Được gia hạn (nếu có):
  - Lần 1 từ tháng 04/2007 đến 10/2009

2. Kinh phí và sử dụng kinh phí:

a) Tổng số kinh phí thực hiện: 2000 tr.đ, trong đó:

+ Kính phí hỗ trợ từ SNKH: 2000 tr.đ.

+ Kinh phí từ các nguồn khác: không

b) Tình hình cấp và sử dụng kinh phí từ nguồn SNKH:

Số TT	***Theo kế hoạch***		***Thực tế đạt được***		***Ghi chú*** *(Số đề nghị quyết toán)*
	Thời gian (Tháng, năm)	Kinh phí (Tr.đ)	Thời gian (Tháng, năm)	Kinh phí (Tr.đ)	
1	2007	740			
2	2008				
3	2009				

c) Kết quả sử dụng kinh phí theo các khoản chi:

Đối với đề tài:

Đơn vị tính: Triệu đồng

Số TT	***Nội dung các khoản chi***	***Theo kế hoạch***			***Thực tế đạt được***		
		Tổng	SNKH	Nguồn khác	Tổng	SNKH	Nguồn khác
1	Trả công lao động (khoa học, phổ thông)	**1361**	**1361**				
2	Nguyên, vật liệu, năng lượng	**90**	**90**				
3	Thiết bị, máy móc	**300**	**300**				
4	Xây dựng, sửa chữa nhỏ						
5	Chi khác	**249**	**249**				
	Tổng cộng	**2000**	**2000**				

- Lý do thay đổi (nếu có):

3. Các văn bản hành chính trong quá trình thực hiện đề tài/dự án:

(Liệt kê các quyết định, văn bản của cơ quan quản lý từ công đoạn xác định nhiệm vụ, xét chọn, phê duyệt kinh phí, hợp đồng, điều chỉnh (thời gian, nội dung, kinh phí thực hiện... nếu có); văn bản của tổ chức chủ trì đề tài, dự án (đơn, kiến nghị điều chỉnh ... nếu có)

Số TT	***Số, thời gian ban hành văn bản***	***Tên văn bản***	***Ghi chú***
1	22/9/2006	Quyết định số 2089/QĐ-BKHCN ngày 22 tháng 9 năm 2006 của Bộ trưởng Bộ khoa học và Công nghệ về việc phê duyệt nội dung và kinh phí các đề tài đã trúng tuyển thuộc Chương trình khoa học và công nghệ trọng điểm cấp Nhà nước giai đoạn 2006-2010, mã số KC.01.02/06-10	
2	14/5/2007	Hợp đồng "*Nghiên cứu, phát triển hệ thống lọc nội dung hỗ trợ quản lý và đảm bảo an toàn – an ninh thông tin trên mạng Internet*", mã số	

		KC.01.02/06-10 thuộc Chương trình KC.01/06-10 theo các nội dung trong Thuyết minh đề tài	
	20/7/2007	Quyết định số 1488/QĐ-BKHCN ngày 20/7 năm 2007 về việc điều chỉnh thời gian thực hiện của các đề tài, dự án thuộc Chương trình Khoa học và Công nghệ trọng điểm cấp Nhà nước giai đoạn 2006-2010 bắt đầu thực hiện năm 2006	
	14/9/2007	Quyết định số 1942/QĐ-BKHCN ngày 14/9/2007 về việc cử đoàn đi công tác nước ngoài	
	30/10/2008	Công văn số 696/E15 ngày 30/10/2008 của Cục Công nghệ Tin học nghiệp vụ về việc đề xuất kế hoạch đấu thầu mua thiết bị năm 2008 của đề tài KC.01.02/06-10	
	24/11/2008	Quyết định số 2597/QĐ-BKHCN ngày 24/11/2008 của Bộ trưởng Bộ Khoa học và Công nghệ về việc phê duyệt kế hoạch đấu thầu mua sắm tài sản đề tài "Nghiên cứu, phát triển hệ thống lọc nội dụng hỗ trợ quản lý và đảm bảo an toàn - an ninh thông tin trên mạng Internet", mã số KC.01.02/06-10	
	27/03/2009	Công văn số 145/E15(P4) ngày 27/03/2009 của Cục Công nghệ Tin học nghiệp vụ về việc xin gia hạn thời gian thực hiện đề tài	
	29/04/2009	Quyết định số 720/QĐ-BKHCN ngày 29/4/2009 của	

		Bộ trưởng Bộ Khoa học và Công nghệ về việc điều chỉnh thời gian thực hiện của đề tài KC.01.02/06-10 thuộc chương trình KH&CN trọng điểm cấp Nhà nước giai đoạn 2006-2010 "Nghiên cứu, phát triển và ứng dụng công nghệ thông tin và truyền thông", mã số KC.01/06-10	

4. Tổ chức phối hợp thực hiện đề tài, dự án:

Số TT	***Tên tổ chức đăng ký theo Thuyết minh***	***Tên tổ chức đã tham gia thực hiện***	***Nội dung tham gia chủ yếu***	***Sản phẩm chủ yếu đạt được***	***Ghi chú****

1	Trường Đại học Công nghệ - Đại học Quốc Gia Hà Nội	Trường Đại học Công nghệ - Đại học Quốc Gia Hà Nội	- Nghiên cứu, phân tích tình hình quản lý Nhà nước về lọc nội dung trên thế giới và các chính sách pháp lý liên quan. - Tìm hiểu, phân tích thực trạng công nghệ lọc Internet theo nội dung trên thế giới theo cả chiều rộng và chiều sâu. - Nghiên cứu đề xuất giải pháp lọc nội dung Internet - Xây dựng, kiến trúc hạ tầng và phát triển các mođun thành phần cơ bản của hệ thống lọc		

2	Công ty điện toán và truyền số liệu, VDC, Tổng Công ty Bưu Chính Viễn Thông Việt Nam	Công ty điện toán và truyền số liệu, VDC, Tổng Công ty Bưu Chính Viễn Thông Việt Nam	- Phân tích, khảo sát các công cụ, kỹ thuật quản lý và giám sát các luồng dữ liệu vào/ra tại một cổng Internet quốc gia. - Phân tích, khảo sát các công cụ, kỹ thuật quản lý và giám sát các luồng dữ liệu vào/ra tại một cổng Internet quốc gia. - Xây dựng hệ thống lọc nội dung Internet tại máy tính cá nhân		

- Lý do thay đổi (nếu có):

5. Cá nhân tham gia thực hiện đề tài, dự án:

(Người tham gia thực hiện đề tài thuộc tổ chức chủ trì và cơ quan phối hợp, không quá 10 người kể cả chủ nhiệm)

Số TT	***Tên cá nhân đăng ký theo Thuyết minh***	***Tên cá nhân đã tham gia thực hiện***	***Nội dung tham gia chính***	***Sản phẩm chủ yếu đạt được***	***Ghi chú****
1	Nguyễn Viết Thế	Nguyễn Viết Thế			Chủ nhiệm đề tài
2	Trần Văn Cầm	Trần Văn Cầm			Thư ký đề tài
3	Lê Văn Toàn	Lê Văn Toàn			
4	Nguyễn Thế Bình	Nguyễn Thế Bình			
5	Hà Quang Thụy	Hà Quang Thụy			
6	Trịnh Nhật Tiến	Trịnh Nhật Tiến			

7	Nguyễn Ngọc Hóa	Nguyễn Ngọc Hóa			
8	Trần Việt Hưng	Trần Việt Hưng			
9	Phạm Anh Chiến	Phạm Anh Chiến			
10	Đỗ Hùng	La Thế Hưng			

- Lý do thay đổi (nếu có):

6. Tình hình hợp tác quốc tế:

Số TT	***Theo kế hoạch*** *(Nội dung, thời gian, kinh phí, địa điểm, tên tổ chức hợp tác, số đoàn, số lượng người tham gia...)*	***Thực tế đạt được*** *(Nội dung, thời gian, kinh phí, địa điểm, tên tổ chức hợp tác, số đoàn, số lượng người tham gia...)*	***Ghi chú****
1	Khảo sát, trao đổi khoa học và tìm hiểu công nghệ kiểm soát Internet tại Trung Quốc	- Khảo sát, trao đổi khoa học và tìm hiểu công nghệ kiểm soát Internet tại Đại học Thanh Hoa - Bắc Kinh và trung tâm kiểm soát mạng thành viên CERNET ở Thượng Hải Trung Quốc từ 22/1/2008 đến 28/1/2008. - Số lượng đoàn, người tham gia: 01 đoàn 6 người	
2			
...			

- Lý do thay đổi (nếu có):

7. Tình hình tổ chức hội thảo, hội nghị:

Số TT	***Theo kế hoạch*** *(Nội dung, thời gian, kinh phí, địa điểm)*	***Thực tế đạt được*** *(Nội dung, thời gian, kinh phí, địa điểm)*	***Ghi chú****
1	Tổ chức hội thảo báo cáo kết quả nghiên cứu	Hội thảo Báo cáo kết quả thực hiện đề tài tổ chức vào ngày 16/09/2009 tại Cục E15 - Bộ Công an	
2			
...			

- Lý do thay đổi (nếu có):

8. Tóm tắt các nội dung, công việc chủ yếu:

(Nêu tại mục 15 của thuyết minh, không bao gồm: Hội thảo khoa học, điều tra khảo sát trong nước và nước ngoài)

Số TT	***Các nội dung, công việc chủ yếu*** *(Các mốc đánh giá chủ yếu)*	***Thời gian*** *(Bắt đầu, kết thúc - tháng ... năm)*		***Người, cơ quan thực hiện***
		Theo kế hoạch	Thực tế đạt được	
1	Nghiên cứu, phân tích và đánh giá tình hình lọc nội dung trên Internet trong nước và trên thế giới	2007	2007	
2	Nghiên cứu, phân tích và đề xuất giải pháp lọc nội dung trên Internet hỗ trợ quản lý và bảo đảm an toàn-an ninh thông tin	2007	2007	
3	Xây dựng, thiết kế kiến trúc hạ tầng hệ thống lọc nội dung trên Internet hỗ trợ quản lý và đảm bảo an toàn-an ninh	2007	2007	
4	Xây dựng, phát triển các thành phần cơ bản trong hệ thống lọc nội dung Internet	2008	2008	
5	Xây dựng, phát triển mô đun lọc văn bản tiếng Việt	2008	2008	
6	Xây dựng, phát triển mô đun lọc văn bản tiếng Anh	2008	2008	
7	Xây dựng, phát triển mô đun	2008	2008	

	lọc hình ảnh			
8	Xây dựng, phát triển mô đun lọc theo URL và chuẩn PICS	2008	2008	
9	Tích hợp các thành phần trong kiến trúc và xây dựng hệ thống lọc nội dung Internet	2008- 2009	2008-2009	
10	Xây dựng hệ thống lọc nội dung Internet tại máy tính cá nhân	2009	2009	

- Lý do thay đổi (nếu có):

III. SẢN PHẨM KH&CN CỦA ĐỀ TÀI, DỰ ÁN

1. Sản phẩm KH&CN đã tạo ra:

a) Sản phẩm Dạng I: *Không đăng ký*

Số TT	***Tên sản phẩm và chỉ tiêu chất lượng chủ yếu***	***Đơn vị đo***	***Số lượng***	***Theo kế hoạch***	***Thực tế đạt được***
1					
2					
...					

- Lý do thay đổi (nếu có):

b) Sản phẩm Dạng II:

Số TT	*Tên sản phẩm*	*Yêu cầu khoa học cần đạt*		*Ghi chú*
		Theo kế hoạch	*Thực tế đạt được*	
1	Báo cáo nghiên cứu	10	10	Nội dung cập nhật các nghiên cứu quốc tế, trong nước
	1. Tài liệu phân tích và đánh giá tình hình quản lý Nhà nước về lọc nội dung trên thế giới (Mỹ, Trung Quốc, Châu Âu, Singapore, …) 2. Tài liệu phân tích và đề xuất chính sách pháp lý tại Việt nam cho vấn đề lọc nội dung thông tin trên mạng Internet. 3. Tài liệu đánh giá tổng quan thực trạng lọc nội dung Internet trên thế giới. 4. Tài liệu đánh giá các thuật toán lọc văn bản theo nội dung (SVM, Neural, Semi-Supervised...). 5. Tài liệu phân tích và đánh giá các giải thuật lọc ảnh (theo màu sắc, text, hình dạng ảnh, …) 6. Tài liệu đánh giá các giải thuật lọc dựa URL, links và chuẩn PICS. 7. Tài liệu khảo sát hạ tầng kỹ thuật tại các cổng Internet quốc gia. 8. Tài liệu nghiên cứu, tìm hiểu và đánh giá các kỹ thuật cho phép quản lý các luồng dữ liệu vào/ra tại một cổng Internet quốc gia. 9. Tài liệu giải pháp lọc nội dung Internet nhằm hỗ trợ quản lý và bảo đảm an toàn-an ninh thông tin. 10. Tài liệu nghiên cứu các đặc trưng của tiếng Việt liên quan đến lọc theo nội dung			
2	Báo cáo giải pháp	5	5	Có phân tích để lựa chọn giải pháp phù hợp với sự tiếp thu các công nghệ tiên tiến
	1. Tài liệu nghiên cứu, thiết kế và xây dựng mô đun chuẩn hoá dữ liệu 2. Tài liệu giải pháp xác định tự động nội dung văn bản tiếng Việt 3. Tài liệu giải pháp lọc văn bản tiếng Anh 4. Tài liệu giải pháp lọc URL và PICS 5. Tài liệu nghiên cứu, đề xuất giải pháp đánh giá hiệu năng bộ lọc Web 6. Tài liệu nghiên cứu, đề xuất giải pháp đánh giá hiệu năng bộ lọc Mail			
3.	Tài liệu thiết kế	13	13	Đảm bảo tính phục tùng các giải pháp đã được lựa chọn
	1. Tài liệu thiết kế bộ lọc Web 2. Tài liệu thiết kế bộ lọc Mail 3. Tài liệu thiết kế chi tiết các thành phần cơ bản của kiến trúc hạ tầng cho toàn bộ hệ thống lọc nội dung 4. Tài liệu thiết kế mô đun kiểm soát các mô đun khác trong kiến trúc hệ thống. 5. Tài liệu thiết kế mô đun ra quyết định xác định chính sách xử lý với từng loại tài liệu cụ thể.			

	6. Tài liệu triển khai tích hợp các mô đun vào hạ tầng kiến trúc, xây dựng bộ lọc Web theo nội dung (Tiếng Anh, Tiếng Việt, lọc ảnh, tài liệu đa cấu trúc Việt+Anh+ảnh) 7. Tài liệu thử nghiệm cục bộ và hoàn thiện thêm bộ lọc Web theo nội dung 8. Tài liệu triển khai tích hợp các mô đun vào hạ tầng kiến trúc, xây dựng bộ lọc Mail theo nội dung (Tiếng Anh, Tiếng Việt, lọc ảnh, tài liệu đa cấu trúc Việt+Anh+ảnh) 9. Tài liệu thử nghiệm cục bộ và hoàn thiện thêm bộ lọc Mail theo nội dung 10. Tài liệu tích hợp các bộ lọc nội dung Web và Mail vào hạ tầng hệ thống 11. Tài liệu thử nghiệm hệ thống lọc 12. Tài liệu đặc tả chi tiết cho phần mềm lọc nội dung tại máy cá nhân (cả client lẫn server) 13. Tài liệu triển khai thử nghiệm hệ thống lọc Web và Mail tại cổng Internet quốc gia ở công ty VDC			
4.	Phần mềm	3 phần mềm với 14 mô đun	3 phần mềm với 15 mô đun	Đáp ứng yêu cầu thử nghiệm đảm bảo độ chính xác theo yêu cầu (90%)
	1. Phần mềm lọc Web theo nội dung - Mô đun chuẩn hoá dữ liệu - Mô đun xác định ngôn ngữ - Mô đun lọc văn bản tiếng Việt - Mô đun lọc văn bản tiếng Anh - Mô đun lọc ảnh - Mô đun lọc URL và PICS - Mô đun kiểm soát - Mô đun ra quyết định - Các mô đun cơ bản trong kiến trúc hạ tầng của hệ thống lọc - Mô đun firewall trong mô hình kiến trúc của hệ thống lọc - Mô đun transparent proxy trong mô hình kiến trúc của hệ thống lọc - Mô đun phân tải phục vụ xử lý thông tin quy mô lớn 2. Phần mềm lọc Mail theo nội dung 3. Phần mềm lọc nội dung cho máy tính cá nhân - Phần mềm lọc nội dung cho máy tính cá nhân phía client - Phần mềm lọc phía server quản lý các danh sách trắng/đen…			
5.	Phần mềm bổ sung **VnGia**	0	1	Phát triển các nội dung lọc nội dung: tự động trích chọn nội dung đúng 90%
	Công bố rộng rãi trên http://vngia.com/			

- Lý do thay đổi (nếu có): *Bổ sung phần mềm VnGia do phát triển được từ các kết quả nghiên cứu của nội dung liên quan tới đề tài và nhận được sự hỗ trợ của đề tài.*

c) Sản phẩm Dạng III:

Số TT	***Tên sản phẩm***	***Yêu cầu khoa học cần đạt***		***Số lượng, nơi công bố*** *(Tạp chí, nhà xuất bản)*
		Theo kế hoạch	Thực tế đạt được	
1	Báo cáo tham gia hội thảo trong nước	10	14	Hai báo cáo đăng kỷ yếu hội nghị trong nước
	1. Trần Mai Vũ, Phạm Thị Thu Uyên, Hoàng Minh Hiền, Hà Quang Thụy (2008). Độ tương đồng ngữ nghĩa giữa hai câu và áp dụng vào bài toán sử dụng tóm tắt đa văn bản để đánh giá chất lượng phân cụm dữ liệu trên máy tìm kiếm VNSEN, *Hội thảo CNTT & TT (ICTFIT08):* 94-102, ĐHKHTN, ĐHQG TP Hồ Chí Minh, Thành phố Hồ Chí Minh, 14/11/2008 2. Lê Diệu Thu, Trần Thị Ngân, Nguyễn Cẩm Tú, Nguyễn Thu Trang (2008). Xây dựng Ontology hỗ trợ tìm kiếm ngữ nghĩa trong linh vực y tế , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã đăng kỷ yếu). 3. Trần Thị Oanh, Lê Hoàng Quỳnh, Lê Anh Cường, Hà Quang Thụy (2009). Một nghiên cứu về gán nhãn từ loại tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày). 4. Trần Nam Khánh, Pham Kim Cuong Nguyễn Thu Trang, Hà Quang Thụy (2009). Finding object-oriented information in unstructured data and adapting to Vietnamese real estate domain, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày). 5. Nguyễn Tiến Thanh, Trần Nam Khánh, Nguyễn Thu Trang, Hà Quang Thụy (2009). Xếp hạng các trường đại học Việt Nam dựa trên "độ đo web" , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày). 6. Nguyễn Thị Thu Chung, Nguyễn Thu Trang, Hà Quang Thụy (2009). Xây dựng danh bạ web tiếng Việt với phân cụm phân cấp văn bản , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày). 7. Trần Mai Vũ, Trần Thị Oanh, Nguyễn Đức Vinh, Phạm Thị Thu Uyên, Nguyễn Đạo Thái, Hà Quang Thụy (2009). Hệ thống hỏi đáp tự động tiếng Việt sử dụng trích rút mối quan hệ ngữ nghĩa trong kho văn bản tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày). 8. Phạm Thị Thu Uyên, Hoàng Minh Hiền, Trần Mai Vũ, Hà Quang Thụy (2008). Độ tương đồng câu và áp dụng vào bài toán tóm tắt đa văn bản tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày). 9. Nguyễn Thị Thu Chung, Nguyễn Thu Trang, Hà Quang Thụy (2008). Đánh giá chất lượng phân cụm trên máy tìm kiếm tiếng Việt VNSEN*, Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*,			

	Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày). 10. Nguyễn Minh Tuấn, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2008). Một mô hình Maximize Entropy phân lớp câu hỏi tiếng Việt . *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày). 11. Nguyễn Thị Thùy Linh, Nguyễn Việt Cường, Hà Quang Thụy (2008). Một mô hình phân lớp đa nhãn SVM đối với văn bản tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày). 12. Đặng Thanh Hải, Trần Thị Oanh, Hà Quang Thụy (2007). Thuật toán Co-training phân lớp Web tiếng Việt sử dụng thông tin liên kết, FAIR 07, Nha Trang, 6-8/8/2007 (Gửi toàn văn và trình bày báo cáo). *13.* Nguyễn Thị Hương Thảo, Nguyễn Thị Thùy Linh, Nguyễn Thu Trang, Hà Quang Thụy (2007). Ứng dụng thuật toán học bán giám sát SVM phân lớp văn bản tiếng Việt, FAIR 07, Nha Trang, 6-8/8/2007 (Gửi toàn văn và trình bày báo cáo) 14. Trần Thị Oanh, Lê Anh Cường, Hà Quang Thụy (2008). Phân đoạn từ tiếng Việt sử dụng Maxent kết hợp nhiều nguồn tri thức, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày)		
2	Báo cáo tham gia hội thảo khoa học quốc tế	2	10	Đăng kỷ yếu Hội nghị quốc tế (có 2-3 phản biện)
	1. Tran Thi Oanh, Le Anh Cuong, Ha Quang Thuy and Quynh Hoang Le (2009). An Experimental Study on Vietnamese POS tagging, *International Conference on Asian Language Processing (IALP 2009)*, Dec 7-9, 2009, Singapore (*accepted,* http://ialp2009.colips.org/index.php?id=14&pg =acceptedlist*)* 2. Vu Tran, Vinh Nguyen, Uyen Pham, Oanh Tran and Quang Thuy Ha (2009). An Experimental Study of Vietnamese Question Answering System, *International Conference on Asian Language Processing (IALP 2009),* Dec 7-9, 2009, Singapore (*accepted,* http://ialp2009.colips.org/index.php?id=14&pg =acceptedlist*)* 3. Huong-Thao Nguyen, Phuong-Thai Nguyen, Quang-Thuy Ha, and Le-Minh Nguyen (2009). Vietnam Noun Phrase Chunking based on Conditional Random Field, *The First International Conference on Knowledge and System Engineering (KSE)*: 172-178, Hanoi, Vietnam, 2009. 4. Dieu-Thu Le, Cam-Tu Nguyen, Quang-Thuy Ha, Xuan-Hieu Phan, and Susumu Horiguchi (2008). Matching and Ranking with Hidden Topics towards Online Contextual Advertising, *The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08):* 888-891, University of Technology, Sydney, Australia, December 9 - 12, 2008. DOI: 10.1109/WIIAT.2008.180 5. Tran Thi Oanh, Le Anh Cuong, Ha Quang Thuy (2008). Improving Vietnamese Word Segmentation by Integrating Different Knowledge Resources, *The 2008 Empirical Methods for Asian Language Workshop (EMALP 2008):* 1-12, Hanoi, Vietnam, Dec. 13, 2008 6. Dang Thanh Hai, Wonjun Lee, Ha Quang Thuy (2008). A pageranking based method for identifying characteristic genes of a disease, *IEEE Proceeding of International Conference on Networking, Sensing and Control, 2008. ICNSC 2008:* 1496-1499, Sanya, China, 6-8 April 2008. DOI:			

	10.1109/ICNSC.2008.4525457 7. Do Thi Minh Viet, Nguyen Hai Chau, Wonjun Lee, Ha Quang Thuy (2008). Using Cross-layer Heuristic and Network Coding to Improve Throughput in Multicast Wireless Mesh Networks, *Information Networking (ICOIN 2008)*, Busan, Korea, January 23-25, 2008 (*The Best paper award*). DOI: 10.1109/ICOIN.2008.4472771 8. Nguyen Viet Cuong, Nguyen Thi Thuy Linh, Ha Quang Thuy and Phan Xuan Hieu (2006). A Maximum Entropy Model for Text Classification, *The International Conference on Internet Information Retrieval 2006:* 134-139, Hankuk Aviation University, Korea, Dec. 6, 2006 9. Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen and Quang-Thuy Ha (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation, *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20):* 215-222, November 1-3, 2006, Wuhan, China. 10. Son Doan, Quang Thuy Ha, and Susumu Horiguchi (2006). A General Fuzzy-based Framework for Text Representation and its Application to Text Categorization*, Lecture Notes on Artificial Intelligence (LNAI),* **4423**: 611-620, 2006 (Springer-Verlag Berlin Heidenberg) form The *Third International Conference on Fuzzy Systems and Knowledge Discovery* - FSKD 2006. DOI: 10.1007/11881599_73			
3	Bài báo đăng tạp chí khoa học quốc tế	0	4	Hai bài tạp chí thuộc ISI
	1. Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha (2009). Classification and Contextual Match on the Web with Hidden Topics from Large Data Collections, *The IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (accepted, 14 pages)*. ISI Journal System 2. Cam-Tu Nguyen, Xuan-Hieu Phan, Susumu Horiguchi, Thu-Trang Nguyen, Quang-Thuy Ha (2009). Web Search Clustering and Labeling with Hidden Topics, *ACM Transactions on Asian Language Information Processing,* **8**(3), 12 (August 2009), 40 pages. DOI=10.1145/1568292.1568295. http://doi.acm.org/10.1145/1568292. 1568295. ISI Journal System 3. Ha Q. Thuy, Nguyen H. Nam, Nguyen Thu Trang (2006). Improve Performance of PageRank Computation with Connected-Component PageRank, *ICMOCCA2006*: 154-158 & *International Journal of Natural Sciences and Technology*, **1**(1): 53-60, 2006 4. Lan N. Bui, Anh Q. Tran, Thuy Q. Ha (2006). User authentic Rating based on Email Networks*, ICMOCCA2006*: 144-148, Seoul, Korea & *International Journal of Natural Sciences and Technology*, **1**(2): 173-180, 2006			

- Lý do thay đổi (nếu có): *Các nội dung nghiên cứu thực hiện và phát triển từ đề tài có giá trị khoa học. Một số công bố quốc tế nhận được sự hỗ trợ từ đề tài.*

d) Kết quả đào tạo:

Số TT	***Cấp đào tạo, Chuyên ngành đào tạo***	***Số lượng***		***Ghi chú*** *(Thời gian kết thúc)*
		Theo kế hoạch	Thực tế đạt được	
1	Thạc sỹ	5	10	10/2009
	1. Phạm Tiến Dũng (2009). Nghiên cứu giải pháp lọc nội dung Internet tại máy tính cá nhân và xây dựng phần mềm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009 2. Lê Đắc Nhường (2009). Tối ưu hóa truy vấn trong máy tìm kiếm thực thể, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009 3. Nguyễn Thu Trang (2009). Học xếp hạng trong tính hạng đối tượng và tạo nhãn cụm tài liệu, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009 4. Trần Thị Oanh (2009). Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009 5. Nguyễn Cẩm Tú (2009). Hidden Topic Discovery Towards Classification and Clustering in Vietnamese Documents*, Luận văn Thạc sỹ (viết bằng tiếng Anh),* Trường ĐHCN, 2008 6. Ngô Thương Huyền (2008). Phân lớp thư điện tử sử dụng máy hỗ trợ vector, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008 7. Nguyễn Thị Thu Hằng (2008). Phương pháp phân cụm tài liệu Web và áp dụng vào máy tìm kiếm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008 8. Nguyễn Việt Cường (2007). Tự động sinh mục lục cho văn bản, *Luận văn Thạc sỹ*, Trường ĐHCN, 2007 9. Đặng Thanh Hải (2007). The Biological Sample Classification Using Gene Expression Data, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHCN, 2007 10. Nguyễn Hoài Nam (2006). The WWW and The PageRank-Related Problems, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHKHTN, 2006			
2	Tiến sỹ	**0**	**3**	Đóng góp nội dung luận án
	1. Nguyễn Cẩm Tú: ĐHCN (2006-2008), ĐH Tohoku-Nhật Bản (2008-2011). 2. Nguyễn Việt Cường: ĐHCN (2006-2007), JAIST-Nhật Bản (2007-2010). 3. Đặng Thanh Hải: ĐHCN (2007-2008), ĐH Antwerp - Bỉ (208-2011).			

- Lý do thay đổi (nếu có): *Số lượng luận văn Thạc sỹ hoàn thành từ hoạt động của đề tài là do nội dung nghiên cứu của đề tài là vấn đề khoa học và công nghệ thời sự nên thu hút được nhiều nghiên cứu sinh và học viên cao học tham gia.*

đ) Tình hình đăng ký bảo hộ quyền sở hữu công nghiệp, quyền đối với giống cây trồng: *Không đăng ký*

Số TT	***Tên sản phẩm đăng ký***	***Kết quả***		***Ghi chú*** *(Thời gian kết thúc)*
		Theo kế hoạch	Thực tế đạt được	
1				

...				

- Lý do thay đổi (nếu có):

e) Thống kê danh mục sản phẩm KHCN đã được ứng dụng vào thực tế:

Số TT	***Tên kết quả đã được ứng dụng***	***Thời gian***	***Địa điểm*** *(Ghi rõ tên, địa chỉ nơi ứng dụng)*	***Kết quả sơ bộ***
1	Phần mềm lọc web		E15	
2	Phần mềm lọc mail		E15	
3	Phần mềm lọc nội dung máy tính cá nhân		Theo VDC	
4	Phần mềm bổ sung http://vngia.com/	Từ tháng 8/2009	http://vngia.com/	Sử dụng rộng rãi

2. Đánh giá về hiệu quả do đề tài, dự án mang lại:

a) Hiệu quả về khoa học và công nghệ:

Về khoa học và công nghệ, xây dựng hệ thống lọc nội dung trên Internet là một đề tài nghiên cứu liên ngành thời sự, đề cập tới các nội dung nghiên cứu về quản lý Nhà nước, về công nghệ thông tin:

- Theo khía cạnh quản lý Nhà nước, thông qua việc khảo sát, phân tích và tổng hợp nội dung các tài liệu liên quan tới vấn đề lọc nội dung trên Internet tại các quốc gia điển hình trên thế giới, đề tài đã chứng tỏ sự cần thiết phải có hệ thống lọc nội dung trên Internet về an ninh quốc gia và thuần phong mỹ tục, tính tất yếu và tính đa dạng hình thức của quản lý Nhà nước về nội dung trên Internet. Đề tài cũng chứng tỏ sự phức tạp của bài toán lọc nội dung trên Internet khi xem xét tới yếu tố tâm lý xã hội, truyền thống, đạo đức, lối sống của từng dân tộc. Như vậy, ngoài các nội dung mang tính quy luật chung của quản lý Nhà nước, hệ thống lọc nội dung trên Internet còn mang đặc thù riêng của mỗi quốc gia. Các nội dung nghiên cứu khoa học về thuần phong mỹ tục, về tâm lý xã hội cũng như về quản lý Nhà nước cũng đã được nhóm thực hiện đề tài quan tâm khi thi hành hệ thống lọc nội dung trên Internet.

Theo khía cạnh công nghệ thông tin, thông qua việc khảo sát, phân tích và tổng hợp một lượng tài liệu phong phú và cập nhật, thông qua quá trình triển khai xây dựng các thành phần và tích hợp hệ thống, nhóm nghiên cứu đề tài đã trình bày các khái niệm cơ bản liên quan tới lọc nội dung trên Internet, phương pháp luận và các giải pháp được lựa chọn để xây dựng các thành phần cũng như tích hợp hệ thống. Bản chất của bài toán lọc nội dung trên Internet là bài toán phân lớp tự động nội dung trang Web, nhóm nghiên cứu đã tập trung nghiên cứu để lựa chọn các giải pháp phân lớp nội dung trang web phù hợp. Đồng thời, đáp ứng yêu cầu lọc nội dung nhanh với luồng dữ liệu với dung lượng lớn, cần kết hợp các giải pháp lọc nội dung với các giải pháp lọc địa chỉ, phân lớp nội dung trang Web theo học máy với phân lớp theo tiêu chí thống kê, phân cấp lọc nội dung theo lọc thô và lọc tinh. Các công trình khoa học được công bố (28 công trình) với một số công bố quốc tế có giá trị và hệ thống phần mềm thử nghiệm là các kết quả khoa học - công nghệ có giá trị của đề tài.

b) Hiệu quả về kinh tế xã hội:

(Nêu rõ hiệu quả làm lợi tính bằng tiền dự kiến do đề tài, dự án tạo ra so với các sản phẩm cùng loại trên thị trường…)

Hệ thống lọc nội dung trên Internet thuộc loại hình quản lý Nhà nước cho nên không tính bằng lợi ích bằng tiền thông qua việc chuyển giao trực tiếp sản phẩm cho các nhà kinh doanh hoặc/và sản xuất.

Hiệu quả kinh tế - xã hội của đề tài được tính gián tiếp thông qua việc so sánh với các nghiên cứu tương đương tại các nước trên thế giới, chẳng hạn tại Cộng đồng chung châu Âu chỉ một sự án nhỏ POESIA (Public Open-source Environment for a Safer Internet Access) đã là 1 triệu 20 nghìn €. Hơn nữa, lợi ích kinh tế - xã hội của đề tài cũng được tính gián tiếp thông qua số công trình công bố quốc tế liên quan (trong đó có các công trình khoa học

thuộc loại có chỉ số ISI) và kết quả đào tạo nhân lực trình độ Thạc sỹ về các nội dung liên quan tới lọc nội dung trên Internet.

3. Tình hình thực hiện chế độ báo cáo, kiểm tra của đề tài, dự án:

Số TT	***Nội dung***	***Thời gian thực hiện***	***Ghi chú*** *(Tóm tắt kết quả, kết luận chính, người chủ trì…)*
I	Báo cáo định kỳ		
	Lần 1: 10/11/2007	5/2007 - 09/2007	- Về số lượng: Hoàn thành về cơ bản theo tiến độ đã đặt ra trong thuyết minh đề tài - Về chất lượng: tuân thủ chặt chẽ những yêu cầu đã được - Do có sự chậm chễ về mặt kinh phí, một số hạng mục của đề tài chưa được triển khai và thực hiện đúng thời điểm
	Lần 2: 7/5/2008	9/2007 - 3/2008	- Về số lượng: Hoàn thành về cơ bản theo tiến độ đã đặt ra trong thuyết minh đề tài - Về chất lượng: tuân thủ chặt chẽ những yêu cầu đã được - Thời gian thực hiện một số hạng mục của đề tài còn chậm so với tiến độ do tổ chức chủ trì đề tài chuyển về địa điểm mới
	Lần 3: 16/10/2008	3/2008 - 9/2008	- Về số lượng: Hoàn thành về cơ bản theo tiến độ đã đặt ra trong thuyết minh đề tài - Về chất lượng: tuân thủ chặt chẽ những yêu cầu đã được đề ra trong thuyết minh
	Lần 4: 1/4/2009	9/2008 - 3/2009	Thời gian thực hiện một số hạng mục còn chậm so với tiến độ đặt ra
II	Kiểm tra định kỳ		
	Lần 1	10/11/2007	- Có sự phối hợp tương đối tốt giữa

			các nhóm nghiên cứu. Đã thực hiên về cơ bản các nội dung đúng tiến độ và yêu cầu đặt ra. - Chủ trì: GS.TS Nguyễn Thúc Hải
	Lần 2	7/5/2008	- Đề tài, gồm các đề tài nhánh đã thực hiện công việc theo tiến độ, kế hoạch - Chủ trì: GS.TS Nguyễn Thúc Hải
	Lần 3	16/10/2008	- Đề tài đã thực hiện thêm một số báo cáo và phần mềm trung gian - Giải ngân còn chậm so với lịch trình chuyên môn. Đề nghị có kế hoạch mua sắm đấu thầu thiết bị - Chủ trì: GS.TS Nguyễn Thúc Hải
	Lần 4	1/4/2009	- Về các nội dung đã thực hiện: tiến độ chậm, đề nghị kéo dài. Cần đề xuất các lý do xin gia hạn bảo vệ logic và khách quan - Chủ trì: GS.TS Nguyễn Thúc Hải
III	Nghiệm thu cơ sở		Dự kiến 30/10/2009

Chủ nhiệm đề tài
(Họ tên, chữ ký)

TS Nguyễn Viết Thế

Thủ trưởng tổ chức chủ trì
(Họ tên, chữ ký và đóng dấu)
TL TỔNG CỤC TRƯỞNG
KT CỤC TRƯỞNG
PHÓ CỤC TRƯỞNG

Nguyễn Văn Thuỷ

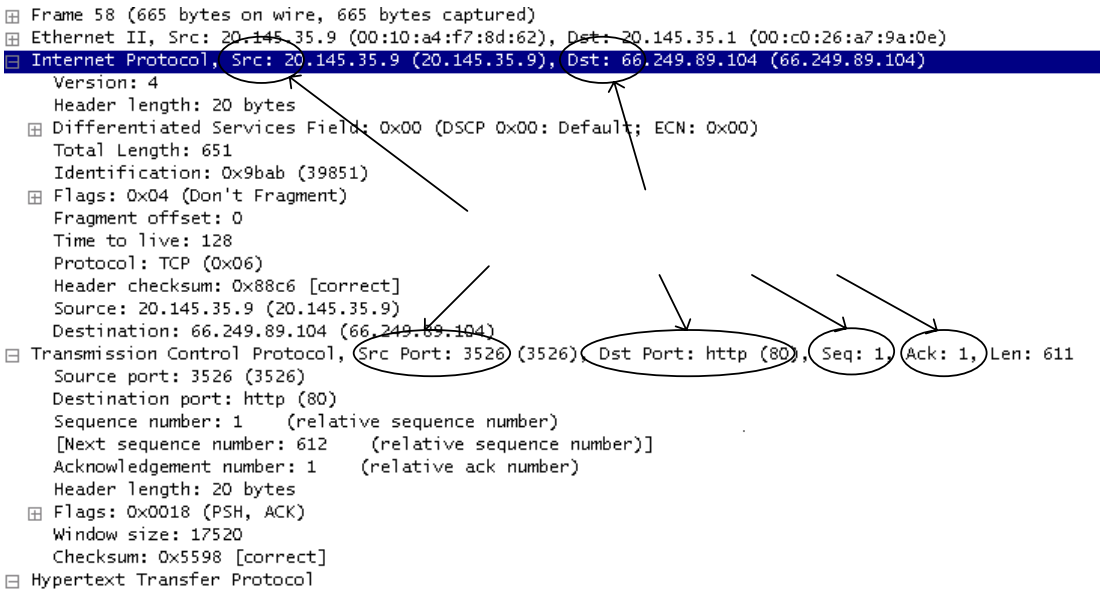
MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

GAPP	*General Administration of Press and Publication*
ONI	The OpenNet Initiative
VNCERT	Trung tâm Ứng cứu khẩn cấp máy tính Việt Nam
VNNIC	Trung tâm Thông tin mạng Internet Việt Nam
PICS	Platform for Internet Content Selection
W3C	World Wide Web Consortium
SPM	Skin Probability Map
LUT	*Normalized lookup table*
FOM	First Order Model
GFE	Global Fit Ellipse
LFE	Local Fit Ellipse
PDF	Portable Document Format
PS	PostScript
JPG/JPEG	Joint Photographic Expert Group
GIF	Graphic Interchange Format
PNG	Portable Network Graphic
BMP	Windows Bitmap
ICAP	Internet Content Adaptation Protocol
ARFF	Attribute-Relation File Format

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ.



MỞ ĐẦU

Căn cứ vào Hợp đồng nghiên cứu khoa học số 02/2006/HĐ-ĐtCT-KC.01/06-10 ký ngày 14/5/2007 giữa Cục Công nghệ tin học nghiệp vụ, Tổng cục kỹ thuật, Bộ Công an với Ban Chủ nhiệm Chương trình KC.01/06-10 và Văn phòng Các chương trình thì đề tài được thực hiện trong 2 năm 6 tháng từ tháng 5/2007 tới tháng 10/2009 (theo hợp đồng) song trên thực tế thì một số nội dung nghiên cứu trong đề tài đã được nhóm thực hiện đề tài tiến hành từ tháng 5/2006 khi Bộ Khoa học và Công nghệ ra thông báo về quyết định triển khai đề tài.

MỤC TIÊU ĐỀ TÀI

Mục tiêu 1: Nghiên cứu và đề xuất giải pháp hỗ trợ công tác quản lý một cách hiệu quả an toàn – an ninh các luồng dữ liệu vào/ra giữa Việt Nam và thế giới qua mạng Internet nói riêng và giữa các mạng diện rộng nói chung

Mục tiêu 2: Phát triển hệ thống thử nghiệm cho phép xử lý khối lượng dữ liệu lớn thời gian thực (tính toán song song, tính toán lưới), có khả năng phát hiện và ngăn chặn thông tin (ảnh, văn bản bằng cả tiếng Việt và tiếng Anh) có nội dung không phù hợp với văn hoá, pháp luật Việt Nam và ảnh hưởng xấu đến trật tự an toàn xã hội

Mục tiêu 3: Triển khai và ứng dụng thử nghiệm tại cổng thông tin vào/ra tại trường Đại học Công nghệ, tại Bộ Công an, và cổng Internet quốc gia tại trung tâm điện toán và truyền số liệu VDC

Đặc điểm chính của đề tài

1. Hệ thống lọc nội dung trên Internet đã và đang được nhiều quốc gia trên thế giới quan tâm đặc biệt trong định hướng an toàn Internet. Đối với nhiều quốc gia, hệ thống này là một bộ phận của hệ thống an ninh quốc gia nói chung. Các quốc gia và tổ chức liên quốc gia đã và đang tiến hành các hoạt động nghiên cứu và triển khai các hệ thống lọc nội

dung trên Internet, điển hình là các dự án của Cộng đồng Châu Âu đã và đang được tiến hành như "Internet Safer" (1999-2004), "Internet Safer Plus" (giai đoạn 2005-2008) và "Chương trình an toàn Internet đối với trẻ em" (giai đoạn 2008-2013). Xây dựng hệ thống lọc nội dung trên Internet là một bài toán phức tạp, đòi hỏi phải thi hành được các giải pháp có tính khoa học và công nghệ cao nhằm phục vụ đắc lực chính sách quốc gia về an toàn, an ninh Internet, khắc phục kịp thời các thủ đoạn vi phạm an toàn, an ninh Internet.

2. *Lọc nội dung Internet* là thuật ngữ được dùng để chỉ các kỹ thuật kiểm soát thông tin trên Internet thông qua việc *phân tích nội dung thông tin (đặc biệt là nội dung trang Web, nội dung thư điện tử)* để sau đó *cho hoặc không cho* người sử dụng Internet nhận được kết quả trả về từ Internet hoặc gửi thông tin lên mạng Internet. Nắm bắt được nội dung thông tin dưới dạng trình bày là văn bản, hình ảnh để sau đó đánh giá, phân loại nó thuộc vào lớp nào trong các lớp nội dung trong chính sách an ninh Internet là bài toán chủ chốt nhất. Đây chính là bài toán phân lớp trang Web, trong đó, hệ thống tiến hành phân tích nội dung một trang web để quyết định trang web đó thuộc lớp nào trong các lớp đã định trước theo chính sách an toàn, an ninh Internet. Bài toán phân lớp trang web đòi hỏi các giải pháp về *xử lý ngôn ngữ tự nhiên* (tiếng Việt, tiếng Anh) và *hình ảnh*, *trích chọn đặc trưng trong nội dung* để *biểu diễn văn bản và hình ảnh* và áp dụng *các thuật toán phân lớp dữ liệu*. Bài toán phân lớp văn bản trang web là một bài toán nghiên cứu, triển khai thời sự, vì vậy, việc phân tích để lựa chọn các giải pháp từ kết quả nghiên cứu trong các lĩnh vực kể trên là một yêu cầu tất yếu. Đáp ứng yêu cầu trên, nhóm thực hiện đề tài KC.01.02/06-10 đã tiến hành các nghiên cứu về các nội dung trên đây, đặc biệt là các nghiên cứu về xử lý tiếng Việt và trích chọn đặc trưng từ nội dung trang Web.

3. Hệ thống lọc nội dung trên Internet cần đảm bảo các yêu cầu là thời gian nhanh đối với luồng thông tin xử lý lớn vì vậy hệ thống cần kết hợp các giải pháp đa dạng. Tiếp thu các kết quả nghiên cứu trên thế giới, nhóm thực hiện đề tài đã thi hành giải pháp lọc nội dung qua hai giai đoạn (lọc thô, lộc sâu), lọc theo nội dung kết hợp với lọc theo địa chỉ, lọc nội dung theo học máy và lọc nội dung theo luật thống kê.
4. Cục Công nghệ Tin học nghiệp vụ - Bộ Công an (có nhiều kinh nghiệm trong quản lý thông tin trên Internet) là tổ chức chủ trì cùng với hai tổ chức phối hợp thực hiện là Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội (có tri thức và kinh nghiệm về khai phá dữ liệu Web, lĩnh vực liên quan trực tiếp tới lọc nội dung) và Công ty điện toán và truyền số liệu VDC (có kinh nghiệm thi hành các giải pháp quản lý cổng Internet quốc gia) đã tập hợp được một lực lượng nghiên cứu - triển khai đề tài gồm 2 PGS, 5 TS, 18 nghiên cứu sinh và học viên cao học thực hiện hoạt động nghiên cứu, đề xuất giải pháp và lập trình thực thi hệ thống lọc nội dung. Tuy nhiên, tiêu chí an toàn – an ninh trong bài toán phân lớp nội dung trang Web là rất mới lạ mà trong một số trường hợp còn bao hàm các yếu tố nhạy cảm, vì vậy việc tập hợp dữ liệu học cho các bộ phân lớp còn có chỗ chưa toàn diện.

Các nguyên tắc tiếp cận khoa học

Nghiên cứu, triển khai xây dựng mô hình và giải pháp lọc nội dung trên Internet được tiếp cận theo các cách thức sau đây:

i. Tìm hiểu sâu rộng hoạt động quản lý Nhà nước liên quan tới lọc nội dung trên Internet tại nhiều quốc gia trên thế giới để nắm bắt được xu thế hoạt động quản lý Nhà nước cả theo phương diện pháp luật, xã hội và kỹ thuật, công nghệ. Nghiên cứu chủ trương, chính sách

của Nhà nước ta liên quan tới các hệ thống lọc nội dung trên Internet.

ii. Khảo sát và phân tích thấu đáo các nội dung về các *công nghệ và kỹ thuật đã và đang được sử dụng trong các hệ thống lọc nội dung thông tin Internet của các quốc gia* (Mỹ, Châu Âu, Trung Quốc, …) cũng như về các sản phẩm thương mại đã có (SmartFilter, R3000G Internet Filter, …) để từ đó phân tích, đánh giá và nghiên cứu đề xuất giải pháp cụ thể cho vấn đề lọc nội dung hỗ trợ công tác quản lý và bảo đảm an toàn-an ninh thông tin trên mạng Internet tại Việt Nam.

iii. Khai thác, phát triển phần mềm mã nguồn mở trong việc xây dựng hệ thống phần mềm lọc nội dung Internet sẽ là một cách tiếp cận quan trọng. Một trong những phần mềm mã nguồn mở mà nhóm cộng tác sẽ chú trọng phân tích nội dung đó là dự án POESIA của Cộng đồng chung Châu Âu.

iv. Luận giải những vấn đề thực tế của những *luồng thông tin luân chuyển* trên mạng Internet liên quan đến đất nước và con người Việt Nam để làm rõ tình hình và những vấn đề liên quan đến việc *bảo đảm an toàn-an ninh luồng thông tin Internet tại Việt nam*.

v. Nghiên cứu các *giải pháp triển khai hệ thống sản phẩm kết quả tại một cổng Internet quốc gia*, nơi có ràng buộc rất lớn về tốc độ xử lý và lưu lượng thông tin chuyển qua và vì vậy các giải pháp lọc nội dung trên Internet cần có độ phức tạp thời gian và không gian phù hợp.

vi. Tham gia các hội thảo khoa học, cả trong nước và quốc tế, liên quan đến lĩnh vực lọc nội dung trên Internet nhằm (1) Tiếp thu các công nghệ mới và mở rộng hợp tác với các cá nhân, tổ chức trong và ngoài nước; (2) Công bố kết quả nghiên cứu về mô hình và giải

pháp liên quan tới nội dung đề tài tại các tạp chí và hội nghị khoa học trong nước và quốc tế để khẳng định tính tin cậy của các mô hình và giải pháp được áp dụng thi hành hệ thống lọc nội dung trên Internet.

Các nghiên cứu trên thế giới liên quan tới các giải pháp lọc nội dung trên Internet, chẳng hạn như [Ayr01, Lanq01, POES04, QD09, Sten04, Zhan05], cho thấy phân lớp (classification) văn bản tự động đã trở thành giải pháp lọc nội dung điển hình. Chính vì lý do đó mà phân lớp văn bản là một trong những nội dung chính của đề tài này với mục tiêu là xác định nội dung trang web thuộc vào lớp văn bản nào theo các tiêu chí đã được xác định, chẳng hạn như văn bản đó có thuộc lớp chứa ***thông tin xấu*** hay không. Phương pháp phân lớp được phân thành một số mức khác nhau, với độ phức tạp tăng dần từ từ khóa, cấu trúc, đến ngữ nghĩa của dữ liệu. Vì thế, công việc phân lớp văn bản trong hệ thống lọc nội dung trên Internet đòi hỏi phải khảo sát các công nghệ mới nhất hiện nay để tìm ra giải pháp thích hợp nhất nhằm đảm bảo đáp ứng cả hai tiêu chí chất lượng và thời gian để đảm bảo tính tức thời của thông tin yêu cầu

Các nguyên tắc tiếp cận về quản lý

- *Nguyên tắc tập trung, thống nhất và cộng tác*: Cục Công nghệ Tin học nghiệp vụ (E15), Bộ Công An là tổ chức chủ trì đề tài phân công các nội dung nghiên cứu tới các đơn vị tham gia thực hiện đề tài (E15, Trường ĐHCN-ĐHQGHN, Công ty VDC) như đã trình bày trong Bản thuyết minh đề tài, trong đó, E15 thi hành phần mềm lọc thư điện tử, Trường ĐHCN thi hành hệ thống lọc nội dung Web và VDC thi hành phần mềm lọc nội dung máy tính cá nhân. E15 cùng với hai đơn vị phối hợp thực hiện tiến hành kiểm tra đánh giá chung và tích hợp hệ thống. E15 là đầu mối của nhóm nghiên cứu trong quan hệ công tác với Ban Chủ nhiệm Chương trình KC.01 và Văn phòng các chương trình cấp Nhà nước.

- *Nguyên tắc phân hoạch trách nhiệm*: Đề tài được phân chia thành các thành phần, mỗi thành phần được phân chia thành các công việc thông qua các hợp đồng công việc được ký kết mà mỗi cá nhân, nhóm nghiên cứu chịu trách nhiệm hoàn thành các công việc được phân công. Việc đánh giá, nghiệm thu nội bộ thực hiện đúng quy định.
- *Nguyên tắc phối hợp, công tác*: Nhóm thực hiện đải tài tiến hành định kỳ các cuộc họp để kiểm tra, đánh giá tiến độ thực hiện công việc của mỗi bộ phận và từng cá nhân.

GIỚI THIỆU NỘI DUNG BÁO CÁO TỔNG HỢP

Mô tả chi tiết về kết quả nghiên cứu và triển khai thực hiện đề tài KC.01.02/06-10 đã được tập hợp thành một hệ thống các báo cáo chuyên đề thuộc các tài liệu sau đây:

- *Báo cáo kết quả thực hiện đề tài KC.01.02/06-10 (Quyển 3): Nghiên cứu chung về lọc nội dung trên Internet,*
- *Báo cáo kết quả thực hiện đề tài KC.01.02/06-10 (Quyển 4): Hệ thống thử nghiệm lọc web,*
- *Báo cáo kết quả thực hiện đề tài KC.01.02/06-10 (Quyển 5): Hệ thống thử nghiệm lọc mail,*
- *Báo cáo kết quả thực hiện đề tài KC.01.02/06-10 (Quyển 6): Hệ thống lọc tại máy tính cá nhân,*
- *Báo cáo kết quả thực hiện đề tài KC.01.02/06-10 (Quyển 7): Kết quả đào tạo, hợp tác quốc tế và sản phẩm bổ sung.*

Báo cáo tổng hợp đề tài tóm lược các nội dung của hệ thống các báo cáo chuyên đề trên đây nhằm cung cấp các thông tin khái quát nhất về các nội dung đã được tiến hành trong đề tài. Nội dung chi tiết và toàn diện mọi khía

cạnh về cơ sở lý thuyết, kỹ thuật nền tảng, thiết kế phần mềm thử nghiệm và đánh giá được trình bày trong các báo cáo chuyên đề.

Báo cáo tổng hợp đề tài bao gồm các chương nội dung chính sau đây:

- *Chương 1. Nghiên cứu và đánh giá tình hình quản lý Nhà nước về lọc nội dung trên Internet* trình bày khái quát về tình hình quản lý Nhà nước về lọc nội dung trên Internet của nhiều quốc gia trên thế giới, các tổ chức liên quốc gia và đưa ra một số nhận định đánh giá. Một số phương hướng giải pháp liên quan tới quản lý Nhà nước về lọc nội dung trên Internet tại Việt Nam cũng được đề cập.
- *Chương 2. Cơ sở lý thuyết và các giải thuật lọc nội dung* trình bày về một số nội dung cơ bản nhất về lọc nội dung trên Internet. Các kỹ thuật học phân lớp (giám sát và bán giám sát) nội dung trang Web - nền tảng của kỹ thuật lọc nội dung, các kỹ thuật lọc địa chỉ trong hệ thống kết hợp với kỹ thuật lọc nội dung, các kỹ thuật lọc ảnh được giới thiệu. Đồng thời, việc phân tích, đánh giá các giải thuật để lựa chọn các giải thuật phù hợp cho hệ thống lọc nội dung được xây dựng đã được trình bày.
- *Chương 3. Xây dựng sản phẩm phần mềm lọc nội dung Internet* trình bày các mô tả về các sản phẩm phần mềm chủ yếu của đề tài, đó là Hệ thống phần mềm lọc nội dung Web (SP.01), Hệ thống lọc thư điện tử - Mail Gateway (SP.02), Phần mềm lọc Web trên máy tính cá nhân (SP.03). Đối với mỗi sản phẩm phần mềm trên đây, các nội dung khái quát về cấu trúc hệ thống, giải pháp và đánh giá thử nghiệm được giới thiệu.
- *Chương 4. Kết quả đào tạo, hợp tác quốc tế và sản phẩm bổ sung* trình bày các kết quả đào tạo đại học, sau đại học, hợp tác quốc tế đã thu nhận được qua quá trình triển khai đề tài, khẳng định tính thời sự của nội dung đề tài khoa học - công nghệ được thực hiện. Một số công bố

liên quan tới trích chọn thông tin góp phần phục vụ lọc nội dung trang Web đã được công bố khoa học trong nước và quốc tế, trong đó đã có công bố khoa học trên tạp chí có chỉ số ISI. Một số nội dung hợp tác quốc tế liên quan tới đề tài cũng được giới thiệu. Đồng thời, một sản phẩm bổ sung trong quá trình thực hiện đề tài đã được giới thiệu.

Phần kết luận và kiến nghị của Báo cáo tổng hợp kết quả thực hiện đề tài trình bày một số nhận định của nhóm thực hiện đề tài (do Cục Công nghệ Tin học nghiệp vụ, Bộ Công an chủ trì) tự đánh giá kết quả thực hiện đề tài và một số đề xuất liên quan. Phần tiếp theo trong báo cáo là *Danh mục các tài liệu tham khảo*.

Kết thúc báo cáo là *Phần phụ lục* bao gồm bìa, lời giới thiệu của 10 luận văn Thạc sỹ và nội dung hai công trình khoa học công bố quốc tế tiêu biểu liên quan tới nội dung đề tài.

CHƯƠNG I

NGHIÊN CỨU VÀ ĐÁNH GIÁ TÌNH HÌNH QUẢN LÝ NHÀ NƯỚC VỀ LỌC NỘI DUNG INTERNET

1.1. Khái quát về hoạt động quản lý Nhà nước về lọc nội dung trên Internet

Jonathan L. Zittrain và John G. Palfrey, Jr. [ZP07] nhận định rằng *tự do ngôn luận*, và cũng tương tự đối với *tự do tôn giáo và tự do đời tư*, *không bao giờ có tính tuyệt đối*. Nhận định nói trên là hoàn toàn xác đáng và phù hợp với bản chất của hoạt động quản lý nhà nước trong các xã hội còn tồn tại các giai cấp khác nhau. Hoạt động quản lý Nhà nước về lọc nội dung trên Internet của các quốc gia vừa tuân theo quy luật phổ biến về quản lý Nhà nước nói chung, vừa có tính đặc thù riêng đối với từng quốc gia vì rằng mỗi quốc gia còn có những đặc điểm riêng tương ứng với đặc trưng của dân tộc về truyền thống, về thuần phong - mỹ tục, về tôn giáo và các đặc trưng khác. Như vậy, vì hoạt động quản lý Nhà nước về lọc nội dung trên Internet là một thành phần trong hệ thống phương tiện đảm bảo lợi ích quốc gia - dân tộc cho nên nội dung và mức độ quản lý Nhà nước đối với hoạt động này cũng có sự khác biệt thực sự giữa các quốc gia.

Ở một góc độ khác, có thể nhận thấy một thực tế là dù rất mong muốn kiểm soát được một cách toàn diện môi trường thông tin trong nước, song đa phần các quốc gia cũng mới chỉ thực thi được mong muốn của họ thông qua việc kiểm soát phương tiện truyền thông và cố gắng ngăn cản bất kỳ phát ngôn nào có chứa các nội dung mang tính lật đổ chính quyền [ZP07]. Điều đó có nghĩa là hoạt động quản lý Nhà nước về lọc nội dung trên Internet không thể được hoàn thiện một cách tuyệt đối và nhu cầu thường xuyên nâng cao chất lượng của hoạt động này là hết sức cần thiết.

1.1.1. Một số đặc điểm chung về hoạt động quản lý Nhà nước về lọc nội dung trên Internet

Hoạt động quản lý về lọc nội dung trên Internet được thể hiện theo các khía cạnh *về pháp luật và tổ chức cơ quan Nhà nước*, *về tổ chức triển khai thực hiện* và *về sự hỗ trợ của Nhà nước đối với hoạt động này*.

1.1.1.1. Pháp luật và tổ chức cơ quan nhà nước

Quản lý xã hội bằng pháp luật là yêu cầu khách quan của một xã hội văn minh, công bằng, dân chủ, và là phương thức rất quan trọng bảo đảm hiệu lực quản lý của Nhà nước [HV04]. Trong thời đại ngày nay, hoạt động quản lý Nhà nước trước hết được thể hiện theo khía cạnh pháp lý. Nhà nước tổ chức xây dựng các văn bản pháp lý mô tả đúng nội dung của hoạt động quản lý Nhà nước và đảm bảo thi hành một cách đúng đắn, toàn diện các nội dung các văn bản pháp lý đã được xây dựng trên phạm vi toàn xã hội.

Đối với hoạt động quản lý Nhà nước về lọc nội dung trên Internet, theo Jonathan L. Zittrain và John G. Palfrey, Jr. [ZP07], thì khi quyết định lọc Internet, tiếp cận chung của các quốc gia là thiết lập một "phòng tuyến" gồm các luật và tiêu chuẩn kỹ thuật để hình thành một khung pháp lý được áp đặt đối với mọi công dân và mọi tổ chức trong quốc gia đó đối với hoạt động truy nhập và công bố thông tin trên Internet. Ở một số quốc gia, thường là các nước phát triển, hình thức phổ biến là mở rộng nội dung các văn bản pháp luật sẵn có về các phương tiện truyền thông đại chúng và viễn thông. Tại các quốc gia này, người ta bổ sung thêm các điều luật, các quy định vào các văn bản pháp luật sẵn có để các văn bản này bao hàm thêm yếu tố Internet. Tại các quốc gia khác, người ta thiết lập các nội dung pháp lý tương ứng thành các đạo luật và quy tắc riêng có phạm vi điều chỉnh là riêng biệt đối với Internet. Nhìn chung, rất ít khi các quốc gia thiết lập hẳn các cách thức kỹ thuật chuyên biệt về lọc nội dung trên Internet mà đa phần người ta thiết lập một khung pháp lý nhằm giới hạn một số kiểu nội dung trực tuyến và ngăn cấm một số

hoạt động trực tuyến [ZP07]. Ở một số nước, đặc biệt là các nước Tây Âu và Bắc Mỹ, các nội dung pháp lý như vậy lại có thể nằm trong khuôn khổ của các điều luật khác (nhiều khi không liên quan tới các phương tiện truyền thông đại chúng và viễn thông), chẳng hạn như ở Mỹ "*Nhiều điều khoản trong Luật yêu nước của Mỹ cho phép nghe lén điện thoại, điều tra hồ sơ cá nhân và đọc email của công dân. FBI được quyền theo dõi người dân đọc gì bằng cách kiểm tra liệt kê các đầu sách họ mượn tại thư viện...*"[1].

Về mặt tổ chức Nhà nước, các quốc gia thường thành lập các cơ quan chuyên trách hoặc liên ngành chịu trách nhiệm về an toàn-an ninh trên Internet, trong đó nhiệm vụ về lọc nội dung. Chẳng hạn, Chính phủ Trung Quốc thành lập tổ chức đa liên ngành quản lý thông tấn và xuất bản *General Administration of Press and Publication* (*GAPP*) thực hiện các chức năng tương ứng với hoạt động này [China02]. Đối với các quốc gia khác, việc thi hành hoạt động như vậy lại do một số cơ quan thi hành mà thường được tương ứng với chức năng, nhiệm vụ cụ thể của các cơ quan đó theo quy định. Như ví dụ đã được giới thiệu ở trên, Cục điều tra liên bang Mỹ FBI là một trong các cơ quan của Nhà nước Mỹ tham gia vào hoạt động lọc nội dung e-mail của người dân hoặc các cơ quan nghiên cứu của Bộ Quốc phòng Mỹ cũng tiến hành các nghiên cứu liên quan [US04].

Về mặt pháp lý, các quốc gia thường quy định về tính không hợp pháp đối với các hoạt động trên Internet theo ba phạm vi chính như sau (được liệt kê theo thứ tự giảm dần về mức độ gắn kết với hoạt động quản lý Nhà nước):

- Về an ninh quốc gia trực tiếp,
- Về đạo đức, truyền thống của dân tộc,
- Về an toàn trẻ em.

[1] http://www.tuoitre.com.vn/Tianyon/Index.aspx?ArticleID=22328&ChannelID=2

Có thể nhận thấy rằng, các phạm vi trên đây là phù hợp với các khía cạnh cần được quan tâm khi đánh giá hệ thống lọc nội dung trên Internet của một quốc gia theo phương pháp luận của ONI [ONI07]. Theo ONI, ba khía cạnh nội dung chính yếu cần được quan tâm là nội dung chính trị (*Political content*), nội dung xã hội (*Social content*), nội dung liên quan đến xung đội và an ninh (*related to conflict & security*).

Từ tính thống nhất của các Nhà nước trên thế giới về việc mong muốn kiểm soát môi trường thông tin trong phạm vi quốc gia dẫn đến có nhiều điểm giống nhau trong quan niệm về các loại hình hoạt động vi phạm trên Internet. Đồng thời, từ đặc trưng truyền thống dân tộc, thuần phong - mỹ tục, đạo đức lối sống có sự khác biệt giữa các quốc gia cũng dẫn đến tồn tại sự khác biệt bộ phận về quan niệm đối với các hoạt động vi phạm. Chính vì lẽ đó, ngoài sự khác biệt về công nghệ và kỹ thuật triển khai, hệ thống lọc nội dung trên Internet của các quốc gia cũng bao chứa những điểm khác biệt về quan niệm về hành vi vi phạm. Các kết quả đánh giá về phạm vi, độ rộng và chiều sâu lọc nội dung trên Internet do ONI tiến hành cung cấp thể hiện cụ thể của sự khác biệt này [ONI07].

Nói riêng về phương diện an ninh quốc gia trực tiếp có thể nhận thấy rằng, có sự thống nhất chung trong quan niệm của hầu hết các Nhà nước trên thế giới về hầu hết các loại hình hoạt động vi phạm an ninh quốc gia trên Internet điển hình (chẳng hạn như hoạt động tuyên truyền lật đổ, chống phá Nhà nước...). Tuy nhiên, do sự khác nhau về phương thức trình bày các văn bản pháp luật liên quan (có Nhà nước công bố luật riêng, có Nhà nước đưa các nội dung đó trong các văn bản pháp luật khác nhau) cho nên cách hiểu về loại hình hoạt động vi phạm an ninh quốc gia trực tiếp trên Internet lại có điểm khác biệt giữa các quốc gia.

Tương tự như vậy đối với phương diện về đạo đức, truyền thống của dân tộc.. cũng có nhiều nét tương tự. Nhìn chung, các quốc gia có sự tương

đồng chung đối với nhiều thành phần cơ bản trong đạo đức, truyền thống và thuần phong mỹ tục của mình. Vì vây, hầu hết các quốc gia đều đưa ra các luật ngăn cấm các hành vi vi phạm thuần phong mỹ tục trên Internet. Tuy nhiên, quan niệm riêng của từng quốc gia về đạo đức, thuần phong mỹ tục cũng như truyền thống của dân tộc mình, cho nên phạm vi liên quan đến đạo đức, truyền thống lại có những điểm khác biệt giữa các quốc gia. Một hành động ứng xử nào đó đối với một dân tộc này không được coi là vi phạm đạo đức, truyền thống song đối với một dân tộc khác thì lại được quan niệm ngược lại, nhiều khi lại được coi là rất nghiêm trọng. Mặt khác, cùng một hành vi được coi là vi phạm đạo đức, truyền thống, thì các quốc gia khác nhau lại có quan niệm khác nhau về mức độ nghiêm trọng [ZP07]. Thông thường, một quốc gia có quốc đạo, chẳng hạn như các nước Hồi giáo, thì một số luật lệ tôn giáo cũng được quy định thành một số điều khoản trong các văn bản luật pháp [ONI04Arap, ONI05Iran, ONI05UAE].

Sự thống nhất cao trên toàn thế giới về phương diện an toàn Internet đối với trẻ em với minh chứng là mọi quốc gia đều đưa ra các quy định pháp luật ngăn cấm việc cung cấp nội dung độc hại trên Internet đến với trẻ em. Một số quyết nghị của Cộng đồng châu Âu đã khuyến cáo về việc cần thiết phải có những văn bản pháp lý với các điều khoản về việc đảm bảo an toàn Internet cho trẻ em. Tồn tại nhiều điều khoản trong các văn bản pháp luật của các nước liên quan tới việc đảm bảo an toàn Internet đối với trẻ em.

1.1.1.2. Khía cạnh tổ chức triển khai và hỗ trợ

Các cơ quan Nhà nước được giao trách nhiệm quản lý hoạt động lọc nội dung trên Internet tiến hành các công tác cần thiết để hoàn thành tốt trách nhiệm được giao. Các cơ quan này (chuyên trách hoặc liên ngành) thường là đầu mối tổ chức xây dựng các văn bản pháp luật, các chính sách, các dự án cùng các giải pháp đảm bảo các chính sách lọc nội dung trên Internet của

quốc gia được thi hành đúng đắn. Ví dụ, Ủy ban về các vấn đề mạng truyền thông và an ninh của Cộng đồng Châu Âu là cơ quan thực hiện các công việc nói trên.

Tất cả các quốc gia tiến hành lọc nội dung trên thế giới đều đưa ra các chính sách hỗ trợ việc tiến hành hoạt động nói trên. Một trong những hỗ trợ điển hình là đầu tư kinh phí hỗ trợ việc nghiên cứu thực trạng, đề xuất các giải pháp công nghệ và xã hội liên quan.

Các quốc gia cũng khuyến khích các nhà cung cấp sản phẩm lọc nội dung đối với các tổ chức và gia đình.

1.1.1.3. Một số xu hướng

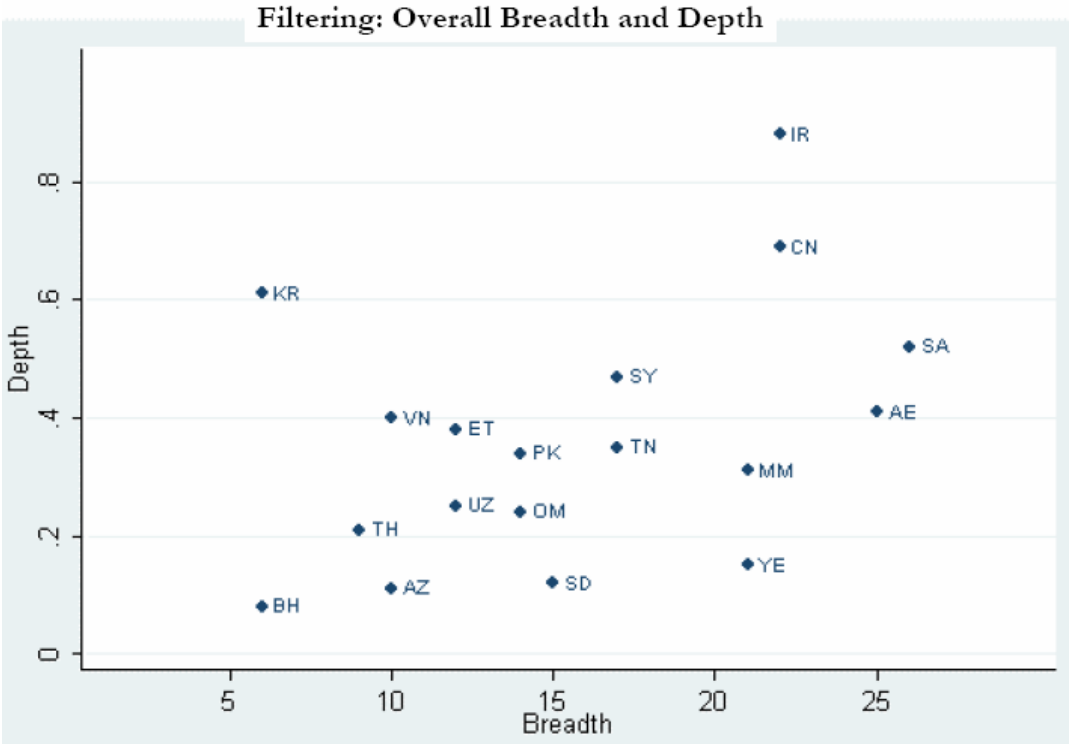
Xu hướng đầu tiên cần được quan tâm là *ảnh hưởng của công nghệ truy nhập và cung cấp thông tin trên Internet đối với hoạt động lọc nội dung trên Internet*. Sự phát triển không ngừng công nghệ Internet cũng như sự hình thành và phát triển các công nghệ truyền thông mới là cơ hội để Internet phát triển và đồng thời hoạt động lọc nội dung trên Internet cũng phải được mở rộng và nâng cấp để đảm bảo mục tiêu của quản lý Nhà nước về lọc nội dung trên Internet. Một minh chứng điển hình là trong thời đại ngày nay truy nhập Internet không chỉ thông qua máy tính mà còn có thể được hoàn thành trên các thiết bị cầm tay, điển hình là điện thoại di động. Chính vì lý do đó, các khía cạnh quản lý Nhà nước về lọc nội dung trên Internet (luật pháp, triển khai và hỗ trợ) đều phải được thích ứng với việc ra đời và phát triển các phương tiện cung cấp và truy nhập thông tin trên Internet hiện đại. Chẳng hạn, Ủy ban Cộng đồng Châu Âu đã quyết nghị những nội dung liên quan đến việc sử dụng thiết bị di động truy nhập Internet.

Xu hướng phát triển các hệ thống lọc nội dung cũng cần được quan tâm. Jonathan L. Zittrain and John G. Palfrey, Jr. [JP07] cho rằng *xu hướng phát triển sắp tới của hệ thống lọc Internet được Nhà nước ủy nhiệm* có thể xuất

hiện ở các khu vực khác trên thế giới song ở mức độ nhẹ nhàng hơn, như hệ thống công cộng lọc Internet tại thư viện và trường học ở Mỹ, hệ thống lọc khiêu dâm ở Bắc Âu, hệ thống lọc của các site Nazi và Holocaust tại Pháp và Đức. Tuy nhiên, có một thực tế là cùng với sự tiến bộ của công nghệ Internet nói chung, cũng như sự phát triển các giải pháp lọc nội dung trên Internet thì các hình thức cung cấp các nội dung độc hại hoặc không có nhu cầu cũng phát triển theo. Lọc nội dung trong cuộc chiến chống các hoạt động phá hoại này cũng phải luôn luôn được quan tâm. Chúng ta có thể nhận thấy rằng thực tế xu hướng này chỉ xuất hiện tại một số nước phát triển, do lịch sử phát triển khoa học - công nghệ, Nhà nước tại các quốc gia này có trình độ cao trong hoạt động kiểm soát môi trường thông tin cả ở khía cạnh luật pháp (hệ thống pháp luật khá tinh xảo và hoàn thiện) lẫn ở khía cạnh công nghệ, kỹ thuật (thường đi trước các nước đang phát triển). Theo đánh giá chung thì trong giai đoạn hiện nay, các nước đang phát triển coi xu hướng này như là một tham khảo trong chiến lược phát triển Internet của mình.

1.1.2. Phương pháp khảo sát của ONI

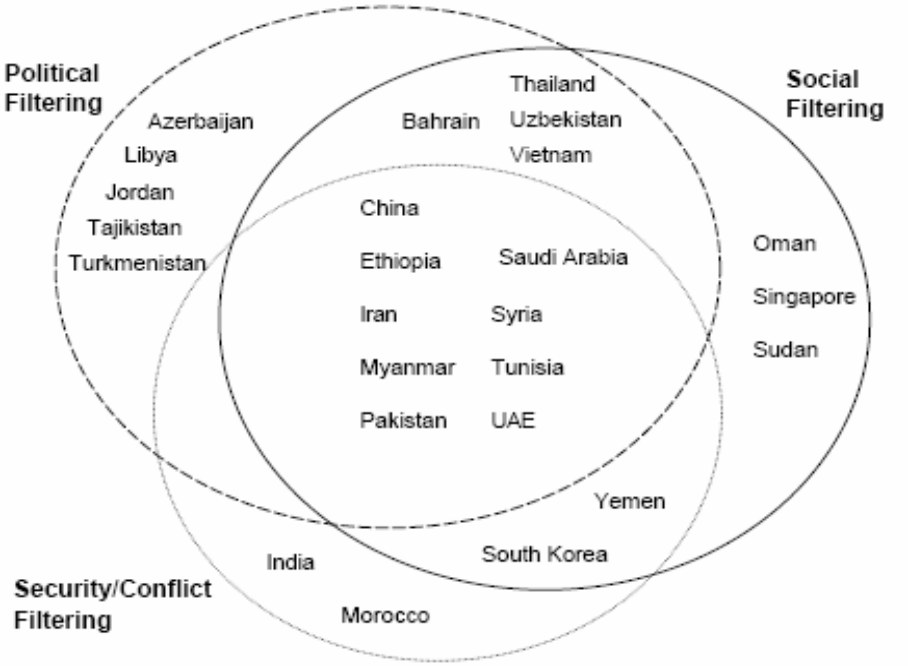
Một số công trình nghiên cứu điển hình về lọc nội dung trên Internet đối với một số nước đã được tổ chức The OpenNet Initiative (ONI) tổng hợp và công bố tại trang Web của tổ chức này. ONI là một tổ chức có nhiệm vụ điều tra nghiên cứu về tình trạng giám sát và lọc thông tin trên thực tế tại các quốc gia, để từ đó tìm ra những ảnh hưởng đến chủ quyền đất nước, các tác động đến người sử dụng... Để đạt được mục đích đó, ONI sử dụng một cách tiếp cận kết hợp *các phương tiện kỹ thuật tiên tiến* (các công cụ giám sát mạng tinh vi, các kỹ thuật đánh giá phù hợp với từng hoàn cảnh cụ thể, …) và năng lực về *tri thức địa phương* dựa trên quan hệ hợp tác giữa các nhà nghiên cứu và chuyên gia trên toàn thế giới.



AE - United Arab Emirates; AZ – Azerbaijan; BH – Bahrain; CN – China; ET – Ethiopia; IR – Iran; JO – Jordan; KR - South Korea; LY – Libya; MM – Burma/Myanmar; OM – Oman; PK – Pakistan; SA - Saudi Arabia; SD – Sudan; SY – Syria; TH – Thailand; TH – Tunisia; UZ – Uzbekistan; VN – Vietnam; YE – Yemen.

Hình 1. 1 Lọc Internet các quốc gia: về bề rộng và chiều sâu [ONI07]

Về nguyên tắc thì cách đặt vấn đề trong phương pháp tiếp cận của ONI là có tính khoa học, đặc biệt về khía cạnh sử dụng các phương tiện kỹ thuật tiên tiến. Tuy nhiên, do nhiều nguyên nhân, đặc biệt là nguyên nhân từ tính không đầy đủ và toàn diện về tri thức địa phương (do một số nhà nghiên cứu và chuyên gia cung cấp cho ONI) đối với lĩnh vực nhạy cảm này. Đối với nhiều nước được ONI khảo sát (trong đó có Trung Quốc và Việt Nam), các nhà nghiên cứu của ONI ít có điều kiện tiếp cận được tri thức địa phương chính thống. Chính vì lý do đó, một số nhận định của ONI đối với không ít quốc gia còn mang tính phiến diện. Nếu bỏ qua một ít nội dung phiến diện nói



Hình 1. 2 Hệ thống lọc nội dung của các nước: theo kiểu nội dung [ONI07]

trên, các kết quả nghiên cứu của ONI là hoàn toàn hữu ích và vì vậy công việc nghiên cứu, khảo sát nội dung các tài liệu được ONI công bố là rất cần thiết.

Trong [ONI07], ONI cung cấp một kết quả tổng quan về các nghiên cứu lọc nội dung trên Internet trên thế giới mà các chuyên gia thuộc tổ chức này đã tiến hành. Nội dung chủ yếu của công trình nghiên cứu này là tổng kết sơ bộ về các kết quả nghiên cứu về an ninh Internet tại hàng chục nước đang phát triển trên thế giới. Lý do ONI không khảo sát hiện trạng lọc nội dung trên Internet đối với nhiều nước phát triển, trong đó có Mỹ, Canada, Úc và một số nước Châu Âu, là ONI coi rằng hoạt động lọc nội dung trên Internet của các quốc gia này coi như đã được biết (hoặc đã được thiết lập hệ thống lọc cấp quốc gia hoặc điều hướng nội dung Internet).

Theo ONI, tồn tại một số hướng tiếp cận khi xem xét chính sách và thực trạng lọc nội dung trên Internet của các quốc gia. Đánh giá hiện trạng lọc nội dung trên Internet của một quốc gia cần được xem xét theo bề rộng và

chiều sâu. *Bề rộng* biểu thị số lượng chủ đề mà chính sách lọc nội dung quan tâm, còn *chiều sâu* biểu thị mức độ kỹ lưỡng ra sao khi xem xét lọc mỗi chủ đề đã được chọn để lọc. Theo hướng tiếp cận này, bốn nước Iran, Trung Quốc, Ả rập Xaudi và Cộng hòa Ả rập thống nhất được xếp vào nhóm các quốc gia có các hệ thống lọc nội dung trên Internet chặt chẽ nhất (rộng và sâu nhất), trong khi đó ba nước Baranh, Azerjian và Thái Lan được xếp vào nhóm các quốc gia có hệ thống lọc nội dung yếu nhất. Theo đánh giá của các tác giả ONI, lọc Internet của Việt Nam tuy không thuộc vào nhóm yếu nhất song cũng được xếp vào nước có giải pháp lọc nội dung ở mức độ thấp cả về bề rộng và chiều sâu. Kết quả đánh giá mức độ khái quát này được trình bày trong Hình 1.1 [ONI07].

Filtering: Theme and Degree

	Political	Social	Conflict & Security	Internet Tools
Azerbaijan	?	–	–	–
Bahrain	? ?	?	–	?
Belarus	?	?	–	–
Burma/Myanmar	? ? ?	? ?	? ?	? ?
China	? ? ?	? ?	? ? ?	? ?
Ethiopia	? ?	?	?	?
India	–	–	?	?
Iran	? ? ?	? ? ?	? ?	? ? ?
Jordan	?	–	–	–
Kazakhstan	?	–	–	–
Libya	? ?	–	–	–
Morocco	–	–	?	?
Oman	–	? ? ?	–	? ?
Pakistan	?	? ?	? ? ?	?
Saudi Arabia	? ?	? ? ?	?	? ?
Singapore	–	?	–	–
South Korea	–	?	? ? ?	–
Sudan	–	? ? ?	–	? ?
Syria	? ? ?	?	?	? ?
Tajikistan	?	–	–	–
Thailand	?	? ?	–	?
Tunisia	? ? ?	? ? ?	?	? ?
Turkmenistan	? ?	–	–	–
United Arab Emirates	?	? ? ?	?	? ?
Uzbekistan	? ?	?	–	?
Vietnam	? ? ?	?	–	? ?
Yemen	?	? ? ?	?	? ?

? ? ? Pervasive filtering; ? ? Substantial filtering; ? Selective filtering;

? Suspected filtering; – no evidence of filtering

Hình 1.3 Hiện trạng lọc nội dung: Chủ đề và phương án lọc [ONI07]

Hướng tiếp cận thứ hai quan tâm tới bốn lĩnh vực chính yếu để đánh giá chính sách lọc nội dung trên Internet của một quốc gia (hay cũng vậy, chính sách lọc nội dung của một quốc gia) là nội dung chính trị, nội dung xã hội, các nội dung liên quan đến xung đột và an ninh và các tiện ích Internet liên quan (*Internet tools*). Như đã giới thiệu, ba nội dung đầu tiên đã được nhận định là tương ứng với ba phạm vi cần quan tâm trong quản lý Nhà nước

về lọc nội dung trên Internet. Kết quả nghiên cứu của ONI được trình bày tại Hình 1.2.

ONI xác định có bốn phương án sau đây mà hệ thống lọc nội dung của các quốc gia tiến hành:

- Lọc tràn lan (pervasive filter),
- Lọc chắc chắn (substantial filter),
- Lọc theo nghi ngờ (suspected filter),
- Không có dấu hiệu của lọc.

Filtering: Other Variations

	Locus	Consistency	Concealed Filtering	Transparency & Accountability
Azerbaijan	D	Low		Medium
Bahrain	D	High	Yes	Low
Burma	D	Low		Medium
China	C & D	Medium	Yes	Low
Ethiopia	C	High	Yes	Low
India	D	Medium		High
Iran	D	Medium		Medium
Jordan	?	High		Low
Libya	C	High		Low
Morocco	C	High	Yes	Low
Oman	C	High		High
Pakistan	C & D	Medium	Yes	High
Saudi Arabia	C	High		High
Singapore	D	High		High
South Korea	D	High		High
Sudan	C	High		High
Syria	D	High		Medium
Tajikistan	D	Low		Medium
Thailand	D	Medium		Medium
Tunisia	C	High	Yes	Low
Turkmenistan	C	High	Yes	Low
United Arab Emirates	D	Low		Medium
Uzbekistan	C & D	High	Yes	Low
Vietnam	D	Low	Yes	Low
Yemen	D	High		Medium

Hình 1.4 Đánh giá hiện trạng lọc nội dung theo các tiêu chí bổ sung [ONI07]

Lọc tràn lan chỉ cách thức mà hệ thống tiến hành việc lọc nội dung đối với mọi trang web, bao gồm hình thức nhất quán (lúc nào cũng lọc đối với mọi trang web) hoặc hình thức ngẫu nhiên (đối với một trang web thì có lúc lọc, có lúc không). *Lọc chắc chắn* là cách thức mà hệ thống lọc nội dung luôn lọc để chặn lại các trang web có nội dung thuộc một trong số kiểu đã được xác định trước. *Lọc theo nghi ngờ* là cách thức mà hệ thống lọc nội dung chặn các trang web thuộc vào một số lớp nội dung nào đó, tuy nhiên, trong đó có các trang web luôn bị chặn song cũng có các trang web chỉ bị chặn lần đầu. *Không có dấu hiệu lọc* dùng để chỉ hiện tượng không thấy có việc chặn trang web. Hình 1.3 cho đánh giá về hiện trạng thi hành các phương án lọc nội dung của các nước tính theo các nhóm chủ đề được quan tâm [ONI07].

Một số chỉ số khác đánh giá hệ thống lọc nội dung của các quốc gia cũng được xem xét bao gồm hoạt động lọc nội dung tập trung ở cấp quốc gia hay phân tán tại các nhà cung cấp dịch vụ Internet - ISP (*locus*), mức độ đồng nhất giữa các phần mềm lọc nội dung tại các ISP (*consistency*), cách hiện thị của hệ thống lọc nội dung (*Concealed Filtering*), mức độ trong suốt của hệ thống lọc nội dung (*Transparency & Accountability*). Về bản chất, tiêu chí trong suốt đã bao hàm tiêu chí về cách hiển thị của hệ thống lọc nội dung.

Hình 1.4 cung cấp kết quả khảo sát của ONI về thực trạng các hệ thống lọc nội dung của một số quốc gia theo các tiêu chí trên đây. Trong hình vẽ, C biểu thị có hệ thống lọc ở cấp quốc gia và D biểu thị có hệ thống lọc tại các ISP. Có ba mức độ nhất quán là cao (*High*), trung bình (*Medium*) và thấp (*Low*). Tương tự, mức độ trong suốt của các hệ thống lọc nội dung cũng có ba mức là cao, trung bình và thấp.Chỉ riêng có Pakistan là quốc gia có hệ thống lọc có mức độ trong suốt cao nhưng lại cho hiển thị về hệ thống lọc nội dung.

1.2. Quản lý Nhà nước về lọc Internet tại Công đồng chung Châu Âu

1.2.1. Về chính sách

Ngày 14/03/2004, Nghị viện Cộng đồng chung Châu Âu (EU) chính thức thông qua quyết định thành lập Cơ quan an toàn thông tin và an toàn mạng toàn Châu Âu (European Network and Information Security Agency[2] - ENISA). Mục đích chính của cơ quan này là tăng cường khả năng của các cộng đồng kinh kế, của các nước thành viên đáp ứng được những vấn đề về an toàn thông tin và an toàn mạng. Do tầm quan trọng của vấn đề nghiên cứu và triển khai các hệ thống lọc nội dung Internet, Cộng đồng chung Châu Âu đã tiến hành nhiều giải pháp theo mức toàn thể cộng đồng và từng quốc gia về vấn đề lọc nội dung trên Internet. Nhiều nghị quyết, văn bản pháp lý của cộng đồng đã nêu ra các khuyến cáo các nước thành viên về việc cần bổ sung các điều khoản pháp lý phù hợp với việc tăng cường chất lượng hoạt động lọc nội dung trên Internet.

Cộng đồng chung Châu Âu đã có nhiều quyết nghị về an toàn - an ninh Internet, trong đó vấn đề lọc nội dung trên Internet được quan tâm đặc biệt [47-60]. Đồng thời, Cộng đồng đã và đang triển khai các chương trình "Safer Internet" trong các giai đoạn 1999-2004, "Safer Internet Plus" 2005-2008 và 2009-2013. Mục dưới đây sẽ trình bày chi tiết về các chương trình giai đoạn của một chương trình hành động dài hạn này.

1.2.2. Các chương trình "Safer Internet"

Hoạt động điển hình nhất của Cộng đồng chung Chấu Âu liên quan tới vấn đề lọc nội dung trên Internet chính là một loạt các quyết nghị triển khai hệ thống các chương trình Internet Safer (1999-2004), Internet Safer Plus (các giai đoạn 2005-2008) và chương trình an toàn Internet đối với trẻ em (2008-2013). Mục đích chung của các chương trình bày là đảm bảo thực thi hành thế

[2] http://www.enisa.eu.int/

chủ động đối với các nội dung trái phép hoặc không thích đáng. Bốn phương châm hành động của dãy dự án này là:

- Tạo ra một môi trường Internet an toàn: khởi tạo một mạng Châu Âu các đường dây nóng mà người dùng cuối có thể thông báo các nội dung bất hợp pháp trên Internet; khuyến khích hoạt động tự điều chỉnh và các chuẩn mực đạo đức.

- Phát triển các hệ thống lọc và phân loại. Phát triển các hệ thống lọc và hệ thống phân loại: giải thích các ích lợi của lọc và sắp xếp, tạo thuận lợi cho việc thỏa thuận trên các hệ thống phân loại.

- Khuyến khích các hoạt động tăng cường nhận thức về an toàn Internet.

Từ các chương trình, dự án được triển khai trên đây và một số chương trình nghiên cứu khác, nhiều dự án về lọc nội dung Internet đã và đang được triển khai tại Cộng đồng chung Châu Âu.

Báo cáo tổng kết giai đoạn 1999-2002 đối với Chương trình "Safer Internet"[3] được tiến hành trong giai đoạn 2003-2004 với kinh phí 14,1 triệu € nhận định:

- Ở mức độ chính sách, Dự án đã thành công khi đưa ra các vấn đề phát triển các tổ chức và công ty an toàn Internet phạm vi Cộng đồng và từng quốc gia.
- Ở mức độ công nghiệp, Dự án đã thành công đưa một số nhà cung cấp dịch vụ Internet (ISP) và cung cấp nội dung Internet (ICP) ngày càng đáp ứng yêu cầu an toàn Internet. Tuy nhiên, vẫn còn một số ít quốc gia thành viên có mức độ tham gia thấp.

[3] http://secretariat.efta.int/EFTASec/Web/EuropeanEconomicArea/programmes/e_Safe_plus/e_Safe

- Ở mức độ chỉ dẫn và tư vấn, Dự án kết nối được các hoạt động thông qua Dự án và các uỷ ban quốc gia. Đã có rất nhiều hoạt động phối hợp với các cơ quan cảnh sát quốc gia và Europol và Interpol.
- Tăng cường liên kết các chuyên gia liên quan tới hoạt động an ninh Internet.

Tiếp tục phát triển các kết quả của chương trình "Safer Internet", vào tháng 12 năm 2004, Cộng đồng chung Châu Âu quyết định tiến hành chương trình "Safer Internet Plus"[4], là kế tục của chương trình "Safer Internet" trong giai đoạn 2005-2008, với tổng kinh phí 45 triệu € được phân bố cho nội dung thực hiện thành phần theo tỷ lệ phân bổ kinh phí như sau:

(1) Đấu tranh chống nội dung vi phạm pháp luật : 25-30 %
(2) Xử trí nội dung không mong muốn và nội dung độc hại : 10-17 %
(3) Nâng cấp môi trường an toàn : 8-12 %
(4) Nâng cao nhận thức : 47-51 %.

Như vậy, qua nội dung đề cập trong các chương trình Safer Internet, chúng ta có thể thấy được tầm quan trọng của công cuộc lành mạnh hoá việc sử dụng Internet và các công nghệ trực tuyến mới, đặc biệt là đối với trẻ em, nhằm chống lại những nội dung không lành mạnh cũng như không mong muốn của người dùng tại cộng đồng chung Châu Âu. Các văn bản pháp lý, các công nghệ cũng như giải pháp về lọc nội dung cũng đã được quan tâm chú trọng cả về nghiên cứu lý thuyết lẫn các hiện thực hoá các kết quả nghiên cứu tại các nước thành viên [EC06c].

Hai mươi bảy quốc gia trong Cộng đồng chung Châu Âu đã tiến hành khảo sát về tình hình sử dụng Internet của trẻ em, phân tích về các khía cạnh có hại có thể thâm nhập đối với trẻ em.

[4] http://secretariat.efta.int/EFTASec/Web/EuropeanEconomicArea/programmes/e_Safe_plus/e_Safe_plus

Khi điện thoại di động đã trở thành phương tiện truy nhập Internet thông dụng thì người ta phải quan tâm tới hình thức truy nhập này trong hoạt động lọc nội dung trên Internet. Theo [EC06c], trong giai đoạn 2003-2004, các nghiên cứu liên quan đã được mở rộng sang các công nghệ trực tuyến mới, bao gồm nội dung không dây và quảng bá, trò chơi trực tuyến, chuyển file ngang hàng và mọi hình thức truyền thông thời gian thực như chat, các gói tin thi hành được nhằm mục đích tăng cường bảo vệ trẻ em.

1.3. Mỹ

1.3.1. Về pháp luật

Trong [Smi05], Marcia S. Smith đã trình bày về một số văn bản pháp lý của nước Mỹ có nội dung về việc đảm bảo an toàn sử dụng Internet đối với trẻ em gồm *1996 Communications Decency Act* (CDA), *1998 Child Online Protection Act* (COPA), *2000 Children's Internet Protection Act* (CIPA), *2002 "Dot Kids" Act* (P.L. 107-317), *2003* "Amber Alert" Act (*The 108th Congress passed the PROTECT (Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today) Act* (P.L. 108-21) and *Pending Amendments*.

Trong các văn bản được công bố về vấn đề lọc nội dung trên Internet của nước Mỹ, người ta thường thu hẹp vào vấn đề an ninh truy nhập Internet đối với trẻ em. Tuy nhiên, ngoài hình thức này, *chính quyền Mỹ cũng sử dụng quyền chủ động để tiến hành các giải pháp kiểm soát Internet khác, trong nhiều trường hợp được tiến hành một cách bí mật*.

Là quốc gia khởi thủy của công nghệ Internet, vấn đề bảo đảm an toàn-an ninh trên mạng Internet của nước Mỹ đã được đề cập đến ngay từ những ngày đầu xuất hiện Internet. Đồng thời với các đạo luật an ninh mạng trong các giao dịch điện tử, vấn đề lọc nội dung Internet, đặc biệt đối với việc truy nhập Internet của trẻ em, được quan tâm rất sớm.

Công trình nghiên cứu của Ủy ban Cố vấn Công nghệ và Bộ Quốc phòng [US04] cung cấp một số mô tả về mối liên quan giữa "tự do cá nhân" đến "đảm bảo an toàn trên Internet".

Trong báo cáo được công bố vào tháng 12-2005, Marcia S. Smith [Smith05] đã tổng hợp và phân tích các văn bản pháp lý điển hình của nước Mỹ về vấn đề lọc Internet đối với trẻ em bao gồm các văn bản the 1996 Communications Decency Act (CDA), the 1998 Child Online Protection Act (COPA), the 2000 Children's Internet Protection Act (CIPA - http://www.ala.org/CIPA/), the 2002 "Dot Kids" Act (P.L. 107-317) và the 2003 "Amber Alert" Act (P.L. 108-21).

Thêm nữa, có tới 21 bang của nước Mỹ bổ sung các luật lọc Internet áp dụng cho các trường phổ thông và thư viện công cộng, bao gồm cả đòi hỏi bắt buộc phải sử dụng các bộ lọc Internet [US07][5]. Hầu hết trong số các bang này đơn giản yêu cầu các Hội đồng nhà trường, các thư viện công cộng cần thực hiện các chính sách bảo vệ trẻ em tránh truy nhập vào các trang web kích dục, các tư liệu không mong muốn. Hai bang Texas và Utah còn có các đạo luật riêng đối với nhà cung cấp dịch vụ Internet hoặc nhà cung cấp máy tính về các điều khoản đảm bảo cơ chế lọc Internet. Hai bang này còn yêu cầu các trường công lập phải khởi động các phần mềm lọc tại các trạm cuối truy nhập thư viện hoặc các máy tính của trường.

Việc sử dụng Internet của trẻ em đã đặt ra cho các bậc phụ huynh thêm nhiều mối quan tâm, lo lắng về các hiểm họa mới, trong đó có hiện tượng trẻ em sử dụng "blog" để đưa nhật ký cá nhân trên mạng[6]. Vì vậy, các gia đình Mỹ đã sử dụng các phương tiện kỹ thuật để đảm bảo an toàn truy nhập Internet cho con em mình. Theo kết quả nghiên cứu của Amanda Lenhart[7], số

[5] http://204.131.235.67/programs/lis/cip/filterlaws.htm
[6] http://www.dantri.com.vn/cong-nghe/2006/2/100239.vip
[7] http://www.pewinternet.org/PPF/a/100/about_staffer.asp

lượng gia đình có trẻ vị thành niên kết nối internet trực tuyến đã sử dụng bộ lọc Internet ngày càng tăng và đạt tới 54% vào tháng 3-2005. Hiện nay, hầu hết các trường học và thư viện trên khắp nước Mỹ đều sử dụng những hệ thống lọc Internet theo những ràng buộc được thể hiện trong luật lọc (National Conference of State Legislatures[8]).

1.3.2. Về chính sách liên bang và các bang

Nhiều sản phẩm phần mềm lọc nội dung đã được công bố và được sử dụng từ rất sớm. Dự án The InFoPeople Project kết thúc vào năm 2001 [Ayr01], cung cấp một cái nhìn tổng quát về hoạt động lọc nội dung trên Internet tại nước Mỹ, đặc biệt đã cung cấp các đánh giá xác đáng về các sản phẩm phần mềm lọc nội dung điển hình như CyberPatrol, i-Gear, i-Prism, N2H2, S4F, SmartFilter, Web Inspector, WebSense, X-Stop.

America Online và Earthlink, thừa nhận rằng họ đang hỗ trợ FBI thu thập những thông tin có liên quan tới vụ tấn công Lầu Năm Góc và Trung tâm Thương mại Thế giới. Tuy nhiên, hai ISP này bác bỏ lời đồn rằng họ đã đồng ý cài đặt hệ thống do thám Carnivore[9]. Nhiều điều khoản trong Luật yêu nước (*Act Patriot*) của Mỹ cho phép nghe lén điện thoại, điều tra hồ sơ cá nhân và đọc email của công dân. FBI được quyền theo dõi người dân đọc gì bằng cách kiểm tra liệt kê các đầu sách họ mượn tại thư viện...[10]. Giới chuyên môn gọi những máy tính bị kiểm soát từ xa và phát tán thư rác là các *máy tính Zombie* (sống dở chết dở). Theo dự kiến từ tình trạng chỉ mới xuất hiện tình trạng các máy tính Zombie đơn lẻ đã phát triển nhanh thành các mạng các máy tính

[8] http://www.ncsl.org/programs/lis/CIP/filterlaws.htm

[9] vietnamnet.vn/thegioi/2004/12/354698/

[10] http://www.tuoitre.com.vn/Tianyon/Index.aspx?ArticleID=22328&ChannelID=2

Zombie. Bằng cách này, các spammer sẽ tiết kiệm được nhiều chi phí và ẩn danh tốt hơn[11].

1.4. Trung Quốc

Với số dân trên một tỷ người và nền kinh tế tăng trưởng với tốc độ cao nhất thế giới, việc sử dụng Internet của Trung Quốc trong khoảng mười năm gần đây đã tăng với tốc độ rất cao. Hiện nay Trung Quốc được coi là một trong những nước đi đầu trong việc bảo đảm chặt chẽ an toàn-an ninh quốc gia trên Internet. Kết quả nghiên cứu của ONI [ONI07, ONIChina] và các nghiên cứu bổ sung của nhóm đề tài về nội dung này đã khẳng định nhận định trên đây.

1.4.1. Nghiên cứu của ONI

Như đã trình bày ở phần trên, ONI đánh giá hệ thống lọc nội dung trên Internet tại Trung Quốc vào loại chặt chẽ nhất trong các quốc gia được ONI xem xét [ONI07, ONIChina]. Một mặt, các nghiên cứu của ONI về chính sách lọc nội dung trên Internet tại Trung Quốc căn cứ vào nội dung các chính sách chính thống do Chính phủ Trung Quốc công bố, và mặt khác, lại dựa trên quan điểm riêng của một số cá nhân[12] hoặc Tổ chức về nhân quyền[13] về vấn đề này. Chính vì lý do đó, một số luận điểm của ONI về chính sách của Chính phủ Trung Quốc không phải luôn luôn có tính khách quan.

Theo các nghiên cứu nói trên, chính quyền quan tâm tới một phạm vi rộng các vấn đề nhạy cảm và gây tranh cãi (bề rộng).

1.4.2. Các nghiên cứu khác

Với mục đích đảm bảo an toàn – an ninh cho mọi luồng thông tin công cộng tại nước này, Chính phủ Trung Quốc đã thành lập tổ chức đa liên ngành

[11] *Spam, vẫn là vấn đề "nhức đầu" trong năm 2006*. http://www.tuoitre.com.vn/Tianyon/Index.aspx?ArticleID=120670&ChannelID=16, Thứ Sáu, 27/01/2006, 21:58 (GMT+7)
[12] He Qinglian, *Media Control in China*, http://www.hrichina.org/public/contents/8991 (Nov. 4, 2004).
[13] Human Rights Watch, *World Report 2005: China,* January

có tên gọi General Administration of Press and Publication (GAPP) chịu trách nhiệm theo nhiều chức năng liên quan. Về mặt pháp lý, ngay từ năm 1994, Chính phủ Trung Quốc đã cho phép Bộ An ninh công cộng (the Minister of Public Security) có trách nhiệm toàn diện đối với việc giám sát Internet. Trong thực tế, Trung Quốc đã thi hành việc điều khiển truy nhập đối với các ISP, ICP, người đăng ký Internet và người sử dụng cafe-internet.

Trung Quốc đã đạt được thành công trong việc thi hành nội dung các văn bản pháp lý nói trên [ONIChina]. Hiện nay, hệ thống quản lý lọc Internet ở nước này được coi là phức tạp nhất thế giới, được phân bố thành nhiều cấp tinh vi và hiệu quả [ONIChina, China02]. Hệ thống này kiểm duyệt nội dung được truyền tải qua Internet đối với hầu hết phương thức khác nhau, bao gồm trang Web, Web blog, các diễn đàn thảo luận trực tuyến (online forum), thư điện tử, v.v., với rất nhiều những kỹ thuật và quy tắc luật lọc nội dung khác nhau.

Các giải pháp lọc nội dung đối với các máy tìm kiếm rất được quan tâm ở Trung Quốc. Hầu hết các máy tìm kiếm thông tin đều được kiểm duyệt nội dung kết quả tìm kiếm trước khi trả về cho người dùng tại Trung Quốc. Đối với các máy tìm kiếm của Trung Quốc (chẳng hạn như www.baidu.com hay www.yisou.com), [ONIChina] cho chúng ta thấy ngay tại các máy tìm kiếm này, việc lọc nội dung đã được cài đặt thông qua từ khoá và thường bỏ đi một phần danh sách kết quả tìm kiếm. Hơn thế nữa, các kết quả tìm kiếm của các máy tìm kiếm còn được kiểm duyệt một lần nữa tại các cổng Internet quốc gia và tại đây, các kết quả tìm kiếm có thể bị cắt bỏ đi thêm một lần nữa. Những nghiên cứu của ONI cho thấy việc lọc nội dung chưa được tiến hành ở quá trình thu thập thông tin (crawler) – các thông tin nhạy cảm vẫn được thu thập và đánh chỉ mục. Ngoài ra, khi một yêu cầu tìm kiếm bị cấm, các máy tìm kiếm này sẽ kết thúc kết nối với người yêu cầu thông qua việc gửi một gói tin

TCP RST (connection reset) và theo sau là một TCP ZeroWindow size[14]. Kết quả là người dùng sẽ không thể gửi thêm một yêu cầu nào nữa đến các máy tìm kiếm đó trong một khoảng thời gian nhất định.

Đối với các máy tìm kiếm đặt ngoài Trung Quốc, ví dụ như www.google.com, thì các kết quả tìm kiếm của người dùng tại Trung Quốc đều qua bộ lọc nội dung tại cổng Internet quốc gia và được cắt bỏ đi những mục nhạy cảm và không lành mạnh. Hơn thế nữa, các dữ liệu được cache tại máy tìm kiếm của google đều bị cấm không cho phép truy cập tại nước này (thông qua việc kiểm soát yêu cầu HTTP GET với xâu « search?q=cache »).
Như vậy ta có thể thấy rằng các nhận xét và đánh giá về chính sách, kỹ thuật, ngữ cảnh pháp luật cũng như phương pháp luận lọc nội dung trên Internet tại Trung Quốc [China02, ONIChina] là rất có giá trị trong việc lựa chọn phương án xây dựng hệ thống lọc nội dung trên Internet tại Việt Nam. Hơn nữa, dù rằng vấn đề an toàn – an ninh trên mạng thường mang yếu tố bí mật quốc gia, song nếu thiết đặt được quan hệ với các cơ sở đào tạo, nghiên cứu của Trung Quốc, tổ chức các hội thảo khoa học về nội dung này thì chúng ta có thể tiếp thu được kinh nghiệm từ Trung Quốc.

1.5. Một số nước khác

1.5.1. Một số nước phát triển

Tháng 7/1999, thông qua hoạt động của CSIRO (Commonwealth Scientific and Industrial Research Organisation), Chính phủ Úc đã thông qua điều lệ dịch vụ quảng bá (Broadcasting Services Act) cho phép kiểm soát nội dung thông tin truyền tải trên Internet. Với luật này, từ 01/01/2000, người dân có thể khiếu nại về nội dung trực tuyến trên Internet của một tổ chức, cá nhân,... tại cơ quan ABA - Australian Broadcasting Authority. Vào tháng 9-2001, các kết quả khảo sát đánh giá về một số sản phẩm lọc Internet hoạt

[14] Tham khảo thêm tại trang http://www.ncsa.uiuc.edu/People/vwelch/net_perf/tcp_windows.html

động tại Úc đã được tổng hợp trong [48GRT01]. Các tiêu chí điển hình nhất để đánh giá sản phẩm lọc Internet đã được trình bày trong công trình nghiên cứu này.

1.5.2. Một số nước đang phát triển

Các nghiên cứu của The OpenNet Initiative về lọc Internet tại các nước Singapore, Tunisia, Ả rập Saudi và Iran được trình bày trong các tài liệu [ONI05Sin, ONI04, ONI05, ONI05a, ONI05b]. Các nghiên cứu này cho thấy việc triển khai vấn đề lọc nội dung trên Internet của các nước là đa dạng nhằm mục đích phù hợp với an toàn – an ninh quốc gia.

Vào tháng 3/2008, Nhà nước Indonesia ban hành điều luật về việc *"bất cứ ai bị kết tội truyền bá trái phép văn hóa phẩm đồi trụy, tung tin thất thiệt, hoặc phát tán các thông điệp gây chia rẻ sắc tộc, tôn giáo trên mạng Internet sẽ phải đối mặt với án tù 6 năm, và khoản tiền phạt 100.000USD"*[15].

1.6. Quản lý Nhà nước Việt Nam về lọc nội dung trên Internet

1.6.1. Chính sách Nhà nước

Nhà trường ta chủ trương khuyến khích mọi tổ chức, cá nhân khai thác tốt nhất các lợi ích tiềm tàng từ Internet vào việc hình thành các thành phần trong nền kinh tế tri thức, đóng góp vào công cuộc hiện đại hóa đất nước. Nhà nước rất quan tâm phát triển công nghiệp nội dung số, mà nội dung số trên Internet được coi như một thành phần quan trọng của ngành công nghiệp này. Chương trình phát triển công nghiệp nội dung số Việt Nam đến năm 2010, được phê duyệt theo Quyết định số 56/2007/QĐ-TTg ngày 03 tháng 5 năm 2007 của Thủ tướng Chính phủ, đặt ra mục tiêu tổng quát là "*phát triển công nghiệp nội dung số thành một ngành kinh tế trọng điểm, đóng góp ngày càng nhiều cho GDP, tạo điều kiện thuận lợi cho các tầng lớp nhân dân tiếp cận*

[15] http://www.vnmedia.vn/newsdetail.asp?NewsId=123867&CatId=34

các sản phẩm nội dung thông tin số, thúc đẩy mạnh mẽ sự hình thành và phát triển xã hội thông tin và kinh tế tri thức" [VN56-07].

Vấn đề an toàn – an ninh Internet, bao gồm cả vấn đề lọc nội dung, nhận được sự quan tâm đặc biệt. Nghị định số 55/2001/NĐ-CP ngày 23-8-2001 của Chính phủ về *Quản lý, cung cấp và sử dụng dịch vụ Internet* đã đề cập về vấn đề này (Khoản 2 Điều 2; Điều 6; Khoản 3 Điều 11; Khoản 3 Điều 18; Khoản 8 Điều 28; Điều 33; Điều 35; Mục 2e, Mục 5e, Mục 5g Điều 41 và Điều 45). Những nội dung hoạt động quản lý Nhà nước về Internet (Điều 11, các khoản 7-8-9 Nghị định 55 [VN55-01]...)...

Nhà nước ta nghiêm cấm "*Lợi dụng Internet để chống lại nhà nước Cộng hoà xã hội chủ nghĩa Việt Nam; gây rối loạn an ninh, trật tự; vi phạm đạo đức, thuần phong, mỹ tục và các vi phạm pháp luật khác"* (Điều 11, khoản 3, Nghị định 55 [VN55-01])

Một số nội dung chi tiết hơn được thể hiện trong *Quy định về biện pháp và trang thiết bị kiểm tra, kiểm soát đảm bảo an ninh quốc gia trong hoạt động Internet ở Việt Nam* của Bộ Nội vụ được ban hành kèm theo Quyết định số 848/1997/QĐ-BNV(A11) ngày 23.10.1997 (Mục 2 Khoản 3 Điều 5, Khoản 3 Điều 6). *Quy định về Đảm bảo an toàn, an ninh trong hoạt động quản lý, cung cấp, sử dụng Internet tại Việt Nam* được ban hành kèm theo Quyết định số 71/2004/QĐ-BCA (A11) ngày 29 tháng 1 năm 2004 của Bộ trưởng Bộ Công an quy định toàn diện và chi tiết về các nội dung đảm bảo an toàn – an ninh trên Internet của Nhà nước ta.

1.6.2. Nghiên cứu của ONI

Như đã được giới thiệu, ONI [ONI07] đã cung cấp khái quát về hiện trạng lọc nội dung trên Internet của một số nước trên thế giới, trong đó có Việt Nam. Kết quả nghiên cứu này một cách xem xét về hiện trạng lọc nội dung của nước ta khi đối sánh với các nước khác trên thế giới (Hình 4). *Hệ*

thống lọc nội dung Internet của Việt Nam thuộc vào loại yếu, mới được thực hiện tại một số ISP, có độ nhất quán thấp, không trong suốt, chưa quan tâm đúng mức tới các nội dung xã hội, chưa rõ ràng về việc ngăn chặn các trang web xung đột và an ninh... Điều đó cho thấy việc nghiên cứu đề ra các giải pháp nâng cấp hệ thống lọc nội dung ở mức quốc gia là cần thiết.

Một số ví dụ kiểm chứng về mức độ yếu lọc nội dung xã hội đã được trình bày, bao gồm một thực tế là có hơn 90 % người dùng Internet trẻ (90%) đã truy nhập các trang web tình dục xấu. Kết quả nghiên cứu này cũng phù hợp với một thống kê của Google cho thấy Việt Nam là một trong 10 nước tìm kiếm các thông tin liên quan tới tình dục "sex" nhiều nhất trên thế giới (số câu lệnh tìm kiếm chứa "sex" xuất phát từ các địa chỉ IP Việt Nam là nhiều nhất)[16]. Ngoài ra, nội dung của một số tin trong chuyên mục xã hội hoặc pháp luật với nội dung tình dục trên một số báo điện tử có sự nhập nhằng với nội dung tình dục.

Từ thực tế hệ thống lọc nội dung mới chỉ triển khai ở mức độ nhà cung cấp dịch vụ Internet cho nên, nghiên cứu của ONI mới quan tâm xem xét vấn đề này đối với hai nhà cung cấp dịch vụ Internet là FPT và VNPT. Theo đánh giá của ONI, hoạt động lọc nội dung của FPT có một số điểm tiến bộ hơn so với VNPT. Đồng thời với nhận xét đánh giá hệ thống lọc nội dung trên Internet của nước ta ở trình độ thấp, chưa đáp ứng các yêu cầu của Nhà nước song ONI cũng nhận định rằng kỹ thuật, chiều sâu và tính hiệu quả của các hệ thống lọc này ngày càng được nâng cao theo thời gian.

Về thư điện tử, theo thống kê của hãng bảo mật Symantec, năm 2008, có khoảng 349,6 tỷ thư rác được gửi đi trên khắp thế giới, chiếm đến 94% tổng lượng email được gửi đi và 90% trong số đó là xuất phát từ những chiếc máy tính bị hacker chiếm mất quyền quản trị và bị kiểm soát, điều khiển từ xa

[16] http://www.tienphongonline.com.vn/Tianyon/Index.aspx?ArticleID=99729&ChannelID=46. "10 nước tìm 'sex' trên Google nhiều nhất thế giới"

để gửi thư rác. Thế giới có khoảng 9,4 triệu chiếc máy tính dạng này và gần như toàn bộ chủ nhân của những chiếc máy tính ấy không hề biết rằng họ đã bị mất quyền quản trị. Theo một báo cáo khác của Google cho biết số lượng trung bình thư rác trong quý hai năm 2009 cao hơn quý một 53%.

Cuộc chiến chống lại nạn thư rác nhiều khi khiến các công ty Internet "khốn khổ". Có rất ít công ty có thể định lượng được họ đã tốn bao nhiêu tiền cho cuộc chiến này. Thường thì một ISP lớn phải tuyển dụng từ 5-10 nhân viên chuyên trách việc theo dõi spam. Thêm vào đó là các cơ sở vật chất như máy chủ, bộ định tuyến, bộ lọc thư rác, phần mềm ngăn chặn tấn công web, thành lập các trung tâm trợ giúp khách hàng và nhiều loại thiết bị khác để có thể "tải" nổi lượng thư rác ngày càng khổng lồ.

Tại Hội thảo chống thư rác qua hệ thống email ngày 19/03/2009, VNCERT (Trung tâm Ứng cứu khẩn cấp máy tính Việt Nam) trình bày khảo sát mới đây nhất, lấy ý kiến của hơn 250 người thường xuyên sử dụng thư điện tử. Với 10 câu hỏi khảo sát về thư rác bao gồm các nội dung: Thống kê tình hình sử dụng email, tỷ lệ thư rác và tỷ lệ thư rác có nội dung tiếng Việt; Hành vi, cảm nhận của người dùng đối với thư rác; Các biện pháp ngăn ngừa thư rác; Mong muốn về định hướng bảo vệ người dùng cuối của Nghị định về thư rác. Kết quả cuộc khảo sát cho thấy trong số 250 người thì 53% nhận dưới 15 email rác/ngày, 43% nhận trên 30 email/ngày. Thư rác tiếng Việt hiện đang chiếm khoảng 20-50% lượng thư rác ở Việt Nam. Có đến 90% số lượng người dùng email rất khó chịu đối với thư rác. Trong phần kiến nghị có 63% người dùng muốn có quy định quản lý thư rác; 64,7% người dùng thấy cần có đơn vị chuyên trách; 56% thấy cần nâng cao nhận thức cộng đồng về thư điện tử.

Hiện nay, có nhiều thách thức đối với việc chống thư rác. Việc gửi thư rác có thể thực hiện dễ dàng, nặc danh và rất khó khống chế hoàn toàn bằng kỹ thuật. Bên cạnh đó, kiến thức người dùng cuối còn thấp, lại chưa có các bộ

phận tiếp nhận báo cáo và xử lý thư rác. Về mặt luật pháp vẫn thiếu chế tài hoặc chế tài quá nhẹ đối với các đối tượng vi phạm. Trong khi đó, các tổ chức cung cấp dịch vụ email chưa đảm bảo hạ tầng kiểm soát SPAM hiệu quả; Người dùng, nhà cung cấp dịch vụ email thờ ơ với SPAM; Số địa chỉ IP bị liệt vào Blacklist gia tăng; Chưa có tổ chức chống SPAM chuyên nghiệp

1.6.3. Tình hình phát triển Internet và vấn đề web độc hại

Việc phát triển mạnh Internet ở Việt Nam sau khi Nghị định 55/2001/NĐ-CP ra đời đã khẳng định lộ trình phổ cập và xã hội hoá Internet của nhà nước đã đi đúng hướng. Nhờ những chính sách quản lý hợp lý và phù hợp với xu thế phát triển với quan điểm rất mới "quản lý phải theo kịp yêu cầu của sự phát triển", Internet Việt Nam đã phát triển với tốc độ chóng mặt, giá cước Internet liên tục được giảm xuống, số lượng người sử dụng Internet gia tăng ngày càng nhanh.

Theo số liệu thống kê của Trung tâm Thông tin mạng Internet Việt Nam (VNNIC), tính đến hết tháng 04/2008, số lượng thuê bao Internet Việt Nam đã đạt con số 5,6 triệu thuê bao, khoảng 19,5 triệu người sử dụng Internet, đạt mật độ 23,12%. Mật độ người dùng Internet tại Việt Nam đã vượt mật độ trung bình của khu vực châu Á là 10,33%, và vượt trên cả mật độ trung bình của thế giới là 16,02%. Tuy nhiên song hành với sự phát triển của Internet là những vấn đề tiêu cực: các hoạt động phạm pháp trên Internet có ảnh hưởng xấu đến an ninh trật tự, an ninh quốc gia, vi phạm luật pháp và văn hóa Việt nam ngày càng gia tăng.

Do đó, để kiểm soát tốt hơn nữa tình trạng này, cần phải kết hợp các giải pháp kỹ thuật, hành chính, giáo dục ý thức... Trong đó, giải pháp kỹ thuật được coi là giải pháp trước nhất. Các ISP là các đơn vị có trách nhiệm ở đây. Nhiệm vụ của ISP chủ yếu là giải pháp kỹ thuật. Nhiều ý kiến cho rằng cần phát triển mạnh hơn nữa các hệ thống ngăn chặn web đen ở ngay cổng

Internet quốc gia, yêu cầu các nhà cung cấp dịch vụ đường truyền (IXP) và cung cấp dịch vụ Internet (ISP) thiết lập hệ thống tường lửa hữu hiệu. Tuy nhiên, giải pháp này vẫn không đủ để phong tỏa những web đen, vốn cực kỳ linh động (thường xuyên đổi địa chỉ tên miền và địa chỉ IP..., vượt qua tường lửa). Mặt khác, việc chặn từ cổng quốc gia - nơi tập trung lưu lượng thông tin khổng lồ qua lại - gây ảnh hưởng đến tốc độ của hệ thống, đòi hỏi phải đầu tư nâng cấp trang thiết bị.

1.6.4. Tình hình tại các điểm truy cập Internet công cộng

Các đại lý Internet công cộng đã đóng góp một vai trò rất quan trọng trong việc góp phần phổ cập Internet rộng rãi tới cộng đồng xã hội. Tuy nhiên, bên cạnh vai trò lớn và yếu tố tích cực nêu trên, hoạt động của các đại lý Internet đã bộc lộ một số mặt tiêu cực như liên quan đến tội phạm mạng, để xảy ra tình trạng ăn cắp mật khẩu, tài khoản truy nhập, sử dụng Internet để tiến hành các hoạt động tội phạm như buôn bán ma túy, mại dâm, các hành vi vi phạm trật tự xã hội. Chẳng hạn như tại thành phố Đà Nẵng, qua thanh kiểm tra, các cơ quan chức năng đã nhận thấy phần lớn các đại lý Internet tại đây đều có sai phạm (mức độ sai phạm và hành vi sai phạm có khác nhau). Hoặc tại Hà Nội, thanh tra Sở Văn hoá thông tin (VHTT) đã tiến hành các đợt thanh kiểm tra các đại lý Internet. Kết quả cho thấy, những vi phạm chủ yếu của các đại lý Internet là để khách hàng truy nhập vào những nội dung đồi trụy trên mạng. Những hành vi vi phạm này đã và đang gây ảnh hưởng nghiêm trọng đến tầng lớp thanh thiếu niên, đặc biệt là trẻ vị thành niên, khiến dư luận hết sức bức xúc. Trong thời gian vừa qua, công tác quản lý Internet của Việt Nam đã đạt được nhiều kết quả đáng ghi nhận, nhưng do sự phát triển quá nhanh của Internet nên năng lực quản lý của các cơ quan hiện vẫn chưa theo kịp yêu cầu.

1.6.5. Hoạt động của cơ quan quản lý nhà nước về vấn đề chống truy cập web độc hại

Các cơ quan chức năng rất quan tâm đến tình hình quản lý nội dung truy cập Internet, liên tục có những qui định về quản lý truy cập Internet. Cụ thể hóa sự quan tâm đó là Thông tư 02 (02/2005/TTLT-BCVT-VHTT-CA-KHĐT), có qui định "Quyền và nghĩa vụ của đại lý Internet": các đại lý cần cài đặt chương trình phần mềm quản lý đồng thời thực hiện các giải pháp kỹ thuật đảm bảo ngăn chặn người sử dụng truy cập đến các trang web có nội dung xấu trên Internet. Đại lý Internet chỉ được cung cấp nội dung thông tin về người sử dụng cho các cơ quan Nhà nước có thẩm quyền.

Thông tư liên tịch số 02/2005/TTLT-BCVT-VHTT-CA-KHĐT về quản lý đại lý Internet ra đời tạo điều kiện thúc đẩy phát triển đại lý Internet theo đúng quy định của pháp luật, hướng dẫn tăng cường quản lý việc phát hành, khai thác, sử dụng thông tin qua Internet và các hoạt động khác theo quy định của pháp luật, ngoài ra giúp ngăn ngừa hoạt động vi phạm pháp luật thông qua việc sử dụng dịch vụ Internet tại các đại lý Internet công cộng. Nhiệm vụ đặt ra ở đây cho các cơ quan quản lý là đảm bảo sự phát triển, phát huy tối đa hiệu quả của Internet, mang tri thức nhân loại phụ vụ cho sự nghiệp công nghiệp hóa, hiện đại hóa đất nước, đồng thời hạn chế thấp nhất những ảnh hưởng tiêu cực của Internet. Văn bản này khi đi vào cuộc sống sẽ có tính hiệu lực và hiệu quả quản lý đại lý Internet rất cao. Tuy nhiên với sự phát triển của Internet Việt Nam việc giám sát hoạt động của các đại lý Internet nếu không có công cụ kỹ thuật hiệu quả thì rất khó khăn.

1.6.6. Vấn đề lọc chặn tại các ISP

Hiện tại do tốc độ phát triển nhanh của Internet với nhiều kết nối băng rộng, nhu cầu đường truyền quốc tế gia tăng, hệ thống Firewall Gateway của các ISP không đáp ứng được và thường xuyên bị quá tải dẫn tới bỏ qua việc lọc chặn nhiều trang web độc hại. Vấn đề xử lý hiện tượng web ‘đen’ bị lọt

lưới là yêu cầu bức xúc và là mối quan tâm của toàn xã hội và một phần cũng là trách nhiệm của nhà cung cấp dịch vụ Internet.

Trên thị trường cũng có nhiều sản phẩm hỗ trợ người dùng cho mục đích ngăn chặn truy cập Internet độc hại. Một số sản phẩm là của nước ngoài, một số là trong nước đều có nhược điểm là: hoặc là giá cả quá cao và không phù hợp với đặc thù Việt Nam, hoặc chức năng lọc chặn không hoàn thiện, hoặc không có một chiến lược phát triển, hỗ trợ và duy trì, cập nhật thường xuyên dẫn tới nhiều bất tiện cho người dùng.

Cùng với sự phát triển Internet Việt Nam, nhu cầu phải có một giải pháp phần mềm hỗ trợ cho các vị phụ huynh bảo vệ con em mình khỏi các thông tin độc hại, giúp cho các điểm Internet công cộng cấm các truy cập vào trang web xấu, … là rất bức thiết. Ngay cả khi firewall gateway được nâng cấp với năng lực mạnh hơn hiện nay, giải pháp này cũng rất cần thiết nó có thể giúp hỗ trợ lọc chặn các trang web xấu còn lại mà firewall gateway bỏ sót hoặc chưa cập nhật kịp thời.

Hệ thống phần mềm ngăn chặn người dùng truy cập thông tin độc hại với giải pháp quy mô, được đầu tư, xây dựng và duy trì, cập nhật một cách liên tục sẽ có hiệu quả áp dụng vào thực tế rất lớn. Góp phần giải quyết bức xúc của xã hội về vấn nạn web đen, thể hiện trách nhiệm của nhà cung cấp dịch vụ Internet đối với người dùng, nâng cao uy tín và thương hiệu Internet Việt Nam.

CHƯƠNG II

CƠ SỞ LÝ THUYẾT VÀ CÁC GIẢI THUẬT LỌC NỘI DUNG

2.1. KHÁI NIỆM CƠ BẢN

2.1.1. Một số khái niệm về lọc thông tin trên Internet

Lọc nội dung Internet là một thuật ngữ được dùng để chỉ các kỹ thuật kiểm soát thông tin trên Internet thông qua việc phân tích nội dung thông tin yêu cầu để cho hoặc không cho người sử dụng Internet nhận được kết quả trả về từ Internet hoặc gửi thông tin lên mạng Internet. Tùy theo yêu cầu và ngữ cảnh, người ta có thể sử dụng kết hợp một hoặc nhiều *cách tiếp cận* sau đây:

- **Lọc cụ thể** (inclusion filtering): người dùng chỉ được phép truy cập những thông tin đã được cho phép, nằm trong một danh sách được hiểu theo nghĩa "***danh sách trắng***", thông thường là một danh sách các địa chỉ web được phép truy nhập. Những thông tin nằm ngoài danh sách này đều không được phép truy cập.

  Do tính toàn cầu của Internet, rất khó có thể có một tập điều kiện được chấp nhận rộng rãi. Hầu hết nếu không muốn nói rằng tất cả các nền văn hóa đều đồng thuận rằng những nội dung như lạm dụng tình dục trẻ em nên bị cấm với không chỉ trẻ vị thành niên mà cả người lớn. Tuy nhiên các nền văn hóa khác nhau có những cách nhìn nhận khác nhau. Một vài nhà cung cấp còn cho phép người dùng tùy biến danh sách trắng của họ.

  Điểm hạn chế nhất của lọc bao gồm là ta cần có một danh sách dài các trang web do số lượng thông tin trên Internet. Tại thời điểm tháng 3 năm 1999, số lượng trang Web có thể đánh chỉ số đã lên tới 800 tỉ trang, nén vào khoảng 15 terabytes thông tin hay khoảng 6 terabytes sau khi đã loại bỏ các thẻ HTML, các lời bình luận hoặc khoảng trắng. Vì số lượng trang tốt bao giờ cũng nhiều hơn các trang xấu, vì thế danh

sách bao gồm sẽ lớn hơn danh sách loại trừ được xem xét trong phần sau.

Yaholigans là một trong số những "danh sách bao gồm" điển hình đã được công bố với 3000 địa chỉ web – một phần rất nhỏ của Internet "thực"

- **Lọc loại trừ** (exclusion filtering): người dùng sẽ bị chặn luồng thông tin nằm trong một danh sách, gọi là "***danh sách đen***", thông thường là một danh sách các địa chỉ web không được phép truy nhập. Tất cả những thông tin không liên quan đến danh sách này đều được phép truy cập.

  Vấn đề chính với lọc loại trừ là nó phụ thuộc vào các danh sách đen – các danh sách này cần có độ phủ rộng và cập nhật thường xuyên. Một vài danh sách đen như danh sách Bess của N2H2 có độ phủ lớn với 130000 trang trên 40000 máy chủ. Hầu hết những nhà cung cấp hỗ trợ các sản phẩm lọc loại trừ trên cơ sở những nỗ lực tối đa cho một độ bao phủ rộng.
- **Lọc dựa theo phân tích nội dung**: hạn chế và ngăn chặn người dùng những thông tin chứa những ***nội dung cấm*** theo những tiêu chuẩn đã được đề ra.

Chúng ta có thể nhận thấy rằng hai cách tiếp cận đầu cho khả năng thi hành đơn giản nếu cho trước một danh sách trắng hoặc một danh sách đen. Tuy nhiên, trong thực tế thì khó khăn gặp phải chính là bài toán xác định chính xác các danh sách như vậy và luôn đưa đến một kết quả hoặc là lọc không đầy đủ (xuất hiện liên tục các trang web "đen" mới trên Internet) hoặc hạn chế miền truy cập thông tin Internet (danh sách "trắng"quá hạn chế, không tương thích với sự tăng trưởng không ngừng của Internet). Cách tiếp cận lọc thông qua phân tích nội dung được tiến hành nhờ các kỹ thuật "hiểu" nội dung của thông tin trên web để ngăn chặn các thông tin có nội dung xấu.

Công việc "hiểu" và đánh giá thông tin được tải về cho phép việc lọc Internet có tính công phu và hoàn hảo hơn nhưng lại đòi hỏi khối lượng tính toán lớn để xem xét từng nội dung được tải về hoặc gửi đi. Tuy nhiên, do tính chất công phu của cách tiếp cận lọc nội dung và sự tăng trưởng không ngừng về năng lực tính toán mà cách tiếp cận lọc thông qua phân tích nội dung ngày càng được phát triển mạnh.

2.1.2. Phân loại quy mô lọc thông tin

Việc triển khai các hệ thống lọc thông tin phụ thuộc rất nhiều vào ngữ cảnh và vị trí tiến hành. Tùy thuộc theo những yêu cầu cụ thể mà mức độ lọc thông tin cũng như quy mô của nó có thể được cài đặt khác nhau. Chúng ta có thể chia làm *ba mức lọc thông tin* chính sau:

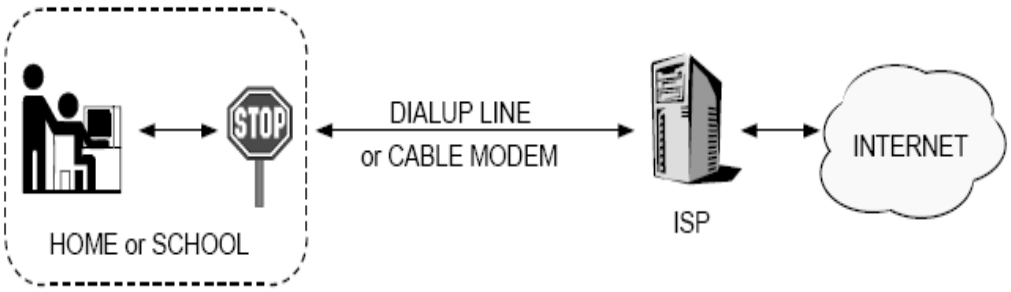
- *Mức cục bộ*: mức này được thể hiện thông qua các phần mềm cài đặt trong các máy tính cá nhân với một mục đích sử dụng trong một phạm vi nhỏ như gia đình, công ty có quy mô nhỏ v.v. (bộ lọc mức này được gọi là *client-based filter*).
- *Mức tổ chức*: mức này cần đến những giải pháp lọc nội dung cho một mạng cỡ vừa, ví dụ như một mạng intranet trong một cơ quan, một trường học, một công ty cỡ lớn, v.v. (bộ lọc mức này được gọi là *server-based filter*).
- *Mức quốc gia*: yêu cầu ở mức này đòi hỏi rất nhiều yếu tố khác nhau về công nghệ và kỹ thuật để đạt được khả năng lọc nội dung ở mạng xương sống (backbone) của việc truy cập Internet của cả một quốc gia.

Tuỳ theo từng mức lọc thông tin nêu trên, việc lọc thông tin sẽ được thực hiện tại những điểm sau:

2.1.2.1. Lọc tại máy tính của người dùng cuối

Người dùng cuối có thể được hỗ trợ bởi các phần mềm trên máy tính riêng của họ, phần mềm này sẽ kiểm tra những yêu cầu của người dùng theo

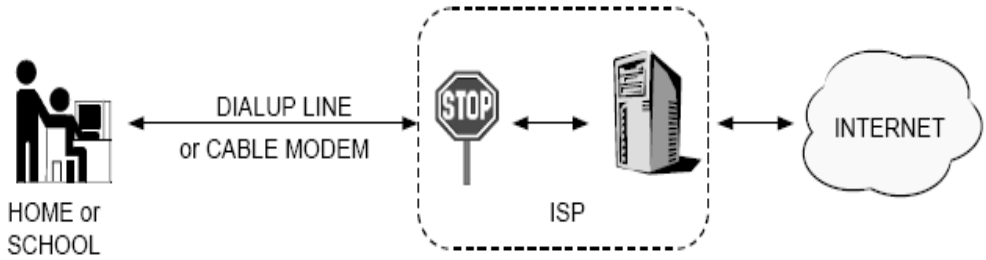
một danh sách chặn hay danh sách trắng được cho trước. Yêu cầu của người dùng hoặc sẽ được cho phép hoặc sẽ bị chặn. Quá trình này được mô tả như trong Hình 2.1. Người dùng có trách nhiệm cập nhật danh sách định kì từ nhà cung cấp phần mềm.



Hình 2. 1. Lọc thông tin tại máy tính của người dùng

2.1.2.2. Lọc thông tin tại ISP

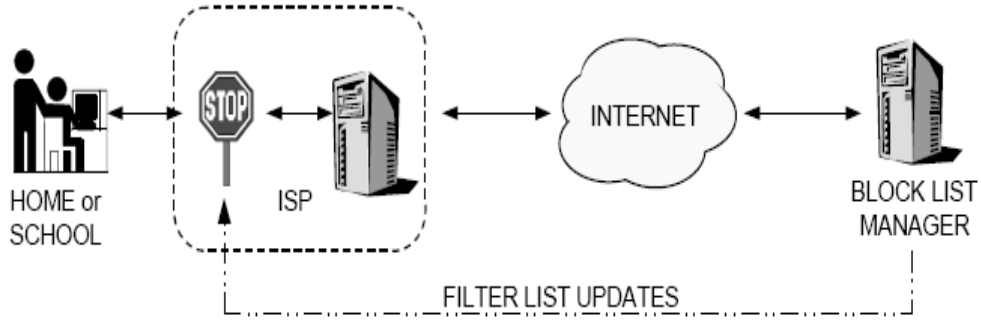
Với loại lọc thông tin này, yêu cầu của người dùng được kiểm tra bởi ISP, bằng cách so sánh nó với một danh sách chặn, như được mô hình trong Hình 2.2.



Hình 2. 2. Lọc thông tin tại ISP

Công ty xây dựng phần mềm lọc cung cấp danh sách lọc cho ISPs nhưng đôi khi những danh sách lọc này có thể được cung cấp bởi bên thứ ba (Hình 2.3). Dù với mô hình cung cấp nào thì danh sách lọc đó cũng phải được cập nhật thường xuyên – lý tưởng nhất là cập nhật hàng ngày – để phát hiện

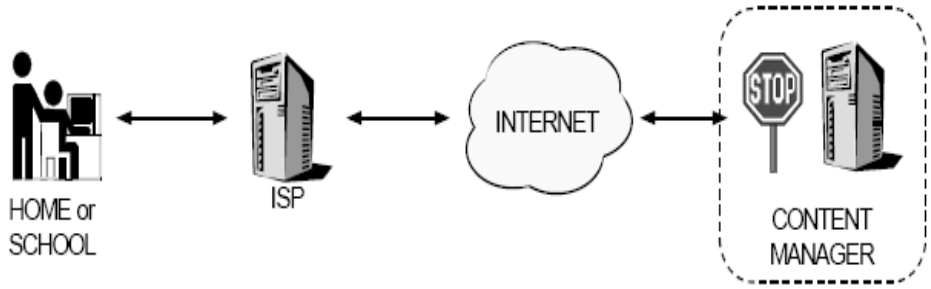
sự xuất hiện của các trang mới hoặc trang mới được đổi tên. Các danh sách lọc của bên ba là tài sản quí giá và cần được bảo vệ, vì thế chúng thường được mã hóa



Hình 2. 3. Danh sách lọc của ISP được cập nhật thường xuyên bởi bên thứ ba

2.1.2.3. Lọc thông tin tại bên thứ ba

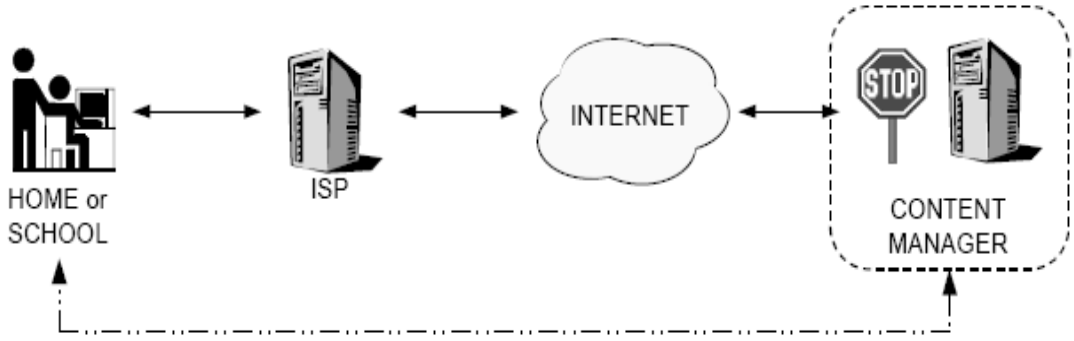
Trong lược đồ này, thông qua ISP, yêu cầu của người dùng được chuyển trực tiếp đến một bên thứ ba xác định, ở đó yêu cầu được kiểm tra theo một danh sách lọc. Để có thể đạt được hiệu quả trong lược đồ này, trình duyệt của người dùng cuối phải được cấu hình để trỏ tới trang web của bên thứ ba với bất cứ yêu cầu nào – không thể truy cập vào Internet mà không thông qua trang của bên thứ ba. Lược đồ này được mô tả trong Hình 2.4.



Hình 2. 4. Lọc thông tin tại bên thứ ba

Ưu điểm của lược đồ này là ta có thể có nhiều tùy chọn về phương pháp Lọc thông tin bao gồm cả phương pháp lọc loại trừ, lọc bao gồm hay giới hạn truy cập đến một số nội dung xác định.

Cũng Lọc thông tin do bên thứ ba nhưng ta có thể sử dụng một lược đồ khác tương tự (như được mô tả trong Hình 2.5), trong đó có yêu cầu phải tải phần mềm về máy người dùng. Phần mềm được tải về (có thể bao gồm một trình duyệt thay thế) sẽ thao tác với danh sách lọc trên webserver của trình quản lý nội dung. Trong trường hợp này, bên thứ ba sẽ giữ điều khiển hòan tòan những truy cập của người dùng.



Hình 2. 5. Lọc thông tin tại bên thứ ba, giao tác với phần mềm của người dùng

2.1.3. Công cụ lọc nội dung

Công cụ sử dụng để lọc nội dung bao gồm cả phần cứng lẫn phần mềm, trong đó nòng cốt là các ứng dụng phần mềm. Lori Bowen Ayre [2.1] và *©2005 TopTenREVIEWS, Inc.*[17] đã cung cấp danh sách các sản phẩm phần mềm lọc Internet thông dụng nhất. Đồng thời, việc thiết đặt cơ chế an ninh mạng, ở mức cao hơn là cơ chế lọc nội dung Internet, cũng đã được tiến hành trên các thiết bị phần cứng, có thể kể đến các thiết bị *Draytek Vigor2900, Planet VRT-311, một số sản phẩm CISCO...*

[17] http://internet-filter-review.toptenreviews.com/

2.1.4. Các kỹ thuật lọc thông tin trên Internet

2.1.4.1. Lọc dựa trên từ khóa

Kỹ thuật lọc thông tin theo từ khóa được xem như một phương pháp đơn giản và tương đối hiệu quả trong việc phát triển các hệ thống lọc. Với phương pháp này, hệ thống lọc sẽ sử dụng danh sách đen, trắng, hoặt kết hợp cả hai để thực hiện việc lọc. Từ những danh sách đó, nội dung thông tin sẽ được quét để tìm những từ đã liệt kê trong danh sách đen/trắng ngay khi nó đang được tải về máy người dùng. Nếu có từ khóa trong danh sách đen, nội dung thông tin đó sẽ bị hệ thống ngăn chặn, không trả về cho người dùng. Trong trường hợp sử dụng danh sách trắng, chỉ những nội dung có từ khóa trong danh sách này mới được trả về cho người dùng.

Với kỹ thuật này, vấn đề mấu chốt đề phát triển hệ thống chính là việc xây dựng các danh sách trắng và đen. Đối với danh sách đen, những từ khóa được chọn phải phản ánh đúng những nội dung mà hệ thống muốn lọc (ví dụ những thông tin có nội dung kích động bạo lực, khiêu dâm, xuyên tạc sự thật, khủng bố….). Đối với danh sách trắng, việc xây dựng sẽ cần phải có những phân tích và lựa chọn những từ khóa thật thận trọng. Các từ khóa này phải đảm bảo được những thông tin chứa từ khóa đó phải là những thông tin có nội dung lành mạnh, không ảnh hưởng đến các tiêu chí lọc thông tin của hệ thống. Nhìn chung, với kỹ thuật lọc theo từ khóa, hệ thống lọc cần ít các tính tóan phụ trội, tốc độ thực hiện quá trình lọc cao.

Tuy nhiên, phương pháp này có một số nhược điểm như sau:

- Chỉ kiểm tra văn bản, nên nó không chặn được những hình ảnh xấu.
- Thông tin ngữ cảnh không được tính đến. Một ví dụ điển hình là từ khóa “breast cancer”, có thể bị chặn với hệ thống lọc từ khóa do từ “breast”, sẽ dẫn đến việc chặn tòan bộ trang. Một ví dụ khác là khi các từ trong danh sách đen nằm trong các từ khác – ví dụ một

bộ lọc từ khóa với “sex” trong danh sách đen sẽ chặn các tài liệu chứa các từ “middlesex”.

- Kỹ thuật lọc này phụ thuộc hoàn toàn vào ngôn ngữ được sử dụng để thể hiện thông tin. Chính vì thế các danh sách trắng/đen bắt buộc phải được cập nhật thường xuyên để tránh lạc hậu so với sự tiến triển của ngôn ngữ sử dụng.

2.1.4.2. Lọc gói tin (packet filtering)

Kỹ thuật lọc thông tin theo hình thức phân tích các gói tin được sử dụng để truyền thông tin đó là một kỹ thuật được thực hiện dựa theo nguyên tắc hoạt động mạng Internet. Thật vậy, thông tin được lưu chuyển trên mạng thông qua việc gửi các gói tin giữa nơi gửi và nhận. Với mạng Internet, mỗi gói tin có địa chỉ IP của máy đích cũng như địa chỉ IP của máy nguồn ngoài những dữ liệu được đóng gói kèm theo trong gói tin đó.

Kỹ thuật lọc dựa trên phân tích gói tin chủ yếu dựa trên việc kiểm tra các địa chỉ IP và cổng nguồn và đích để từ đó có thể xây dựng các luật sử dụng trong các hệ thống lọc. Thông thường, việc lọc các gói tin được thực hiện tại các router. Các router là những máy tính chuyên dụng, phân phối các gói tin trên Internet từ nguồn đến đích: chúng tạo thành đường backbone cho Internet

Các routers tốt có thể tiến hành lọc gói tin mà không làm giảm hiệu suất mạng, vấn đề chính của phương pháp này là nó không được mịn. Một địa chỉ IP là đại diện của một máy tính – không phải là của một website – và một nỗ lực nhằm lọc một trang bằng cách sử dụng địa chỉ IP có thể sẽ chặn luôn một số lượng lớn các trang hợp lệ cùng được đặt trên máy chủ đó.

Vì những vấn đề này, lọc gói tin không được sử dụng rộng rãi như là một thành phần cơ bản trong các sản phẩm lọc

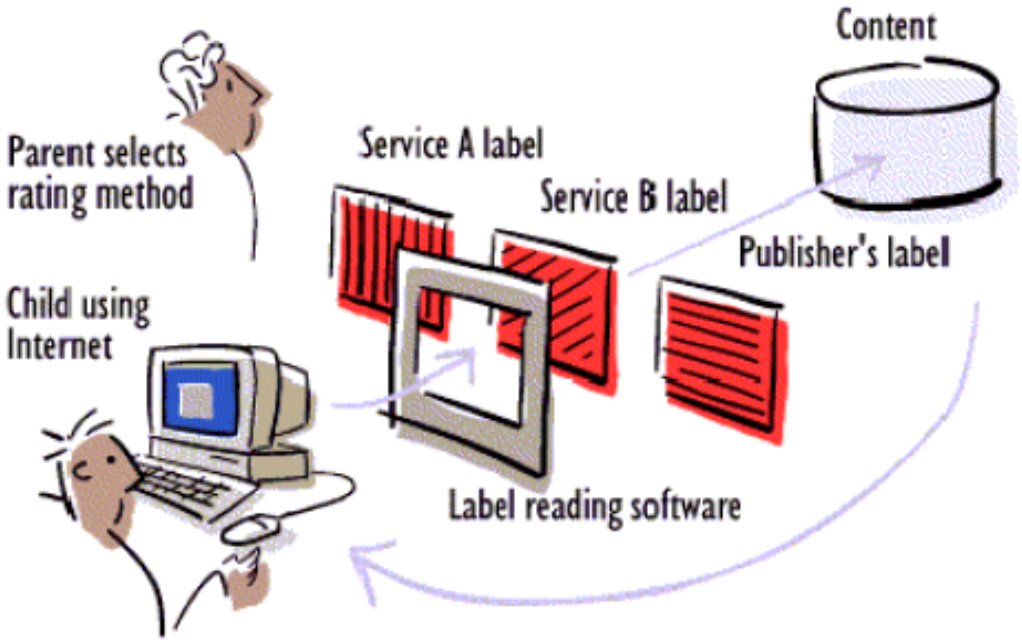
2.1.4.3. Lọc dựa trên URLs và PICS

Đây là phương pháp lọc phổ biến nhất, việc lọc các trang được dựa trên URL (hay địa chỉ trang), cung cấp một phương thức "mịn" hơn so với lọc gói tin (chặn địa chỉ), do một URL có thể xác định một trang trong một trang lớn hơn, thay vì bắt buộc phải chặn tất cả các trang tại cùng một địa chỉ IP. Các hệ thống lọc dựa theo URL có thể kết hợp với kỹ thuật lọc theo từ khóa để nâng cao hiệu năng của hệ thống lọc

Một lượng lớn URLs có thể được lọc theo phương pháp này bằng cách cho phép các kí tự đặc biệt (wildcard) trong các URLs, ví dụ: danh sách lọc có thể chứa thuật ngữ dạng như http://www.pornosite.com/*.* , ở đây *.* khớp với tất cả các trang trong một domain xác định.

Một vấn đề của lọc URLs là người ta phải mất nhiều công sức hơn trong việc dò vết những thay đổi về mặt tổ chức nội dung. Các tổ chức thường xuyên thay đổi lại các trang web, trong khi địa chỉ ở mức cao có thể không thay đổi thì URLs của các mức thấp hơn có thể thay đổi tương đối thường xuyên.

Một phương pháp khác cũng cần được đề cập đến là lọc dựa theo những thông tin đã được gán nhãn đánh giá, phân hạng tại các nhà cung cấp nội dung thông qua chuẩn PICS (Platform for Internet Content Selection – 1999, http://www.w3c.org/PICS). Người dùng hoặc các nhóm người dùng của PICS có thể xác định thang đánh giá riêng cho mình theo chuẩn PICS.



Hình 2. 6. Lựa chọn nội dung theo chuẩn PICS

Các nội dung cung cấp trên Internet bởi các nhà cung cấp thông tin khi sử dụng chuẩn PICS sẽ phải chú ý đến những tiêu chí sau đây:

- Tự xếp loại thông tin cung cấp và gán nhãn cho những trang thông tin đó trước khi phân phối.
- Có thể cho phép một cơ quan khác thực hiện việc xếp loại và gán nhãn những trang thông tin đó
- Có thể cho phép phụ huynh/ giáo viên thực hiện các thao tác phân loại và gán nhãn thông tin.

Việc lọc thông tin theo các nhãn đã được gán theo chuẩn PICS, lúc bấy giờ sẽ trở nên đơn giản và tăng hiệu quả của hệ thống lọc.

2.1.4.4. Lọc theo nội dung

Các hệ thống lọc thông tin sử dụng các kỹ thuật vừa đề cập ở các phần trên đều bộc lộ rất nhiều nhược điểm. Quá trình lọc thông thường với các kỹ

thuật này thường chịu ảnh hưởng lớn từ khâu xác định/đánh giá nội dung cần lọc với hình thức thủ công và nhiều khi mang tính chủ quan.

Kỹ thuật lọc thông tin dựa trên phương pháp phân tích tự động nội dung thông tin đó, đánh giá xếp hạng và từ đó tự quyết định trả về hay không cho người sử dụng cho phép giải quyết tốt và khắc phục những những nhược điểm của các phương pháp trên.

Việc lọc nội dung văn bản thường dựa trên những thuật toán trong khai phá văn bản và xử lý ngôn ngữ tự nhiên. Do đặc thù của mỗi ngôn ngữ, việc lọc nội dung văn bản trong các thứ tiếng khác nhau cần phải được tách biệt. Đối với mỗi ngôn ngữ, tùy theo kết quả nghiên cứu và thử nghiệm có thể sử dụng các phương pháp khác nhau. Một số phương pháp điển hình là dựa trên biểu diễn bag-of-words và dựa trên học máy.

- *Phương pháp bag-of-words (túi từ)*: biểu diễn văn bản dưới dạng bag-of-words kết hợp với loại bỏ từ dừng và lấy các từ gốc. Việc đánh chỉ mục các thuật ngữ được lực chọn thông qua ngưỡng tối thiểu đối với tần suất văn bản (document frequency) trong corpus huấn luyện. Ta có thể sử dụng các độ đo tương tự như độ tương tự vector cosine, entropy chéo... để đánh giá mức độ tương tự của một tài liệu mới với tập tài liệu đã bị chặn để đưa ra quyết định xem tài liệu này có nằm trong một trong các lớp tài liệu bị chặn hay không.
- *Phương pháp dựa trên học máy*: Với phương pháp này ta cần phải tạo ra một corpora cho từng loại văn bản cần chặn đối mỗi ngôn ngữ đích, có thể kết hợp với lựa chọn một tập các từ chính xác (bằng tay và phương pháp thống kê) để cung cấp nền tảng cho các nhiệm vụ lọc. Tiếp đó, ta trích rút ra các mẫu ngữ cảnh từ các tài liệu học và kết hợp với một phương pháp học máy như ANN, SVMs, CRFs,... để xác định ra đặc trưng của mỗi lớp tài liệu. Các

thông tin ngữ cảnh được quan tâm không chỉ là các từ khóa, từ gốc mà có thể kết hợp cả các thông tin ngữ pháp (POS) hay nhãn thực thể (NE) như tên người , tên địa danh, tên tổ chức, ... Với mỗi tài liệu mới không nằm danh sách lọc, hệ thống cũng trích rút các thông tin ngữ cảnh theo cách thức tương tự và sử dụng mô hình đã được huấn luyện để quyết định xem tài liệu mới này có nằm trong các lớp cần chặn hay không.

Với mục đích lọc được nhiều ngôn ngữ, việc nhận dạng ngôn ngữ trong một văn bản cũng là công việc cần phải thực hiện trước khi tiến hành phân tích nội dung. Thông thường người ta có thể sử dụng mô hình thống kê đơn giản là n-gram cũng có thể đem lại hiệu quả cao.

Với lọc nội dung thời gian thực, trước khi có thể phân tích và xác định xem nội dung đó là tốt hay xấu, phần lớn nội dung cần phải được lấy về. Trong lúc đó, một phần hoặc tất cả trang đó đã được hiển thị lên màn hình của người dùng, đây là hiệu ứng không mong muốn nếu trang này chứa nội dung độc hại. Vì lý do này, lọc dựa trên nội dung thường được dùng kết hợp với các phương pháp khác như lọc URL, và được sử dụng để đánh giá nội dung của những site chưa nằm trong danh sách lọc.

2.1.4.5. Lọc hình ảnh

Việc lọc thông tin dựa trên các kỹ thuật phân tích nội dung hình ảnh cũng là một trong những phương pháp quan trọng trong các hệ thống lọc nội dung. Cách tiếp cận chính của vấn đề này cũng tương tự như vấn đề lọc văn bản text: phân loại, nhận dạng ảnh dựa trên các kỹ thuật học máy và thống kê. Có rất nhiều kỹ thuật khác nhau được đề xuất cho việc phân loại cũng như nhận dạng loại ảnh, chẳng hạn:

- Lọc dựa trên màu sắc: đối tượng cần xác định trong trường hợp này là các hình ảnh khỏa thân, có màu và kết cấu chung. Khi đó

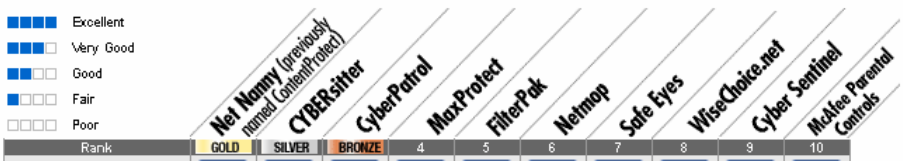
cần phải xác định vùng điểm ảnh tương ứng với da người cũng như phân bố màu sắc và kết cấu.

- Lọc dựa vào màu sắc, văn bản và sử dụng trí tuệ nhân tạo: Kỹ thuật này yêu cầu phải có bước học tự động đặc trưng từ các mẫu. Khi có ảnh mới, các đặc trưng của nó sẽ được chắt lọc và so sánh với các mẫu để quyết định lọc hay không. Đồng thời, việc phân tích văn bản kèm theo trang web và kết hợp cả hai phân tích lại cũng sẽ góp phần làm tăng hiệu quả của quá trình lọc ảnh.
- Lọc dựa vào màu sắc và hình dạng: kỹ thuật này kết hợp các đặc điểm của cả hình dạng và màu sắc để phát hiện ảnh có nội dung không lành mạnh.

2.1.5. Đánh giá một số hệ thống lọc Internet

Hình dưới đây là xếp hạng đánh giá một số hệ thống lọc thông tin trên Internet được thực hiện bởi TopTenReview (http://internet-filter-review.toptenreviews.com/) năm 2008. Một số tiêu chí đánh giá được đưa ra bao gồm:

- Ease of Use – thuộc tính đầu tiên một phần mềm lọc thông tin có thể cung cấp là tính "dễ dàng sử dụng", người dùng dù ở mức độ thành thạo máy tính nào cũng có thể dễ dàng cài đặt và sử dụng bộ lọc với tính năng cao nhất của nó.
- Lọc hiệu quả - Các sản phẩm lọc nằm trong top này cung cấp một sự cân bằng giữa việc lọc các tài liệu không được pháp và không lọc quá nhiều nội dung. Khía cạnh quan trọng khác đó là tính năng tùy biến bộ lọc cho những đối tượng khác nhau.



Overall Rating										
Ratings										
Feature Set										
Ease of Use										
Ease of Installation										
Filtering Effectiveness										
Uses										
Home	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Small Business	Read Review		✓	✓				✓		
Enterprise	Read Review									
Filtering Algorithm										
Object Analysis	✓									✓
URL Based	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Keyword Based	✓	✓	✓	✓		✓	✓	✓		
Dynamic Categorization	✓	✓	✓	✓	✓				✓	✓
Image Recognition										
Filtering Capabilites										
Filter Categories	29	30	13	8	15	19	35	23		41
Editable Filter Lists	✓	✓	✓	✓			✓		✓	✓
Chat Filtering		✓	✓	✓						
Chat Monitoring	✓	✓	✓	✓			✓			
Chat Blocking	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Newsgroup Blocking	✓	✓			✓		✓		✓	✓
IM Port Blocking	✓	✓		✓	✓	✓	✓	✓	✓	✓
Peer-to-Peer (P2P) Blocking	✓	✓		✓	✓	✓	✓	✓	✓	
FTP Blocking		✓			✓	✓	✓	✓	✓	
Customizable Port Blocking		✓								
Email Filtering		✓		✓						
Email Blocking	✓	✓		✓	✓	✓	✓	✓	✓	
Popup Blocking		✓		✓			✓			
Predator Blocking		✓		✓					✓	
Personal Info Blocking		✓	✓	✓		✓	✓		✓	✓
Reporting Capabilities										
Remote Reporting	✓			✓		✓	✓	✓		
Notification Alerts by E-mail	✓	✓		✓			✓		✓	
Log Reports Sent by E-mail	✓	✓								
Summary History Reporting	✓			✓		✓	✓	✓		
Detailed History Reporting	✓	✓		✓			✓	✓		✓
Graphical Reporting	✓							✓		
Logging of Security Violations	✓	✓	✓						✓	
Management Capabilities										
Individual User Profiles	✓	✓	✓	✓	✓	✓	✓	✓		✓
Password Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Remote Management	✓	✓		✓		✓				
Stealth Options		✓	✓	✓	✓				✓	
Other Features										
Immediate Overriding of Blocks	✓	✓		✓	✓			✓	✓	✓
Warning, Not Just Blocking	✓			✓					✓	
Daily Time Limits	✓	✓	✓	✓			✓		✓	✓
Negligible Surfing Time Impacts	✓	✓			✓				✓	✓
Updated URL/Filtering Rules	✓	✓	✓	✓	✓					✓
Blocking Sensitivity Settings	✓	✓		✓						✓
Help/Support Options										
Help	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Product Documentation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Technical Support Available	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Hình 2. 7. Đánh giá một số hệ thống lọc thông tin

- Thuật toán lọc – Các chương trình lọc tốt nhất sử dụng sự kết hợp các phương pháp lọc như lọc URL, lọc từ khóa, …
- Báo cáo hoạt động – Các phần mềm lọc Internet hữu ích cũng đưa ra báo cáo về những hoạt động mà một người dùng thực hiện trên máy

tính (bao gồm duyệt web, các hoạt động trên chat room…) – hữu ích cho việc quản lý cha mẹ đối với con cái.

- Lọc dựa trên Client-Server – mềm dẻo đối với platforms, cho phép người dùng thực hiện các giải pháp lọc ở client-based (máy ở nhà) hay server –based (Proxy hay ISP) hay là kết hợp của cả hai phương pháp đó.
- Lọc ngôn ngữ nước ngoài – Lọc trên nhiều ngôn ngữ (ít nhất là tiếng Anh và một ngôn ngữ địa phương).
- Lọc và chặn cổng – Các chương trình lọc có thể chặn và thao tác với hầu hết các giao thức trên Internet như truy cập web, chat rooms, email, mạng ngang hàng và các cửa số popup.

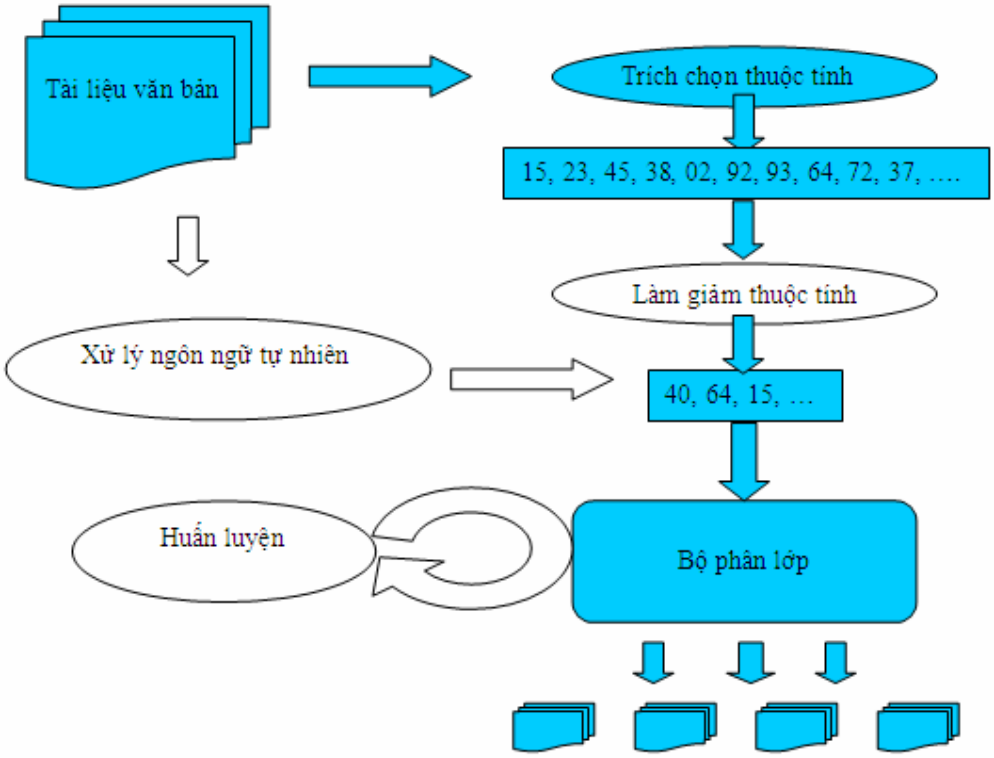
2.2. Bài toán phân lớp văn bản

Phân lớp văn bản là một lĩnh vực nghiên cứu được nghiên cứu khá kỹ, bắt đầu từ những năm 1980. Ngày nay, với tốc độ phát triển của Internet và tầm quan trọng của các máy tìm kiếm, bài toán phân lớp văn bản ngày càng được quan tâm hơn. Quá trình phân lớp văn bản được tập trung nghiên cứu nhiều.

Phân lớp văn bản là quá trình chia một tập hợp các tài liệu vào hai hay nhiều lớp cho trước [2.133]. Ngày nay với sự phát triển không ngừng của mạng Internet đã tạo ra một khối lượng khổng lồ các tài liệu điện tử, do đó là động lực cho sự phát triển của bài toán phân lớp văn bản tự động. Sự phát triển của phần cứng máy tính đã tạo ra sức mạnh tính toán, cho phép quá trình phân lớp văn bản tự động được sử dụng trong các ứng dụng thực tế. Bài toán phân lớp văn bản được sử dụng rộng rãi để loại bỏ thư rác, phân loại các tập hợp văn bản vào các chủ đề cho trước, quản lý tri thức và tìm kiếm thông tin trên Internet. Quá trình phân lớp tự động văn bản được chia thành hai giai đoạn chính sau:

1. Giai đoạn tìm kiếm thông tin, là giai đoạn mà dữ liệu số được trích rút từ nội dụng các văn bản
2. Giai đoạn phân lớp, là giai đoạn mà một thuật toán xử lý dữ liệu được sử dụng để đưa ra quyết định về chủ đề mà văn bản thuộc về.

Các giai đoạn khác có thể được thêm vào quá trình phân lớp [2.117] để làm giảm chi phí tính toán và huấn luyện các thuật toán với dự liệu huấn luyện trước khi phân lớp chính thức các văn bản (Hình 2.8).



Hình 2. 8. Lược đồ quá trình phân lớp tài liệu văn bản

Giai đoạn trích chọn thuộc tính được sử dụng để lấy ra các thuộc tính đặc trưng cho văn bản [2.133]. Thuộc tính hay được sử dụng nhất trong bài toán phân lớp tự động văn bản là phân bố tần số xuất hiện của các từ khóa hoặc các cụm từ khóa trong nội dung văn bản, chẳng hạn [2.28]. Nếu chúng ta có thể phát hiện ra bất kì loại cấu trúc nào xuất hiện trong nội dung văn bản

được phân lớp (Ví dụ cấu trúc HTML), các thuộc tính của cấu trúc đấy có thể được trích rút và đưa vào các thuật toán phân lớp [2.29, 2.160].

Quá trình xử lý ngôn ngữ tự nhiên cũng có thể được sử dụng theo các cách trong đó bao gồm cả quá trình trích rút các thuộc tính và làm giảm thuộc tính; ví dụ, các tiện ích có thể được sử dụng để trích chọn đặc trưng (xác định các từ khóa) [2.147] từ một văn bản hoặc thậm chí tạo ra một tổng hợp bán cấu trúc của nội dung văn bản đó. Các thuộc tính được trích rút ra từ nội dung được tổng hợp này của tài liệu gốc sẽ thực sự làm giảm số chiều của vector biểu diễn mà không làm giảm hiểu quả của quá trình phân lớp.

Quá trình trích rút các cụm từ khóa được mô tả ở trên và các phương pháp làm giảm thuộc tính được mô tả ở dưới đây đều sử dụng các tri thức từ khoa học về ngôn ngữ và xử lý ngôn ngữ tự nhiên. Tuy nhiên việc càng ngày càng có nhiều sự hợp tác giữa các nhà ngôn ngữ học và các nhà khoa học nghiên cứu về bài toán phân lớp tự động văn bản có thể mang lại các phương pháp mới cho quá trình tích hợp các tri thức về ngôn ngữ vào quá trình thiết kế chương trình trích rút thuộc tính, làm giảm thuộc tính và thậm chí là bộ phân lớp.

Phương pháp biểu diễn tài liệu dựa trên tần suất xuất hiện các từ khóa sẽ tạo ra các vector biểu diễn với số chiều rất lớn, có thể lên tới hàng triệu thành phần [2.115]. Độ phức tạp tính toán sẽ tỉ lệ thuận với kích thước của vector biểu diễn [2.19, 2.59], do vậy bất kỳ một phương pháp làm giảm kích thước của vector biểu diễn mà không ảnh hưởng nghiêm trọng đến chất lượng của quá trình phân lớp sẽ rất có ý nghĩa trong các ứng dụng thực tiễn. Bên cạnh đó, có một số từ đặc biệt trong một số ngôn ngữ gây ra nhiễu trong quá trình tính toán, vì vậy việc loại bỏ chúng ra khỏi vector biểu diễn sẽ thực sự làm tăng sự hiệu quả của quá trình phân lớp [2.59]. Hầu hết các phương pháp giảm thuộc tính là sự kết hợp của ba cách tiếp cận sau [2.79]:

- Loại bỏ từ dừng,

- Lấy từ gốc,

- Lọc thống kê.

Từ dừng (Stop words) là những từ không mang nghĩa, nhưng mang chức năng ngữ pháp của ngôn ngữ. Có thể thấy rằng trong hầu hết các ngôn ngữ, một từ gốc có thể có nhiều thể hiện. Do vậy chúng ta có thể thay thế các thể hiện này bằng từ gốc tương ứng của chúng. Ví dụ, nếu một tài liệu có từ khóa "house" xuất hiện 7 lần, từ "houses" xuất hiện 3 lần và từ "housing" xuất hiện 2 lần, khi đó ba từ riêng biệt này có thể được thay thế bằng một từ gốc chung của chúng, đó là "house" với số lần xuất hiện là 12 lần (7+3+2). Thông thường thuật toán của Porter [2.61] vẫn đang được sử dụng để thực hiện quá trình lấy từ gốc trong tiếng Anh và các thuật toán khác [2.8, 2.134] được phát triển cho các ngôn ngữ khác.

Các kỹ thuật lọc thống kê được sử dụng để lấy ra các từ khóa có ý thống kê cao. Có rất nhiều phương pháp thống kê khác nhau đang được nghiên cứu và sử dụng cho bài toán lọc vector thuộc tính, nhưng sự khác nhau giữa các phương pháp chủ yếu là về lượng thông tin về dữ liệu nguồn được sử dụng trong thuật toán. Mức độ ý nghĩa về mặt thống kê của một từ khóa có thể được tính toán thông qua sự khác nhau giữa tần số sử dụng của nó trong các tài liệu khác nhau, tuy nhiên các thuật toán phức tạp hơn có thể sử dụng nhãn được gán cho các tài liệu đó và qua đó tính toán mức độ ý nghĩa của các từ khóa một cách hiệu quả trong các thể loại.

Các phương pháp tiếp cận cho bài toán lọc thống kê mới nhất là tỉ lệ bất thường (odds ratio), độ đo tương hỗ (mutual information), entropy chéo (cross entropy), độ đo thông tin lấy được (information gain), trọng số của chứng cứ (weight of evidence), kiểm chứng X2 (X2 test), hệ số tương quan (correlation coefficient [2.125], cực đại thông tin tương hỗ có điều kiện (conditional mutual information maximin [2.129], tiêu chuẩn hợp lý (conformity/uniformity criteria) [2.23], Yang [2.168] và Mladenic [2.107,

2.108]. Một cách đơn giản, hầu hết các công thức sẽ gán trọng số cao cho các từ khóa xuất hiện nhiều lần trong một thể loại và các từ xuất hiện với tần số thấp ở bên ngoài một thể loại (conformity) hoặc ngược lại (non-conformity). Ngoài ra trọng số cao có thể được gán cho các từ khóa xuất hiện trong hầu hết các tài liệu của một thể loại nhất định (uniformity).

Một cách khác để giảm vector thuộc tính là thông qua việc sử dụng thuật toán di truyền [2.169]. Tuy nhiên, trong hầu hết các ứng dụng việc sử dụng thuật toán di truyền sẽ tốn nhiều tài nguyên, do đó quá trình làm giảm thuộc tính có thể trong trạng thái không hoạt động trong quá trình hệ thống chạy.

Sau quá trình ứng dụng các thuật toán làm giảm thuộc tính chúng ta có thể hi vọng rằng kích thước thông thường của vector thuộc tính có thể giảm từ hàng trăm ngàn chiều xuống còn một vài ngàn [2.21]. Tuy nhiên một số nghiên cứu [2.109] chỉ ra rằng trong quá trình làm giảm thuộc tính một số thông tin hữu ích cho quá trình phân lớp có thể bị loại bỏ.

Quá trình xử lý ngôn ngữ tự nhiên có thể được áp dụng cho cả giai đoạn trích chọn thuộc tính và làm giảm thuộc tính của quá trình phân lớp văn bản. Các thuộc tính ngôn ngữ có thể được lấy ra từ nội dung văn bản và được sử dụn như là một phần của vector thuộc tính [2.62]. Ví dụ như các phần nội dung được viết theo lối trường thuật trực tiếp, độ dài của các câu, các tỉ lệ giữa các chức năng ngôn ngữ khác nhau trong các câu (ví dụ như các cụm danh từ, giới từ hoặc cụm động từ) có thể được phát hiện và sử dụng như là một vector thuộc tính hoặc là một phần bổ sung cho vector thuộc tính các tần suất xuất hiện các từ khóa.

2.2.1. Phân lớp dựa vào thống kê

Hiện nay tồn tại rất nhiều thuật toán phân lớp văn bản, nhưng phương pháp phân lớp được sử dụng rộng rãi nhất là phương pháp phân lớp Bayesian.

Trong số các phương pháp phân lớp dựa vào thống kê [2.148], bộ phân lớp Bayesian là đơn giản nhất, nhưng vẫn rất hiệu quả. Do tính đơn giản của nó nên Bayesian là bộ phân lớp thường được sử dụng để nghiên cứu, nó xuất hiện trong hầu hết các bài báo liên quan đến các phương pháp phân lớp văn bản. Thuật toán Naïve Bayesian giả sử rằng các đặc trưng của vector đặc trưng là độc lập [2.110, 2.115]. Rất nhiều nghiên cứu [2.26, 2.111, 2.149] đã đề xuất cải tiến nhằm nâng cao độ chính xác của thuật toán Naïve Bayesian. Naïve Bayesian cũng được coi là bộ phân lớp chuẩn [2.110] mà mọi hướng tiếp cận về phân lớp văn bản đều kiểm tra thuật toán này đầu tiên [2.117].

2.2.2. Bộ phân lớp chức năng

Nếu chúng ta xem vector đặc trưng có một toạ độ nào đấy thì mỗi tài liệu có thể được biểu diễn như một điểm trong không gian, trong đó số chiều của không gian bằng số đặc trưng của vector. Cách hiểu này cho phép chúng ta sử dụng phương pháp hình học để biểu diễn bài toán một cách trực quan hơn.

Một trong những thuật toán hình học (hoặc chức năng) đơn giản nhất là k người láng giếng gần nhất (kNN) [2.74]. Ý tưởng của bộ phân lớp này rất đơn giản – trong không gian nhiều chiều, chúng ta tìm điểm biểu diễn tài liệu được phân lớp và tìm kiếm xung quanh những điểm gần với điểm biểu diễn tài liệu đó. Chỉ k láng giềng gần nhất được xem xét. Nếu chúng cùng thuộc một lớp thì tài liệu mới cũng được phân vào lớp đó. Nếu không, ta xem xét phân phối lớp của k người láng giềng đó để tính xác suất tài liệu mới thuộc vào lớp nào. Nói cách khác, nếu trong 5 láng giềng gần nhất, 4 láng giềng thuộc lớp A và 1 láng giềng thuộc lớp B thì tài liệu mới được phân vào lớp A với độ tin cậy là 80%. Có thể cải tiến thuật toán này bằng cách tính khoảng cách tới các hàng xóm để quyết định lớp của tài liệu mới. Mặc dù thuật toán

rất đơn giản nhưng nó đem lại hiệu quả rất ngạc nhiên và cũng nhận được nhiều sự quan tâm của các nhà nghiên cứu [2.133].

Hiện nay, một trong số những bộ phân lớp được nghiên cứu nhiều nhất là bộ phân lớp máy vector hỗ trợ (Support Vector Machine, SVM) [2.127, 2.146]. Nếu ta coi mỗi tài liệu là một điểm trong không gian, ý tưởng của thuật toán SVM là tìm một siêu phẳng phân tách những điểm này thành hai phần. Định nghĩa biên *M* của bộ phân lớp là khoảng cách giữa các siêu phẳng và các dữ liệu học gần nhất. Siêu phẳng tối ưu nhất là siêu phẳng có biên lớn nhất, điều đó có nghĩa là chúng ta cần tìm siêu phẳng sao cho khoảng cách từ siêu phẳng đến những điểm gần nhất là lớn nhất. Những điểm nằm trên hai đường biên này gọi là các vector hỗ trợ. Về mặt toán học, bằng cách lấy trung bình hai vector hỗ trợ của hai lớp, chúng ta có thể xác định được một vector nằm giữa hai lớp và vector này có thể được sử dụng để phân lớp các tài liệu mới.

Có nhiều phương pháp khác nhau để xác định tập các vector hỗ trợ. Hiện nay lựa chọn kernel tối ưu và thiết lập các tham số phù hợp cho kernel là một trong những thách thức lớn nhất của SVM. Bộ phân lớp SVM [2.52, 2.142] nhận được nhiều sự quan tâm của các nhà nghiên cứu trong lĩnh vực phân lớp văn bản.

2.2.3. Bộ phân lớp mạng nơron

Mạng nơron là một bộ xử lý phân tán song song quan trọng, được tạo bởi nhiều đơn vị xử lý đơn giản (nơron). Mỗi nơron chứa tri thức, tri thức này thu được từ quá trình học. Về mặt lý thuyết, sử dụng mạng nơron có thể giải quyết bất kỳ bài toán phân lớp nào. Tuy nhiên, vấn đề cơ bản trong việc sử dụng mạng nơron là thiết kế một mạng thực sự. Về mặt lý thuyết, có thể xây dựng mạng nơron có độ phức tạp bất kỳ; nhưng rất khó có thể dự đoán về mặt toán học nếu cho trước một mạng nơron, mạng đó cho độ chính xác cao với

những vấn đề cụ thể nào. Vì thế, các nhà nghiên cứu thường tập trung vào thiết kế các mạng nơron đơn giản và có thể dự đoán được cho hầu hết các bài toán thực tế trong lĩnh vực phân lớp văn bản; và chỉ sử dụng thiết kế phức tạp hơn trong các lĩnh vực mới và phức tạp trong xử lý ảnh và nhận dạng tiếng nói.

Các mô hình chuẩn được sử dụng khi có ít tri thức về vấn đề cần giải quyết, nhưng đôi khi một số kiến thức của chuyên gia về dữ liệu hoặc về cấu trúc các lớp có thể được kết hợp để thiết kế bộ phân lớp. Ví dụ, nếu chúng ta cần phân lớp một số lượng lớn các bài báo vào một cấu trúc phân cấp chủ đề cho trước, thì bộ phân lớp phải được thiết kế để thích ứng với cấu trúc đó.

Nghiên cứu chỉ ra rằng, với bài toán phân lớp phân cấp, một hệ thống perceptron phân cấp [2.21] thu được kết quả tốt hơn (độ đo F cao hơn) so với một perceptron đơn hoặc phương pháp Bayesian. Phân lớp phân cấp văn bản theo chiến lược top-down định hướng vào bài toán phân lớp lớn ban đầu theo phương pháp chia nhỏ và đệ quy. Với phương pháp này, ta cần xây dựng nhiều bộ phân lớp và phân lớp một tài liệu mới được thực hiện bằng cách bắt đầu từ gốc và duyệt qua cây phân cấp cho đến khi tìm được các lớp phù hợp. Đệ quy tại mỗi nút, và các bộ phân lớp tại các nút trong sẽ quyết định nhánh nào, cạnh nào của cây phân cấp sẽ được đi xuống sâu hơn [2.49, 2.129].

2.2.4. Đánh giá bộ phân lớp.

Hai phương pháp chính để học bộ phân lớp là học giám sát (học có thầy) và học bán giám sát.

- **Học có giám sát**: là phương thức học máy trực tiếp nhất, được nghiên cứu nhiều hơn cả. Hệ thống học được cung cấp tập ví dụ huấn luyện trực tiếp. Sau đó, hệ thống phải có khả năng cho kết luận khi nhận được các dữ liệu mới

- **Học bán giám sát**: Quá trình học sử dụng cả dữ liệu huấn luyện được gán nhãn và tập dữ liệu rất lớn gồm các tài liệu chưa được gán nhãn.

Để đánh giá tính hiệu quả của thuật toán, cách trực tiếp là áp dụng độ hồi tưởng *(recall)* và độ chính xác (*precision)* [2.53]

Lớp C_i		**Dữ liệu thực**	
		Thuộc lớp C_i	Không thuộc lớp C_i
Dự đoán	Thuộc lớp C_i	TP_i	TN_i
	Không thuộc lớp C_i	FP_i	FN_i

Trong đó

a. **TP_i** (true positives): số lượng ví dụ dương được thuật toán phân đúng vào C_i.
b. **TN_i** (true negatives): số lượng ví dụ âm được thuật toán phân đúng vào C_i.
c. **FP_i** (false positives): số lượng ví dụ dương được thuật toán phân sai vào C_i.
d. **FN_i** (false negatives): số lượng ví dụ âm được thuật toán phân sai vào C_i.

Độ chính xác Pr_i của lớp C_i là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ được thuật toán phân lớp vào lớp C_i :

$$\mathrm{Pr}_i = \frac{TP_i}{TP_i + TN_i}$$

Độ hồi tưởng Re_i của lớp C_i là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ dương thực sự thuộc lớp C_i:

$$\mathrm{Re}_i = \frac{TP_i}{TP_i + FP_i}$$

Ngoài ra, người ta còn sử dụng là độ đo F_1 [2.40] nó là độ đo đơn giản được tính từ độ chính xác và độ hồi tưởng

Năm 2006, nhóm tác giả Rich Caruana và Alexandru Niculescu-Mizil thuộc bộ môn Khoa học máy tính, trường Đại học Cornell, Ithaca, NY 14853 USA đã tiến hành thực nghiệm để so sánh giữa 10 thuật toán học giám sát điển hình: *SVMs, Neural nets, Logistic regression, Naïve Bayes, Memory-based learning, Random-forest, Decision Tree, Bagged Trees, Boosted trees, boosted stumps* [2.113]. Mỗi thuật toán học trên được tiến hành thực nghiệm trên 11 tập dữ liệu nhị phân được mô tả chi tiết ở ***Bảng 2.1***.

Bảng 2. 1. Tập dữ liệu nhị phân được sử dụng trong thực nghiệm của Rich et. al.[2.113]

PROBLEM	#ATTR	TRAIN SIZE	TEST SIZE	%POZ
ADULT	14/104	5000	35222	25%
BACT	11/170	5000	34262	69%
COD	15/60	5000	14000	50%
CALHOUS	9	5000	14640	52%
COV_TYPE	54	5000	25000	36%
HS	200	5000	4366	24%
LETTER.P1	16	5000	14000	3%
LETTER.P2	16	5000	14000	53%
MEDIS	63	5000	8199	11%
MG	124	5000	12807	17%
SLAC	59	5000	25000	50%

Bảng 2. 2. Tập dữ liệu nhị phân được sử dụng trong thực nghiệm của Rich et. al.[2.113]

Kết quả thực nghiệm được đánh giá dựa trên 8 tiêu chí sau: Accuracy, F-Score, Lift, ROC Area, Average precision, Precision/recall, Break-event point, Squarred error, Cross-entropy.

Kết quả thực nghiệm (*Bảng 2.2*) cho thấy nếu không thực hiện calibration thì phương pháp *bagged trees, random forests, và mạng neural* là

hiệu quả nhất. Nếu thực hiện calibration bằng phương pháp Platt, phương pháp *boosted tree* sẽ là phương pháp học tốt nhất. Thuật toán Naïve Bayes và memory-based learning là hai phương pháp học kém hiệu quả nhất. Tuy nhiên, *Rich et. al.* cũng khẳng định rằng không có thuật toán nào là tối ưu cho mọi bài toán. Một vài thuật toán có thể được sử dụng hiệu quả trong bài toán này nhưng trong bài toán khác thì nó lại không thực sự hiệu quả (*Bảng 2.3*).

Bảng 2. 3. Kết quả thực nghiệm của các thuật toán học giám sát (Có thực hiện calibration hoặc không) theo từng độ đo [2.113].

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	**.939**	**.963**	**.938**	.929*	**.880**	**.896**	**.896**	**.917**
RF	PLT	.872*	.805	.934*	.957	.931	**.930**	.851	.858	.892	.898
BAG-DT	–	.846	.781	.938*	.962*	.937*	.918	.845	.872	.887*	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	–	**.872**	.790	.934*	.957	.931	**.930**	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	**.861**	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	.843*	.933*	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	–	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	.836*	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	–	.834*	.816	**.939**	**.963**	**.938**	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	–	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	–	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	–	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774
DT	–	.647	.639	.824	.843	.762	.777	.562	.607	.708	.763
DT	PLT	.651	.618	.824	.843	.762	.777	.575	.594	.706	.761
LR	–	.636	.545	.823	.852	.743	.734	.620	.645	.700	.710
LR	ISO	.627	.567	.818	.847	.735	.742	.608	.589	.692	.703
LR	PLT	.630	.500	.823	.852	.743	.734	.593	.604	.685	.695
NB	ISO	.579	.468	.779	.820	.727	.733	.572	.555	.654	.661
NB	PLT	.576	.448	.780	.824	.738	.735	.537	.559	.650	.654
NB	–	.496	.562	.781	.825	.738	.735	.347	-.633	.481	.489

Bảng 2. 4. Kết quả thực nghiệm của các thuật toán học giám sát (Có thực hiện calibration hoặc không) trên các tập dữ liệu mẫu [2.113]

MODEL	CAL	COVT	ADULT	LTR.P1	LTR.P2	MEDIS	SLAC	HS	MG	CALHOUS	COD	BACT	MEAN
BST-DT	PLT	**.938**	.857	**.959**	**.976**	.700	.869	**.933**	.855	**.974**	**.915**	.878*	**.896***
RF	PLT	.876	.930	.897	.941	**.810**	.907*	.884	.883	.937	.903*	.847	.892
BAG-DT	–	.878	.944*	.883	.911	.762	.898*	.856	**.898**	.948	.856	**.926**	.887*
BST-DT	ISO	.922*	.865	.901*	.969	.692*	.878	.927	.845	.965	.912*	.861	.885*
RF	–	.876	.946*	.883	.922	.785	.912*	.871	.891*	.941	.874	.824	.884
BAG-DT	PLT	.873	.931	.877	.920	.752	.885	.863	.884	.944	.865	.912*	.882
RF	ISO	.865	.934	.851	.935	.767*	**.920**	.877	.876	.933	.897*	.821	.880
BAG-DT	ISO	.867	.933	.840	.915	.749	.897	.856	.884	.940	.859	.907*	.877
SVM	PLT	.765	.886	.936	.962	.733	.866	.913*	.816	.897	.900*	.807	.862
ANN	–	.764	.884	.913	.901	.791*	.881	.932*	.859	.923	.667	.882	.854
SVM	ISO	.758	.882	.899	.954	.693*	.878	.907	.827	.897	.900*	.778	.852
ANN	PLT	.766	.872	.898	.894	.775	.871	.929*	.846	.919	.665	.871	.846
ANN	ISO	.767	.882	.821	.891	.785*	.895	.926*	.841	.915	.672	.862	.842
BST-DT	–	.874	.842	.875	.913	.523	.807	.860	.785	.933	.835	.858	.828
KNN	PLT	.819	.785	.920	.937	.626	.777	.803	.844	.827	.774	.855	.815
KNN	–	.807	.780	.912	.936	.598	.800	.801	.853	.827	.748	.852	.810
KNN	ISO	.814	.784	.879	.935	.633	.791	.794	.832	.824	.777	.833	.809
BST-STMP	PLT	.644	**.949**	.767	.688	.723	.806	.800	.862	.923	.622	.915*	.791
SVM	–	.696	.819	.731	.860	.600	.859	.788	.776	.833	.864	.763	.781
BST-STMP	ISO	.639	.941	.700	.681	.711	.807	.793	.862	.912	.632	.902*	.780
BST-STMP	–	.605	.865	.540	.615	.624	.779	.683	.799	.817	.581	.906*	.710
DT	ISO	.671	.869	.729	.760	.424	.777	.622	.815	.832	.415	.884	.709
DT	–	.652	.872	.723	.763	.449	.769	.609	.829	.831	.389	.899*	.708
DT	PLT	.661	.863	.734	.756	.416	.779	.607	.822	.826	.407	.890*	.706
LR	–	.625	.886	.195	.448	.777*	.852	.675	.849	.838	.647	.905*	.700
LR	ISO	.616	.881	.229	.440	.763*	.834	.659	.827	.833	.636	.889*	.692
LR	PLT	.610	.870	.185	.446	.738	.835	.667	.823	.832	.633	.895	.685
NB	ISO	.574	.904	.674	.557	.709	.724	.205	.687	.758	.633	.770	.654
NB	PLT	.572	.892	.648	.561	.694	.732	.213	.690	.755	.632	.756	.650
NB	–	.552	.843	.534	.556	.011	.714	-.654	.655	.759	.636	.688	.481

2.3. Bài toán phân lớp trang web

Phân lớp trang Web, ví dụ như tổ chức các trang web hoặc các kết quả trả về của máy tìm kiếm dưới cấu trúc cây thư mục [Yahoo!, DMOZ]. Bài toán phân lớp trang Web là một trường hợp đặc biệt của phân lớp văn bản do có sự tồn tại của các siêu liên kết. Các siêu liên kết này sẽ cung cấp một nguồn thông tin hữu ích, vì chúng có thể được hiểu như là sự khẳng định về sự liên quan của các trang được liên kết tới với các trang liên kết. Nghiên cứu trong ngành khoa học về nội dung và thông tin đã chỉ ra rằng các phân tích định lượng của các chỉ dẫn tới sách, tài liệu khác (siêu liên kết) có thể hữu ích trong quá trình quyết định sự liên quan của nội dung. Việc phân tích nội dung các trang Web khó khăn hơn rất nhiều so với việc so sánh nội dung văn bản

chuẩn, vì một vài lý do như nội dung các trang web thường sử dụng ngôn ngữ gợi tả thay vì ngôn ngữ miêu tả, sự xuất hiện nhiễu). Do vậy cần thiết phải sử dụng các kỹ thuật phức tạp để đạt được mức hiểu quả hợp lý. Chúng ta có thể tận dụng các tri thức sẵn có trong các trang Web, ví dụ như sử dụng nhãn của các trang web láng giếng theo nghĩa siêu liên kết như là các thuộc tính và sử dụng kỹ thuật lặp gán nhãn lỏng nếu các nhãn này không được biết trước [2.15].

2.3.1. Các ứng dụng của bài toán phân lớp trang Web

2.3.1.1. Quá trình xây dựng, duy trì và mở rộng các thư mục Web (Cấu trúc phân cấp các trang Web)

Các thư mục Web, ví dụ như của Yahoo! (2007) và dự án thư mục mở dmoz (ODP) (2007), cho phép duyệt các thông tin trong một tập hợp các thể loại cho trước một cách tiện lợi nhất. Hiện tại những thư mục này được xây dựng và duy trì chủ yếu bằng công sức lao động của những người biên soạn. Theo báo cáo vào tháng 7 năm 2006, có khoảng 73,354 người biên soạn làm việc cho dự án ODP dmoz. Do sự phát triển quá nhanh của Internet nên phương pháp tiếp cận thủ công sẽ kém hiểu quả. Một bộ phân lớp tự động các trang web có thể được sử dụng để tự động cập nhật và mở rộng các thư mục này. Ví dụ, một phương pháp tiếp cận đã được đề xuất để xây dựng bộ phân lớp từ kho dữ liệu Web dựa trên các cấu trúc cây phân cấp được xây dựng bởi người dùng [2.59, 2.60].

2.3.1.2. Quá trình nâng cao chất lượng kết quả tìm kiếm

Sự nhập nhằng trong câu truy vấn là một trong những vấn đề quyết định đến chất lượng của các kết quả trả về của máy tìm kiếm. Ví dụ từ "bank" có thể mang nghĩa là bờ biển hoặc ngân hàng. Nhiều phương pháp tiếp cận đã được đề xuất để nâng cao chất lượng tìm kiếm bằng cách phát hiện sự nhập nhằng của các từ khóa tìm kiếm. Các phương pháp phân lớp trang Web đã

được nghiên cứu nhằm mục đích nâng cao độ chính xác của quá trình tìm kiếm trang Web [2.18]. Một bộ phân lớp thống kê, được học từ các thư mục web sẵn có, được sử dụng để phân loại các trang web mới và tạo ra một danh sách có thứ tự của các thể loại mà nó thuộc về. Các kết quả trả về thường được hiển thị theo một danh sách được xếp hạng. Tuy nhiên, việc hiển thị theo thể loại, hoặc theo các cụm tài liệu, sẽ thực sự hữu ích hơn cho người dùng [2.19]. Bài toán phân lớp web còn được sử dụng trong các thuật toán xếp hạng các trang web trả về bởi máy tìm kiếm [2.55, 2.92] dựa trên thuật toán PageRank [2.98, 2.126].

2.3.1.3. Ứng dụng trong các hệ thống trả lời câu hỏi

Quá trình phân lớp web đã được sử dụng trong giải pháp tìm câu trả lời cho các câu hỏi danh sách (Là câu hỏi có câu trả lời là một danh sách các thực thể, ví dụ như "*Tên của tất cả các nước ở Châu Âu*") [2.136, 2.137]. Các trang web kết quả được phân lớp bởi một bộ phân lớp dựa trên cây quyết định vào một trong bốn thể loại sau: "Các trang chứa danh sách các nước", "Các trang chứa một câu trả lời", "Các trang liên quan", và "Các trang không liên quan". Để tăng độ tổng quát, sẽ có thêm nhiều "trang web chứa câu trả lời" được thêm vào bằng cách đi theo các liên kết ra từ các "trang web chứa danh sách các nước".

Ngoài ra đã có rất nhiều phương pháp tiếp cận để tăng độ chính xác của câu trả lời bằng cách phân loại các câu hỏi [2.56, 2.73, 2.145].

2.3.1.4. Quá trình xây dựng bộ crawler tập trung hoặc máy tìm kiếm dọc (theo miền)

Khi các câu truy vấn chỉ liên quan đến một chủ đề nào đó, việc dò tìm tất cả các trang web trên mạng Internet là không hiệu quả. Để khắc phục vấn đề này, quá trình dò tìm (được gọi là bộ dò tìm tập trung) chỉ những tài liệu nào liên quan đến một tập hợp được định nghĩa trước các chủ đề mới được

quan tâm và tải về [2.17]. Trong trường hợp này, một bộ phân lớp sẽ được sử dụng để đánh giá sự liên quan của một trang web tới các chủ đề cho trước để hướng dẫn quá trình dò tìm (crawling).

2.3.1.5. Bài toán lọc nội dung trang web

Việc áp dụng kỹ thuật phân lớp văn bản và phân lớp web vào bài toán lọc nội dung trên trên mạng Internet đã được nghiên cứu và triên khai thành công [2.54, 2.116]. Bộ phân lớp sử dụng mạng noron nhân tạo đã được xây dựng để lọc các trang web khiêu dâm [2.151]. Các tác giả đã sử dụng một tập hợp các trang web khiêu dâm và trang web không có nội dung khiêu dâm để huấn luyện mạng noron nhân tạo. Tuy nhiên phương pháp này yêu cầu khối lượng tính toán lớn, do đó không phù hợp cho các ứng dụng thực tế. Ngoài ra một hệ thống lọc các trang Web có nội dung khiêu dâm cũng đã được xây dựng dựa việc cải tiến công thức tính độ đo độ tương tự giữa hai trang web [2.100, 2.106]. Phương pháp này cho kết quả nhanh hơn và chính xác hơn kết quả từ phương pháp của Lee và các đồng nghiệp.

2.3.2. Các đặc trưng (thuộc tính) của trang web

2.3.2.1. Sử dụng các thuộc tính của chính trang web

Nội dung văn bản web là đặc trưng đầu tiên cần phải xem xét trong quá trình phân lớp web. Tuy nhiên, do các thường có rất nhiều nhiễu xuất hiện trong các trang web, nên việc sử dụng biểu diễn túi các từ khóa sẽ không thu được hiệu quả cao. Một trong những phương pháp phổ biến để tăng tính hiệu quả là quá trình lọc, chọn thuộc tính. Phương pháp biểu diễn N-gram là một phương pháp tỏ ra hiểu quả. Năm 1998, Mladenic đã đưa ra một phương pháp tiếp cận để tự động phân lớp web dựa trên câu trúc phân cấp Yahoo!. Trong hướng tiếp cận này, mỗi tài liệu được biểu diễn bằng một vector các thuộc tính, trong đó không chỉ bao gồm các từ khóa đơn, mà còn chứa những cụm

từ có độ dài tối đa là năm. Việc sử dụng biểu diễn N-gram cho phép thu nhận được những khái niệm được thể hiện bằng một chuỗi các từ.

Một đặc trưng chỉ xuất hiện trong các tài liệu HTML mà không xuất hiện trong các tài liệu văn bản là các thẻ HTML. Những đặc trưng này có thể được sử dụng để nâng cao tính hiệu quả của quá trình phân lớp. Các nội dung văn bản trong các thẻ khác nhau có thể được gắn các dấu hiệu nhận biết khác nhau[2.50]. Bốn loại thẻ khác nhau đã được các tác giả sử dụng, đó là thẻ title, headings, metadata và main text. Kết quả cho thấy rằng kết quả thu được sẽ tốt nhất khi kết hợp tuyến tính bốn phần tử trên một cách hiệu quả. Kwon và Lee [2.74, 2.75] đã đề xuất một phương pháp phân lớp web sử dụng thuật toán k-NN, trong đó các từ khóa trong các thẻ khác nhau được gán trọng số khác nhau. Họ đã chia các thẻ HTML vào ba nhóm và gán mỗi nhóm một trọng số ngẫu nhiên.

Do vậy, việc sử dụng các thẻ này có thể tận dụng được thông tin về cấu trúc tồn tại trong tệp HTML. Tuy nhiên, bởi vì hầu hết các thẻ HTML thường hướng tới việc biểu diễn (thể hiện) hơn là ngữ nghĩa. Do vậy, việc sử dụng thông tin về các thẻ HTML trong bài toán phân lớp có thể chịu ảnh hưởng của các thông tin không thống nhất trong tài liệu HMTL.

Một bản tóm tắt có chất lượng tốt của tài liệu có thể biểu diễn các chủ đề chính xuất hiện trong một trang web. Năm 2004, Shen và các đồng nghiệp đã đề xuất phương pháp phân lớp trang web thông qua quá trình tổng hợp văn bản, và thu được kết quả tốt hơn 10% so với những bộ phân lớp dựa trên nội dung ban đầu của trang web.

Thay vì lấy thông tin từ nội dung trang web, một trang web có thể được phân loại dựa trên chính URL của nó [2.67, 2.68]. Mặc dù nó không được chính xác tuyệt đối, nhưng phương pháp này không cần phải tải nội dung trang Web về. Bởi vậy nó thực sự hữu ích khi việc tải nội dung trang web về tốn nhiều thời gian.

2.3.2.2. Sử dụng các đặc trưng của các láng giềng

Mặc dù các trang web chứa các đặc trưng hữu ích như đã được thảo luận ở trên, tuy nhiên trong một số các trang web cụ thể những đặc trưng này thường bị thiếu, hoặc không thể phát hiện được vì rất nhiều lí do. Ví dụ, một số trang web chứa các ảnh kích thước lớn hoặc các đối tượng flash nhưng lại chứa rất ít nội dung văn bản.

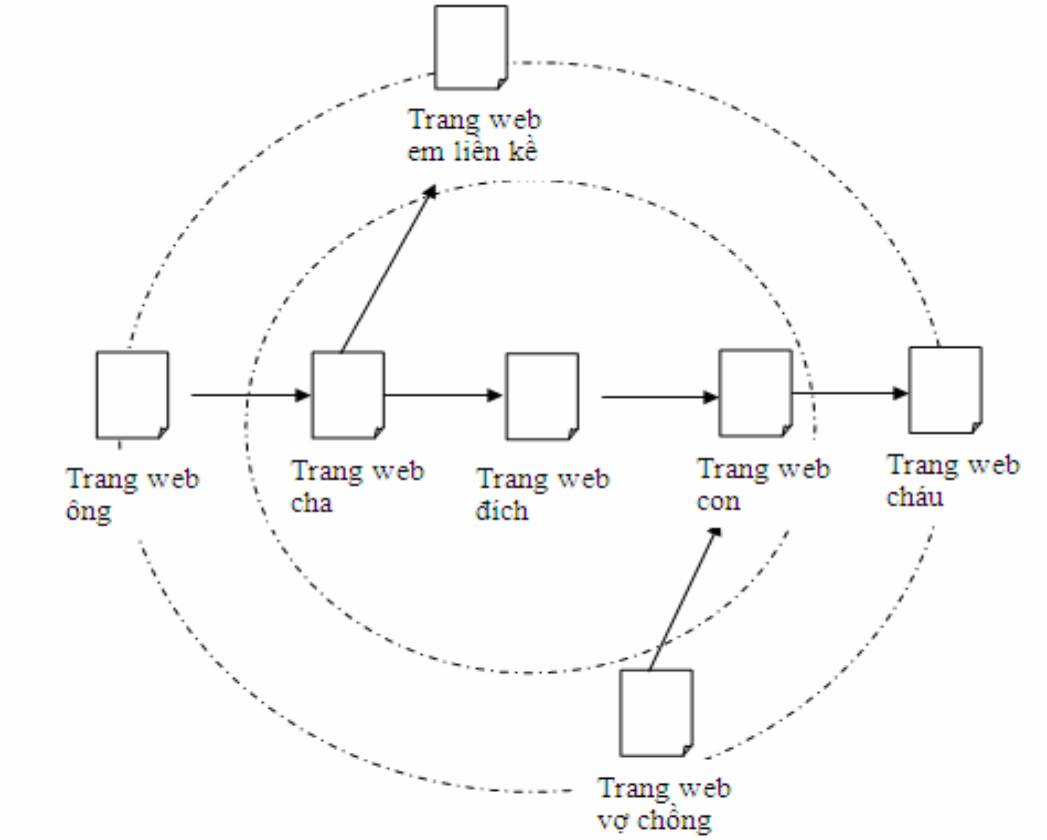
Để khắc phục điều này, các đặc trưng có thể được lấy từ các trang web láng giềng có liên quan ở một cách nào đấy tới trang web đang được phân lớp. Có rất nhiều cách để thu được các mối quan hệ giữa các trang web. Một trong những mối quan hệ đó là siêu liên kết.

Trong quá trình sử dụng các đặc trưng của láng giềng, có một số giả thuyết phải được chấp nhận. Thông thường, người ta giả sử rằng nếu trang web pa và pb thuộc về cùng một thể loại, các trang web láng giềng của chúng trong đồ thị web sẽ có chung các đặc trưng. Giả thuyết này không yêu cầu các trang láng giềng thuộc về cùng một thể loại như của pa và pb. Giả thuyết này được gọi là giả thuyết yếu.

Trong quá trình phân lớp theo chủ đề, một giả thuyết mạnh hơn phải được thỏa mãn đó là, một trang web có các trang láng giềng chủ yếu thuộc về cùng một chủ đề. Nói cách khác, sự xuất hiện của nhiều trang web "sports" trong các trang láng giềng của pa sẽ làm tăng khả năng pa thuộc về chủ đề "sports". Các trang web được liên kết tới thường có các từ khóa giống nhau [2.27]. Các trang web có xu hướng liên kết tới các trang web khác có cùng chủ đề [2.16]. Tương tự, giữa các liên kết và nội dung các trang web tồn tại mối liên hệ mật thiết với nhau [2.83]. Giả thuyết mạnh đã được chứng minh là thực sự hiệu quả trong bài toán phân lớp theo chủ đề trong các chủ đề rộng. Tuy nhiên dấu hiệu cho sự đúng đắn của nó cho các chủ đề cụ thể (hẹp) thì vẫn còn thiếu. Dưới giả thuyết mạnh này, người ta có thể xây dựng các bộ

phân lớp để dự đoán chủ đề cho một trang web đơn giản dựa vào chủ đề chiếm phần đa trong các trang láng giềng của nó.

Một câu hỏi khác được đặt ra khi sử dụng các đặc trưng của các láng giềng là những láng giềng nào sẽ được sử dụng. Các nghiên cứu hiện tại đều tập trung vào các trang láng giềng với bán kính bằng 2 tính từ trang web đang được phân lớp.



***Hình 2**. Các trang web láng giềng trong phạm vi bán kính bằng 2*

Hình 2. 9. Các trang web láng giềng trong phạm vi bán kính bằng 2

Nhìn chung, việc tích hợp các nội dung từ các trang cha và trang con vào trang đích sẽ phần nào làm giảm tính hiệu quả vì các trang web cha và con này thường chứa nhiều chủ đề hơn trang web đích đang được phân lớp [2.15, 2.47, 2.139]. Tuy nhiên, điều này không có nghĩa là nội dung các trang

web cha và con là không hữu ích. Nhiễu tồn tại trong các láng giềng có thể được loại bỏ bằng ít nhất hai phương pháp: sử dụng một số lượng phù hợp các láng giềng, và sử dụng một phần thích hợp của nội dung các trang láng giềng. Cả hai phương pháp này đều tỏ ra khá hiệu quả.

Việc sử dụng tiêu đề, các text liên kết, và một phần của văn bản xung quanh text liên kết trên trang cha trong quá trình phân loại trang đích đã đạt được kết quả rất khả quan [2.6]. Các đặc trưng trên trang web cha như text liên kết, lân cận của text liên kết, và các đầu đề phía trước liên kết trong quá trình phân lớp đã được sử dụng trong quá trình phân lớp web và thu được kết quả tốt hơn so với kết quả của bộ phân lớp chỉ sử dụng nội dung văn bản trên trang đích [2.37]. Sau đó vào năm 2001, một phương pháp tiếp cận thú vị đã được đề xuất bởi Furnkranz, trong đó nội dung văn bản trên các trang cha xung quanh liên kết được sử dụng để huấn luyện bộ phân lớp thay vì sử dụng nội dung của chính trang web này. Với phương pháp này, một trang web sẽ được gán với nhiều nhãn, mỗi nhãn cho một liên kết vào. Những nhãn này sau đó được kết hợp bằng lược đồ bỏ phiếu để đưa ra quyết định cuối cùng về nhãn của tài liệu này [2.38]. Rất nhiều phương pháp tiếp cận cho bài toán phân lớp trang web đã được nghiên cứu. Kết quả nghiên cứu đã chỉ ra rằng việc xác định các quy tắc siêu liên kết và phương pháp biểu diễn thích hợp là thiết yếu cho tính hiệu quả của bài toán phân lớp web [2.139]. Tuy nhiên các kết quả nghiên cứu đó cũng cho thấy rằng rằng " các thuật toán tập trung vào quá trình tự động khám phá các phần có liên quan của các láng giềng siêu liên kết sẽ còn mở ra rất nhiều khó khăn và triển vọng hơn so với phương pháp tiếp cận thông thường". Bộ phân lớp SVM sử dụng nội dung văn bản của trang đích, tiêu đề trang, và text siêu liên kết từ các trang cha có thể nâng cao tính hiệu quả của bộ phân lớp [2.123]. Tương tự việc sử dụng các text siêu liên kết và các lân cận xung quanh chúng từ các trang web cha có thể làm tăng độ chính xác của bộ phân lớp [2.49, 2.128].

Các trang web em liền kề thậm chí còn hữu ích hơn các trang web cha và trang con [2.15, 2.104]. Các quan hệ anh em liền kề như vậy còn có thể hữu ích trong quá trình học quan hệ giữa các chủ đề về chức năng [2.121].

Việc sử dụng nhiều kiểu láng giềng có thể cung cấp thêm nhiều lợi ích. Việc sử dụng các độ đo tương tự giữa các siêu liên kết trong bài toán phân lớp web cũng đã được nghiên cứu, trong đó độ đo về sự tương tự giữa các siêu liên kết được tính toán từ các láng giềng được liên kết theo hai bước. Sự kết hợp hợp lý giữa độ tương tự của các siêu liên kết và bộ phân lớp dựa trên nội dung văn bản có thể tạo ra sự cải thiện lớn về độ chính xác của bộ phân lớp [2.11].

Năm 2006, một phương pháp [2.104] đã được đề xuất để cải thiện độ chính xác của quá trình phân lớp web bằng cách tận dụng thông tin về nội dung và lớp từ các trang láng giềng trong đồ thị liên kết. Các chủ đề thể hiện bởi bốn loại láng giềng được kết hợp để trợ giúp quá trình phân lớp. Mặc dù các nghiên cứu đã chỉ ra rằng các trang em liền kề là hữu ích nhất để sử dụng, nhưng các loại láng giềng khác cũng đóng một vai trò nhất định [2.104].

Thay vì việc xem xét mỗi trang đích cùng với các láng giềng của nó, một số thuật toán có thể xem xét một cách có chọn lọc các nhãn của tất cả các nút trong đồ thị. Ý tưởng sử dụng thông tin từ các láng giềng cũng có thể được áp dụng trong bài toán học không có giám sát. Các cộng đồng người dùng trong dữ liệu được liên kết có thể được tìm ra bằng việc sử dụng ma trận đo độ tương tự dựa trên các quan hệ đồng trích dẫn cũng như sự tương tự về nội dung [2.32]. Một phương pháp tiếp cận [2.5] cho bài toán phân cụm các tài liệu được liên kết đã được đề xuất bằng phương thức lặp lỏng lẻo qúa trình gán vào các cụm trên đồ thị liên kết.

Như đã đề cập ở trên, có rất ít nghiên cứu về ảnh hưởng của các trang web nằm ngoài bán kính bằng hai tính từ trang web đích. Có ít nhất hai lý do giải thích cho việc này: thứ nhất, vì số lượng các trang láng giềng phát triển

rất nhanh cho nên việc sử dụng các trang láng giềng quá xa sẽ rất tốn kém; thứ hai, các trang láng giềng càng xa, thì nó càng chứa ít chủ đề liên quan đến các chủ đề tồn tại trong trang web đang được phân loại [2.16], do đó chúng không hữu ích nhiều trong quá trình phân lớp.

a) Lựa chọn các đặc trưng trên các trang láng giềng

Các đặc trưng đã được sử dụng từ các trang láng giềng bao gồm các nhãn, một phần nội dung (text liên kết, vùng văn bản lân cận xung quanh text liên kết, tiêu đề, đầu đề), và toàn bộ nội dung.

Nhiều nhà nghiên cứu đã tận dụng các trang láng giềng đã được gán nhãn bởi con người và sử dụng các nhãn trong quá trình phân lớp web [2.15, 2.121, 2.11]. Điểm nội bật của phương pháp sử dụng nhãn trực tiếp này là sự chính xác của quá trình gán nhãn được thực hiện bởi con người là tốt hơn so với bộ phân lớp tự động. Điểm yếu của phương pháp này là những nhãn này thường không có sẵn. (Các trang được gán nhãn bởi con người có sẵn trên mạng Internet thườn chỉ chiếm một phần rất nhỏ). Khi các nhãn không sẵn có, những phương pháp tiếp cận này sẽ phải đối mặt với độ bao phủ (Không thể đưa ra quyết định cho một số trang) hoặc làm giảm tính hiệu quả của bộ phân lớp dựa trên nội dung.

Có thể thấy rằng text liên kết là một loại nội dung cục bộ trên các trang cha, thường có khả năng mổ tả cô đọng nội dung của web đích [2.28, 2.27]. Đây là những thông tin thực sự hữu ích trong quá trình phân lớp trang web cũng như tìm kiếm web. Tuy nhiên các text liên kết thường ngắn, do đó nó có thể không chứa đủ các thông tin cần thiết cho bài toán phân lớp [2.49]. Điều này có thể được khắc phục bằng việc sử dụng các vùng nội dung văn bản xung quang text liên kết này, bao gồm cả nội dung text liên kết [2.49, 2.24]. Với những trang web được tạo một cách cận thận, thông tin ở tiêu đề và các đầu đề thường là những phần quan trọng nhất. Bởi vậy, sẽ là hợp lý khi sử

dụng các tiêu đề và text siêu liên kết thay vì sử dụng toàn bộ nội dung của các trang láng giềng. Khi so sánh với việc sử dụng nhãn của các láng giềng, việc sử dụng một phần nội dung của các trang láng giềng sẽ không bị phụ thuộc vào các trang láng giềng được gán nhãn bởi người dùng. Lợi ích của việc sử dụng một phần nội dung của các trang láng giềng là chỉ phụ thuộc vào chất lượng của các trang được liên kết.

Việc sử dụng nhãn của các trang láng giềng đã được thực nghiệm và làm tăng độ đo F lên 12% so với cách tiếp cận chỉ sử dụng duy nhất nội dung của trang web [2.97]. Ngoài ra việc sử dụng toàn bộ nội dung của các trang láng giềng sẽ tăng độ đo F lên 1.5%. Thực nghiệm cho thấy rằng việc sử dụng quá trình gán nhãn cho các trang láng giềng trong bài toán phân lớp web sẽ cho kết quả tốt hơn hai phương pháp trên, chỉ sử dụng nhãn của các trang láng giềng [2.104].

Bảng 2.4 mô tả các phương pháp tiếp cận cho bài toán phân lớp web có sử dụng các đặc trưng của các trang láng giềng. Từ ***Bảng 2.4***, ta có thể thấy rằng nhãn các trang láng giềng là một đặc trưng được sử dụng thường xuyên. Một điều thú vị là text liên kết lại được sử dụng ít thường xuyên hơn so với dự kiến.

Bảng 2. 5. Các phương pháp tiếp cận cho bài toán phân lớp web có sử dụng các đặc trưng của các trang láng giềng

Approach	Assumption based upon	Types of neighbors used	On-page features utilized?	Types of features used	Combination method
Chakrabarti et al. (1998)	weak	sibling	no	label	N/A
Attardi et al. (1999)	weak	parent	no	text	N/A
Fürnkranz (1999)	weak	parent	no	anchor text, extended anchor text, & headings	multiple voting schemes
Slattery & Mitchell (2000)	weak	sibling	no	label	N/A
Fürnkranz (2001)	weak	parent	no	anchor text, extended anchor text, & headings	multiple voting schemes
Glover et al. (2002)	weak	parent	yes	extended anchor text	N/A
Sun et al. (2002)	weak	parent	yes	anchor text, plus text & title of target page	
Cohen (2002)	weak	parent	no	text & anchor	N/A
Calado et al. (2003)	strong	all six types types within two steps	yes	label	Bayesian Network
Angelova & Weikum (2006)	weak	"reliable" neighbors in a local graph	yes	text & label	N/A
Qi & Davison (2006)	strong	parent, child, sibling, spouse	yes	text & label	weighted average

b) Các liên kết nhân tạo

Mặc dù siêu liên kết là một kiểu liên kết giữa các trang web dễ nhận thấy nhất, tuy nhiên nó không phải là lựa chọn duy nhất. Một người có thể hỏi câu hỏi như những trang web nào có thể được liên kết/kết nối với nhau (thậm chí siêu liên kết giữa chúng không tồn tại). Trong khi sự tương tự về nội dung là một xuất phát điểm hợp lý, thì độ đo về sự đồng xuất hiện trong kết quả câu truy vấn có thể xem là một độ đo mạnh [2.36, 2.7, 2.48, 2.28]. Trong mô hình này, hai trang web được xem xét là tương tự bởi một máy tìm kiếm trong một

ngữ cảnh cụ thể, và nhìn chung có thể bao gồm các trang web có chứa những nội dung giống nhau hoặc cùng chung độ tương tự. Dựa trên ý tưởng này, một phương pháp đã được đề xuất để tạo ra mối liên kết giữa các trang web xuất hiện trong kết quả trả về của cùng một câu truy vấn và cả hai đều được xem bởi người dùng, chúng được gọi là các liên kết suy diễn [2.120]. Các kết quả so sánh của họ cho thấy rằng các liên kêt suy diễn này cũng hữu ích cho quá trình phân lớp web. Một phương pháp tiếp cận tương tự cho bài toán phân lớp web bằng cách sử dụng mối quan hệ nội tại giữa các trang web và các câu truy vấn đã được đề xuất bởi Xue và các đồng nghiệp [2.133]. Ý tưởng chính của phương pháp là thực hiện lặp quá trình truyền bá thông tin về chủ đề của một đối tượng (các trang web hoặc câu truy vấn) tới các đối tượng liên qua. Phương pháp này đã làm tăng độ đo F lên 26% so với quá trình phân lớp web chỉ dựa vào nội dung.

Các liên kết được suy từ sự tương tự về nội dung có thể có ích. Dựa trên một tập hợp các vector thuộc tính được sinh ra từ các thư mục web, việc sử dụng các vector thuộc tính tương tự với nội dung của trang web cần phân lớp đã được nghiên cứu [2.42], mặc dù những liên kết này không được sinh ra một cách rõ ràng. Cho trước một thư mục web như là dmoz hoặc Yahoo!, một vector thuộc tính được xây dựng cho mỗi nút bằng việc lựa chọn các từ khóa có khả năng có khả năng đại diện tốt trong các mô tả về chủ đề, các URL và các trang web trong nút đó. Sau đó, bộ sinh các thuộc tính so sánh văn bản đầu vào với các vector thuộc tính của tất cả các nút trong thư mục, và những vector nào có độ tương tự vượt qua một ngưỡng cho trước sẽ được chọn để làm giàu phương pháp biểu diễn túi các từ khóa của trang web đầu vào. Một cách tiếp cận tương tự được mô tả trong [2.43], trong đó các tri thức bách khoa toàn thư được sử dụng cho quá trình sinh các thuộc tính tự động.

Có một số phương pháp khác được giới thiệu để tạo ra các liên kết nhân tạo nhưng không được kiểm tra với bài toán phân lớp web, ví dụ như các liên

kết sinh được giới thiệu bởi Kurland và Lee [2.72], trong đó các liên kết được tạo ra giữa các tài liệu nếu mô hình ngôn ngữ được suy diễn từ một tài liệu sẽ gán xác suất cao cho tài liệu còn lại [2.71, 2.72]. Một ví dụ khác về loại liên kết này đã được giới thiệu bởi Luxenburger và Weikum, trong đó các liên kết được tạo ra từ những liên kết bậc cao trong các log của câu truy vấn và sự tương tự giữa nội dung [2.81].

Hướng tiếp cận sử dụng các liên kết nhân tạo này sẽ còn mở ra nhiều cơ hội cho những nghiên cứu trong tương lai. Ví dụ, dữ liệu chứa các trang web được click bởi cùng một người dùng trong cùng một lần duyệt một trang web hiện tại còn thưa. Một câu hỏi thú vị là với độ thưa của dữ liệu này như thế nào thì sẽ ảnh hưởng đến quá trình phân lớp web. Ngoài ra, sẽ là thực sự hữu ích khi chúng ta quan tâm nghiên cứu về vấn đề kết hợp giữa các liên kết nhân tạo và siêu liên kết.

2.3.2.3. Các thuật toán phân lớp trang web

Ngoài việc quyết định loại đặc trưng nào sẽ được sử dụng, việc gán trọng số cho các đặc trưng này cũng đóng một vai trò quan trọng trong quá trình phân lớp. Việc nhấn mạnh các thuộc tính có năng lực phân biệt cao thường là quá trình phân lớp boost. Quá trình lựa chọn thuộc tính có thể xem là một trường hợp đặc biệt của quá trình đánh trọng số cho các thuộc tính, trong đó các thuộc tính bị loại bỏ sẽ được gán trọng số bằng 0.

Có rất nhiều độ đo đã được sử dụng cho quá trình lựa chọn thuộc tính. Một số phương pháp tiếp cận đơn giản đã chứng minh tính hiệu quả của nó. Ví dụ, việc chỉ sử dụng phần đầu tiên của mỗi tài liệu thường tạo ra bộ phân lớp nhanh và chính xác cho các bài báo mới [2.117]. Phương pháp này dựa trên giả thuyết rằng đoạn tổng kết nội dung của văn bản thường ở đầu văn bẳn, điều này thường đúng đối với các bài báo, nhưng thường không đúng với

các tài liệu khác. Tuy nhiên phương pháp này sau đó đã được áp dụng vào bài toán phân lớp phân cấp các trang web và cho kết quả tốt [2.130,2.131].

Bảng 2. 6. Các thuật toán phân lớp web có sử dụng các đặc trưng trên trang láng giềng

Approach	Task	Baseline classifier	Reported improv.	Document represent.	Feature selection criteria	Evaluation dataset
Chakrabarti et al. (1998)	Topical	A term-based classifier built on TAPER	From 32% to 75% (accuracy)	Bag-of-words	A score-based function derived from Duda & Hart (1973)	Yahoo directory
Mladenic (1999)	Topical	N/A	N/A	N-gram	Gram frequency	Yahoo directory
Fürnkranz (1999)	Functional	A text classifier	From 70.7% to 86.6% (accuracy)	Set-of-words	Entropy (for baseline classifier)	WebKB
Slattery & Mitchell (2000)	Functional	N/A	N/A	Relations	No feature selection	WebKB
Kwon & Lee (2000)	Topical	A kNN classifier with traditional cosine similarity measure	From 18.2% to 19.2% (micro-averaging breakeven point)	Bag-of-words	Expected mutual information and mutual information	Hanmir
Fürnkranz (2001)	Functional	A text classifier	From 70.7% to 86.9% (accuracy)	Set-of-words	Entropy (for baseline classifier)	WebKB
Sun et al. (2002)	Functional	A text classifier	From 0.488 to 0.757 (F-measure)	Set-of-words	No feature selection	WebKB
Cohen (2002)	Topical	A simple bag-of-words classifier	From 91.6% to 96.4% (accuracy)	Bag-of-words	No feature selection	A custom crawl on nine company web sites
Calado et al. (2003)	Topical	A kNN textual classifier	From 39.5% to 81.6% (accuracy)	Bag-of-words	Information gain	Cade directory
Golub & Ardi (2005)	Topical	N/A	N/A	Word/phrase	No feature selection	Engineering Electric Library
Qi & Davison (2006)	Topical	An SVM textual classifier	From 73.1% to 91.4% (accuracy)	Bag-of-words	No feature selection	ODP directory

Ngoài những độ đo đơn giản này, có một số phương pháp lựa chọn thuộc tính đã được phát triển cho bài toán phân lớp văn bản, ví dụ như độ đo lượng thông tin thu được, độ đo thông tin tương hỗ. Nhưng phương pháp tiếp cận này cũng hữu ích cho bài toán phân lớp web. Một phương pháp dựa trên thuật toán k-NN, trong đó các thuộc tính được lựa chọn thông qua việc sử

dụng hai ma trận: thông tin tương hỗ kỳ vọng và thông tin tương hỗ [2.74]. Các từ khóa được gán trọng số theo các thẻ HTML mà các từ khóa này xuất hiện trong đó. Trong bài toán phân lớp văn bản, tồn tại một lớp các vấn đề trong đó các chủ đề có thể được phân biệt bằng một số ít các đặc trưng trong khi một số lượng lớn các thuộc tính còn lại chỉ đóng vai trò rất ít trong việc phân biệt giữa các chủ đề. Những nghiên cứu về bài toán phân lớp kiểu này chỉ ra rằng bộ phân lớp SVM có thể cải thiện vấn đề này bằng quá trình lựa chọn thuộc tính [2.41]. Bên cạnh đó, một độ đo cho phép dự đoán tính hiệu quả của quá trình lựa chọn thuộc tính mà không cần huấn luyện và kiểm tra bộ phân lớp cũng đã được đề xuất [2.41].

Phương pháp đánh chỉ mục các ngữ nghĩa tiềm ẩn (LSI) [2.29] là một phương pháp giảm số chiều thông dụng. Tuy nhiên, độ phức tạp tính toán của nó làm cho nó không hiệu quả khi kích thước bài toán lớn. Bởi vậy, các thực nghiệm sử dụng LSI trông bộ phân lớp web [2.107] đều được tiến hành trên các bộ dữ liệu nhỏ. Một số nghiên cứu đã đề xuất cải tiến trên LSI [2.57, 2.58] để nó có thể hoạt động hiệu quả trên tập dữ liệu lớn. Các nghiên cứu đã chỉ ra tính hiệu quả của những cải tiến này [2.25, 2.35].

a). Thuật toán học quan hệ

Bởi vì các trang web có thể được xem như là các thể hiện được liên kết với nhau bởi các quan hệ siêu liên kết, bài toán phân lớp Web có thể được xem là bài toán học mối quan hệ, là một chủ đề thông dụng trong ngành học máy. Bởi vậy, sẽ có ý nghĩa nếu phương pháp học quan hệ được áp dụng vào bài toán phân lớp web. Quá trình gán nhãn lỏng là một trong những thuật toán hiệu quả cho bài toán phân lớp web.

Quá trình gán nhãn lỏng [2.112] ban đầu là một thủ tục được đề xuất cho quá trình phân tích hình ảnh. Sau đó, nó được sử dụng rộng rãi trong phân tích hình ảnh, trí tuệ nhân tạo, nhận dạng mẫu, và khai phá web. Trong ngữ

cảnh của bài toán phân lớp, thuật toán gán nhãn lỏng đầu tiên sử dụng bộ phân lớp để gán nhãn cùng với xác suất cho mỗi nút (trang web). Sau đó nó lần lượt xem xét mỗi trang web và đánh giá lại các xác suất gán nhãn của nút đó dựa trên đánh giá mới nhất về phân lớp nhãn của các láng giềng [2.14]. Gán nhãn lỏng là một kỹ thuật phân lớp hiệu quả [2.15, 2.80, 2.4]. Dựa trên một framework mới cho quá trình mô phỏng phân bố liên kết thông qua các thông kê liên kết, Lu và Getoor đã giới thiệu một cải tiến của thuật toán gán nhãn lỏng, trong đó một bộ phân lớp logic kết hợp sẽ được sử dụng dựa trên thông tin về nội dung và liên kết. Phương pháp này [2.80] không chỉ chứng tỏ được tính hiệu quả so với bộ phân lớp dựa trên nội dung, mà còn thực hiện tốt hơn bộ phân lớp phẳng dựa trên cả nội dung lẫn các đặc trưng liên kết. Trong hai cải tiến được giới thiệu bởi Angelova và Weikum [2.4], không phải tất cả các láng giềng đều được xem xét. Thay vào đó, chỉ những trang láng giềng nào có độ tương tự đủ gần với nội dung của trang web đich mới được sử dụng.

Bên cạnh thuật toán gán nhãn lỏng, các thuật toán học quan hệ khác cũng được áp dụng vào bài toán phân lớp web. Sen và Getoor đã so sánh và phân tích thuật toán gán nhãn lỏng với hai thuật toán phân lớp phổ biến dựa trên liên kết, đó là thuật toán lặp truyền bá độ tin tưởng và phân lớp lặp. Kết quả thực nghiệm trên dữ liệu web tốt hơn so với kết quả trên nội dung [2.117]. Macskassy và Provost đã xây dựng một bộ công cụ phục vụ cho bài toán phân loại dữ liệu mạng [2.82], trong đó có sử dụng một bộ phân lớp quan hệ và một tập hợp các thủ tục suy luận trên tập hợp. Kết quả thực nghiệm cho thấy bộ công cụ này thực hiện tốt trên một vài tập dữ liệu, bao gồm cả dữ liệu web.

b) Sự cải tiến các thuật toán truyền thống

Để nâng cao hiệu quả phân lớp, bên cạnh phương pháp lựa chọn thuộc tính và đánh trọng số thuộc tính, chúng ta còn có thể cải tiến các thuật toán truyền thống như k người láng giềng gần nhất, máy vector hỗ trợ để phù hợp với bài toán phân lớp trang web.

Bộ phân lớp kNN đánh giá vào mức độ khác nhau giữa tài liệu kiểm tra với các tài liệu khác bằng cách xác định khoảng cách từ tài liệu kiểm tra tới mỗi tài liệu học. Hầu hết các bộ phân lớp kNN đã có sử dụng độ đo tương tự cosin hoặc tích vô hướng. Những độ đo này chưa thể hiện được mức độ liên quan giữa các từ nên một độ đo tương tự mới đã được đề xuất, sử dụng những từ đồng xuất hiện trong tài liệu [2.74, 2.75]. Tần suất của những từ đồng xuất hiện có ràng buộc về ngữ ngữ nghĩa với các từ khác. Nếu hai tài liệu có nhiều từ cùng xuất hiện thì chúng có mối quan hệ gần gũi hơn. Thực nghiệm của họ chỉ ra rằng kết quả của phương pháp này cao hơn so với sử dụng độ đo cosin và tích vô hướng. Với thuật toán kNN sử dụng các tính toán xác suất [2.51], xác suất một tài liệu d thuộc lớp c được xác định bằng khoảng cách từ tài liệu đó tới các hàng xóm của nó và xác suất của những hàng xóm thuộc lớp c.

Hầu hết các thuật toán học giám sát chỉ học mô hình từ dữ liệu huấn luyện. Co-training, được giới thiệu bởi Blum và Mitchell là một hướng tiếp cận khác, sử dụng cả dữ liệu gán nhãn và chưa gán nhãn để thu được độ chính xác cao hơn. Trong bài toán phân lớp nhị phân, hai bộ phân lớp được huấn luyện trên hai tập thuộc tính khác nhau, được sử dụng để phân lớp các ví dụ chưa gán nhãn. Dự đoán của bộ phân lớp này được sử dụng để huấn luyện cho bộ phân lớp kia. So sánh với các thuật toán khác chỉ sử dụng dữ liệu đã gán nhãn, phương pháp co-training có thể giảm tỉ lệ lỗi xuống còn một nửa. Ghani tổng quát thuật toán để giải quyết bài toán trong trường hợp nhiều lớp [2.45, 2.46, 2.47]. Kết quả thực nghiệm cho thấy, sử dụng co-training, độ chính xác không tăng khi số lớp nhiều. Họ cũng đề xuất phương pháp nâng cao độ chính xác, áp dụng kỹ thuật của Dietterich và Bakiri [2.30] sử dụng nhiều hơn số

bộ phân lớp cần thiết. Thông tin về nội dung và cấu trúc của trang web cũng được xem xét trong thuật toán co-training áp dụng vào bài toán phân lớp web [2.102].

Để học bộ phân lớp cần rất nhiều dữ liệu được gán nhãn thủ công thuộc lớp hai lớp âm và dương. Hướng tiếp cận dựa vào SVMs để loại bỏ tập dữ liệu âm mà không làm giảm độ chính xác của bộ phân lớp đã được đề xuất [2.142]. Với dữ liệu dương và dữ liệu chưa gán nhãn, thuật toán của họ có thể xác định những đặc trưng dương quan trọng nhất. Sử dụng những đặc trưng dương này, nó sẽ lọc và đưa ra những ví dụ dương có thể từ tập ví dụ chưa gán nhãn. Bộ phân lớp SVM có thể được huấn luyện trên tập ví dụ dương đã được gán nhãn và lọc các ví dụ âm.

c) Phân lớp trang web theo cây phân cấp chủ đề

Trong các nghiên cứu phân lớp văn bản, hầu hết đều tập trung vào bài toán phân lớp mà các lớp cho trước được xem là tách biệt nhau và không có cấu trúc xác định mối quan hệ giữa chúng. Những bài toán phân lớp như vậy được gọi là bài toán phân lớp phẳng (flat classification). Tuy nhiên, trong trường hợp có số lượng khá lớn các lớp, bài toán sẽ phức tạp hơn rất nhiều và khi thực hiện các giải pháp phân lớp thường cho kết quả không chính xác. Vì vậy, một vấn đề được đặt ra là cần phân lớp các văn bản sử dụng cấu trúc phân cấp.

Theo chiến lược chia nhỏ và đệ quy, việc sử dụng cấu trúc phân cấp cho bài toán phân lớp trang web đã được đề xuất [2.31]. Bằng cách việc chia bài toán phân lớp ban đầu thành những bài toán con nhỏ hơn tại mỗi mức của cây phân cấp, thực nghiệm của họ cho thấy độ chính xác thu được cao hơn so với không phân cấp. Wibowo và Williams [2.130] cũng nghiên cứu về bài toán phân lớp trang web theo cấu trúc phân cấp và đề xuất phương pháp làm giảm tối đa lỗi phân lớp [2.131]. Một phương pháp hiệu quả cũng đã được đề

xuất [2.103] để phân lớp trang web vào cấu trúc phân cấp các chủ đề và cập nhật thông tin các lớp khi cây phân cấp được mở rộng.

Liu và các đồng nghiệp đã nghiên cứu sử dụng SVM một cách hiệu quả để phân lớp tài liệu vào một taxonomy lớn [2.77]. Kết quả nghiên cứu cho thấy, mặc dù SVMs phân cấp hiệu quả hơn SVMs phẳng, nhưng cả hai cách tiếp cận đều chưa đủ năng lực để giải quyết bài toán trong trường hợp taxonomies lớn. Bên cạnh đó, cấu trúc phân cấp không đem lại hiệu quả hơn cấu trúc phẳng với bộ phân lớp kNN và Bayesian [2.78].

Để đánh giá hiệu quả của bộ phân lớp phân cấp, độ đo về mức độ phân lớp sai đã được đề xuất [2.120], bằng cách đo khoảng cách giữa lớp mà bộ phân lớp dự đoán với lớp thực sự chứa tài liệu đó. Kiritchenko cũng xem xét một cách chi tiết về bài toán phân lớp phân cấp [2.69].

d) Kết hợp thông tin từ nhiều nguồn tài nguyên

Tận dụng thông tin từ nhiều nguồn tài nguyên khác nhau có thể giúp nâng cao độ chính xác, đặc biệt khi thông tin được xem xét là trực giao. Trong phân lớp web, kết hợp các liên kết và nội dung trang web là khá phổ biến [2.11, 2.104].

Phương pháp phổ biến để liên kết nhiều thông tin là xem xét thông tin từ các nguồn khác nhau như những tập đặc trưng khác nhau. Mỗi tập đặc trưng này được huấn luyện bằng một bộ phân lớp. Sau đó, các bộ phân lớp này được kết hợp với nhau để đưa ra quyết định cuối cùng. Có một vài phương pháp để liên kết các bộ phân lớp này [2.70], bao gồm một số phương pháp đã được phát triển khá tốt trong học máy như bỏ phiếu (voting) và sắp xếp (stacking) [2.132]. Bên cạnh những phương pháp kết hợp trực tiếp này, co-training - được thảo luận ở phần trên, cũng là phương pháp hiệu quả sử dụng nhiều bộ phân lớp, mỗi bộ phân lớp được huấn luyện trên tập đặc trưng trích chọn từ các nguồn thông tin khác nhau.

Dựa trên giả thuyết mỗi nguồn thông tin cung cấp một khung nhìn khác nhau, phương pháp kết hợp nhiều nguồn thông tin có tiềm năng cho kết quả tốt hơn so với bất cứ phương pháp đơn nào. Tuy nhiên, liên kết hai nguồn thông tin không phải lúc nào cũng tốt hơn sử dụng từng tài nguyên. Kết hợp độ đo tương tự giữa hai thư mục với bộ phân lớp dựa vào nội dung kNN cho kết quả thấp hơn sử dụng một mình bộ phân lớp kNN [2.11].

2.3.2.4. Phân lớp website

Một nhánh nghiên cứu của bài toán phân lớp Website là sử dụng nội dung của site. Năm 2001 Pierre đã giới thiệu một phương pháp tiếp cận cho bài toán phân loại website vào các chủ đề về công nghiệp bằng việc sử dụng các thẻ HTML [2.103].

Một hướng nghiên cứu khác tập trung vào việc sử dụng các tính chất về cấu trúc của website. Người ta đã chứng minh được rằng có một mối quan hệ mật thiết giữa cấu trúc liên kết của website và chức năng của nó. Tông tin về cấu trúc của một site đã được sử dụng để xác định chức năng của nó (ví dụ như là máy tìm kiếm, các thư mục web, website hợp tác) [2.3]. Cũng xuất phát từ tính chất trên, mối quan hệ giữa thông tin về cấu trúc và nội dung cũng đã được phân tích rõ hơn [2.76].

Đã có một số nghiên cứu về việc sử dụng cả thông tin cấu trúc và nội dung vào bài toán phân lớp website. Ester và các đồng nghiệp đã khảo sát ba phương pháp tiếp cận để xác định chủ đề cho một website dựa trên các phương pháp biểu diễn website khác nhau [2.34]. Các website được biểu diễn như là một trang ảo bao gồm tất cả các trang trong site đó, một vector tần suất các chủ đề, hoặc một cây gồm các trang web và chủ đề. Thực nghiệm đã chứng tỏ rằng phương pháp phân lớp cây cấu trúc cho kết quả tốt. Một phương pháp biểu diễn website như là một mô hình cây hai tầng đã được giới thiệu [2.124], trong đó mỗi trang được mô hình bởi một mô hình DOM và

một site được biểu diễn như là một cây phân cấp được xây dựng dựa trên các siêu liên kết giữa các trang. Bộ phân lớp được các tác giả sử dụng là bộ phân lớp dựa trên mô hình cây Markov ẩn.

Bài toán phân lớp các trang web sẽ rất hữu ích cho bài toán phân lớp website, do việc biết được chủ đề của các trang web trong site có thể hỗ trợ quá trình quyết định chủ đề của site đấy [2.34].

2.3.2.5. Phân lớp blog

Ngày nay blog đang trở nên phổ biến trên mạng Internet nên việc nghiên cứu bài toán phân loại blog cũng trở nên cấp thiết. Nghiên cứu về phân lớp blog được chia làm ba loại: xác định blog (để xem một trang web có phải là blog hay không), phân lớp trạng thái tình cảm và phân lớp theo danh mục.

Nanno và các đồng nghiệp đã giới thiệu một hệ thống tự động sưu tập và quản lý các bộ sưu tập các blog [2.91]. Sự hiểu quả của các thuật toán thông dụng vào bài toán xác định blog đã được khảo sát [2.33]. Sử dụng một số lượng các thuộc tính được lựa chọn bởi người dùng (một số trong đó là các từ đặc trưng cho blog như “Comments” và “Archives”), các tác giả đã cho thấy rằng nhiều thuật toán học máy hiện có đều có thể đạt được độ chính xác thỏa mãn (khoảng 90%).

Lĩnh vực nghiên cứu thứ hai trong bài toán phân lớp web là việc phát hiện trạng thái tình cảm của blog. Hầu hết các phương pháp hiện tại đều phát hiện trạng thái tình cảm ở mức các lần viết bài của cá nhân. Các bài viết trên blog thể hiện hai thái cực của trạng thái tình cảm, hạnh phúc và buồn, có thể được phân biệt bằng nội dung ngôn ngữ [2.84]. Một bộ phân lớp Bayes được huấn luyện trên 10,000 bài viết và độ chính xác thu được là 79%. Trên thực tế, trạng thái tình cảm trên blog không chỉ đơn giản là hai thái cực là hạnh phúc và buồn [2.22, 2.85]. Khi thực hiện quá trình phân lớp nhị phân các bài viết trên blog vào hơn một trăm trạng thái tình cảm được định nghĩa trước, bộ

phân lớp SVM chỉ cải tiến được khoảng 8% độ chính xác. Trong khi việc phát hiện trạng thái tình cảm cho các bài viết riêng lẻ rất khó, thì những nghiên cứu về sau này [2.86] đã chỉ ra rằng việc quyết định trạng thái tình cảm trung bình trên một tập hợp lớn các bài viết trên blog lại có thể đạt được độ chính xác cao.

Loại nghiên cứu thứ ba của bài toán phân lớp blog thường được thực hiện ở mức blog. Nowson đã thảo luận về sự khác nhau của ba loại blogs: tin tức, bình luận, và tạp chí [2.96]. Một phương pháp đã được đề xuất [2.105] để tự động phân loại các blog vào một trong bốn thể loại: Nhật ký cá nhân, tin tức, chính trị và thể thao. Việc sử dụng biểu diễn tài liệu 2-gram tfidf và bộ phân lớp naïve Bayes, phương pháp tiếp cận của Qu và các đồng nghiệp [2.105] có thể đạt được độ chính xác là 84%. Cho đến nay, dường như các nghiên cứu về hai loại nghiên cứu thứ hai và thứ ba ở trên đang phải đối mặt với việc thiếu một taxonomy được xây dựng tốt.

Bài toán lọc nội dung trang web có thể xem là sự kết hợp các đặc điểm của hai ứng đặc biệt của bài toán phân lớp tự động văn bản, đó là

1. Phân lớp trang Web, ví dụ như tổ chức các trang web hoặc các kết quả trả về của máy tìm kiếm dưới cấu trúc cây thư mục [Yahoo!, DMOZ].
2. Lọc nội dung trang Web, ví dụ như phân loại mỗi một dòng các tài liệu nhận được vào một trong hai thể loại hoặc *Phù hợp với người dùng thứ i* hoặc (*Không phù hợp với người dùng thứ i*) dựa trên sở thích hoặc sự phù hợp của các tài liệu đối với người dùng. Bài toán lọc nội dung có hai điều đặc biệt so với các ứng dụng phân lớp văn bản

Tỉ lệ tài liệu được gán sai nhãn là dương và gán sai nhãn là âm thường có độ quan trọng khác nhau.

Khối lượng dữ liệu huấn luyện được lấy về tăng liên tục, thông qua quá trình tương tác với người dùng. Điều này có nghĩa rằng các kỹ thuật/ thuật toán có khả năng học bộ phân lớp tăng dần (Thuật toán học trực tuyến) phải được sử dụng. Gần đây việc sử dụng thuật toán học trực tuyến làm cực đại hóa đường biên [2.10] đã được sử dụng và thu được kết quả tốt so với các ứng dụng phân lớp văn bản chuẩn.

Phát hiện các nội dung không phù hợp, ví dụ như các trang web có nội dung khiêu dâm hoặc thư rác. Đây là một thể hiện của Bài toán quản trị nội dung trong môi trường quảng cáo; nơi mà các ứng dụng quản trị nội dung phải đối mặt với sự xuất hiện của các tác giả thường là do mục đính kinh tế mà không hợp tác với người tìm kiếm nội dung.

2.3.3. Lựa chọn giải pháp phân lớp trang web trong bài toán lọc nội dung

Qua quá trình khảo sát bài toán phân lớp web, do số lượng các trang web được phân lớp là rất lớn, thời gian và khối lượng tính toán yêu cầu là rất lớn. Để hệ thống lọc có thể được thực thi trong môi trường Internet, các thuật toán phân lớp web phải đủ đơn giản về độ phức tạp tính toán nhưng đồng thời phải đảm bảo tính hiệu quả (độ chính xác/độ hồi tưởng) của kết quả phân lớp. Lược đồ áp dụng bài toán phân lớp web vào hệ thống lọc nội dung trên Internet sẽ bao gồm hai giai đoạn: Giai đoạn lọc đơn giản và giai đoạn lọc phức tạp.

Ở giai đoạn lọc đơn giản, phương pháp tiếp cận thống kê cho bộ phân lớp, sử dụng phương pháp biểu diễn túi từ khóa, kết hợp với quá trình loại từ dừng và lấy từ gốc được sử dụng. Các từ khóa được đánh chỉ mục được lựa chọn thông qua một ngưỡng cực tiểu cho tần số tài liệu trong tập dữ liệu huấn luyện. Mô hình được xây dựng cho mỗi chủ đề sẽ bao gồm danh sách các tần số được xếp hạng của các từ khóa. Quá trình phân lớp được sử dụng thông qua độ đo sai lệch vị trí trên các danh sách xếp hạng các từ khóa.

Giai đoạn thứ lọc thứ hai (giai đoạn lọc phức tạp) sẽ tập trung vào các trang web bị phân loại sai vào các thể loại không phù hợp bởi bộ lọc đơn giản ở trên. Thay vì sử dụng các từ khóa đơn trong việc biểu diễn các tài liệu, tập hợp các từ khóa xung quanh từ khóa đấy sẽ được sử dụng. Tập hợp các từ khóa ngữ cảnh này có thể được lấy bằng một của sổ trượt, và được biểu diễn như là một danh sách hoặc một tập hợp. Ngoài ra các kỹ thuật của quá trình xử lý ngôn ngữ tự nhiên (NLP) như POS, gán nhãn tên thực thể có thể, được áp dụng để làm giàu ngữ nghĩa của từ khóa. Đối với tiếng Việt, quá trình tách từ [2.9] sẽ được áp dụng để tăng độ chính xác của quá trình phân lớp. Thuật toán phân lớp web được lựa chọn trong giai đoạn này là thuật toán gán nhãn lỏng [2.15], dựa trên thuật toán học trực tuyến [2.10].

2.4. Phương pháp cập nhật danh sách lọc URL

Như đã được giới thiệu, để có được một hệ thống lọc nội dung hiệu quả thì cần thi hành các giải pháp đa dạng, trong đó có cả các giải pháp lọc nội dung và các giải pháp lọc theo địa chỉ. Danh sách trắng (các website lành mạnh) và danh sách đen (các website độc hại) vừa có tác dụng giảm nhẹ công việc lọc nội dung, vừa cung cấp các tập trang web nhân để loang tới các trang web trắng và đen mới. Các danh sách lọc URL (trắng và đen) nhận được trên cơ sở ý kiến chuyên gia và thường xuyên được cập nhật cả theo giải pháp ý kiến chuyên gia hoặc loang tự động.

Để tự động cập nhật danh sách URL, một số phương pháp được đưa ra như: các phương pháp dựa vào liên kết để xác định link spam, hay phân lớp để xác định các URL chứa nội dung xấu/tốt,...từ đó cập nhật các URL tốt/xấu vào danh sách lọc.

Các phương pháp link spam: sử dụng các thuật toán tính hạng dựa vào liên kết như PageRank để đánh giá độ tốt/xấu của các URL, với quan niệm: các trang có nội dung xấu thường có xu hướng liên kết tới nhau và

ngược lại các trang có nội dung tốt thường liên kết với nhau và hiếm có liên kết tới các trang xấu. Từ tập nhân các URL được gán trọng số tốt/xấu ban đầu, thông qua liên kết, trọng số được phân phối cho các URL được loang tới. Những URL có trọng số càng cao càng được đánh giá tốt.

Vấn đề đặt ra là cần xác định ngưỡng (threshold) để quyết định một URL thực sự tốt/xấu: ***blackThreshold*** – xác định những URL có trọng số nhỏ hơn blackThreshold được coi là URL xấu và ***whileThreshold*** – xác định những URL có trọng số lớn hơn whileThreshold được coi là URL tốt. Sau đó các URL mới này được cập nhật vào danh sách lọc tương ứng (tốt hoặc xấu). Các URL có trọng số nằm trong khoảng black/whileThreshold có thể không cần cập nhật vào danh sách lọc bởi được coi là chưa xác định.Phương pháp cần sử dụng cấu trúc của đồ thị web - các liên kết vào ra từ mỗi URL, và do tính chất hội tụ, tính gia tăng (sử dụng các trọng số URL ở lần đánh giá trước để cập nhật cho lần đánh giá sau) của các phương pháp tính hạng trang nên để tăng nhận diện nhanh cần lưu lại cấu trúc web và trọng số của các URL đã được duyệt qua. Hơn nữa việc đánh giá này có thể làm độc lập với hệ thống, được thực hiện ngoại tuyến (offline) và chỉ cập nhật lại danh sách lọc sau những khoảng thời gian nhất định.

Phương pháp kết hợp bộ lọc nội dung: dựa vào nội dung của URL và nội dung của trang tài liệu ứng với URL để phân loại tài liệu tốt/xấu. Tận dụng bộ phân lớp nội dung tài liệu để xác định các URL chứa nội dung tốt/xấu và cập nhật vào danh sách lọc nếu cần.

Khi ứng dụng vào hệ thống lọc nội dung, cần quan tâm tới một số vấn đề sau đây:

- Vấn đề xây dựng các mẫu lọc URL,
- Vấn đề cập nhật danh sách lọc URL ban đầu, tự động tìm kiếm các trang web “tương tự” với các trang web trong danh sách trắng và danh sách đen,

- Xây dựng module cập nhật danh sách lọc có thể áp dụng thuật toán PageRank, TrustRank, trong đó lưu ý các giải pháp línhpam.
- một số vấn đề liên quan khác

Dưới đây trình bày các nội dung về lọc theo chuẩn PICS.

2.4.1. Giới thiệu lọc theo chuẩn PICS

Chuẩn PICS (Platform for Internet Content Selection) được W3C (World Wide Web Consortium) đưa ra vào năm 1995, chuẩn này cung cấp tập từ điển các nhãn nhằm xác định nội dung các trang tài liệu trên Internet. Ban đầu chuẩn được đưa ra giúp người dùng dễ dàng tìm các trang web có nội dung phù hợp và tránh những trang có nội dung không mong muốn, nhằm kiểm soát việc truy cập Internet của trẻ nhỏ. Sau đó, dần trở thành chuẩn định dạng để đánh giá nội dung các tài liệu.

Chuẩn PICS là cung cấp một khung (framework) cho việc phát triển các lược đồ gán nhãn và cho phép người dùng lựa chọn linh động các tiêu chuẩn lọc theo mong muốn.

2.4.2. Đánh giá và gán nhãn

Chuẩn PICS đưa ra 2 loại nhãn: *self-label* và *third-party label*. Mỗi tài liệu Web được gán nhãn phù hợp với nội dung bởi người tạo tài liệu, gọi là self-label. Tuy nhiên có thể tác giả không gán nhãn hoặc không gán nhãn đúng cho trang tài liệu. Do vậy giải pháp gán nhãn của bên thứ ba, third-party label, được đưa ra để đánh giá và gán nhãn cho các tài liệu Web. Và khi đó người dùng có thể lựa chọn sử dụng nhãn do tác giả hoặc của bộ gán nhãn bên thứ 3 đưa ra. Một nhãn bao gồm 3 phần: *service identifier*, *label options*, và *rating*. *Service identifier*: khai báo URL của server được dùng để đánh giá. *Label options*: đưa ra các thuộc tính của tài liệu được đánh giá. *Rating*: trọng số được đánh giá. Tương ứng có các cách đánh giá (thực hiện gán nhãn):

Self-rating:

Người cung cấp (tạo tài liệu)tự gán nhán cho tài liệu mà họ tạo ra.

***Third-party rating**:*

Dịch vụ thực hiện gán nhãn độc lập với phía tạo và phân phối tài liệu, và có hệ thống gán nhãn riêng. Do vậy cùng một nội dung, các hệ thống khác nhau có thể đưa ra các nhãn khác nhau.

***Ease-of-use**:*

Người dùng được phép tự đánh giá, gán nhãn các tài liệu.

Do sự đa dạng của các tiêu chuẩn đạo đức, văn hóa của người dùng nên chuẩn PICS được triển khai trên nhiều hệ thống gán nhãn để có thể đáp ứng được các nhu cầu sử dụng trên Net. Không một hệ thống gán nhãn đơn nào phù hợp với tất cả mọi người trên cộng đồng web.

Vấn đề gán nhãn cần đảm bảo tính toàn vẹn nội dung của các tài liệu Web và trong suốt đối với các browser. Các hệ thống gán nhãn dựa trên PICS cần linh động trong các tiêu chuẩn được sử dụng để đánh giá, và các tiêu chẩn cần mở để người dùng có thể kiểm tra.

Các nhãn được thêm vào mã HTML được coi như thông tin của thẻ META, và ở phần đầu (header) của trang tài liệu.

```
<META http-equiv="PICS-Label" content='labellist'>
```

Ví dụ một trang HTML có header chứa thông tin gán nhãn PICS:

```
<head>
    <META http-equiv="PICS-Label" content='
    (PICS-1.1 "http://www.gcf.org/v2.5"
       labels on "1994.11.05T08:15-0500"
           until "1995.12.31T23:59-0000"
         for "http://w3.org/PICS/Overview.html"
       ratings (suds 0.5 density 0 color/hue 1))
    '>
</head>
```

2.4.3. Cấu trúc PICS

2.4.3.1. Cấu trúc của PICSRules

Các luật của PICS (PICSRule)[2b.3] dựa trên các biểu thức S-expression, gồm cặp giá trị và thuộc tính. Các giá trị là xâu được đặt trong dấu nháy kép hoặc danh sách các cặp thuộc tính – giá trị được đặt trong dấu ngoặc đơn, và được phép lồng nhau. Khi giá trị của một thuộc tính là một danh sách các cặp thì được gọi là thuộc tính chính, và có thể được bỏ qua không cần khai báo.

Bộ phân tích cú pháp nhận diện giá trị của các thuộc tính: các giá trị bắt đầu là dấu nháy kép hoặc dấu mở ngoặc đơn; và những giá trị không ghép cặp với một tên của thuộc tính nào thì nó là giá trị của thuộc tính chính.

Ví dụ về các khai báo các cặp thuộc tính – giá trị:

```
attrvalpair:: attribute whitespace value | value
attribute:: alphanumstr
value:: quotedstring |'(' attrvalpair+ ')'
quotedstring:: '"'notdoublequotechar*'"' |
            "'"notsinglequotechar*"'"
alphanumstr:: (alphanum | '.')+
whitespace:: ' ' | '\t' | '\r' | '\n'
alphanum:: '0' - '9' | 'A' - 'Z' | 'a' - 'z'
notdoublequotechar :: bất kì ký tự Unicode nào trừ "
notsinglequotechar :: bất kì ký tự Unicode nào trừ '
```

Để xác định một xâu có thể dùng dấu nháy đơn hoặc dấu nháy kép. Tuy nhiên dấu nháy đơn còn có ý nghĩa như dấu bắt đầu kết thúc một xâu, do vậy khi trong xâu có chứa dấu nháy kép thì cần được đặt trong cặp dấu nháy đơn. Dấu nháy đơn, nháy kép và phần trăm có thể được đưa vào xâu dưới dạng mã (%27, %22, %25) và ngoài ra không có mã nào khác được chấp nhận.Bảng dưới là ví dụ cách biểu diễn xâu trong PICS.

Xâu biểu diễn trong PICS Rule	Xâu phân tích được
"string"	string
'string'	string
'This is "quoted" text.'	This is "quoted" text.
"It's nice to quote."	It's nice to quote.
"It%27s nice to %22quote.%22"	It's nice to "quote."
"50%25 of test scores are above the median"	50% of test scores are above the median
"50% are below the median"	<xâu lỗi>

PICSRules còn cung cấp cấu trúc cho phép đưa chú thích (comment):

```
comment:: '{' comment-text* '}'
comment-text:: bất kì ký tự nào trừ '}'
```

Với cấu trúc này, chú thích không được phép lồng nhau. Chú thích có thể xuất hiện ở bất kì đâu trong PICSRules.

Trong định dạng của PICSRules, có thể có một số thành phần như "policy-expression" và "URLpattern" được sử dụng trước và khai báo sau như ví dụ:

```
rule:: '(' 'PicsRule-'verMajor'.'verMinor rule-body ')'
      verMajor :: integer
      verMinor :: integer
      rule-body :: '(' rule-clauses ')'
      rule-clauses :: rule-clause+
      rule-clause :: policy-clause |
                     name-clause |
                     source-clause |
                     service-clause |
                     opt-extension-clause |
                     req-extension-clause |
                     extension-aval
      policy-clause :: 'Policy' '(' policy-attribute+ ')'
      policy-attribute :: ['Explanation'] quotedstring |
                          'RejectByURL' URL-strings |
                          'AcceptByURL' URL-strings |
```

```
                    'RejectIf' policy-string |
                    'RejectUnless' policy-string |
                    'AcceptIf' policy-string |
                    'AcceptUnless' policy-string |
                    extension-aval
URL-strings :: URL-string | '(' ['patterns'] URL-string+ ')'
URL-string :: '"'URLpattern'"'
policy-string :: '"'policy-expression'"'
name-clause :: 'name' '(' name-attribute+ ')'
name-attribute :: ['Rulename'] quotedstring |
                  'Description' quotedstring |
                  extension-aval
source-clause :: 'source' '(' source-attribute+ ')'
source-attribute :: ['SourceURL'] quotedURL |
                    'CreationTool' quotedstring |
                    'author' quoted-address |
                    'LastModified' quoted-date |
                    extension-aval
service-clause :: 'serviceinfo' '(' service-attribute+ ')'
service-attribute :: ['Name'] quotedURL |
                     'shortname' quotedstring |
                     'BureauURL' quotedURL |
                     'UseEmbedded' yes-no |
                     'Ratfile' quotedstring |
                     'BureauUnavailable' pass-fail |
                     extension-aval
yes-no :: '"'Y-N'"'
Y-N :: 'Y' | 'N'
pass-fail :: '"'P-F'"'
P-F :: 'PASS' | 'FAIL'
opt-extension-clause :: 'optextension' '(' extension-name+ ')'
extension-name :: ['extension-name'] quotedURL | 'shortname'
                  quotedstring |
                  extension-aval
```

```
req-extension-clause :: 'reqextension' '(' extension-name+ ')'
extension-aval :: attrvalpair
quotedURL :: '"'URL'"'
URL :: as defined in RFC-1738 for URLs.
quoted-address :: '"'e-mail-address'"'
e-mail-address :: as defined in RFC-822 for addresses.
quoted-ISO-date :: '"'YYYY'-'MM'-'DD'T'hh':'mmStz'"'
```

2.4.3.2. Các ví dụ về PICS Rules

Ví dụ 1: Cấm truy cập theo URL

```
(PicsRule-1.1
    (
    Policy (RejectByURL ("http://*@www.grody.com:*/*"
                         "http://*@www.gross.net:*/*"))
    Policy (AcceptIf "otherwise")
    )
)
```

Với ví dụ này, bất kì URL nào thuộc host www.grody.com hay www.gross.net sẽ bị khóa, kể cả khi có thêm tên tài khoản truy cập, cổng hay các đường dẫn sâu hơn được xác định.

Ví dụ 2: Cấm truy cập dựa trên nhãn PICS

```
(PicsRule-1.1
   (
    serviceinfo (
        "http://www.coolness.org/ratings/V1.html"
        shortname "Cool"
        bureauURL "http://labelbureau.coolness.org/Ratings"
        UseEmbedded "N"
        )
    Policy (RejectIf "((Cool.Coolness <= 3) or (Cool.Graphics >=
                                              3))")
    Policy (AcceptIf "otherwise")
   )
)
```

Luật này kiểm tra các tài liệu theo dịch vụ đánh giá "Cool" ("http://www.coolness.org/ratings/V1.html"). Các nhãn sẽ được lấy từ tập nhãn tại "http://labelbureau.coolness.org/Ratings". Thuộc tính UseEmbedded

đặt "N" tức các nhãn do tác giả gán cho tài liệu sẽ bị bỏ qua. Những tài liệu không phù hợp với Cool và có quá nhiều ảnh sẽ bị khóa, còn mọi tài liệu khác được chấp nhận.

Ví dụ 3: Cho phép truy cập dựa trên nhãn PICS

```
(PicsRule-1.1
    (
    ServiceInfo (
        name "http://www.coolness.org/ratings/V1.html"
        shortname "Cool"
        bureauURL "http://labelbureau.coolness.org/Ratings"
        )
    Policy (RejectUnless "(Cool.Coolness)")
    Policy (AcceptIf "((Cool.Coolness > 3) and (Cool.Graphics < 3))")
    Policy (RejectIf "otherwise")
   )
)
```

Luật này kiểm tra tài liệu theo dịch vụ đánh giá "Cool". Trong trường hợp này do không xác định UseEmbedded nên mặc định sử dụng nhãn đã được gán cùng với nhãn được lấy từ bureauURL.

Dòng Policy (RejectUnless "(Cool.Coolness)") chỉ ra rằng những tài liệu không được đánh giá trên "Coolness" của hệ thống "Cool" sẽ bị khóa. Dòng tiếp theo xác định chính sách: những tài liệu thỏa mãn Cool, có ít hơn 3 ảnh sẽ được chấp nhận, còn các tài liệu khác sẽ bị khóa.

Ví dụ 4: Cho phép truy cập dựa trên nhãn PICS

```
(PicsRule-1.1
    (
       name    (rulename "Example"
                 description "Example from PICSRules spec; ")
       source (sourceURL
          "http://www1.raleigh.ibm.com/pics/PICSRulz/Example1.html")

       ServiceInfo (name "http://www.coolness.org/ratings/V1.html"
                    shortname "Cool"
                    bureauURL
                "http://labelbureau.coolness.org/Ratings" )
       ServiceInfo ("http://www.kid-protectors.org/ratingsv01.html"
                   shortname "KP")

       Policy (RejectByURL ("http://*@www.badnews.com:*/*"
                           "http://*@www.worsenews.com:*/*"
                           "*://*@18.0.0.0!8:*/*"))
       Policy (AcceptByURL "http://*rated-g.org/movies*")
```

```
        Policy (AcceptIf "(KP.educational = 1)"
                    Explanation "Always allow educational content.")
        Policy (RejectIf "(KP.violence >= 3)"
                    Explanation "Blood's a %22scary%22 thing.")
        Policy (RejectUnless "(Cool.Graphics < 4)" )
        Policy (AcceptIf "otherwise")
    )
)
```

`Source` : cho biết luật này lấy từ đầu,

`ServiceInfo`: khai báo dịch vụ đánh giá với tên viết tắt **Cool** /**KP** và xác định nơi lấy tập nhãn.

Các chính sách được thiết lập bằng thuộc tính Policy:

RejectByURL: chặn tất cả các HTTP URL tới các host *www.badnews.com* và *www.worsenews.com*, và tất cả các URL với địa chỉ IP bắt đầu là 18 (tương ứng với các địa chỉ thuộc mit.edu).

AcceptByURL: chấp nhận những tên miền kết thúc là *rated-g.org* với đường dẫn bắt đầu là "movies" và không sử dụng tài khoản người dùng, cổng. Ví dụ "http://www.mystuff.rated-g.org/movies/hello" được chấp nhận, còn http://joe@www.mystuff.rated-g.org/movies/hello hay "http://www.mystuff.rated-g.org:8009/movies/hello" không được chấp nhận.

Tất cả những tài liệu được hệ thống KP xác nhận là educational được chấp nhận. Những tài liệu do hệ thống KP đánh giá có giá trị violence lớn hơn hay bằng 3 hoặc những tài liệu được Cool đánh giá có giá trị Graphics không nhỏ hơn 4 thì sẽ bị khóa.

2.4.4. Lấy nhãn PICS cho các tài liệu

2.4.4.1. Sử dụng HTTP để lấy nhãn của tài liệu

Client sử dụng HTTP để gửi yêu cầu lấy nhãn trong phần header của một tài liệu nào đó. Ví dụ:

Client gửi tới HTTP server www.greatdocs.com (một server hỗ trợ PICS):

```
GET /foo.html HTTP/1.0
```

***Protocol-Request**: {PICS-1.1 {params full*

{services "http://www.gcf.org/v2.5"}}}

Server trả về cho client nội dung với phần header:

```
HTTP/1.0 200 OK
Date: Thu, 30 Jun 1995 17:51:47 GMT
Last-modified: Thursday, 29-Jun-95 17:51:47 GMT
Protocol: {PICS-1.1 {headers PICS-Label}}
PICS-Label:
 (PICS-1.1 "http://www.gcf.org/v2.5" labels
  on "1994.11.05T08:15-0500"
  exp "1995.12.31T23:59-0000"
  for "http://www.greatdocs.com/foo.html"
  by "George Sanderson, Jr."
  ratings (suds 0.5 density 0 color/hue 1))
Content-type: text/html
...nội dung của foo.html...
```

Client yêu cầu trang tài liệu foo.html, với nhãn tài liệu được gán bởi đánh giá tại "http://www.gcf.org/v2.5". Server gửi trả nội dung tài liệu và phần header bao gồm nhãn PICS của tài liệu (như trên).

Nếu client chỉ muốn lấy phần header chứa thông tin về nhãn của tài liệu thì trong lời gọi thay vì dùng GET sử dụng HEAD. Khi đó server chỉ trả về phần header mà không kèm nội dung của trang tài liệu phía sau.

Cấu trúc chi tiết của một HTTP Request để lấy nhãn tài liệu

Để yêu cầu lấy thông tin nhãn tài liệu, cần thêm vào phần header của HTTP Request các thành phần tương ứng theo cấu trúc sau:

```
request-header ::
       'Protocol-Request: {PICS-1.1 {params ' [completeness]
                                              extension*
                                              services '}}'
completeness :: 'minimal' | 'short' | 'full' | 'signed'
extension :: '{' token-or-quoted-string+ '}'
```

```
token-or-quoted-string :: token | quotedname
token :: alphanumpm+
services :: '{' 'services' quotedURL+ '}'
```

Trong yêu cầu, giá trị của thuộc tính **completeness** cho biết chỉ lấy những nhãn cơ bản tối thiểu nhất, hay những nhãn server thấy phù hợp hoặc tất cả các nhãn. Với yêu cầu *signed*: mọi thông tin về nhãn được gửi cùng với ký số trên chính nhãn đó- đảm bảo tính chính xác khi truyền trên mạng. Các nhãn được ký số có các thành phần MICs và Digital Signature.

Phần *extensions* được đưa ra nhằm hỗ trợ việc mở rộng giao thức trong tương lai.

Mỗi *quotedURL* xác định một dịch vụ đánh giá mà client muốn lấy thông tin về nhãn tài liệu. Tức là có thể yêu cầu lấy nhãn tài liệu từ nhiều dịch vụ đánh giá thông qua một HTTP Request.

2.4.4.2. Yêu cầu lấy các nhãn tài liệu từ nhiều nguồn khác nhau

Client có thể liên kết với một tập nhãn để yêu cầu các dịch vụ đánh giá riêng rẽ từng loại nhãn đó.

Cấu trúc của một truy vấn

```
request :: get | post
get :: 'get' url-fragment '?' [opt] [format]
                                extension* url+ service
post :: 'post' url-fragment crlf crlf formencodeddata
url-fragment :: the part of the original URL after the host
    name, as specified in HTTP 1.0.
crlf :: carriage return (hex D) followed by line feed (hex A)
opt :: 'opt=' option
option :: 'generic' | 'normal' | 'tree' | 'generic+tree'
format :: [and] 'format=' form
form :: 'minimal' | 'short' | 'full' | 'signed'
extension :: token '=' token-or-quoted-string
token-or-quoted-string :: token | quotedname
```

token :: alphanumpm+

url :: [and] 'u=' encodedURL

service :: [and] 's=' encodedURL

boolean :: 't' | 'f' | 'true' | 'false'

and :: '&'

encodedURL :: dạng chuyển mã của *quotedURL. Ví dụ các ký tự đặc biệt như* $_-.+!*'(), cần đặt trong ‘’, (:) chuyển sang mã *%3A,* (/) chuyển thành *%2F*.

formencodeddata :: Xác định encode của MIME được mô tả trong HTML 2.0.

2.4.4.3. MICs và Digital Signatures

Vấn đề đặt ra là khi một tài liệu được kiểm tra và gán nhãn, sau đó tài liệu bị sửa đổi mà nhãn lại không được cập nhật lại. Chuẩn PICS cung cấp các lựa chọn hỗ trợ giải quyết vấn đề đó:

At : Khi nhận thấy sự thay đổi của đối tượng, ngày giờ sửa đổi cuối cùng của tài liệu sẽ được cập nhật. Giả thiết rằng thời điểm sửa đổi cuối cùng luôn được cập nhật chính xác, do đó nhãn của tài liệu sẽ được cập nhật tương ứng.

Until / **exp** : Khi tài liệu được cập nhật thường xuyên, theo định kì, thì nhãn gán cho tài liệu cũng cần có thời hạn sử dụng tương ứng.

MIC-md5/ **md5** : Khi nhãn được gán cho tài liệu thì một mã kiểm tra checksum tương ứng được sinh ra. Mã này được chuyển sang dạng ký tự ASCII và lưu trong trường **MIC-md5** (hay **md5)**. Khi đọc tài liệu, thuật toán sinh mã được thực hiện và so sánh mã mới với mã được lưu trong **md5**. Thuật toán MD5 đảm bảo sinh mã hoàn toàn khác nhau giữa 2 tài liệu gần giống nhau (bản gốc và bản sửa đổi). Đây gọi là kỹ thuật mã hóa. Với các tài liệu HTML, mã được sinh ra sau khi đã loại bỏ tất cả các thành phần thuộc thẻ META (phần bao gồm các nhãn PICS)

Với MICS, ta giải quyết vấn đề nội dung bị thay đổi và không phù hợp với nhãn đã được gán. Nhưng thay vì sửa đổi nội dung, còn có hiện tượng giả mạo nhãn. Do đó cần có giải pháp đảm bảo nhãn mà người dùng nhận được từ dịch vụ đánh giá là chính xác từ server mà họ mong muốn. PICS giải quyết vấn đề này bằng cách thực hiện ký số (*Digital Signatures)* cho các nhãn và kỹ thuật RSA được lựa chọn:

- Với mỗi nhãn, dịch vụ đánh giá cần một cặp khóa công khai: một khóa bí mật do bên cung cấp dịch vụ giữ riêng (bí mật), còn một khóa được cung cấp cho phía sử dụng nhãn, muốn kiểm tra tính đúng đắn của nhãn.
- Sau khi nhãn được tạo ra, nó được chuyển về dạng mã để sinh mã MD5 tương ứng. Sau đó mã MD5 của nhãn được mã hóa với RSA, rồi được chuyển lại về dạng ký tự ASCII và lưu lại vào trường **signature-rsa-md5** của nhãn để chuyển cho phía client.
- Khi client nhận được một nhãn và muốn kiểm tra tính đúng đắn của nhãn đó cần thực hiện các bước: chuyển nhãn về dạng mã xác định và tính MD5 tương ứng. Sau đó lấy nội dung của trường **signature-rsa-md5** chuyển về dạng mã gốc (nhị phân) và giải mã nó với khóa công khai do dịch vụ cung cấp. So sánh kết quả thu được với mã MD5 đã sinh ra trước đó, nếu không giống nhau tức đó là nhãn giả và ngược lại tức đó là nhãn đúng.

Ngoài ra với phương pháp ký số cũng nảy sinh vấn đề khó khăn: khi dịch vụ đánh giá hỗ trợ cả phương pháp self-rating (tác giả gán nhãn cho tài liệu). Lúc đó cần cả phía dịch vụ và tác giả tài liệu cùng ký cho nhãn. Điều này trở nên rắc rối khi cần phân phối khóa công khai cho người dùng để kiểm tra. Tuy nhiên PICS không đề xuất giải pháp cho vấn đề này.

2.4.5. Áp dụng vào bộ lọc nội dung

Nhóm đề tài đã tiến hành thử nghiệm bộ lọc PICS sử dụng trong Poesia và bộ lọc PICS cơ bản (pics980520, picsle101) từ W3C.

Để xây dựng bộ lọc theo chuẩn PICS cần:

- Sử dụng thư viện Dsig của W3C [2b.2] – cung cấp một chuẩn định dạng thực hiện mã hóa và giải mã.
- Tạo bộ phân tích cấu trúc PICS
- Lựa chọn rating-system, rating-service (ví dụ: rsac.rat, safesurf.rat)

2.5. Học bán giám sát trong lọc nội dung

Trong lý thuyết học máy có hai loại nhiệm vụ phổ biến, đó là học có giám sát và học bán giám sát.

Học không giám sát: Cho trước X=(x1, x2, …, xn) là một tập hợp gồm n ví dụ mẫu (hoặc điểm dữ liệu), trong đó $x_i \in X$ với mọi i∈[n]:= {1,2, ..., n}. Thông thường các điểm dữ liệu này được giả sử phân bố giống nhau và độc lập (iid) trên không gian X. Mục đích của học không giám sát là tìm ra cấu trúc tồn tại tiềm ẩn bên trong tập dữ liệu đó. Vấn đề học không có giám sát đã được thảo luận như là bài toán ước lượng một mật độ xác suất có khả năng sinh ra không gian X. Tuy nhiên tồn tại nhiều dạng đơn giản hơn của quá trình học không giám sát, ví dụ như ước lượng tỉ lệ phần trăm, phân cụm, phát hiện biên, làm giảm số chiều.

Học giám sát: Mục đích là học có giám sát là tìm một phép ánh xạ từ x tới y, khi cho trước một tập huấn luyện gồm các cặp (x_i, y_i), trong đó $y_i \in Y$ gọi là các nhãn hoặc đích của các mẫu x_i. Nếu nhãn là các số, $y = (y_i)^T_{i\in[n]}$ biểu diễn vector cột của các nhãn. Như đã nêu, các cặp (x_i, y_i) phải tuân theo các phân bố giống nhau và độc lập (iid) giả thiết i.i.d trên $X \times \mathcal{Y}$. Nhiệm vụ

của quá trình học có giám sát được định nghĩa rất rõ ràng, do đó một ánh xạ có thể được đánh giá thông qua hiệu quả thực hiện việc dự đoán của nó trên tập dữ liệu kiểm tra. Nếu các nhãn lớp là liên tục, bài toán phân lớp được gọi là hồi quy (regression). Có hai loại thuật toán cho quá trình học có giám sát. Các thuật toán sinh sẽ tìm ra mô hình cho mật độ xác suất có điều kiện p(x|y) bởi một số thủ tục học không có giám sát. Hàm mật độ xác suất có thể được ước lượng thông qua định lý xác suất Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{\int_y p(x|y)p(y)dy}.$$

Trong thực tế, p(x|y)p(y) = p(x,y) là mật độ phân bố xác suất chung cho dữ liệu, mà từ đó các cặp dữ liệu học x_i, y_i được sinh ra. Các thuật toán *phân biệt* sẽ không ước lượng việc dữ liệu xi được sinh ra như thế nào, thay vào đó chúng tập trung vào việc ước lượng trực tiếp p(y|x). Một số thuật toán loại này chỉ giới hạn trong việc mô hình xem p(y|x) lớn hơn hay bé hơn 0.5; một ví dụ của những thuật toán này là SVM. Đã có rất nhiều thảo luận xung quanh vấn đề rằng mô hình phân biệt là phù hợp hơn với mục tiêu cùa quá trình học có giám sát và do đó có xu hướng hiệu quả hơn trong thực tế.

Học bán giám sát

Quá trình học bán giám sát nằm giữa quá trình học có và không có giám sát. Tập dữ liệu X làm đầu vào cho thuật toán học bán giám sát được chia làm hai phần, $X_l = (x_1, \ldots, x_l)$ với các nhãn đã được gán là $Y_l = (y_1, \ldots, y_l)$ và tập các điểm dữ liệu chưa được gán nhãn $X_u = (x_{l+1}, \ldots, x_{l+u})$. Thuật toán học bán giám sát có thể được định nghĩa theo hai cách nhìn sau:

- Học giám sát cộng thêm dữ liệu chưa gán nhãn (Supervised learning + additional unlabeled data).
- Học không giám sát cộng thêm dữ liệu gán nhãn (Unsupervised learning + additional labeled data).

Có một vấn đề liên quan đến học bán giám sát được gọi là *học có lý* đã được giới thiệu bởi Vapnik [2c.19] cách đây vài thập kỉ. Trong quá trình này, một tập hợp dữ liệu huấn luyện (đã được gán nhãn) và tập kiểm tra (chưa được gán nhãn) được cho trước. Ý tưởng của quá trình học có lý là thực hiện việc dự đoán chỉ cho các dữ liệu kiểm tra cho trước này. Điều này trái ngược với *học quy nạp*, mà mục tiêu là sinh ra một hàm dự đoán được định nghĩa trên toàn bộ không gian X.

Dễ dàng nhận thấy rằng, các thuật toán học có giám sát đòi hỏi tập dữ liệu huấn luyện lớn được gán nhãn từ trước để tạo ra bộ phân lớp có độ chính xác cao. Các dữ liệu huấn luyện được gán nhãn bằng tay bởi các chuyên gia trong lĩnh vực phân lớp. Trong bài toán nhận dạng tiếng nói, việc gán nhãn cho một số lượng khổng lồ các dữ liệu âm thanh đã được thu đòi hỏi con người phải nghe và ghõ lại lời thoại. Một ví dụ khác là quá trình phân lớp hàng tỷ trang web. Nhưng để phân loại tốt, con người phải đọc trước phần lớn các trang web đó. Quá trình gán nhãn dữ liệu này rõ ràng là tốn thời gian, công sức, tiền bạc và nhàm chán. Do vậy số lượng dữ liệu được gán nhãn thường ít hơn rất nhiều so với số lượng dữ liệu chưa được gán nhãn.

Liệu các phương pháp học bán giám sát có ích hay không?. Chính xác hơn, khi so sánh với học giám sát chỉ sử dụng dữ liệu gán nhãn, ta có thể hy vọng vào việc kết quả dự đoán sẽ chính xác hơn khi xét thêm các điểm dữ liệu không có nhãn. Nhìn chung câu trả lời là "có", tuy nhiên có một ràng buộc tiên quyết mà các dữ liệu không có nhãn buộc phải tuân theo, đó là sự phân bố các ví dụ dữ liệu phải phù hợp với bài toán phân lớp.

Về mặt công thức toán học, các tri thức trên $p(x)$ thu được thông quá việc phân tích dữ liệu không có nhãn phải chứa các thông tin hữu ích cho việc suy luận $p(x|y)$. Nếu không, quá trình học bán giám sát sẽ không mang lại cải thiện nào đối với quá trình học có giám sát, thậm chí có thể làm giảm độ

chính xác. Sau đây là một số giả thiết phải được thỏa mãn để thuật toán học bán giám sát thực sự hiệu quả hơn so với thuật toán học có giám sát.

Giả thiết tính trơn trong học bán giám sát: *Nếu hai điểm* x_1, x_2 *thuộc vùng có mật độ cao là gần nhau thì đầu ra tương ứng của chúng là* y_1, y_2 *cũng gần nhau..*

Giả thiết này ngụ ý là nếu hai điểm được liên kết bởi một đường dẫn trên vùng mật độ cao thì đầu ra của chúng nên gần nhau. Đối với bài toán phân lớp văn bản, ta hình dung như sau: Dữ liệu chưa gán nhãn sẽ cung cấp thông tin về phân phối xác suất đồng thời (*joint probability distribution*) của các từ khóa. Ví dụ với bài toán phân lớp trang web với hai lớp: trang chủ của một khoá học và không phải trang chủ của một khoá học. Ta coi trang chủ của một khoá học là hàm đích. Vì vậy, trang chủ của một khoá học sẽ là mẫu dương *(positive example)*, và các trang còn lại là các mẫu âm *(negative example)*. Giả sử chỉ sử dụng dữ liệu gán nhãn ban đầu ta xác định các văn bản có chứa từ "bài tập"*("homework")* thường thuộc lớp dương. Nếu sử dụng quan sát này để gán nhãn các dữ liệu chưa gán nhãn, chúng ta lại xác định được từ "bài giảng"*("lecture"*) xuất hiện thường xuyên trong các văn bản chưa gán nhãn mà được dự đoán là thuộc lớp dương. Sự xuất hiện của các từ "bài tập" và "bài giảng" trên một tập lớn các dữ liệu huấn luyện chưa gán nhãn có thể cung cấp thông tin hữu ích để xây dựng một bộ phân lớp chính xác hơn – xem xét cả "bài tập" và "bài giảng" như là các thể hiện của các mẫu dương.

Giả thuyết về tính đồng nhất trong các cụm: Nếu các điểm thuộc về cùng một cụm, thì chúng sẽ có khả năng thuộc về cùng một lớp là cao. Ngoài ra giả thuyết này có thể được phát biểu bởi giả thuyết tương đương về khả năng tách biệt của vùng phân bố mật độ thấp; đó là ranh giới quyết định có nhiều khả năng nằm ở vùng có mật độ phân bố thấp.

Giả thuyết về bề mặt tiếp xúc: Dữ liệu nhiều chiều nằm trên bề mặt có ít chiều hơn.

2.5.1. Một số phương pháp học bán giám sát

Có rất nhiều phương pháp học bán giám sát đã được đề xuất và nghiên cứu, nên trước khi quyết định lựa chọn phương pháp học cho một bài toán cụ thể cần phải xem xét các tính chất/cấu trúc của tập dữ liệu đối với miền bài toán đang được áp dụng. Chúng ta nên sử dụng phương pháp học mà giả thiết của nó phù hợp với cấu trúc của bài toán. Việc lựa chọn này có thể là khó khăn trong thực tế, tuy nhiên có một số gợi ý sau: Nếu các lớp tạo ra dữ liệu *có tính co cụm cao* thì EM là một sự lựa chọn tốt; nếu *các features có sự phân chia tự nhiên thành hai tập riêng biệt* thì co-training có thể phù hợp; nếu hai mẫu dữ liệu với *các feature tương tự nhau hướng tới thuộc về cùng một lớp* thì có thể sử dụng các phương pháp dựa trên đồ thị; nếu *các bộ phân lớp giám sát được xây dựng từ trước là phức tạp và khó sửa đổi* thì self-training sẽ là một lựa chọn ưu tiên. Phần tiếp theo giới thiệu các phương pháp học bán giám sát điển hình.

2.5.1.1. Thuật toán cực đại kỳ vọng toán

Thuật toán cực đại kỳ vọng (Expectation Maximization - EM) [2c.2] là một thuật toán tổng quát đánh giá sự cực đại khả năng*(ML – Maximum Likelihood)* mà dữ liệu là không hoàn chỉnh *(incomplete data)* hoặc hàm likelihood liên quan đến các biến ẩn*(latent variables)*. Ở đây, hai khái niệm "*incomplete data*" và "*latent variables*" có liên quan đến nhau: Khi tồn tại biến ẩn, thì dữ liệu là không hoàn chỉnh vì ta không thể quan sát được giá trị của biến ẩn; tương tự như vậy khi dữ liệu là không hoàn chỉnh, ta cũng có thể liên tưởng đến một vài biến ẩn với dữ liệu thiếu.

Thuật toán EM gồm hai bước lặp: E-step và M-step. Khởi đầu, nó gán giá trị ngẫu nhiên cho tất cả các tham số của mô hình. Sau đó, tiến hành lặp hai bước lặp sau:

E-step (Expectation step): Trong bước lặp này, nó tính toán likelihood mong muốn cho dữ liệu dựa trên các thiết lập tham số và *incomplete data*.

M-step (Maximization step): Tính toán lại tất cả các tham số sử dụng tất cả các dữ liệu. Khi đó, ta sẽ có một tập các tham số mới.

Tiến trình tiếp tục cho đến khi likelihood hội tụ, ví dụ như đạt tới cực đại địa phương. EM sử dụng hướng tiếp cận leo đồi, nên chỉ đảm bảo đạt được cực đại địa phương. Khi tồn tại nhiều cực đại, việc đạt tới cực đại toàn cục hay không là phụ thuộc vào điểm bắt đầu leo đồi. Nếu ta bắt đầu từ một đồi đúng *(right hill)*, ta sẽ có khả năng tìm được cực đại toàn cục. Tuy nhiên, việc tìm được *right hill* thường là rất khó. Có hai chiến lược được đưa ra để giải quyết bài toán này: Một là, chúng ta thử nhiều giá trị khởi đầu khác nhau, sau đó lựa chọn giải pháp có giá trị likelihood hội tụ lớn nhất. Hai là, sử dụng mô hình đơn giản hơn để xác định giá trị khởi đầu cho các mô hình phức tạp. Ý tưởng là: một mô hình đơn giản hơn sẽ giúp tìm được vùng tồn tại cực đại toàn cục, và ta bắt đầu bằng một giá trị trong vùng đó để tìm kiếm tối ưu chính xác khi sử dụng mô hình phức tạp hơn.

Thuật toán EM rất đơn giản, ít nhất là về mặt khái niệm. Nó được sử dụng hiệu quả nếu dữ liệu có tính phân cụm cao. Thuật toán EM trên mô hình trộn các phân bố đa thức đã được áp dụng cho bài toán phân lớp văn bản, và thu được bộ phân lớp tốt hơn khi chỉ được huấn luyện trên tập dữ liệu có nhãn [2c.14,2c.15].

2.5.1.2. Thuận toán tự học (self-traininig)

Thuật toán self-training (tự học) là một kỹ thuật thường được dùng cho quá trình học bán giám sát. Với thuật toán này đầu tiên, một bộ phân lớp sẽ

được huấn luyện từ một tập dữ liệu nhỏ đã được gán nhãn. Bộ phân lớp này sau đó được sử dụng để phân lớp các dữ liệu chưa có nhãn. Các điểm dữ liệu chưa có nhãn với độ tin cậy dự đoán cao được thêm vào tập dữ liệu huấn luyện, cùng với nhãn được dự đoán của chúng. Bộ phân lớp ban đầu sẽ được huấn luyện lại và quá trình này được lặp lại. Phương pháp tiếp cận EM ở trên có thể được xem như là một trường hợp đặc biệt của quá trình tự học mềm. Một số thuật toán cố gắng tránh việc làm tăng thêm lỗi cho bộ phân lớp bằng cách không học các dữ liệu với độ tin cậy của nhãn được gán thấp hơn một ngưỡng nào đấy.

Thuật toán tự học đã được áp dụng vào một số bài toán xử lý ngôn ngữ tự nhiên, ví dụ như bài toán phát hiện nhập nhằng nghĩa của các từ [2c.21], bài toán xác định các danh từ [2c.16], phân loại các đoạn hội thoại thành hai chủ đề, "có tình cảm" hoặc "không có tình cảm" [2c.13].

Bên cạnh đó thuật toán self-training còn được áp dụng vào quá trình phân tích và dịch máy và đã cho thấy tính hiệu quả cao [2c.18].

2.5.1.3. Thuật toán học SVM suy diễn (TSVM)

TSVM là một mở rộng của SVM chuẩn. Trong SVM chỉ có dữ liệu gán nhãn được sử dụng, mục đích là tìm siêu phẳng cực đại dựa trên các mẫu dữ liệu huấn luyện. Với TSVM, các điểm dữ liệu chưa gán nhãn cũng được sử dụng. Mục đích của TSVM là gán nhãn cho các điểm dữ liệu chưa gán nhãn để cho biên tuyến tính có lề phân cách là lớn nhất trên cả dữ liệu gán nhãn và dữ liệu chưa gán nhãn [2c.19] (xem *Hình 1*). TSVMs xây dựng liên kết giữa p(x) và đường biên quyết định phân lớp bằng việc không đặt đường biên này trong những vùng có mật độ phân bố cao. Nói cách khác, các điểm dữ liệu không có nhãn sẽ hướng dẫn đường phân cách tuyến tính ra xa các miền có mật độ cao.

Tuy nhiên việc tìm giải pháp bằng TSVM là bài toán có độ phức tạp NP. Hầu hết các nghiên cứu đều tập trung vào các thuật toán ước lượng đường phân tách. Các thuật toán đề xuất lúc đầu [2c.3, 2c.7, 2c.9] không thể xử lý khi dữ liệu không có nhãn lớn hơn vài trăm. SVM-light là một bộ công cụ thực hiện thuật toán TSVM [2c.11] khá hiệu quả và được sử dụng rộng rãi.

Một phương pháp huấn luyện bộ phân lớp TSVM dựa trên kỹ thuật lập trình bán định nghĩa (đã được áp dụng thành công cho bộ phân lớp SVM) đã được giới thiệu bởi Xu và Schuurmans [2c.20]. Quan trọng hơn, trong nghiên cứu này các tác giả đã đề xuất thành công một phiên bản DSP cho bài toán nhiều lớp.

Ngoài ra, TSVM có thể được xem như SVM với một số hạng điều khiển được thêm vào trên các dữ liệu không có nhãn [2c.5]. Đặt f(x) = h(x)+b với h thuộc H_K. TSVM sẽ được chuyển về bài toán tìm tối ưu sau:

$$\min_f \sum_{i=1}^{l}(1-y_i f(x_i))_+ + \lambda_1 \|h\|^2_{\mathcal{H}_K} + \lambda_2 \sum_{i=l+1}^{n}(1-|f(x_i)|)_+$$

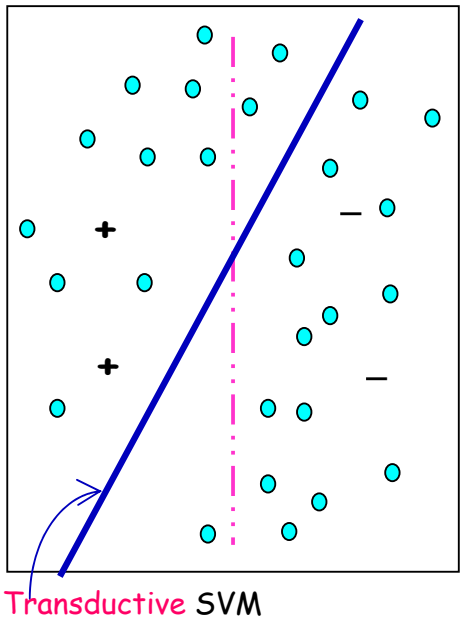
Trong đó $(z)_+ = \max(z,0)$. Thành phần cuối cùng tăng từ việc gán nhãn là sign(f(x)) cho các điểm dữ liệu chưa có nhãn x. Đường biên trên các điểm dữ liệu chưa có nhãn sẽ là sign(f(x))f(x)=|f(x)|. Hàm đo độ mất mát thông tin $(1-|f(xi)|)_+$ có đồ thị không lồi hình nón, và là nguyên nhân chính gây ra khó khăn trong việc giải bài toán tối ưu. (Xem *Hình 2.11*)

Phương pháp học TSVM:

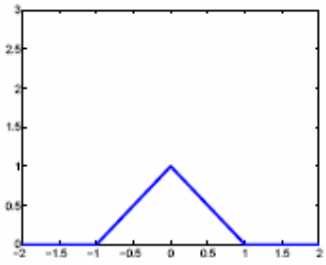
Qui ước:

+, -: các mẫu âm, dương

- : các mẫu chưa gán nhãn



Hình 2.10. Siêu phẳng cực đại. Đường chấm chấm là kết quả của bộ phân lớp SVM quy nạp, đường liên tục chính là phân lớp SVM truyền dẫn



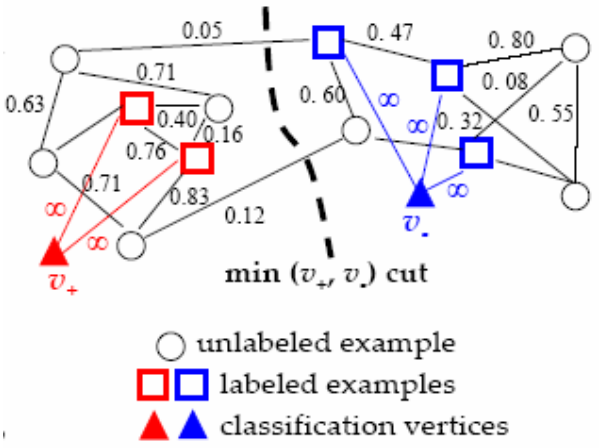
Hình 2. 110. Đồ thị Hàm đo độ mất mát thông tin $(1-|f(xi)|)_+$

2.5.1.4. Phân hoạch đồ thị quang phổ

Phương pháp phân hoạch đồ thị quang phổ (Spectral Graph Partitioning) xây dựng một đồ thị có trọng số dựa trên các mẫu gán nhãn và các mẫu chưa gán nhãn. Trọng số của các cạnh tương ứng với một vài mối quan hệ giữa các mẫu như độ tương tự hoặc khoảng cách giữa các mẫu.

Mục đích là tìm ra một nhát cắt cực tiểu (v_+, v_-) trên đồ thị (như hình 2). Sau đó, gán nhãn dương cho tất cả các mẫu chưa gán nhãn thuộc đồ thị con chứa v_+, và gán nhãn âm cho tất cả các mẫu chưa gán nhãn thuộc đồ thị

con chứa v_-. Phương pháp này đưa ra một thuật toán có thời gian đa thức để tìm kiếm tối ưu toàn cục thực sự của nó.



Hình 2. 12. Đồ thị trọng số dựa trên các mẫu dữ liệu gán nhãn và dữ liệu chưa gán nhãn

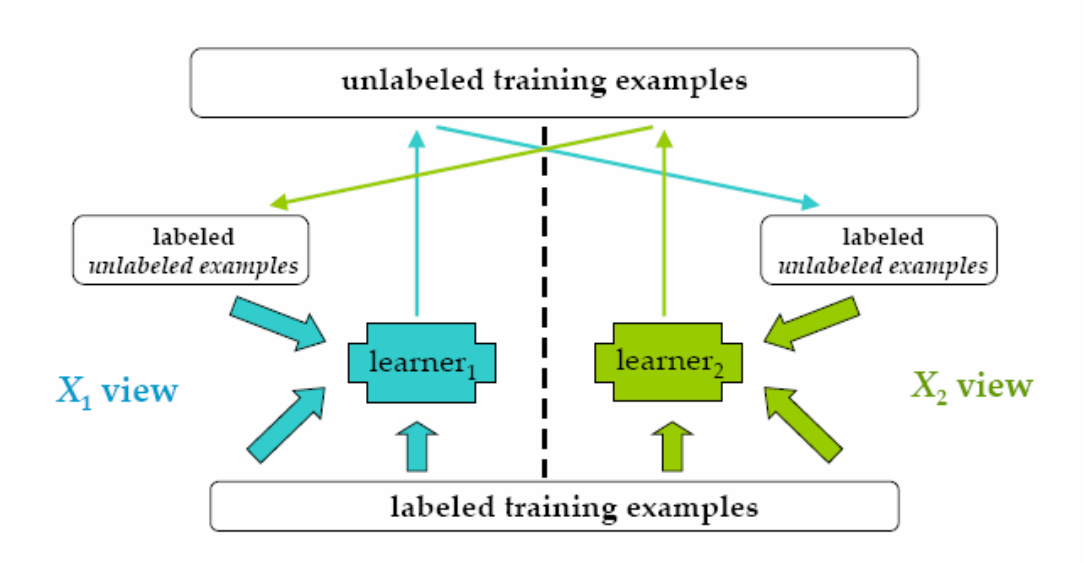
2.5.2. Thuật toán co-training

Thuật toán co-training dựa trên giả thiết rằng các *features* có thể được phân chia thành 2 tập con; Mỗi tập con phù hợp để huấn luyện một bộ phân lớp tốt. Hai tập con đó phải thoả *mãn tính chất độc lập điều kiện (conditional independent)* khi cho trước phân lớp. Thủ tục học được tiến hành như sau:

- Học 2 bộ phân lớp riêng rẽ bằng dữ liệu đã được gán nhãn trên hai tập thuộc tính con tương ứng.
- Mỗi bộ phân lớp sau đó lại phân lớp các dữ liệu *unlabel data*. Sau đó, chúng lựa chọn ra các *unlabeled data + nhãn dự đoán* của chúng (các mẫu có độ tin cậy cao) để dạy cho bộ phân lớp kia.
- Sau đó, mỗi bộ phân lớp được học lại *(re-train)* với các mẫu huấn luyện được cho bởi bộ phân lớp kia và tiến trình lặp bắt đầu.

Ràng buộc đối với thuật toán co-training là hai bộ phân lớp phải dự đoán trùng khớp trên tập dữ liệu lớn chưa được gán nhãn cũng như dữ liệu đã được gán nhãn.

Những ý tưởng về sử dụng sự dư thừa *feature* đã được thi hành trong một vài nghiên cứu. Yarowsky đã sử dụng co-training để phát hiện sự nhập nhằng về nghĩa cho các từ vựng [2c.21], ví dụ quyết định xem từ "plant" trong một ngữ cảnh cho trước có nghĩa là một sinh vật sống hay là một xí nghiệp. Yarrowsky tiến hành tìm nghĩa của từ bằng cách xây dựng một bộ phân lớp nghĩa *(sense classifier)* sử dụng ngữ cảnh địa phương của từ và một bộ phân lớp nghĩa dựa trên nghĩa của những lần xuất hiện khác trong cùng một văn bản; Riloff và Jones [2c.17] phân lớp cụm danh từ chỉ vị trí địa lý bằng cách xem xét chính cụm danh từ đó và ngữ cảnh ngôn ngữ mà cụm danh từ đó xuất hiện; Collin và Singer [2c.4] thực hiện phân lớp tên thực thể định danh sử dụng chính từ đó và ngữ cảnh mà từ đó xuất hiện. Sơ đồ co-training đã được sử dụng trong rất nhiều lĩnh vực như phân tích thống kê và xác định cụm danh từ. Hình vẽ 2.13 dưới đây cho chúng ta một cái nhìn trực quan của thiết lập co-training.



Hình 2. 11. Sơ đồ biểu diễn trực quan thiết lập co-training [2c.1]

Blum và Mitchell [2c.1] đã công thức hoá hai giả thiết của mô hình co-training và chứng minh tính đúng đắn của mô hình dựa trên thiết lập học giám sát theo mô hình PAC chuẩn. Cho trước một không gian các mẫu $X = X_1 \times X_2$, ở đây X_1 và X_2 tương ứng với hai khung nhìn *(views)* khác nhau của cùng một mẫu *(examples)*. Mỗi mẫu x vì vậy có thể được biểu diễn bởi một cặp (x_1, x_2). Chúng ta giả thiết rằng mỗi khung nhìn là phù hợp để phân lớp chính xác. Cụ thể, nếu D là một phân phối trên X, và C_1, C_2 là các lớp khái niệm *(concept classes)* được định nghĩa tương ứng trên X_1 và X_2; giả thiết rằng tất cả các nhãn trên các mẫu với xác suất lớn hơn không dưới phân phối D là trùng khớp với một hàm đích *(target function)* $f_1 \in C_1$, và cũng trùng khớp với hàm đích $f_2 \in C_2$. Nói cách khác, nếu f biểu diễn khái niệm đích kết hợp trên toàn bộ mẫu, thì với bất kỳ mẫu $x = x_1 \times x_2$ có nhãn l, ta có $f(x) = f_1(x_1) = f_2(x_2) = l$. Nghĩa là D gán xác suất bằng không cho mẫu (x_1, x_2) bất kỳ mà $f_1(x_1) \neq f_2(x_2)$

- ***Giả thiết thứ nhất***: Tính tương thích *(compatibility)*

  Với một phân phối D cho trước trên X, ta nói rằng hàm đích $f = (f_1, f_2) \in C_1 \times C_2$ là tương thích *(compatible)* với D nếu thoả mãn điều kiện: D gán xác suất bằng không cho tập các mẫu (x_1, x_2) mà $f_1(x_1) \neq f_2(x_2)$. Nói cách khác, mức độ tương thích của một hàm đích $f = (f_1, f_2)$ với một phân phối D có thể được định nghĩa bằng một số $0 \leq p \leq 1$: $p = 1 - \Pr_D\left[(x_1, x_2) : f_1(x_1) \neq f_2(x_2)\right]$.

- ***Giả thiết thứ hai***: Độc lập điều kiện *(conditional independence assumption)*

Ta nói rằng hàm đích f_1, f_2 và phân phối D thoả mãn giả thiết độc lập điều kiện nếu với bất kỳ một mẫu $(x_1, x_2) \in X$ với xác suất khác không thì,

$$\Pr_{(x_1,x_2)\in D}\left[x_1 = \hat{x}_1 \middle| x_2 = \hat{x}_2 \right] = \Pr_{(x_1,x_2)\in D}\left[x_1 = \hat{x}_1 \middle| f_2(x_2) = f_2\left(\hat{x}_2\right) \right]$$

và tương tự,

$$\Pr_{(x_1,x_2)\in D}\left[x_2 = \hat{x}_2 \middle| x_1 = \hat{x}_1 \right] = \Pr_{(x_1,x_2)\in D}\left[x_2 = \hat{x}_2 \middle| f_1(x_1) = f_1\left(\hat{x}_1\right) \right]$$

Kết quả nghiên cứu của Blum và Mitchell [2c.4] đã chỉ ra rằng, cho trước một giả thiết độc lập điều kiện trên phân phối D, nếu lớp đích có thể học được từ nhiễu phân lớp ngẫu nhiên theo mô hình PAC chuẩn, thì bất kỳ một bộ dự đoán yếu ban đầu nào cũng có thể được nâng lên một độ chính xác cao tuỳ ý mà chỉ sử dụng các mẫu chưa gán nhãn bằng thuật toán co-training. Hai ông cũng đã chứng minh tính đúng đắn của sơ đồ co-training bằng định lý sau:

Định lý (A.Blum & T. Mitchell).

Nếu C_2 có thể học được theo mô hình PAC với nhiễu phân lớp, và nếu giả thiết độc lập điều kiện thoả mãn, thì (C_1, C_2) có thể học được theo mô hình co-training chỉ từ dữ liệu chưa gán nhãn, khi cho trước một bộ dự đoán yếu nhưng hữu ích ban đầu $h(x_1)$.

Cho trước:

- L là tập các mẫu huấn luyện đã gán nhãn.
- U là tập các mẫu chưa gán nhãn.

Tạo một tập U' gồm u mẫu được chọn ngẫu nhiên từ U

Lặp k vòng

- Sử dụng L huấn luyện bộ phân lớp h_1 trên phần x_1 của x.
- Sử dụng L huấn luyện bộ phân lớp h_2 trên
- phần x_2 của x.
- Cho h_1 gán nhãn p mẫu dương và n mẫu âm từ tập U'.
- Cho h_2 gán nhãn p mẫu dương và n mẫu âm từ tập U'.
    - Thêm các mẫu tự gán nhãn này vào tập L.
- Chọn ngẫu nhiên $2p+2n$ mẫu từ tập U bổ sung vào tập U'.

Hình 2. 14. Sơ đồ thiết lập co-training gốc cho vấn đề hai lớp [1]

Việc so sánh tính hiệu quả của thuật toán co-training với thuật toán xây dựng mô hình trộn và thuật toán EM đã đã tiến hành thực nghiệm [2c.15]. Kết quả thực nghiệm đã cho thấy thuật toán co-training thực hiện hiệu quả nếu giả thiết về độc lập điều kiện được thỏa mãn. Ngoài ra, sẽ là tốt hơn gán nhãn thống kê tất cả các dữ liệu trong U, thay vì chỉ một số ít các dữ liệu có độ tin cậy cao. Phương pháp tiếp cận này được gọi là co-EM [2c.12]. Cuối cùng, nếu không tồn tại một phân tách tự nhiên các thuộc tính, một số tác giả đã tạo ra các phân tách nhân tạo bằng cách chia ngẫu nhiên tập thuộc tính thành hai tập con riêng biệt. Phương pháp tạo phân tách nhân tạo này cũng mang lại hiệu quả, tuy nhiên là không nhiều.

Có nhiều thể hiện của thuật toán co-training đã được nghiên cứu và thử nghiệm. Năm 2000, hai bộ phân lớp thuộc hai loại khác nhau đã được huấn luyện trên toàn bộ tập thuộc tính [2c.10]. Các dữ liệu trong U được gán nhãn

với độ tin cậy cao bởi một bộ phân lớp sẽ được sử dụng để huấn luyện bộ phân lớp còn lại, và ngược lại. Sau đó, Zhou và Goldman [2c.22] đã đề xuất một bộ phân lớp co-training được huấn luyện trên một khung nhìn. Tập hợp các bộ phân lớp với các độ lệch quy nạp khác nhau được huấn luyện một cách độc lập trên toàn bộ thuộc tính của dữ liệu đã được gán nhãn. Sau đó chúng được dùng để phân loại tất cả các dữ liệu chưa được gán nhãn. Nếu phần lớn bộ phân lớp đều thống nhất nhãn của một dữ liệu chưa được gán nhãn x_u với độ tin cậy cao, thì nhãn đó cùng với x_u sẽ được thêm vào tập dữ liệu huấn luyện. Tất cả bộ phân lớp sẽ được huấn luyện lại trên tập dữ liệu huấn luyện mới. Kết quả phân lớp cuối cùng sẽ được quyết định dựa trên quy tắc bầu cử có trọng số giữa các bộ phân lớp.

2.5.3. Thuật toán co-training áp dụng cho bài toán phân lớp web

Tính chất của trang web phù hợp cho việc áp dụng thuật toán co-training, trong đó nội dung của trang web sẽ tạo ra khung nhìn nhứ nhất và các text liên kết trên tất cả các trang web trỏ đến nó sẽ tạo ra khung nhìn thứ hai.

Blum và Mitchell đã tiến hành thực nghiệm co-training trong phân lớp trang web [2c.1]. Dữ liệu thực nghiệm gồm 1051 trang Web lấy từ 4 trường Đại học gồm: Cornell, University of Washington, University of Wisconsin, và University of Texas. Những trang web này được gán nhãn bằng tay vào một trong hai lớp xem nó có là trang chủ của một trường hay không? Để so sánh với phương pháp học giám sát, hai ông sử dụng bộ phân lớp Naïve Bayes trong đó lấy tập dữ liệu huấn luyện chỉ gồm 12 mẫu và huấn luyện hai bộ phân lớp tương ứng dựa trên nội dung trang web và text trong các hyperlink trỏ tới chính trang web đó. Ở đây, hai ông thực hiện một bộ phân lớp thứ 3 kết hợp đầu ra của hai bộ phân lớp trên.

	Page-based classifier	Hyperlink-based classifier	Combined classifier
Supervised training	12.9	12.4	11.1
Co-training	6.2	11.6	5.0

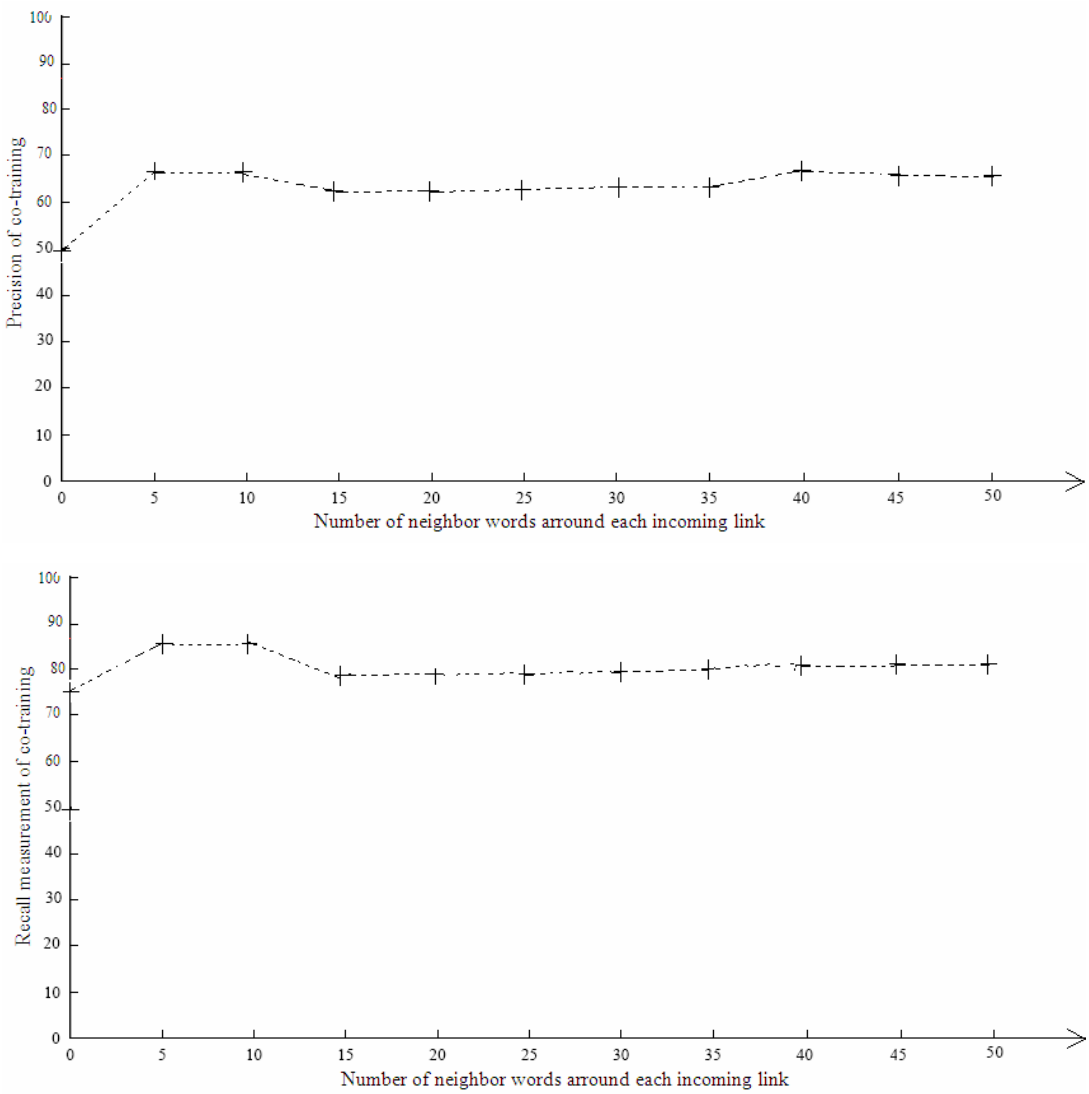
Hình 2. 15. Kết quả thực nghiệm thuật toán co-training với bài toán phân lớp web [2c.1]

Kết quả thực nghiệm được cho trên bảng trên, trong đó các con số thể hiện tỷ lệ lỗi trung bình trên 5 tập phân chia train/test ngẫu nhiên. Ta thấy rằng, đối với cả 3 bộ phân lớp thì phương pháp phân lớp dựa trên co-training đều cho kết quả tốt hơn so với phương pháp bán giám sát. Trên thực tế thì phương pháp dựa trên nội dung trang web và bộ phân lớp liên kết có tỷ lệ lỗi chỉ bằng một nửa so với phương pháp học giám sát.

Kết quả thực nghiệm của Blum và Mitchel cho thấy rằng phương pháp co-training thực sự hiệu quả trong vấn đề phân lớp tự động các trang web.Tuy nhiên, khi xem xét các kết quả thực nghiệm hai ông cũng nhân thấy rằng bộ phân lớp dựa trên anchor text không mấy hiệu quả dựa trên co-training. Điều này được giải thích do các anchor text này chứa ít thông tin hơn so với bộ phân lớp dựa trên nội dung của trang web. Thêm nữa, xuất phát từ thực tế rằng vùng văn bản lân cận anchor text thực sự có quan hệ ngữ nghĩa với nội dung của trang web. Trên cơ sở đó, đã xây dựng bộ phân lớp trang web dựa vào cotraing với chỉnh sửa là: Bộ phân lớp thứ nhất dựa vào nội dung trang web được giữ nguyên. Còn bộ phân lớp thứ hai, ngoài việc sử dụng thông tin về anchortext, vùng văn bản lân cận anchor text đó cũng được xem xét. Bộ phân lớp Naïve Bayes [2c.8] được sử dụng trên hai khung nhìn này. Những text này sẽ làm gia tăng lượng thông tin thu được cho bộ phân lớp và mở rộng khả năng dự đoán của nó.

Thực nghiệm được tiến hành trên tập dữ liệu WEBKB và trích chọn ra các text lân cận xung quang các liên kết đến. Kết quả thu được 39 và 95 trang web tương ứng thuộc vào các lớp “course home page” và “non-course home

page" tương ứng. Mỗi lần thực nghiệm 8 và 10 trang web tương ứng cho course và non-course được lấy ngẫu nhiên để huấn luyện, 8 và 75 trang web tương ứng để làm dữ liệu kiểm tra kết quả. Phần trang web còn lại được sử dụng làm dữ liệu chưa gán nhãn gồm 63 trang web. Kết quả thực nghiệm với độ đo Precision và Recall được cho ở dưới đây.



Hình 2. 12. Kết quả thực nghiệm phân lớp web sử dụng co-training và text lân cận

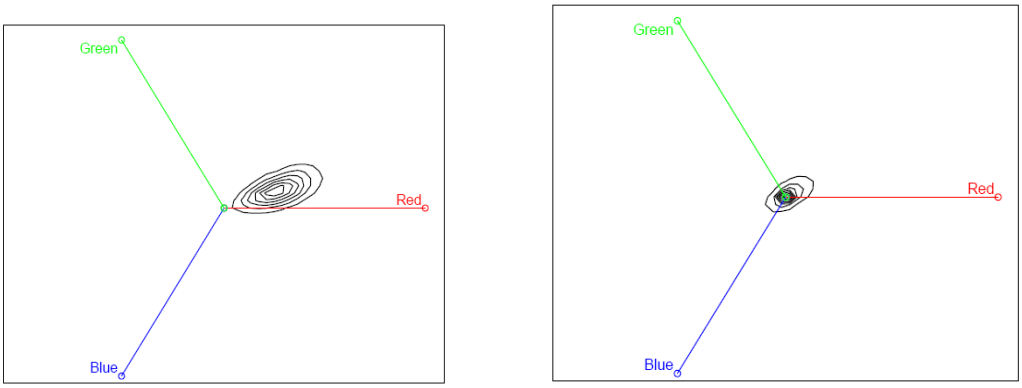
2.6. KỸ THUẬT LỌC ẢNH

2.6.1. Phát hiện màu sắc da người trong ảnh

Nhận thấy giữa các ảnh chứa các vùng da lớn và các ảnh khiêu dâm có mối tương quan với nhau. Do đó phát hiện màu sắc da trong ảnh là bước xử lý đầu tiên trong quá trình phân tích nội dung ảnh của các kỹ thuật lọc ảnh. Trong quá trình phát hiện màu da trong ảnh có những khó khăn như:

- Sự đa dạng của màu sắc da người, ví dụ: da trắng, da vàng, da màu.
- Sự đa dạng của các điều kiện chụp ảnh, ví dụ như: cường độ chiếu sáng, bản thân máy ảnh, các thuật toán nén ảnh, các nhiễu.

Không gian màu RGB là cách thuận tiện để mô tả màu sắc. Trong không gian này, màu sắc như là tổ hợp của 3 thành phần màu (red, green, blue). RGB là không gian màu được sử dụng rộng rãi để xử lý và lưu trữ dữ liệu ảnh màu. Tuy nhiên do có tương quan liên hệ cao giữa các kênh màu, RGB không phải là sự lựa chọn tốt nhất cho các thuật toán nhận dạng.



Hình 2. 13. Đường bao cho mô hình màu da và không phải là da trong không gian màu

Trong đánh giá của [182] trên tập dữ liệu gồm 18,696 ảnh. Tập dữ liệu chứa gần 2 tỷ điểm ảnh chỉ ra rằng quá trình phát hiện da có thể gặp khó khăn do có sự chồng lên nhau giữa các mô hình da và không phải là da. Tuy nhiên

sự chồng lên nhau này chỉ có ảnh hưởng lớn nếu số trường hợp quan sát được cả là da và không phải là da là đáng kể. Hình vẽ 2.17 thể hiện điều đó.

Để giảm ảnh hưởng do thay đổi của điều kiện rọi sáng, không gian màu thường được sử dụng phổ biến là RGB chuẩn hóa (normalized RGB) hoặc HSV. Chuẩn hóa RGB là một cách thể hiện màu sắc dễ dàng đạt được từ các giá trị RGB qua thủ tục chuẩn hóa:

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \qquad b = \frac{B}{R+G+B}$$

Tổng của 3 thành phần chuẩn hóa $r + g + b = 1$. Sự phụ thuộc của các thành phần *r, g* vào độ sáng của màu sắc nguồn RGB được loại bỏ. Một thuộc tính quan trọng của cách thể hiện này là loại bỏ được ảnh hưởng của các nguồn sáng xung quanh và sự thay đổi hướng của bề mặt đối tượng so với nguồn sáng. Điều này cộng với sự đơn giản trong phép chuyển nên không gian màu này được dùng phổ biến trong các nghiên cứu.

Để đánh giá khả năng phát hiện điểm da trong ảnh, tiểu chuẩn thường được sử dụng là tỷ lệ dương sai (*false positive*) và tỷ lệ phát hiện (*detection*). Tỷ lệ dương sai là tỷ lệ các điểm không phải là da được phân loại là da. Tỷ lệ phát hiện là tỷ lệ các điểm da được xác định là da. Người dùng thường mong muốn kết hợp cả hai tỷ lệ này trong đánh giá.

Các tiếp cận phân loại da thường được phân thành hai loại cơ bản: các tiếp cận xây dựng mô hình phát hiện điểm ảnh da dựa trên định nghĩa trực tiếp và các tiếp cận dựa trên thống kê. Các tiếp cận dựa trên thống kê có thể chia xa hơn thành cách tiếp cận thống kê dựa vào tham số (parametric) và thống kê không dùng tham số (non parametric).

Phần sau trình bày phương pháp phát hiện da dựa trên điểm ảnh, các điểm ảnh được xem xét độc lập với các điểm ảnh lân cận. Bên cạnh đó, cách tiếp cận khác là điểm ảnh được xem xét là da hay không có tính tới các điểm

lân cận của điểm ảnh, với nhận xét rằng quyết định liệu một điểm ảnh là da hoặc không phải là da chỉ dựa trên các giá trị màu của riêng điểm ảnh đó thường dẫn đến các điểm ảnh cô lập được phát hiện là da trong một vùng không phải là da. [185] đề xuất cách tiếp cận phát hiện da dựa trên vùng. Đoạn cuối trình bày phương pháp phát hiện da dựa trên vùng.

2.6.2. Phát hiện da dựa trên điểm ảnh

Trong cách tiếp cận này, mỗi điểm ảnh được phân loại là da hoặc không phải là da một cách riêng biệt, độc lập với các điểm ảnh lân cận của nó. Có 3 cách tiếp cận thường được áp dụng là: định nghĩa mô hình da trực tiếp, sử dụng mô hình không tham số histogram và mô hình định nghĩa da sử dụng tham số. Các cách tiếp cận lần lượt được trình bày ở dưới đây.

2.6.2.1. Mô hình da được định nghĩa trực tiếp

Trong mô hình này bộ phân loại da được định nghĩa trực tiếp thông qua một số luật thể hiện các vùng biên của da trong không gian màu.

Ví dụ: Peer, et al 2003 [2d.4]:

(R, G, B) được phân loại là da nếu :

$R > 95$ và $G > 40$ và $B > 20$ và

$max\{R,G,B\} - min\{R,G,B\} > 15$ và

$|R - G| > 15$ và $R > G$ và $R >$ B

Sự đơn giản của phương pháp này đã thu hút khá nhiều nghiên cứu. Ưu điểm của phương pháp phát hiện da được định nghĩa trực tiếp là: đơn giản và tính trực giác *của* các luật phân loại. Điều này giúp cho quá trình phân loại màu sắc da đạt được tốc độ cao. Tuy nhiên khó khăn của các phương pháp này là cần yêu cầu tìm mô hình không gian mầu tốt và các luật quyết định qua thống kê. Phương pháp đề xuất gần đây sử dụng các thuật toán học máy để tìm ra không gian màu thích hợp và các luật đơn giản như của [2.210] đã cho kết quả khá tốt.

2.6.2.2. Mô hình phân phối da không tham số

Ý tưởng chính của các mô hình sử dụng phương pháp này là đánh giá phân phối da từ các dữ liệu đào tạo mà không phải suy diễn một mô hình rõ ràng cho màu da. Trong cách tiếp cận này, các lược đồ histogram được sử dụng để thể hiện mật độ trong không gian màu. Điều kiện rọi sáng đặc biệt trong các ảnh trên web thường không biết trước và được ghi lại dưới các điều kiện khác nhau. Tuy nhiên, với một số lượng lớn dữ liệu điểm ảnh đào tạo được gán nhãn bao gồm tất cả các loại da người (Châu Á, Châu Âu, Châu Phi), vẫn có thể mô hình phân phối của các màu sắc da và không phải da trong không gian màu.

Kết qủa của các phương pháp này còn được đề cập tới như là quá trình xây dựng Bản đồ xác suất da (Skin Probability Map – SPM) – gán một giá trị xác suất cho mỗi điểm của không gian màu đã được rời rạc hóa.

Một phương pháp phổ biến trong mô hình không tham số là sử dụng bảng tra cứu đã được chuẩn hóa (*Normalized lookup table*) (LUT) [2d.12]. Cách tiếp cận này dựa trên *histogram* để phân vùng màu da. Không gian màu được lượng tử hóa thành dãy các ô, mỗi ô tương ứng với một dải màu. Mỗi ô sẽ lưu số lần các màu tương ứng có trong tập các ảnh đào tạo có chứa da. Sau quá trình đào tạo, histogram được chuẩn hóa (normalized)- chuyển các giá trị của histogram thành các phân phối xác suất rời rạc:

$$P_{skin}(c) = \frac{skin[c]}{Norm}$$

Trong đó $skin[c]$ là giá trị của ô trong histogram tương ứng với vector màu c. $Norm$ là tổng giá trị của tất cả các ô [2.182].

Các giá trị sau khi chuẩn hóa của các ô tạo nên các xác suất trước (likelihood) Các giá trị $P_{skin}(c)$ thực sự là một xác suất có điều kiện $P(c|skin)$ – xác suất quan sát thấy màu c, khi chúng ta thấy một điểm màu da.

$$P(skin \mid c) = \frac{P(c \mid skin)P(skin)}{P(c \mid skin)P(skin) + P(c \mid \neg skin)P(\neg skin)}$$

$P(skin|c)$ và $P(skin|\neg c)$ được tính trực tiếp từ các lược đồ màu da và không phải da

Các xác suất trước $P(skin)$ và $P(\neg skin)$ có thể được ước lượng từ số mẫu là da và không phải là da trong tập mẫu đào tạo.

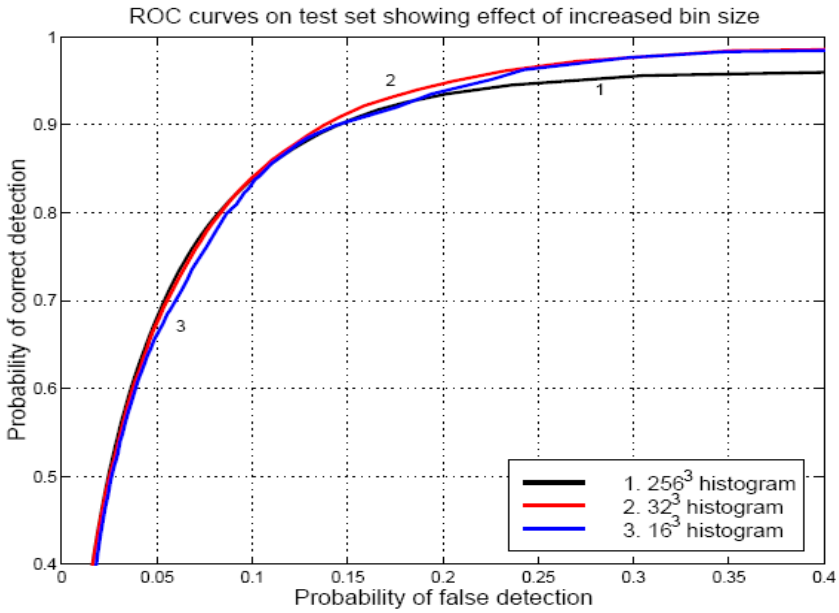
Có thể tránh phải tính toàn bộ biểu thức trên bằng cách chỉ cần so sánh $P(skin|c)$ và $P(\neg skin|c)$

$$\frac{P(skin|c)}{P(\neg skin|c)} = \frac{P(c|skin)P(skin)}{P(c|\neg skin)P(\neg skin)}$$

So sánh biểu thức trên với ngưỡng sẽ tạo ra luật quyết định:

$$\frac{P(c|skin)}{P(c|\neg skin)} > \Theta$$

Các phương pháp không tham số (non-parametric) cho kết quả nhanh trong cả quá trình đào tạo và phân loại, nhưng yêu cầu không gian lưu trữ lớn và không có khả năng tổng quát hóa dữ liệu đào tạo. Ví dụ nếu RGB sử dụng 8 bit cho một màu, chúng ta cần 2^{24} phần tử để lưu trữ các xác suất màu. Để giảm số lượng không gian lưu trữ, các kỹ thuật lấy mẫu thô không gian màu thường được sử dụng, ví dụ: 128x128x128, 64x64x64 và 32x32x32. Đánh giá các cách lấy mẫu khác nhau, kết quả nghiên cứu trong [2.182] cho thấy cách lấy mẫu 32x32x32 cho kết quả tốt nhất. Hình 2 thể hiện kết quả so sánh giữa các cách lấy mẫu khác nhau.



Hình 2. 18. Mô hình ROC so sánh kết quả với các kích cỡ khác nhau của histogram

Hiệu năng của bộ phân loại điểm da là khá tốt trong điều kiện các ảnh Web thường biến đổi. Trong đánh giá của Michael J. Jones và James M. Rehg [2.180] trên tập dữ liệu gồm 18,696 ảnh. Tập dữ liệu chứa gần 2 tỷ điểm. Kết quả tốt nhất được báo cáo là bộ lọc có thể phát hiện gần 80% các điểm ảnh là da với tỷ lệ sai FP là 8.5%, hoặc 90% phát hiện đúng với tỷ lệ sai FP là 14.2% false positives.

2.6.2.3. Mô hình phân phối có tham số

Với yêu cầu mô hình thể hiện gọn nhẹ, cùng với khả năng tổng quát hóa dữ liệu đào tạo các mô hình phân phối có tham số như Gausian đơn, Gausian trộn (*mixture*) đã được phát triển. Mô hình này có đặc điểm là không sử dụng nhiều dữ liệu trong quá trình đào tạo như trong mô hình không tham số. Với mô hình Gaussian đơn: Phân phối màu sắc da có thể được mô hình bởi một hàm mật độ phân phối xác suất Gausian Elliptic.

$$p(c \mid skin) = \frac{1}{2\pi \mid \Sigma_s \mid^{1/2}} e^{-\frac{1}{2}(c-\mu_s)^T \Sigma_s^{-1}(c-\mu_s)}$$

Ở đây *c* là một vector màu, μ_s và Σ_s là các tham số phân phối được ước lượng từ tập dữ liệu đào tạo.

$$\mu_s = \frac{1}{n}\sum_{j=1}^{n} c_j \quad , \quad \Sigma_s = \frac{1}{n-1}\sum_{j=1}^{n}(c_j - \mu_s)(c_j - \mu_s)^T$$

Với *n* là tổng số các mẫu màu da c_j

Xác suất *p*(*c*|*skin*) có thể được sử dụng trực tiếp để đánh giá khả năng là màu da của một màu *c*

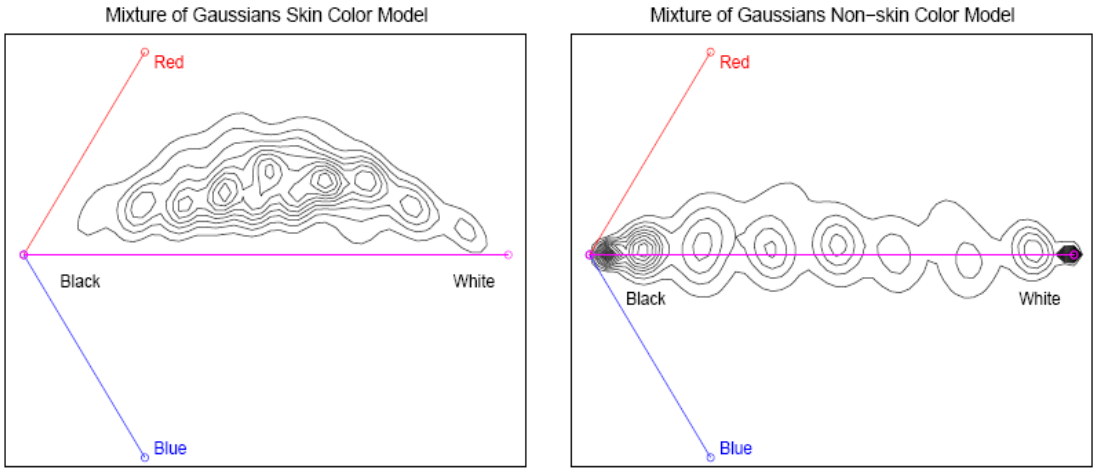
Mô hình Gaussian trộn được thể hiện là tổng của các gausian nhân

$$P(\mathbf{x}) = \sum_{i=1}^{N} w_i \frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)}$$

Đóng góp của Gaussian thứ i được thể hiện thông qua trọng số nhân wi Lee và Yoo [2d.11] bằng cách xem xét các phân phối màu sắc là da và không phải da trong một số không gian màu đã kết luận rằng vùng màu da không đủ tốt để xấp xỉ bởi mô hình Gaussian do sự không đối xứng của các vùng màu. Sử dụng mô hình Gaussian đối xứng thường dẫn đến tỷ lệ *false positive* cao. Lee và Yoo đề xuất mô hình biên elliptical (elliptical boundary model). Mô hình này cho kết quả phát hiện da tốt hơn so mô hình Gaussian đơn và trộn.

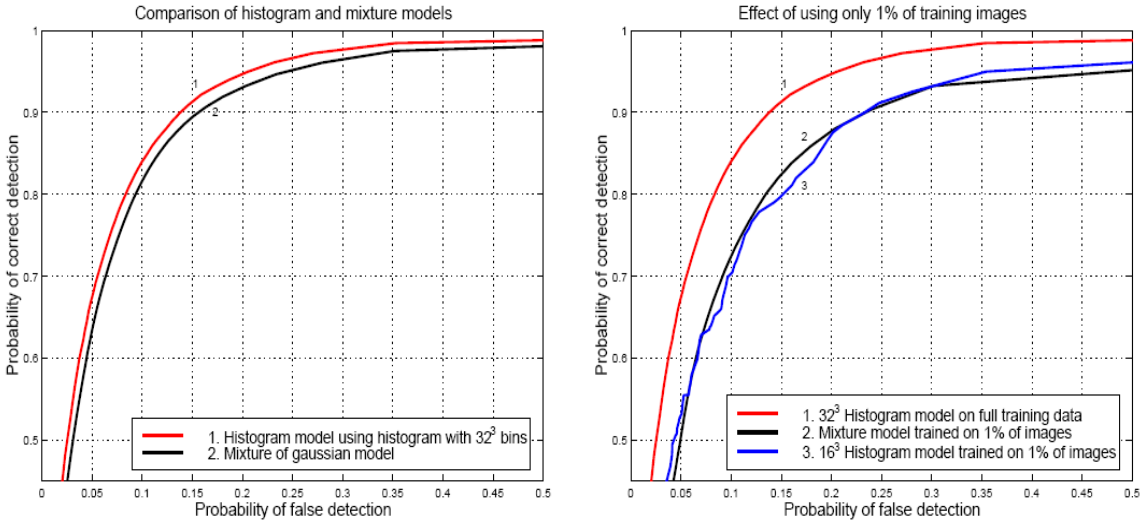
Đặc điểm của các phương pháp thống kê sử dụng tham số (parametric) là có khả năng tổng quát hóa dữ liệu đào tạo không đầy đủ. Chúng được thể hiện thông qua một số lượng nhỏ các tham số và yêu cầu không gian lưu trữ nhỏ. Tuy nhiên chúng thực sự chậm trong cả hai quá trình đào tạo và làm việc. Bên cạnh đó, hầu hết các mô hình tham số bỏ qua các thống kê về màu sắc không phải là da, do đó tỷ lệ *false positive* của các phương pháp này cao hơn so với các phương pháp không tham số. Tỷ lệ *false positive* là số phần trăm ảnh điểm ảnh không phải là da được phân loại là da. Tỷ lệ phát hiện là số phần trăm điểm ảnh là da được phân loại đúng

Để đánh giá kết quả của mô hình Gaussian trộn, Jones và Reg [2.180] đã đào tạo hai mô hình trộn riêng biệt cho các lớp da và không phải là da. Các mô hình được đào tạo sử dụng một phiên bản song song của thuật toán EM chuẩn. Các mô hình sử dụng cùng tập dữ liệu như trong mô hình histogram không tham số đề cập ở phần trên.



Hình 2. 19. Đường bao cho mô hình da và không da sử dụng Gaussian trộn

Hình 2.19 thể hiện mô hình phân phối màu sắc da và không phải là da sau quá trình đào tạo. Kết quả cho thấy có sự chồng lên nhau giữa các mô hình. Kết quả đánh giá trên cùng một tập dữ liệu cho thấy mô hình *histogram* cho kết quả tốt hơn một chút so mô hình Gaussian trộn (*mixture*). Hình 2.20 so sánh kết quả đánh giá của hai phương pháp trên cùng tập dữ liệu



Hình 2. 20. So sánh mô hình sử dụng histogram và mô hình gaussian trộn

Kết quả càng khác nhau nếu số lượng ảnh dùng trong quá trình đào tạo là nhỏ (ví dụ 1%). Bên cạnh đó có thể so sánh mô hình histogram và gaussian trộn trên khía cạnh thời gian tính toán và yêu cầu lưu trữ. Mô hình gaussian trộn mất nhiều thời gian đánh kể trong quá trình đào tạo khi so sánh với mô hình histogram.

Mô hình Gaussian trộn cần khoảng 24 giờ để đào tạo cả hai mô hình trộn da và không phải da sử dụng 10 Alpha trạm làm việc song song. Ngược lại, mô hình histogram có thể được tạo trong khoảng thời vài phút trên một máy trạm làm việc. Mô hình gaussian trộn cũng chậm hơn mô hình histogram trong quá trình phân loại do tất cả các mô hình gausian nhân cần được đánh giá trong quá trình tính toán xác suất của một điểm màu. Ngược lại, mô hình histogram cho kết quả phân loại nhanh do để tính xác suất là da của một điểm màu chỉ cần hai lần tra bảng.

Tuy nhiên về không gian lưu trữ, mô hình gaussian trộn gọn nhẹ hơn nhiều trong cách thể hiện dữ liệu. Ngược lại, mô hình histogram kích thước 32x32x32 cần 262 Kbytes không gian lưu trữ (giả sử cần 4 byte cho 1 ô).

2.6.3. Phát hiện da dựa trên vùng

Với nhận xét rằng, quyết định liệu một điểm ảnh là da hoặc không phải là da chỉ dựa trên các giá trị màu của riêng điểm ảnh đó thường dẫn đến các điểm ảnh cô lập được phát hiện là da trong một vùng không phải là da. Các điểm này là *false positives* và cần loại trừ. Một cách tiếp cận khác trong quá trình phát hiện điểm da trong ảnh là không coi các điểm ảnh là độc lập, cách tiếp cận này có xem xét tới phân bố không gian của các điểm trong quá trình phát hiện da.

2.6.3.1. Một số mô hình phát hiện da dựa theo vùng

Jedynak và cộng sự [2.184, 2.185] xây dựng một mô hình entropy cực đại áp dụng vào quá trình phát hiện da. Mô hình entropy cực đại (MaxEnt) có lịch sử lâu dài, được xem như là một tổng quát hóa của "Nguyên lý suy diễn không đầy đủ" [2d.14]. Mô hình MaxEnt được áp dụng vào quá trình phát hiện da như sau:

(1) Thuộc tính là ràng buộc màu sắc giữa các điểm ảnh kề nhau và xác suất của chúng. (2) Tính toán lược đồ histogram cho các thuộc tính này dựa vào CSDL Compaq đã phân mảnh sẵn. (3) Mô hình được gọi là mô hình đạo hàm bậc 1 (First Order Model-FOM). Mô hình này bao gồm các ràng buộc về biến đổi màu sắc giữa các điểm ảnh lân cận. (4) Quá trình ước lượng tham số tối ưu cho mô hình được giải quyết bằng xấp xỉ cây Bethe (*Hình 2*). Sử dụng thuật toán Belief Propagation cho phép đạt được kết quả nhanh, chính xác cho, nhanh trong tính xác suất là da tại vị trí điểm ảnh. Kết quả đầu ra của quá trình phát hiện da là bản đồ đa mức xám, mức độ xám tỷ lệ với xác suất là da. Người dùng có thể áp dụng ngưỡng để đạt được ảnh nhị phân. Phương pháp phát hiện da này cho kết quả tốt hơn so với kết quả đạt được trong [2.180] .

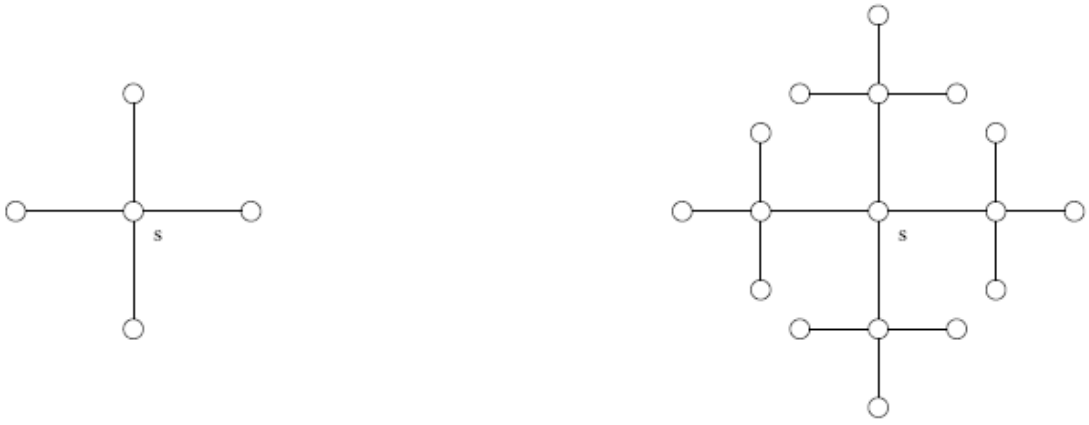
Cụ thể trong mô hình MaxEnt, quan tâm đến xác suất phân phối của màu sắc và mức độ là da của các điểm lân cận trong CSDL đào tạo.

Với hai điểm ảnh lân cận s và t , $p(x_s, x_t, y_s, y_t)$ nên gần với các quan sát trong tập dữ liệu đào tạo. Do đó mô hình MaxEnt cho phân phối $p(x, y)$ trong phát hiện da tuân theo tập các ràng buộc:

$$C : \forall s \in S, \forall t \in V(s), \forall x_s \in C, \forall x_t \in C,$$
$$\forall y_s \in \{0,1\}, \forall y_t \in \{0,1\},$$
$$p(x_s, x_t, y_s, y_t) = q(x_s, x_t, y_s, y_t)$$

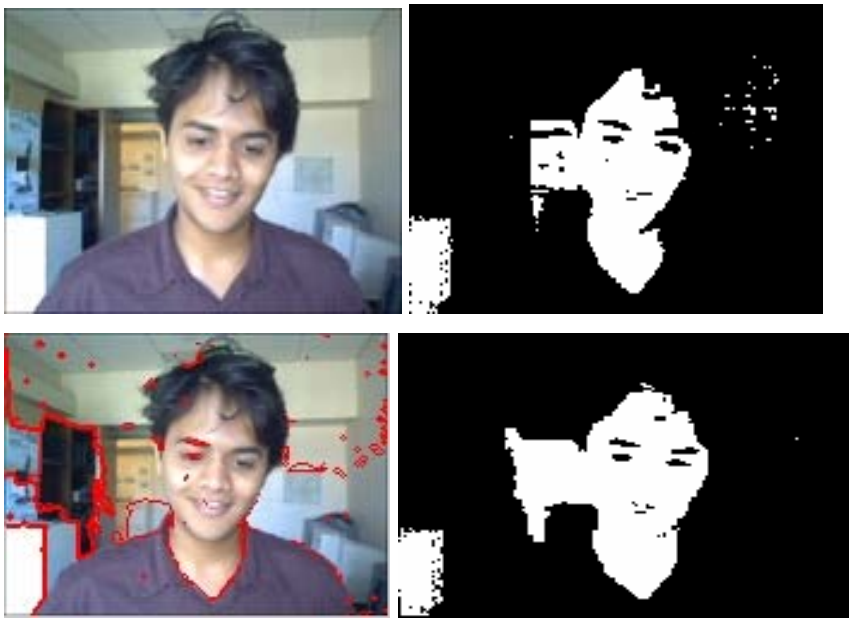
Số lượng $q(x_s, x_t, y_s, y_t)$ tỷ lệ với số lần quan sát tập giá trị (x_s, x_t, y_s, y_t) cho các cặp điểm lân cận, không tính đến hướng của các điểm ảnh s và t trong tập dữ liệu đào tạo.

Với S là tập các điểm ảnh của một ảnh, $C = \{0,...,255\}^3$. Màu sắc của một điểm ảnh $s \in S$ là $x_s \in C$. Mức độ da của điểm ảnh s là y_s, với $y_s = 1$ nếu s là một điểm ảnh da và $y_s = 0$ nếu điểm ảnh không phải là da. Tập các điểm lân cận của s là $V(s)$. Hệ thống sử dụng 4 điểm lân cận, không tính đến hướng. Hình 2.21 thể hiện cây Bethe với gốc là điểm ảnh s, các điểm lân cận không tính tới hướng, có độ xâu lần lượt là 1 và 2



Hình 2. 21. Cây Bethe có gốc tại s, với độ sâu là lần lượt là 1 và 2

Có thể thấy trong Hình 2.22, bằng cách xem xét tới các điểm lân cận của điểm ảnh, số lượng các điểm FP đã được giảm xuống. Một lượng lớn vùng FP bên phải của mặt đã được loại trừ.

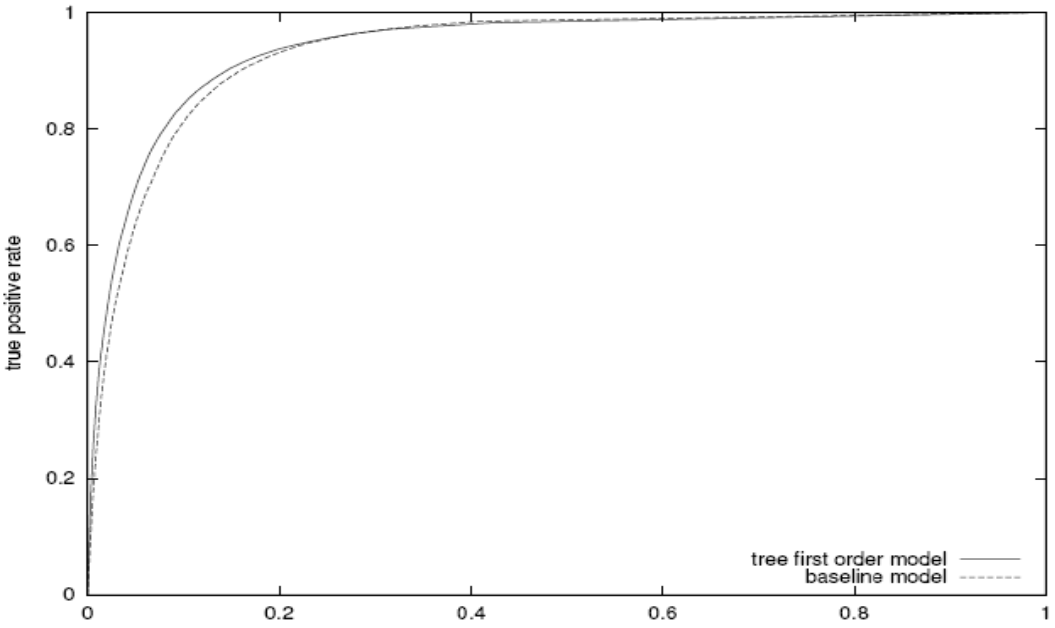


Hình 2. 22. So sánh phương pháp phát hiện da dựa trên điểm và dựa trên vùng. Hàng trên là kết quả dựa trên điểm, Hàng dưới là kết quả dựa trên vùng

Một số kết quả phát hiện da của thuật toán TFOM được thể hiện trong Hình 2.23. Hầu hết các vùng da được phát hiện chính xác. Tuy nhiên còn một số lỗi khi một số đối tượng có màu giống như da xuất hiện ở nền. Ví dụ, các phần của tường gạch được đánh giá là da. Một số vùng da bị thiếu do bị thay đổi bởi ánh sang mạnh. Ví dụ như tay của người trong hình thứ 2.

Hình 2. 23. Hàng trên là ảnh gốc. Hàng dưới: bản đồ da tương ứng sau khi sử dụng phương pháp TFOM

Hình 2.24 so sánh so sánh khả năng phát hiện da dựa trên tỷ lệ TP và FP. Đánh giá dựa trên cơ cở dữ liệu Compaq chứa 18,696 ảnh. Tập dữ liệu được chia ngẫu nhiên thành hai phần bằng nhau. Phần đầu chứa hơn 2 tỷ điểm ảnh và được sử dụng là tập dữ liệu đào tạo. Kết quả cho thấy mô hình TFOM cho kết quả tốt hơn so với mô hình đường cơ bản (baseline) được triển khai trong [2.180].



Hình 2. 144. Đường ROC cho mô hình cây đạo hàm bậc 1 (TFOM) và mô hình đường cơ bản (baseline)

2.6.3.2. Đánh giá hiệu năng của các phương pháp phát hiện da

Để đánh giá hiệu năng của các phương pháp mô hình hóa màu da khác nhau, điều kiện đánh giá được sử dụng là giống nhau. CSDL nổi tiếng nhất cho đào tạo và đánh giá phát hiện da là CSDL Compaq [2.182]. Bảng 2.6 là

các kết quả tốt nhất được thông báo bởi các tác giả qua thống kê của Vehznevets và cộng sự [2.183].

Bảng 2.6. Kết quả của các bộ phát hiện da khác nhau được thông báo bởi các tác giả

Phương pháp	**TP**	**FP**
Bayes SPM in RGB	80%	8.5%
J. Jones and James M. Rehg [2.182]	90%	14.2%
Bayes SPM in RGB (Brand and Mason [2.212])	93.4%	19.8%
Maximum Entropy Model in RGB (Jedynak và cộng sự [2.184])	80%	8%
Gaussian Mixture models in RGB (Jones and Rehg [2.182])	80% 90%	9.5% 15.5%
SOM in TS (Brown et al. 2001 [2.211])	78%	32%
Elliptical boundary model in CIE-xy, (Lee and Yoo [2d.11])	90%	20.9%
Single Gaussian in CbCr [Lee and Yoo 2002]	90%	33.3%
Gaussian Mixture in IQ [Lee and Yoo 2002]	90%	30.0%
Thresholding of I axis in YIQ [Brand and Mason 2000]	94.7%	30.2%

Bayes SPM và kế thừa của nó là mô hình entropy cực đại được đề xuất bởi Jedynak và cộng sự [2.184] cho kết quả tốt nhất (*false positive* thấp hơn đối với cùng tỷ lệ phát hiện đúng). Các mô hình tham số (Gausian, mixture Gausian, và mô hình biên Elliptic) cho kết quả thấp hơn.

2.6.3.3. Trích chọn các thuộc tính từ bản đồ điểm da

Kết quả ra của quá trình phát hiện ra là một bản đồ da đa mức xám, mức độ xám chỉ độ mức độ tin tưởng điểm ảnh đó là da. Bước tiếp theo là phân loại ảnh dựa theo hình dáng của vùng da.

Các thuộc tính ở mức cao dựa trên các vùng da có thể giúp phân biệt ảnh khiêu dâm và không. Để có tính áp dụng thực tế, các thuộc tính được chọn nên đơn giản, dễ tính toán. Đầu tiên ảnh được nhị phân hóa bằng cách sử dụng phân ngưỡng. Tiếp theo các phép biến đổi hình thái *đóng/mở* được áp dụng để loại bỏ các nhiễu và nối các vùng bị dời. Các vùng da nhỏ được coi là có tác động không đáng kể và bị loại bỏ.

Nhóm nghiên cứu của Forsyth và Fleck [2.186] đã thiết kết và triển khai thuật toán tìm người khỏa thân trong ảnh. Thuật toán sử dụng các vùng da được phát hiện là đầu vào của của bộ lọc dựa trên cấu trúc không gian. Theo nội dung của bài báo trên 60% ảnh có chứa người khỏa thân được phát hiện tuy nhiên thời gian cần để phát hiện nhóm người là khoảng 6 phút khi chạy chương trình trên máy tính làm việc. Do quá trình tính toán của thuộc tính liên quan đến hình dạng và kết cấu tiêu tốn thời gian.

Jones và Regh [2.180] tính toán một vector thuộc tính đơn giản từ kết quả ra của bộ phát hiện da. Các thuộc tính được sử dụng bao gồm:

- Phần trăm các điểm ảnh được phát hiện là da
- Xác suất trung bình của các điểm ảnh
- Số lượng các điểm ảnh của vùng da lớn nhất
- Số lượng các vùng da

....

Các thuộc tính này có thể được tính toán nhanh chóng từ ảnh đầu vào, do đó giúp quá trình phát hiện nội dung ảnh nhanh chóng.

Với cách tiếp cận trong [2.184], để phát hiện nhanh ảnh khiêu dâm, một số thuộc tính đơn giản được trích ra từ ảnh bản đồ da. Hầu hết các thuộc tính mô tả đặc tính của các vùng da được phát hiện dựa trên các hình ellipse bao do chúng đáp ứng yêu cầu: đơn giản và nắm bắt được các thông tin quan trọng về hình dáng. Các quan sát chỉ ra rằng với cách tiếp cận dựa trên phát hiện da, các ảnh chân dung thường có xu hướng bị phát hiện là ảnh khiêu dâm

do các ảnh chân dung có một lượng lớn da. Các hình ellipse được hy vọng sẽ giúp phân biệt ảnh chân dung và ảnh khiêu dâm. Hai ellipse bao được sử dụng là ellipse bao tất cả các vùng da (Global Fit Ellipse-**GFE**) Ellipse bao toàn cục và ellipse bao vùng da lớn nhất (Local Fit Ellipse-**LFE**) Ellipse bao địa phương [2d.8].

Hình 2. 15. Ảnh bên trái: ảnh gốc, Ở giữa: GFE trong bản đồ da. Bên phải: LFE trong bản đồ da

Trong nghiên cứu này, Jedynak và cộng sự [184-185] sử dụng 9 thuộc tính đơn giản được trích ra từ bản đồ da để tạo nên vector thuộc tính của ảnh:
3 thuộc tính đầu là các thuộc tính toàn cục:

- Xác suất trung bình của da trong toàn bộ ảnh
- Xác suất trung bình của da trong GFE
- Số lượng vùng da trong ảnh

6 thuộc tính tiếp theo được tính dựa trên LFE:

- Khoảng cách từ tâm của LFE tới tâm của ảnh
- Góc của trục lớn của LFE với trục ngang
- Tỷ lệ của trục nhỏ với trục lớn của LFE
- Tỷ lệ diện tích của LFE với toàn bộ ảnh
- Xác suất trung bình của da trong LFE
- Xác suất trung bình của da ngoài LFE

2.6.3.4. Nhận dạng mẫu

Các bước trích chọn thuộc tính được mô tả ở phần trước tạo ra một vector thuộc tính cho mỗi ảnh. Công việc tiếp theo là tìm ra luật quyết định tối ưu để phân biệt ảnh khiêu dâm hay không dựa trên vector thuộc tính này. Bosson et al. [2d.6] đề xuất một hệ thống phát hiện ảnh khiêu dâm và đã được tích hợp vào trong hệ thống thương mại. Hệ thống cũng phân loại ảnh dựa trên phát hiện da. Bosson đã so sánh các mô hình phân loại và chỉ ra rằng phương pháp phân loại sử dụng mạng noron đa tầng (MLP) cho kết quả thống kê tốt hơn một số tiếp cận khác như: mô hình tuyến tính tổng quát hóa, mô hình k-điểm lân cận gần nhất và mô hình máy vector hỗ trợ (SVM).

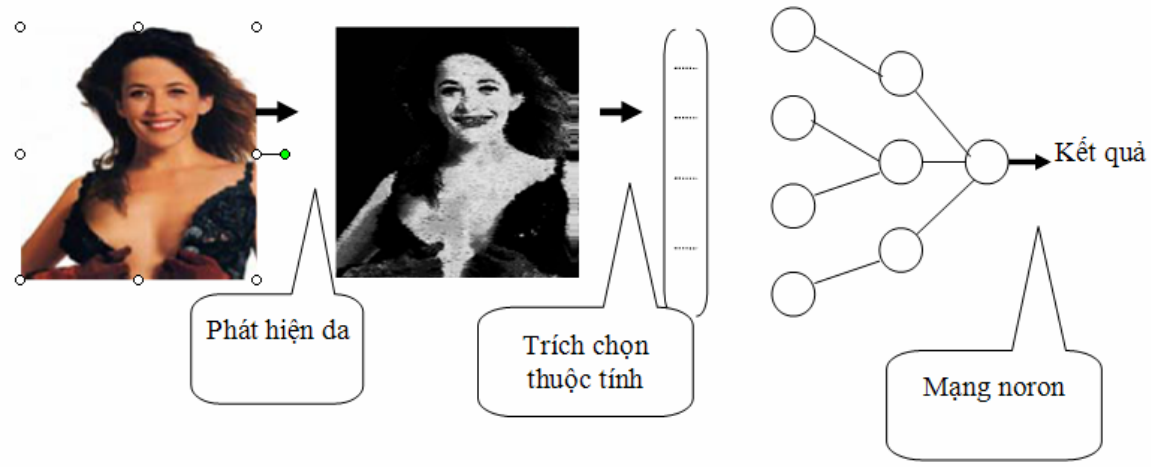
Mạng bán tuyến tính lan truyền ngược được trình bày bởi Rumelhart, Hinton và Williams [2d.15] được coi là một hệ thống hiệu quả cho quá trình học mẫu phân biệt.

Mô hình MLP gồm tầng vào i , tầng ẩn j và tầng ra k. Thủ tục học bắt đầu với tập giá trị tham số ngẫu nhiên w_{ij} . Các mẫu đào tạo p được sử dụng như là giá trị vào để đánh giá kết quả ra. Nếu kết quả ra có lỗi lớn, cần điều chỉnh $\Delta_p w_{ij}$ cho tất cả các tham số w_{ij} trong mạng. Thủ tục này tiếp tục được áp dụng cho tất cả các mẫu khác trong tập đào tạo để tạo ra kết quả $\Delta\, w_{ij}$ cho tất cả các tham số trong mạng. Sau quá trình điều chỉnh này, lỗi trong quá trình đào tạo sẽ giảm xuống với quá trình lặp. Thủ tục này sẽ hội tụ tới một tập các giá trị w ổn định.

Kết quả ra của mạng là một số nằm giữa 0 và 1. Kết quả càng gần 1, càng tăng khả năng ảnh đưa vào có nội dung khiêu dâm. Ngưỡng được sử dụng để quyết định nội dung ảnh trong quá trình phân loại dựa trên xem xét đến tỷ lệ FP và TP.

Quá trình phát hiện nội dung ảnh khiêu dâm về tổng quan có thể chia làm ba bước như trong Hình 2.28.

Bước đầu tiên là phát hiện bản đồ da từ ảnh gốc. Bước tiếp theo là trích chọn các thuộc tính là ảnh bản đồ da. Bước cuối cùng là phân loại ảnh bằng cách đưa các thuộc tính đã tìm ra ở bước trước thành đầu vào bộ phân loại ảnh (sử dụng mạng noron đa tầng lan truyền ngược).

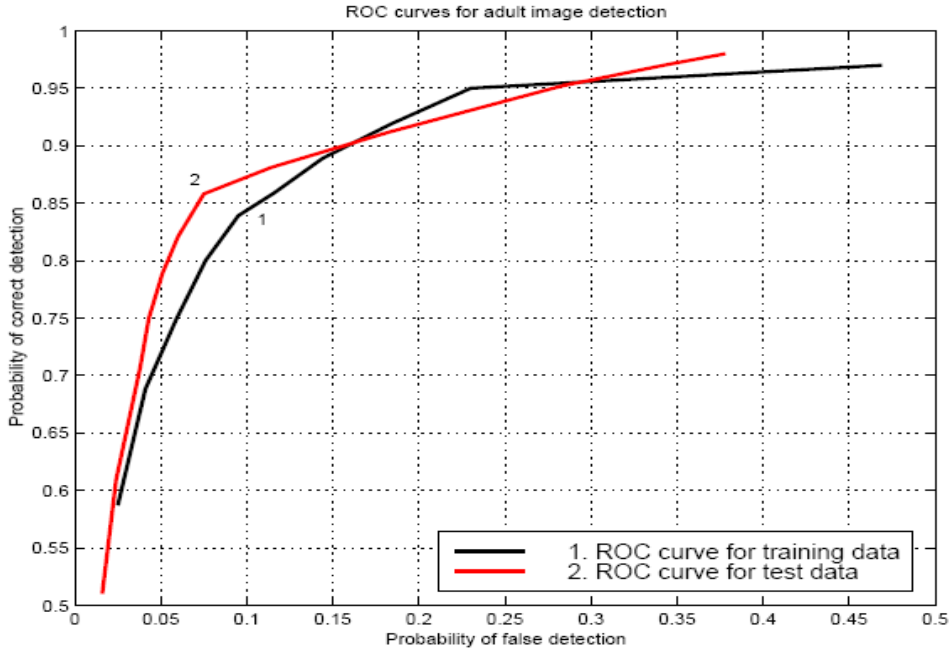


Hình 2. 16. Các bước trong quá trình phát hiện nội dung ảnh

2.6.3.5. Đánh giá kết quả phân loại ảnh của một số phương pháp

Trong thử nghiệm tiến hành trong [2.182], Jones, Rehg đã sử dung 10679 bức ảnh đã được phân loại thành các tập ảnh khiêu dâm và không khiêu dâm dùng trong quá trình đào tạo bộ phân loại sử dụng mạng neural. Có 5453 ảnh khiêu dâm và 5226 ảnh không khiêu dâm. Kết quả ra của mạng neural là một giá trị giữa 0 và 1, với giá trị 1 chỉ ảnh khiêu dâm. Kết quả được phân ngưỡng để đạt được kết định nhị phân. Bằng cách sử dụng các ngưỡng khác nhau. Để đánh giá khả năng phát hiện ảnh khiêu dâm của bộ phân loại, Michael J. Jones và James M. Rehg [2.180, 2.182] đã thu thập các ảnh đánh giá từ web. Một tập A thu thập sử dụng các site khiêu dâm như làm điểm bắt đầu và đã thu thập được khá nhiều ảnh. Tập B sử dụng các site không khiêu dâm như làm điểm bắt đầu và đã thu được một ít ảnh khiêu dâm. Tập A bao gồm 2365 trang web chứa 5241 ảnh khiêu dâm và 6082 ảnh không khiêu dâm. Tập

B bao gồm 2692 trang web chứa 3 ảnh khiêu dâm và 13970 ảnh không khiêu dâm. Jones và Rehg [2.180, 2.182] Sử dụng các ảnh khiêu dâm từ tập A và các ảnh không khiêu dâm từ tập B để đánh giá bộ lọc. Đường ROC của kết quả đánh giá được chỉ ra trong Hình 2.29. Bộ phát hiện đạt được, ví dụ 85.8% phát hiện đúng trên tập A với 7.5% tỷ lệ sai FP trên tập ảnh không khiêu dâm từ tập B.



Hình 2. 17. Kết quả đánh giá bộ lọc trên tập dữ liệu đào tạo và kiểm tra

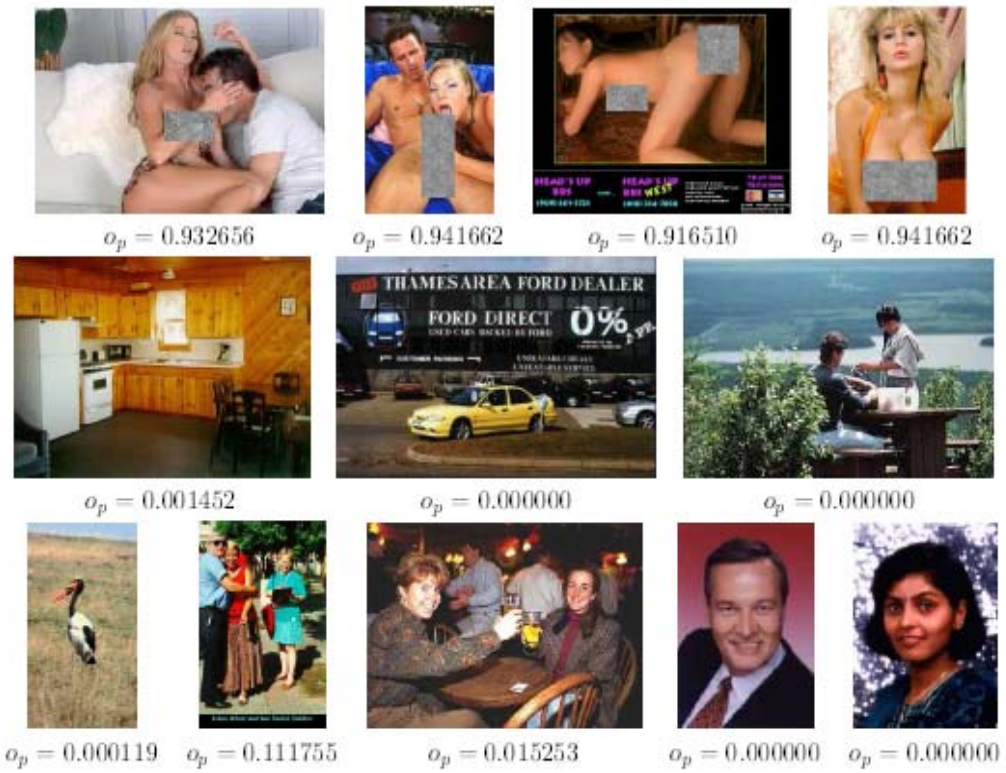
Kết quả này khá tốt khi xem xét đến sự đơn giản của các vector thuộc tính được sử dụng trong quá trình phân loại. Trong các hệ thống trước, ví dụ Forsyth [2d.10], quá trình phát hiện ảnh khiêu dâm thường dựa vào chủ yếu vào hình dạng và kết cấu... Kết quả cho thấy màu sắc da đóng vai trò quan trọng trong quá trình phát hiện nội dung ảnh. Bảng 2.7 cho kết quả so sánh trực tiếp giữa phương pháp của Jones và Jehg [2.180, 2.182] và phương pháp của Forsyth [2.186].

Bảng 2. 7. Kết quả so sánh khả năng phát hiện ảnh khiêu dâm [182]

Phương pháp	**TP**	**FP**
Forsyth [2.186]	79.3%	11.3%
Jones, Rehg [2.180]	88.0%	11.3%
Forsyth [2.186]	42.7%	4.2%
Jones, Rehg [2.180]	75.0%	4.2%

Bảng so sánh cho thấy cùng với tỷ lệ FP, phương pháp đề xuất của Jones, Rehg [2.180,2.182] cho kết quả phát hiện đúng tốt hơn so với kết quả được báo cáo trong [2.186]. Ngoài ra, phương pháp đề xuất bởi Jones và Rehg cũng cho tốc độ xử lý nhanh hơn nhiều do sử dụng phương pháp tra bảng trong quá trình phát hiện màu da, cộng với chỉ sử dụng các thuộc tính đơn giản trong quá trình trích chọn thuộc tính. Thời gian để xử lý được thông báo là ít hơn thời gian đọc file ảnh từ đĩa cứng lưu trữ, dưới 1 giây trên 1 máy trạm làm việc.

Dựa trên các đánh giá so sánh các phương pháp hiện nay, phương pháp của Jedynak và cộng sự [2.184-2.185] cho kết quả tốt hơn các phương pháp khác. Các đánh giá trong [2.185] được tiến hành với CSDL bao gồm 10.168 ảnh chụp từ CSDL Compaq và Poesia. CSDL được chia làm 2 phần bằng nhau một cách ngẫu nhiên. 2 phần này sau đó được sử dụng như là tập đào tạo và kiểm tra tương ứng. Hình 14 chỉ ra một số kết quả với ảnh là khiêu dâm và không. Kết quả ra của MLP được chỉ ra phía dưới các ảnh tương ứng.

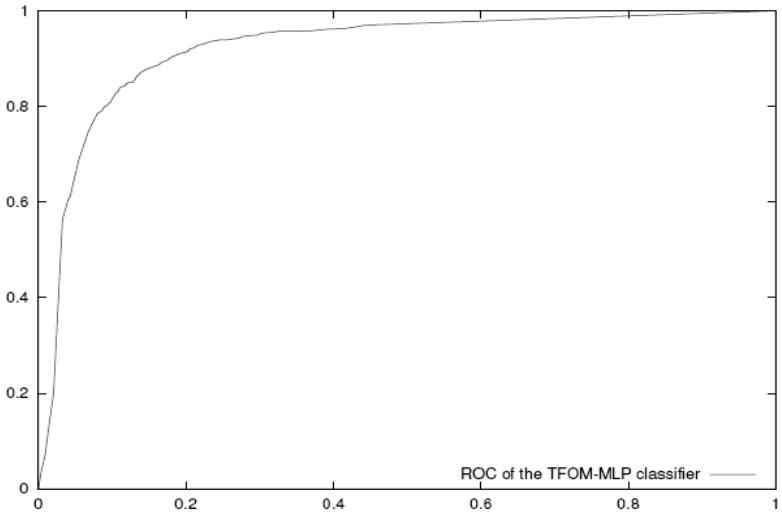


Hình 2. 18. Một số kết quả sau khi phân loại [2.185]

Có một số trường hợp bộ phân loại ảnh làm việc chưa thật tốt. Ví dụ một số trường hợp phân loại sai trong Hình 2.29.

Hình 2. 19. Một số phân loại sai [2.184, 2.185]

Trong ảnh khiêu dâm thứ nhất chứa nhiều ảnh nhỏ nối với nhau. LFE của ảnh này là rất lớn, nhưng xác suất da trung bình trong LFE lại nhỏ. Ảnh thứ hai bị phát hiện sai là do ảnh có màu sắc như da và xác suất da trung bình trong GFE và LFE là rất cao. Ảnh thứ ba là một ảnh chân dung nhưng bị xác định là khiêu dâm do có quá nhiều da và quần áo có màu như da. Tuy nhiên các ảnh khiêu dâm trong các trang web có xu hướng xuất hiện cùng nhau và được bao quanh bởi chữ, các yếu tố này giúp tăng khả năng phát hiện nội dung khiêu dâm.



Hình 2. 20. Đường cong ROC của phương pháp [2.185]

Đường cong ROC kết quả của đánh giá được thể hiện trong Hình 2.32. Thời gian phân loại khoảng 1.51×10^{-5} giây/điểm ảnh, khoảng 1 giây cho phân loại 1 ảnh 256x256.

Thuật toán lọc ảnh theo thuật toán đề xuất bởi Jedynak và cộng sự [2.184-2.185] đã được triển khai trong module lọc ảnh, là một phần trong dự án Poesia (www.**poesia**-filter.org).

Qua các đánh giá cho thấy phương pháp tiếp cận của Jones và Rehg [2.180, 2.182] và Jedynak và cộng sự [2.184-2.185] cho kết quả lọc ảnh khá tốt.

CHƯƠNG III

XÂY DỰNG SẢN PHẨM PHẦN MỀM LỌC NỘI DUNG INTERNET

3.1. SẢN PHẨM LỌC NỘI DUNG WEB (SP.01)

3.1.1. Sơ lược về thông tin được cung cấp trên Web

Các tài liệu được cung cấp trên Internet qua những phân tích đánh giá từ [2.53] cung cấp một số đánh giá tổng thể như sau:

- Độ lớn trung bình của một trang HTML khoảng 16.3 Kbytes, trung bình theo logarithmic khoảng 9.8Kbytes
- Khoảng 1.7% số trang có độ lớn nhỏ hơn 400 bytes
- 10.2% các trang HTML có độ lớn nhỏ hơn 2000 bytes
- 51% các trang HTML nhỏ hơn 12800 bytes (như vậy kích thước median khoảng 12Kbytes)
- 90.5% các trang HTML nhỏ hơn 38000 bytes
- 99% các trang HTML nhỏ hơn 74000 bytes
- Các trang HTML có kích thước lớn hơn 1 Mbytes rất ít, chỉ khoảng 0.003%

Trong mỗi một trang HTML, số lượng liên kết đến các trang khác cũng mang đến nhiều thông tin khác. Chính vì thế, số lượng này cũng cần được các hệ thống lọc coi trọng. Theo [2.53], những liên kết được trỏ bằng thẻ <a href= … > có những đánh giá như sau:

- Số liên kết trung bình trong một trang HTML là 39.3
- 37% các liên kết này được đề cập với một địa URL tường minh bắt đầu bằng http:.
- Khoảng 0.25% các liên kết đó sử dụng giao thức HTTPS
- Khoảng 0.44% liên kết đó sử dụng giao thức truyền tệp FTP

- 58.7% các liên kết đó là cục bộ, chỉ trỏ đến những tài liệu được cung cấp trong cùng một nhà cung cấp

Việc phân bố các liên kết trong một trang cũng tương đối phẳng:

- 8.8% các trang không có liên kết
- 11.1% trang có nhiều nhất một liên kết
- 50% các trang có không quá 21 liên kết
- 90.3% các trang không có quá 80 liên kết
- 1% các trang có số lượng liên kết lớn hơn 240

Liên quan đến các dữ liệu hình ảnh trong một tài liệu, dựa trên việc phân tích các thẻ <img src=…>, [2.53] đánh giá:

- 29.9% các trang không chứa hình ảnh
- Số lượng ảnh trung bình trong một trang HTML là 18.8
- 51% các trang chứa nhiều nhất là 6 ảnh
- 9% các trang HTML không chứa quá 140 ảnh
- 22% các ảnh được trỏ đến bằng từ khoá http:, tức thông thường trỏ đến site khác
- 77% các ảnh được trỏ đến là cục bộ, nằm trên cùng site
- Các ảnh được trỏ đến bằng giao thức FTP là vô cùng ít, dưới 0.01%

3.1.2. Yêu cầu của hệ thống lọc web

Hệ thống lọc web trên Internet phải đảm bảo lọc được những tài liệu có nội dung không lành mạnh thông qua những hình thức truyền thông và thể hiện được miêu tả cụ thể dưới đây:

3.1.2.1. Nội dung cần lọc/kiểm duyệt

Hệ thống lọc nội dung phải có khả năng lọc và xử lý theo từng ngữ cảnh những loại tài liệu có nội dung không lành mạnh. Việc phân loại nội dung không lành mạnh sẽ phải dựa trên những tiêu chí sau đây:

- Danh sách những miền nội dung ảnh hưởng đến vấn đề lối sống, văn hoá, chính trị, kinh tế của người Việt Nam cũng như đất nước Việt Nam. Những miền nội dung này sẽ được xác định và cập nhật thông qua sự phối hợp liên bộ giữa hai bộ Công An và bộ Văn hoá.
- Những tài liệu có nội dung đồi truy, hình ảnh khiêu dâm phi nghệ thuật
- Những tài liệu có nội dung kích động bạo lực, gây thù hằn dân tộc, phân biệt chủng tộc
- …

3.1.2.2. Cơ chế lọc nội dung

Hệ thống web sẽ được xây dựng dựa theo những yêu cầu cụ thể sau :

Giao thức và kênh truyền

Hệ thống lọc web phải có khả năng lọc được những tài liệu có nội dung xấu được truyền tải qua những kênh và giao thức trên mạng Internet sau:

- Giao thức:
  - HTTP: giao thức truyền dữ liệu siêu văn bản

- Kênh truyền tải nội dung :
  - Web : phương tiện chia xẻ những nội dung văn bản text/hình ảnh dựa trên giao thức HTTP. Mỗi tài liệu Web được xác định nguồn cung cấp thông qua địa chỉ URL – Uniform Resource Locator.

Kiểu lọc

Hệ thống lọc nội dung Internet phải cho phép thực hiện lọc thông tin truyền tải trên mạng Internet theo cả ba hình thức: lọc cụ thể, lọc ngoài trừ và lọc theo nội dung.

- Lọc cụ thể và lọc ngoại trừ (Implicite & Explicite Filtering, source-based filtering): kiểu lọc này phải sử dụng hai danh sách chứa các địa chỉ cần phải ngăn chặn (danh sách đen – black list) hay những địa chỉ mà luôn cung cấp những nội dung tốt (danh sách trắng – white list)

- Lọc dựa theo nội dung (content based filtering): kiểu lọc này cần có sự phân tích nội dung của tài liệu yêu cầu để từ đó xác định mức độ xấu/tốt của tài liệu đó và từ đó ra quyết định chặn hay không đối với tài liệu đó. Việc phân tích nội dung tài liệu sẽ được thực hiện chủ yếu với các kỹ thuật trong hai ngành xử lý ngôn ngữ tự nhiên (NLP based techniques) và phân tích hình ảnh (Images analysis techniques).

Định dạng tài liệu

Các chuẩn khuôn dạng tài liệu mà hệ thống lọc nội dung có thể sử lý được bao gồm những kiểu sau :

- Text format :
  - HTML
  - Microsoft Word files
  - PDF (Portable Document Format) and PS (PostScript)
- Image format :
  - JPG/JPEG (Joint Photographic Expert Group):
  - GIF (Graphic Interchange Format)
  - PNG (Portable Network Graphic)
  - BMP (Windows Bitmap)

Phân loại tài liệu theo nội dung

Thông qua hệ thống lọc, các modules lọc văn bản tiếng Việt, Anh, lọc ảnh phải có nhiệm vụ phân loại mức độ xấu/tốt của một tài liệu mà người dùng yêu cầu hoặc truyền tải qua cổng Internet quốc gia. Việc phân loại nội dung phải được tuân thủ theo những yêu cầu/nguyên tắc đã được đề cập trong tài liệu báo cáo [2].

Độ đo nội dung của một tài liệu, được ký hiệu là M_c, sẽ là một số tự nhiên, có khoảng thay đổi từ 0 đến 100 tuỳ theo mức độ tốt, xấu của tài liệu đó. Giá trị M_c của một tài liệu đạt cực đại, 100, khi nó được đánh giá là có nội

dung rất xấu (bạo lực, khủng bố, đồi truỵ, …), và có giá trị cực tiểu khi tài liệu đó thể hiện nội dung được đánh giá là có không nằm trong những phổ nội dung xấu.

3.1.2.3. Người sử dụng hệ thống

Hệ thống lọc nội dung Internet hướng đến sự phân biệt ba đối tượng người dùng sau:

- Người quản trị hệ thống: người quản trị hệ thống phải có những nhiệm vụ sau:
    - Cài đặt hiệu chỉnh hệ thống
    - Theo dõi và quản lý toàn bộ quá trình hoạt động của hệ thống
    - Cập nhật những danh sách trắng/đen từ phía thiết lập chính sách
- Người thiết lập chính sách lọc
    - Cập nhật các chính sách lọc, các miền nội dung cần phải lọc, các danh sách trắng/đen.
    - Xác định, cập nhật các ngưỡng cấm, ngưỡng cần kiểm soát của một tài liệu.
- Người kiểm soát lại nội dung trong trường hợp có nghi vấn
    - Phân tích, đánh giá lại mức độ xấu của tài liệu trong trường hợp hệ thống đánh giá độ xấu của tài liệu nằm trong ngưỡng kiểm soát lại.
    - Cập nhật lại hai danh sách trắng và đen dựa và kết quả phân tích đánh giá trên
- Người sử dụng
    - Những người truy nhập Internet đế tìm kiếm thông tin, đặt biệt là trẻ em, học sinh, sinh viên,

3.1.2.4. Ngữ cảnh lọc

Hệ thống lọc nội dung phải đảm bảo lọc được thông tin tại những cổng Internet quốc gia. Ngoài ra, hệ thống cũng phải có những tiện ích khác cho phép lọc thông tin theo nội dung tại những nơi công cộng, trường học, thư viện, ISP, …

3.1.2.5. Hiệu năng hệ thống lọc

Việc đánh giá hiệu năng của hệ thống lọc nội dung phải được xác định theo những chuẩn quốc tế như ISO/IEC 9126 và những tham số phổ dụng được diễn tả trong hai bảng dưới đây :

Bảng 3. 1. Các chỉ tiêu đánh giá chất lượng theo chuẩn ISO/IEC 9126

Quality framework as defined by ISO/IEC 9126	
Quality Factor	Quality Subcharacteristics
Functionality	Suitability – Accurateness – Interoperability – Compliance – Security
Reliability	Maturity – Fault Tolerance – Recoverability
Usability	Understandability – Learnability – Operability
Efficiency	Time Behaviour – Resource behaviour
Maintainability	Analysability- Changeability – Stability – Testability
Portability	Adaptability – Installability – Conformance – Replaceability

Bảng 3. 2. Các chỉ tiêu đánh giá dựa theo benchmark

Quality factors as defined by the Benchmarking Study		
Attribute	**Quality Factor**	**Quality Criteria**
Quality	Functionality	Usefulness Flexibility Interoperability
	Reliability	Maturity Stability Security
	Usability	Understandability Resource requirements Friendliness Operability
	Effectiveness	Blocking performance Over-blocking Localisation

3.1.2.6. Giải pháp lọc web

Để có thể quản lý và bảo đảm an toàn-an ninh thông tin qua cổng Internet quốc gia thì điều kiện tiên quyết là ta phải kiểm soát được các luồng dữ liệu qua đó, cụ thể đối với bộ lọc web là các luồng dữ liệu sử dụng giao thức HTTP. Điều này sẽ được thực hiện thông qua cả các thiết bị phần cứng (router, switch) và phần mềm (firewall, transparent proxy, phần mềm kiểm soát và lọc nội dung).

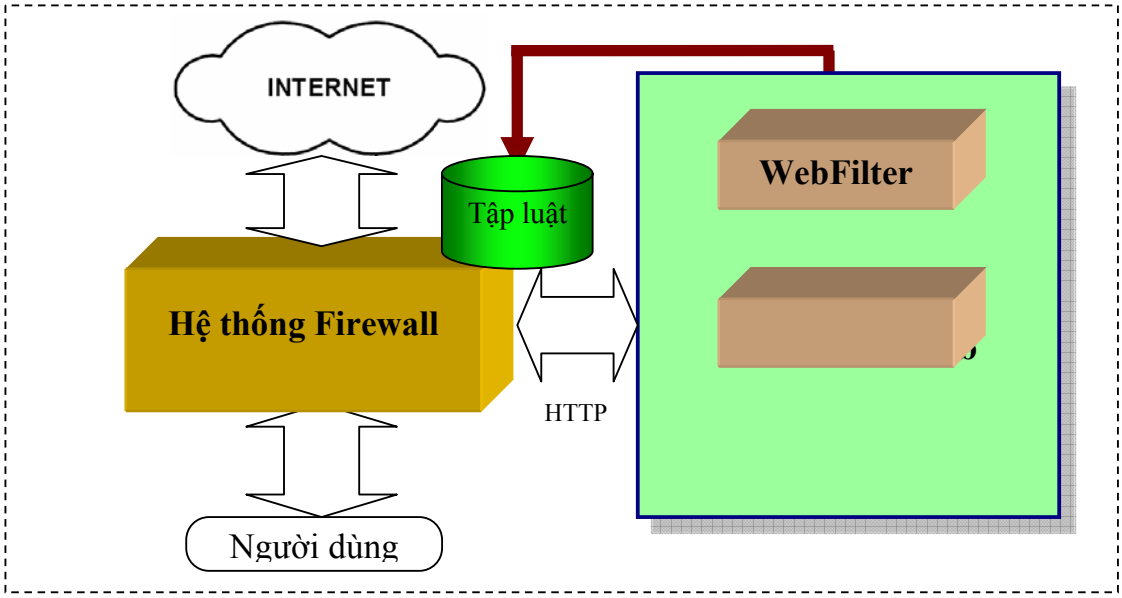
3.1.2.7. Nguyên tắc xử lý các luồng dữ liệu không lành mạnh

Nguyên tắc xử lý các luồng dữ liệu không lành mạnh được thực hiện theo một trong các tính năng kỹ thuật của bộ lọc web dưới đây:

- Cấm không cho truy vấn đến nơi yêu cầu (ví dụ như các thông tin khiêu dâm, văn hoá phẩm đồi truỵ, …)
- Thay đổi một phần nội dung dữ liệu yêu cầu (ví dụ như sửa đổi lại một phần nội dung các trang kết quả tìm kiếm từ các dịch vụ tìm kiếm như google.com, yahoo, …, loại bỏ SPAM, virut và các đoạn mã (script nguy hiểm) trong các tệp đi kèm nội dung thư điện tử ...)
- Định hướng lại luồng dữ liệu đến những đích an toàn hơn (các website quảng cáo thông qua các popup,…)
- Kiểm soát và ngăn chặn cookies, các đoạn mã code có hại viết bằng các ngôn ngữ như: java, ActiveX, javascript, Vbscript.

3.1.2.8. Kỹ thuật quản lý các luồng thông tin vào/ra

Các luồng thông tin vào/ra tại một cổng Internet sẽ được kiểm soát bởi một hệ thống firewall. Tại đó, các danh sách trắng (white list) và danh sách đen (black list) các địa chỉ sẽ được sử dụng trong việc xây dựng tập luật các chính sách (policy) liên quan đến việc quản lý các luồng thông tin vào/ra. Nếu địa chỉ yêu cầu nằm trong hai danh sách đó, hệ thống firewall này sẽ tự quyết định ngăn chặn hay không tuỳ thuộc vào địa chỉ đó có nằm trong danh sách đen hay trắng tương ứng. Nếu không, luồng dữ liệu yêu cầu sẽ được định hướng đế hệ thống lọc nội dung để hệ thống này quyết định cơ chế kiểm duyệt thích đáng. Sơ đồ quản lý các luồng thông tin được minh hoạ như *Hình 3.1*.



Hình 3. 1. Sơ đồ đề xuất kiến trúc giải pháp lọc web

Như vậy, hệ thống firewall phải đảm bảo quản lý được các luồng thông tin vào/ra theo thời gian thực và thực hiện tác vụ *phát hiện và xử lý sớm*. Các kỹ thuật *lọc cụ thể* cũng như *lọc loại trừ* sẽ được thực hiện tại bước này, nhằm giảm tải cho hệ thống lọc. Trong trường hợp không thể áp dụng hai kỹ thuật này, tất các các yêu cầu của người dùng liên quan đến các giao thức HTTP sẽ được "*bẻ ghi*" đi qua hệ thống lọc nội dung trước khi đi đến đích. Vì vậy việc tích hợp các phần cứng cần thiết (switchs, routers) cũng cần được nghiên cứu kỹ ở mức này.

3.1.3. Kiến trúc tổng quan hệ thống lọc nội dung

Hệ thống *lọc web* sẽ được thiết kế với sự sử dụng hệ thống proxy *trong suốt đối với người sử dụng* (transparent proxy) có khả năng cache cao nhằm giảm thời gian truy cập trực tiếp đến địa chỉ yêu cầu cũng như giảm khối lượng tính toán lọc. Một yêu cầu từ người dùng đến một địa chỉ nào đó sẽ được hệ thống lọc web phân tích các yếu tố sau:

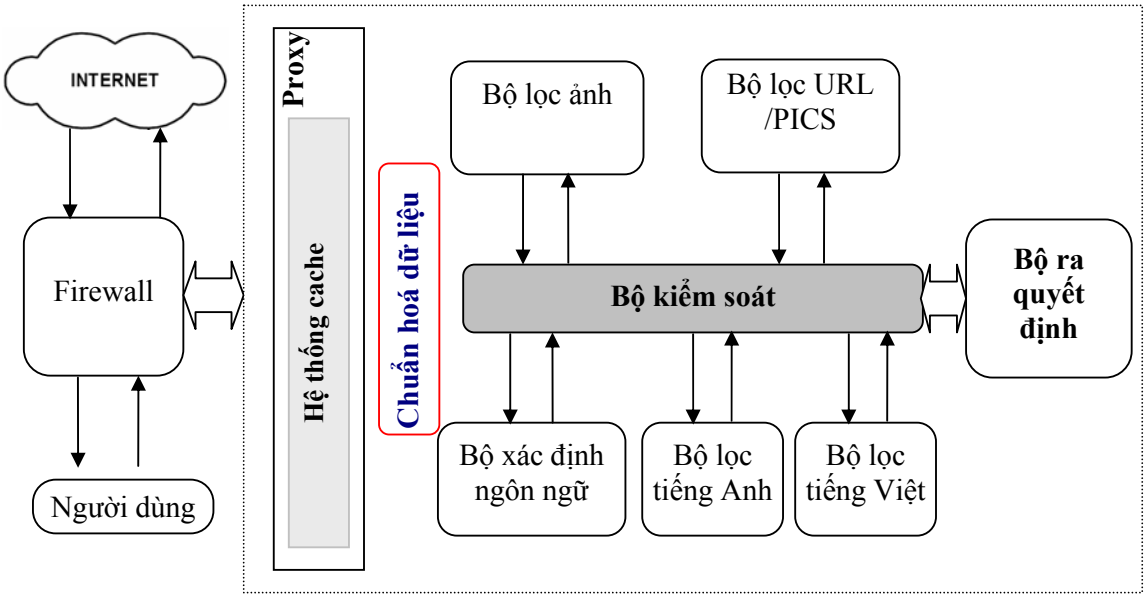
i. Độ lớn của dữ liệu yêu cầu: yếu tố này sẽ ảnh hưởng đến quyết định xử lý trực tuyến (online) hay không (offline).

ii. Kiểu dữ liệu: việc nhận dạng kiểu dữ liệu cũng phải được thực hiện triệt để để có thể phát hiện các loại dữ liệu dạng văn bản và hình ảnh (hai loại dữ liệu chính được nhắm đến trong đề tài) dưới các hình thức khác nhau (chẳng hạn như tệp nén, tệp có định dạng đặc biệt như Word, PDF, v.v.).

Các yếu tố này sẽ cho phép hệ thống quyết định lọc trực tuyến (với dữ liệu dạng văn bản hay hình ảnh và có độ lớn thích hợp cho việc xử lý thời gian thực[18]), lọc offline (văn bản/hình ảnh có dung lượng lớn) hay không lọc (tệp âm thanh, video clips, film, …).

Khi hệ thống lọc trực tuyến, những thông tin không lành mạnh sẽ áp dụng nguyên tắc xử lý các luồng thông tin không lành mạnh. Việc lọc nội dung dữ liệu (cả online lẫn offline) sẽ giúp cho việc cập nhật lại các danh sách đen và trắng, làm tăng hiệu năng phát hiện và xử lý sớm của hệ thống firewall.

Kiến trúc một hệ thống lọc Web được mô tả theo sơ đồ trong *Hình 3.2*.



Hình 3. 2. Kiến trúc tổng quát mô hình lọc web trên Internet

[18] Theo thống kê được đề cập trong báo cáo cuối cùng của dự án POESIA (Deliverable 3.1) thì độ lớn trung bình của một trang Web là 16,3 KB.và có đến 99% số trang có độ lớn bé hơn 7,4KB. Cũng theo báo cáo này thì số lượng ảnh trung bình trong một trang Web là 18,8.

Một hệ thống cache có hiệu quả cao là module Web proxy sẽ được tích hợp vào transparent proxy nhằm tăng thời gian truy cập dữ liệu khi những dữ liệu yêu cầu này đã có lưu tạm thời tại hệ thống cache này.

Khi có một yêu cầu mới (dữ liệu yêu cầu không tồn tại trong cache) được gửi đến mô đun lọc Web, mô đun con "*Chuẩn hoá dữ liệu*" sẽ thực hiện những xử lý cần thiết để chuẩn hoá dữ liệu yêu cầu về dạng văn bản và hình ảnh. Nếu bước chuẩn hoá dữ liệu này không thành công, dữ liệu yêu cầu có thể được lưu lại tạm thời để xử lý offline sau này (tạo tiền đề để mở rộng hệ thống cho việc lọc nội dung với các loại dữ liệu khác).

3.1.3.1. Kỹ thuật lọc nội dung văn bản

Việc lọc nội dung văn bản dạng text sẽ dựa trên những công nghệ hiệu quả nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên NLP. Do đặc thù mỗi ngôn ngữ, việc lọc nội dung văn bản tiếng Việt và tiếng nước ngoài (cụ thể là tiếng Anh) cần phải được tách biệt. Đối với mỗi ngôn ngữ, tuỳ thuộc vào kết quả nghiên cứu và thử nghiệm các phương pháp học không có giám sát (unsupervised learning: chẳng hạn như *K-means/Fuzzy K-mean/Dynamic clouds, Self Organising Maps, Hierarchical classification*) hay các phương pháp học có giám sát (*Naive Bayes, k-Nearest Neighbour, Maximum Entropy, Support Vector Machine*), v.v. để chọn ra giải pháp tối ưu nhất cho việc lọc nội dung văn bản.

Qua quá trình khảo sát bài toán phân lớp web, do số lượng các trang web được phân lớp là rất lớn, thời gian và khối lượng tính toán yêu cầu là rất lớn. Để hệ thống lọc có thể được thực thi trong môi trường Internet, các thuật toán phân lớp web phải đủ đơn giản về độ phức tạp tính toán nhưng đồng thời phải đảm bảo tính hiệu quả (độ chính xác/độ hồi tưởng) của kết quả phân lớp. Lược đồ áp dụng bài toán phân lớp web vào hệ thống lọc nội dung trên

Internet sẽ bao gồm hai giai đoạn: Giai đoạn lọc đơn giản và giai đoạn lọc phức tạp.

Ở giai đoạn lọc đơn giản, phương pháp tiếp cận thống kê cho bộ phân lớp, sử dụng phương pháp biểu diễn túi từ khóa, kết hợp với quá trình loại từ dừng và lấy từ gốc được sử dụng. Các từ khóa được đánh chỉ mục được lựa chọn thông qua một ngưỡng cực tiểu cho tần số tài liệu trong tập dữ liệu huấn luyện. Mô hình được xây dựng cho mỗi chủ đề sẽ bao gồm danh sách các tần số được xếp hạng của các từ khóa. Quá trình phân lớp được sử dụng thông qua độ đo sai lệch vị trí trên các danh sách xếp hạng các từ khóa.

Giai đoạn thứ lọc thứ hai (giai đoạn lọc phức tạp) sẽ tập trung vào các trang web bị phân loại sai vào các thể loại không phù hợp bởi bộ lọc đơn giản ở trên. Thay vì sử dụng các từ khóa đơn trong việc biểu diễn các tài liệu, tập hợp các từ khóa xung quanh từ khóa đấy sẽ được sử dụng. Tập hợp các từ khóa ngữ cảnh này có thể được lấy bằng một của sổ trượt, và được biểu diễn như là một danh sách hoặc một tập hợp. Ngoài ra các kỹ thuật của quá trình xử lý ngôn ngữ tự nhiên (NLP) như POS, gán nhãn tên thực thể có thể, được áp dụng để làm giàu ngữ nghĩa của từ khóa. Đối với tiếng Việt, quá trình tách từ [51] sẽ được áp dụng để tăng độ chính xác của quá trình phân lớp. Thuật toán phân lớp web được lựa chọn trong giai đoạn này là thuật toán gán nhãn lỏng [50], dựa trên thuật toán học trực tuyến [49].

3.1.3.2. Kỹ thuật lọc dựa trên URL và PICS

Việc thực hiện lọc dựa trên URL và các liên kết đơn giản hơn nhiều so với các kỹ thuật lọc nội dung văn bản và hình ảnh. Các kỹ thuật mang lại hiệu quả cao ở đây có thể kể đến như lọc dựa vào từ khoá, theo tên miền, v.v.

Một kỹ thuật nữa cần được nói đến là kỹ thuật lọc dựa theo những thông tin đã được gán nhãn đánh giá, phân hạng tại các nhà cung cấp nội dung thông qua chuẩn PICS (Platform for Internet Content Selection – 1999,

http://www.w3c.org/PICS). Rõ ràng việc đánh giá của các nhà cung cấp thông tin đôi khi không hoàn toàn chính xác, vì thế việc nghiên cứu và đánh giá lại các nhãn tại bộ lọc PICS là cần thiết và phải được nghiên cứu cụ thể hơn nữa.

Việc lọc tài liệu theo chuẩn PICS sẽ chú trọng đến những vấn đề sau :

- Sử dụng thư viện Dsig của W3C [DSig] – cung cấp một chuẩn định dạng thực hiện mã hóa và giải mã.
- Tạo bộ phân tích cấu trúc PICS
- Lựa chọn rating-system, rating-service để đánh giá chính xác nội dung các tài liệu được cung cấp.

3.1.4. Kỹ thuật lọc ảnh

Có rất nhiều kỹ thuật khác nhau được đề xuất cho việc phân loại cũng như nhận dạng loại ảnh, chẳng hạn:

- Lọc dựa vào màu sắc: đối tượng cần xác định trong trường hợp này thường là hình ảnh khoả thân, có màu và kết cấu chung. Khi đó cần phải phải xác vùng điểm ảnh tương ứng với da người cũng như phân bố màu sắc và kết cấu.
- Lọc dựa vào màu sắc, text và sử dụng trí tuệ nhân tạo: Kỹ thuật này yêu cầu cần phải có bước học tự động các đặc trưng từ các mẫu. Khi có một ảnh mới, các đặc trưng của nó sẽ được chắt lọc và so sánh chúng với các mẫu để quyết định lọc hay không. Đồng thời, việc phân tích text kèm theo trang web và kết hợp cả hai phân tích lại cũng sẽ góp phần làm tăng hiệu quả của quá trình lọc ảnh.
- Lọc dựa vào màu sắc và hình dạng: kỹ thuật này kết hợp các đặc điểm của cả hình dạng và màu sắc để có thể phát hiện ảnh có nội dung không lành mạnh.

Hệ thống lọc web sẽ chú trọng đến các kỹ thuật phát hiện biên (skin detection) và phân tích hình dạng (form analysis) kết hợp với việc so sánh các

hình tiêu biểu đối với những ảnh có kích thước bé (symbol detection) và nội dung text kèm theo, nếu có, nhằm đạt được hiệu quả cao nhất trong việc phát hiện và lọc ảnh có nội dung không lành mạnh.

Kỹ thuật lọc hình ảnh có nội dung xấu sẽ dựa trên một số vấn đề sau:

- Phát hiện da, sau đó trích trọn thuộc tính và sử dụng bộ phân loại ảnh theo mô hình thống kê
- Thu thập các mẫu có nội dung xấu (phản động, kích động bạo lực, khủng bố), sử dụng các mô hình thống kê để từ đó có thể phân loại những hình ảnh có nội dung phản động, kích động bạo lực, khúng bố, …

3.1.5. Kỹ thuật quyết định

Mô đun quyết định nhận từ đầu vào nội dung đã được xếp hạng và phải đưa ra quyết định cho phép trả về trọn vẹn cho người yêu cầu hay áp dụng nguyên tắc xử lý các luồng dữ liệu không lành mạnh.

Việc ra quyết định của mô đun này không cần dựa vào toàn bộ kết quả của các bộ lọc. Chẳng hạn như nó có thể quyết định chặn luồng dữ liệu nếu bộ lọc ảnh đưa lại một đánh giá cao về khả năng chứa thông tin xấu trong khi chưa nhận được kết quả của bộ lọc text.

Nhằm đảm bảo ràng buộc xử lý thời gian thực, mô đun này có thể đưa ra quyết định sau một khoảng thời gian nào đó mà không cần chờ kết quả phân tích của các bộ lọc.

Nội dung tài liệu nếu được phân tích trong các bộ lọc sẽ được đánh giá theo thang điểm từ 0 đến 100. Sau đó, tuỳ thuộc vào những chính sách lọc đề ra, các ngưỡng cấm/chặn hay để xem xét sau sẽ được đưa ra. Tài liệu có nội dung được đánh giá lớn hơn ngưỡng cấm sẽ được hệ thống ngăn chặn và thông báo lại cho người dùng, nhỏ hơn ngưỡng xem xét thì sẽ được trả lại ngay cho người yêu cầu. Trong trường hợp nằm trong ngưỡng cấm và lớn hơn

ngưỡng xem xét thi tài liệu đó vẫn được trả lại người dùng, tuy nhiên, nội dung đó sẽ được các nhà chức trách liên quan kiểm duyệt lại và cập nhật trong danh sách đen sau đó.

3.1.6. Các thành phần trong hệ thống lọc web

Dữ liệu đầu vào xuất phát từ hệ thống proxy sẽ được xử lý ở mô đun chuẩn hoá dữ liệu. Mô đun này làm nhiệm vụ chuyển tất cả các tài liệu Web về hai dạng cơ bản : plaintext và hình ảnh. Sau khi chuẩn hoá dữ liệu về hai dạng cơ bản xong, dữ liệu này sẽ được xem như dữ liệu đầu vào của mô đun kiểm soát. Ngoài dữ liệu này ra, địa chỉ URL của tài liệu đó cũng được chuyển đến bộ kiểm soát.

Mô đun kiểm soát làm nhiệm vụ điều phối quá trình phân tích nội dung tài liệu. Trong trường hợp tài liệu này là văn bản plaintext, nó sẽ được chuyển đến mô đun xác định ngôn ngữ, nếu là hình ảnh nó sẽ chuyển trực tiếp đến mô đun lọc hình ảnh.

Sau khi xác định được ngôn ngữ của tài liệu, kết quả này sẽ được mô đun xác định ngôn ngữ chuyển lại cho bộ kiểm soát. Mô đun này sẽ từ đó chuyển tài liệu đó đến bộ lọc tương ứng với ngôn ngữ vừa xác định được (Việt/Anh).

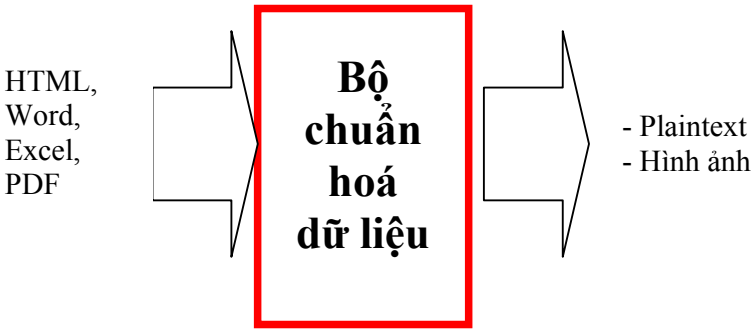
Kết quả của các mô đun lọc văn bản plaintext tiếng Việt, tiếng Anh, lọc hình ảnh và URL+PICS sẽ được mô đun kiểm soát thu nhận và chuyển đến bộ ra quyết định.

Bộ ra quyết định sẽ dựa vào những chính sách lọc đã được quy định, tiến hành ra quyết định cấm, không cấm hay xét duyệt lại nội dung tài liệu. Từ đó mô đun kiểm soát có thể tương tác với hệ thống firewall/proxy để thực hiện những quyết định đưa ra.

Đặc tả chi tiết từng thành phần trong hệ thống lọc tài liệu này sẽ được trình bày cụ thể ở phần dưới đây:

Bộ chuẩn hoá dữ liệu

Mô đun này có nhiệm vụ chuyển những định dạng khác nhau của tài liệu người dùng yêu cầu thành hai dạng cơ bản: hình ảnh và những chuỗi ký tự (plaintext).



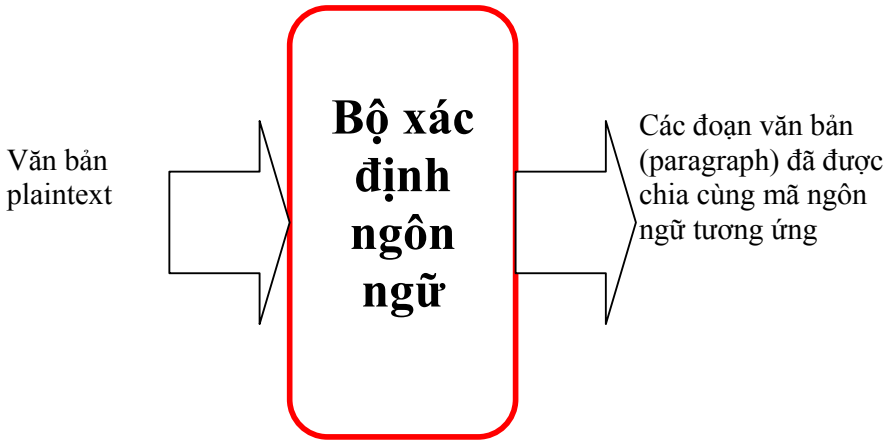
Hình 3. 3. Mô hình bộ chuẩn hoá dữ liệu

Input: các tệp tài liệu có định dạng HTML, Word, Excel, PDF, JPEG, PNG, GIF

Output: Các đoạn văn bản ở dạng plaintext và những tệp hình ảnh

Algorithm: Giải được

Bộ xác định ngôn ngữ



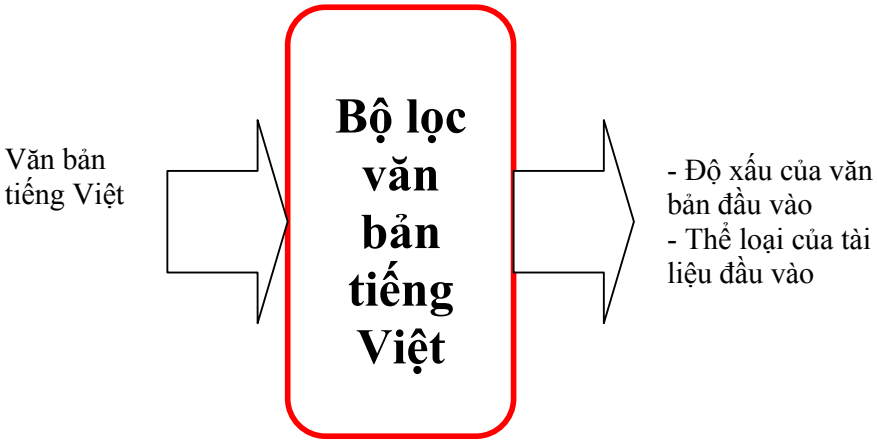
Hình 3. 4. Mô hình bộ xác định ngôn ngữ

Input: tài liệu dạng plaintext

Output: mã ngôn ngữ của từng đoạn văn bản (paragraph) trong tài liệu đầu vào. Mã này bắt buộc phải có khả năng xác định 2 ngôn ngữ chính là tiếng Việt và tiếng Anh

Bộ lọc văn bản tiếng Việt

Bộ lọc này có nhiệm vụ phân tích, phân loại, đánh giá mức độ xấu của một tài liệu (đoạn – paragraph) được xây dựng từ ngôn ngữ tiếng Việt.



Hình 3. 5. Mô hình bộ lọc văn bản tiếng Việt

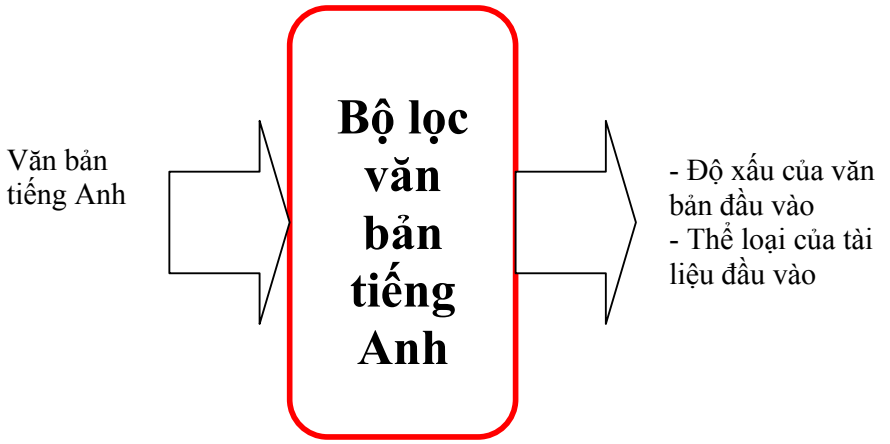
Input: tài liệu văn bản tiếng Việt

Output:

- Độ xấu của văn bản đầu vào, đánh giá trên thang điểm 0-100 tương ứng từ tốt đến xấu
- Thể loại của văn bản đầu vào

Bộ lọc văn bản tiếng Anh

Bộ lọc này có nhiệm vụ phân tích, phân loại, đánh giá mức độ xấu của một tài liệu (đoạn – paragraph) được xây dựng từ ngôn ngữ tiếng Anh.



Hình 3. 6. Mô hình bộ lọc văn bản tiếng Anh

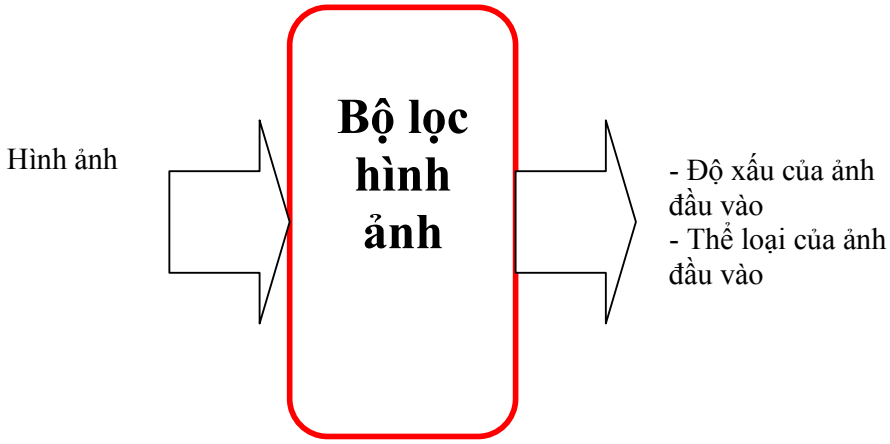
Input: tài liệu văn bản tiếng Anh

Output:

- Độ xấu của văn bản đầu vào, đánh giá trên thang điểm 0-100 tương ứng từ tốt đến xấu
- Thể loại của văn bản đầu vào

Bộ lọc ảnh có nội dung xấu

Bộ lọc này có nhiệm vụ phân tích, phân loại, đánh giá mức độ xấu của một ảnh có định dạng bitmap-BMP, JPEG, GIF, hay PNG.



Hình 3. 7. Mô hình bộ lọc hình ảnh

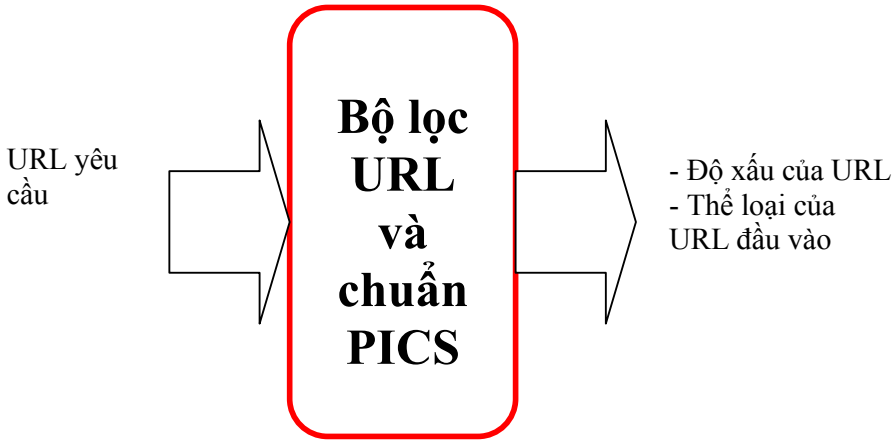
Input: Ảnh có dạng BMP, JPEG, GIF, PNG

Output:

- Độ xấu của ảnh đầu vào, đánh giá trên thang điểm 0-100 tương ứng từ tốt đến xấu
- Thể loại của ảnh đầu vào

Bộ lọc URL và PICS

Bộ lọc này có nhiệm vụ phân tích cú pháp và cả ngữ nghĩa một địa chỉ URL, từ đó đánh giá mức độ xấu của URL đó. Trong trường hợp URL đó đã được đánh giá, phân loại theo chuẩn PICS, bộ lọc sẽ ghi nhận những đánh giá phân loại đó, tiến hành phân tích và ra cập nhật lại độ xấu của địa chỉ URL yêu cầu.



Hình 3. 8. Mô hình bộ lọc địa chỉ URL và chuẩn PICS

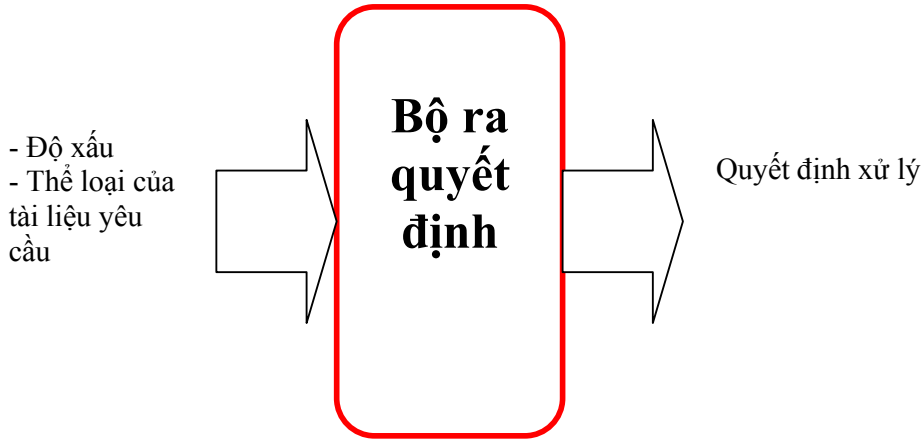
Input: địa chỉ URL của tài liệu yêu cầu

Output:

- Độ xấu của URL đầu vào, đánh giá trên thang điểm 0-100 tương ứng từ tốt đến xấu
- Thể loại của URL yêu cầu

Bộ ra quyết định

Bộ ra quyết định giữ vai trò phân tích các kết quả đến từ các bộ phận khác trong tổng thể hệ thống để từ đó tổng hợp và ra quyết định xử lý với tài liệu mà người dùng yêu cầu.



Hình 3. 9. Mô hình bộ ra quyết định

Quyết định xử lý đối với một yêu cầu sẽ chủ yếu dựa trên độ xấu Đx của yêu cầu đó và bao gồm các trường hợp sau:

i. Cấm không cho người sử dụng truy cập đến địa chỉ yêu cầu và thông báo lý do cấm cho người đó. Đây là trường hợp độ xấu của tài liệu yêu cầu lớn hơn ngưỡng xấu Nx cho phép của một tài liệu Đx > Nx ,
ii. Cho phép người sử dụng truy cập đến địa chỉ yêu cầu. Đây là trường hợp tài liệu yêu cầu có độ xấu nhỏ hơn ngưỡng cần kiểm soát Nk : Đx < Nk.
iii. Cho phép truy cập đến tài liệu yêu cầu, tuy nhiên vết của tài liệu này sẽ được lưu lại trong phần lưu vết của hệ thống. Trong trường hợp này, độ xấu của tài liệu sẽ nằm trong hai ngưỡng cấm và kiểm soát : Nk <= Đx <= Nx. Người phụ trách đảm bảo an toàn/an ninh thông tin

sẽ có nhiệm vụ phân tích đánh giá lại những tài liệu này và cập nhật lại vào trong hai danh sách trắng và đen.

iv. Cho phép truy cập đến tài liệu yêu cầu trong trường hợp độ lớn của tài liệu yêu cầu vượt quá ngưỡng **Nl** cho phép của hệ thống. Trong trường hợp này, tài liệu này cũng sẽ được lưu lại trong vết của hệ thống và sẽ được tiến hành phân tích, đánh giá theo hình thức ngoại tuyến để cập nhật lại hai danh sách trắng và đen.

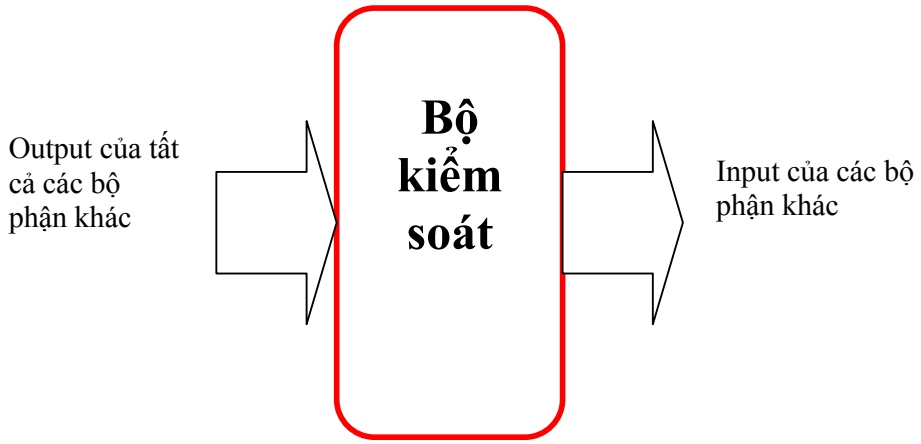
Như vậy, bộ ra quyết định sẽ có đầu vào và đầu ra cụ thể như sau:

Input: Các kết quả (độ xấu + thể loại tài liệu yêu cầu) đến từ các bộ phận khác

Output: Quyết định xử lý (một trong 3 loại quyết định đã nêu ở trên)

Bộ kiểm soát

Bộ kiểm soát có nhiệm vụ điều phối các luồng dữ liệu bên trong hệ thống lọc nội dung.



Hình 3. 10. Mô hình bộ kiểm soát

Từ các dữ liệu đầu ra của mỗi bộ phận trong hệ thống, bộ kiểm soát có nhiệm vụ truyền những kết quả đó đến bộ phận có trách nhiệm xử lý. Các luồng dữ liệu mà bộ kiểm soát phải đảm nhiệm bao gồm:

- Chuyển kết quả của bộ chuẩn hoá dữ liệu đến các bộ lọc ảnh và bộ xác định ngôn ngữ tương ứng với những kết quả text hay ảnh ở bộ chuẩn hoá.
- Chuyển kết quả của bộ xác định ngôn ngữ đến bộ lọc tiếng Việt hoặc/và bộ lọc tiếng Anh tương ứng với kết quả xác định là tiếng Việt hay tiếng Anh của tài liệu yêu cầu
- Chuyển địa chỉ URL yêu cầu từ hệ thống proxy đến bộ lọc URL và PICS
- Chuyển các kết quả từ các bộ lọc URL&PICS, lọc ảnh, lọc tiếng Việt, lọc tiếng Anh đến bộ ra quyết định
- Chuyển kết quả quyết định xử lý đến hệ thống firewall và proxy.

Phần mềm WEBFILTER (SP.01-1)

Phương pháp cập nhật danh sách lọc URL

Việc tạo danh sách lọc URL ban đầu dựa vào ý kiến chuyên gia. Và thường xuyên được cập nhật lại dựa vào chuyên gia hoặc tự động.

Để tự động cập nhật danh sách các URL một số phương pháp được đưa ra như: các phương pháp dựa vào liên kết để xác định link Spam, hay phân lớp để xác định các URL chứa nội dung xấu/ tốt… từ đó cập nhật các URL tốt xấu vào danh sách lọc.

Các phương pháp link Spam:

Sử dụng các thuật toán tính hạng dựa trên liên kết như PageRank[] để đánh giá độ tốt xấu của các URL, với quan niệm: các trang có nội dung xấu thường có xu hướng liên kết với nhau và ngược lại các trang có nội dung tốt thường liên kết với nhau và hiếm có liên kết tới các trang xấu. Từ tập nhân các URL được gán trọng số tốt xấu ban đầu, thông qua liên kết trọng số được phân phối cho các URL được loang tới. Những URL có trọng số càng cao càng được đánh giá tốt.

Vấn đề đặt ra là làm sao xác định được ngưỡng (threshold) để quyết định một URL thực sự tốt/ xấu: ***blankThreshold***- xác định những URL có trọng số lớn hơn blankThreshold được coi là URL xấu và ***whileThreshold***- xác định những URL có trọng số lớn hơn whileThreshold được coi là URL tốt. Sau đó các URL mới này được cập nhật vào danh sách lọc tương ứng(tốt hoặc xấu). Các URL có trọng số nằm ở trong khoảng black/ whileThreshold có thể không cần cập nhật vào danh sách lọc bởi được coi là chưa xác định.

Phương pháp cần sử dụng cấu trúc của đồ thị web- các liên kết vào ra từ mỗi URL và do tính chất hội tụ, tính gia tăng(sử dụng các trọng số URL ở lần đánh giá trước để cập nhật cho lần đánh giá sau) của các phương pháp tính hạng trang nên để tăng nhận diện nhanh cần lưu lại cấu trúc web và trọng số của các URL đã được duyệt qua. Hơn nữa việc đánh giá này có thể làm độc lập với hệ thống, được thực hiện ngoại tuyến (offline) và chỉ cập nhật lại danh sách lọc sau khi những khoảng thời gian nhất định

Phương pháp kết hợp với bộ lọc nội dung: dựa vào nội dung của URL và nội dung của trang tài liệu ứng vơi URL để phân loại tài liệu tốt/ xấu. Tận dụng bộ phân lớp nội dung tài liệu để xác định các URL chứa nội dung tốt/ xấu và cập nhật vào danh sách lọc nếu cần.

Ứng dụng vào hệ thống lọc nội dung

- Xây dựng các mẫu lọc URL
- Cập nhật danh sách lọc URL ban đầu
- Module cập nhật danh sách lọc: áp dụng thuật toán PageRank của trường Stanford University và TrustRank của Yahoo! Research.

YÊU CẦU ĐẶT RA

Đối với modun lọc dựa theo URL

- Ngăn chặn các trang web có vị trí liên qua tới một URL- httt://www.xyz.com và tất cả các trang web có cùng một tên domain.

- Ngăn chặn một tập hợp của trang web- http://www.xyz.com.au/magazineA/…
- Ngăn chặn một trang web với link dạng sau đặc biệt - http://www.xyz.com.au/fileA.html.
- Ngăn chặn các truy xuất file từ các ftp site có liên quan bởi các đường địa chỉ đặc biệt của trang web, hoặc liên quan tới file(hoặc các mẫu) trong phạm vi một trang web. Ví dụ: ftp://ftp.xyz.com/fileB.jpg
- Ngăn chặn các cách truy xuất dữ liệu đặc biệt như việc dùng địa chỉ các cổng trực tiếp để truy xuất nội dung: www.xyz.com:6969/
- Ngăn chặn một nhóm các tin tức liên quan, ví dụ như các mục quảng cáo
- Ngăn chặn hoặc hủy bỏ một mục liên quan trong một tin tức.
- Cho phép những Url thuộc danh sách trắng đi qua mà không bị lọc.
- Khóa hoặc giới hạn một số trang Upload file chia sẻ.<DansGuardian>

LÝ THUYẾT LIÊN QUAN

Nghiên cứu các giải pháp lọc theo URL

Lọc theo tên URL dạng Text

Thành phần phổ biến nhất, và hiệu quả nhất, dạng nền tảng của nguồn lọc là cơ bản URLs, người đọc địa chỉ của trang web. URLs đưa ra nhiều cách điều khiển làm mịn hơn lọc theo gói như là chỉ lọc theo tên trang web. Một trang web ví dụ như www.playboy.com sẽ chứa đựng hàng trăm hoặc nhiều hơn tới hàng nghìn trang riêng lẻ, mỗi cái với địa chỉ URL của họ. Cấu trúc có thứ bậc (trái sang phải) của trang web cho phép lọc tự động theo từng tập dữ liệu của toàn bộ sites hoặc chỉ là một phần của sites bởi phần đặc biệt của URLs. Tất cả các trang được đăt trên máy chủ của trang Playboy sẽ bình thường có địa chỉ bắt đầu với www.playboy.com và ở đây tất cả những thứ ta

cần là lọc đặc biệt danh sách các khối dữ liệu của toàn bộ site, bao gồm cả www.playboy.com/June2001/centrefold.html và http://www.playboy.com/May2001/centrefold.html .

Có một phần tên có thể có một cơ chế đặc biệt cho phép truy nhập dữ liệu tới một phần của site trong khi dang truy xuất dữ liệu theo khối tới phần khác. Ví dụ như một cách tự động có thể được truy cập theo khối tới www.bigpond.com.au/users/~ahacker không có khối dữ liệu nào tới người sử dụng khác với trang chủ của web được được đặt trên máy chủ của Big Pond Web.

Sử dụng nhiều "domain name"

Phần URL đang được trả lời làm cho phù hợp với 'domain names', như là www.playboy.com , với URLs. Nhiều domain names là rất động và một site có thể có nhiều domain names, và đưa ra mới một chỉ định rất nhanh. Một site, như là Playboy có thể sử dụng www.playboy.com để quy về nội dung chính trên Web server nhưng sử dụng cộng thêm nhiều servers với cùng một domain names của họ, như là www2.playboy.com và www3.playboy.com, để cung cấp nhiều hơn, nhanh hơn nội dung thực tế đươc yêu cầu. Một vài trang web lớn như là HotMail, DejaNews và Ebay có một nghìn hoặc nhiều server hơn trên một tên duy nhất của domain name. Ở đây đã sử dụng nhiều server để chạy với sự tăng thêm về dung lượng, không phải để tránh bộ lọc, nhưng tất cả tiềm năng tên hoặc địa chỉ IP có thể chờ để hiện ra trong khối của danh sách lọc trở thành hiệu quả.

Lọc dựa trên URL dạng số 32 bit

Domain names, như là www.playboy.com chỉ tồn tại vì sự tiện lợi cho con người khi sử dụng nó để dàng xem trọn vẹn trang đó. Ngoài ra hệ thống máy tính kết nối tới Internet là thật sự được biết bởi địa chỉ IP của nó. Một số 32-bit là hoàn toàn có thể được chấp nhận là URL trong domain names. Ví dụ, địa chỉ URLs htttp://www.playboy.com/centrefolds/ và

http://206.15.29.10/centrefolds/ là hoàn toàn tương đương và có thể thay thế được cho nhau. Kết quả của nó là bộ lọc có thể khóa có cả 2 dạng của một URL, bởi vì phải cộng thêm tới cỡ của danh sách đen hoặc ngày càng tăng dần sự chậm trễ của bộ lọc nếu domain names có thể được chuyển sang tới địa chỉ IP trước khi có thể được kiểm tra.

Internet làm việc với các địa chỉ IP. Domain names, như là www.playboy.com chỉ tồn tại đối với con người. Con người không thể nhớ những số khó sử dụng, cồng kềnh, và sự thật để thân thiện hơn với con người 'domain names' được sử đụng để truy cập nội dung. Người sử dụng không cần sử dụng dạng số của một URL và để có thể quyết định tốt nhất một bộ lọc phải có thêm cả ở dạng thân thiện của domain name. Ví dụ: URLs http://www.playboy.com/centrefolds/ và http://206.251.29.10/centrefolds/ là được hiểu như nhau và bộ lọc phải làm được cả 2 dạng này.

MÔ HÌNH ĐỀ XUẤT

Để làm được những yêu cầu trên ta cần phải chia nhỏ modun ra thành các thành phần khác nhỏ hơn đảm nhiệm từng nhiệm vụ cụ thể như sau:

- Xây dựng modun lọc URL theo dạng URL text.
- Xây dựng modun lọc URL theo dạng số 32 bit.

} Dựa trên danh sách đen

- Xây dựng modun lọc URL theo việc kiểm soát việc truy xuất các cổng trực tiếp qua trình duyệt web của người sử dụng.
- Xây dựng modun kiểm soát trực tiếp việc truy xuất trực tiếp tới các file cụ thể trên từng trang web.
- Xây dựng danh sách các danh sách đen ở cả dạng URL text và cả dạng URL dạng số 32 bit.
- Modun lọc những đường link có dạng phức tạp, thường xuất hiện ở các trang tin tức.

➔ ngăn chặn luôn domain name của những địa chỉ đó, hay chỉ ngăn chặn một phần nội dung của domain name đó.

THIẾT KẾ MÔ ĐUN

Các giải pháp lọc đối với URL

/*- Đặc tả những ý chính trong kiến trúc mô đun yêu cầu.

- Các giải thuật cho phép giải quyết những vấn đề đặt ra trong mô đun đích cần phải được trình bày trong phần này.

*/

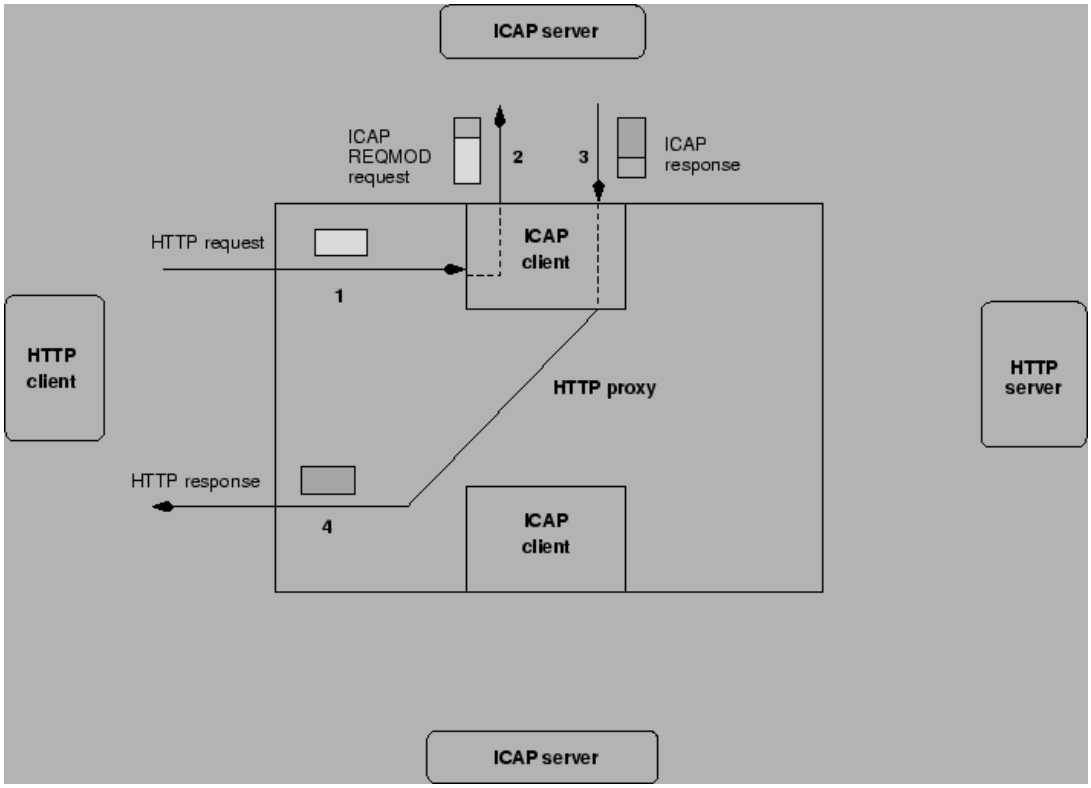
ICAP Communication(Internet Content Adaptation Protocol)

Giao thức ICAP là được lý giải đầy đủ trong RFC 3507. Khi máy ủy nhiệm nhận một HTTP yêu cầu hoặc một HTTP trả lời, nó là được gói gọn trong ICAP yêu cầu và gửi nó tới bộ kiểm tra(the monitor), cái có vai trò của ICAP server. Khi bộ kiểm tra trả lời bởi một thông điệp CAP và gói gọn một sửa đổi HTTP yêu cầu hoặc sửa đổi..

Mục đích của giao thức ICAP là nội dung của tài liệu phỏng theo, nhưng chúng ta chỉ sử dụng cho mục đích lọc trong phần mềm POESIA. Bộ kiểm tra sẽ không "sửa đổi" ICAP yêu cầu, nó chỉ chấp nhận hoặc không chấp nhận. Ở đây có 2 ví dụ:

Ví dụ:

Ví dụ 1: Khối có nội dung xấu cùng được yêu cầu trong cùng một thời gian.



Khi client yêu cầu một trang wed khiêu dâm, ví dụ như www.sex.com . Máy trình duyệt client sẽ gửi yêu cầu tới proxy(bước 1)

```
GET /index.html HTTP/1.1
Host: www.sex.com
User-Agent: Mozilla
Connection: keep-alive
```

Máy proxy sẽ gói gọn HTTP yêu cầu trong một ICAP yêu cầu và gửi nó tới bộ kiểm tra(bước 2)

```
REQMOD icap://monitor.school.com/filter ICAP/1.0
Host: monitor.school.com
Encapsulated: req-hdr=0, null-body=68

GET /index.html HTTP/1.1
Host: www.sex.com
User-Agent: Mozilla
```

Khi đó, ICAP yêu cầu sẽ bắt đầu với REQMOD, là thành phần của ICAP được sử dụng. REQMOD nghĩa là yêu cầu sửa đổi(request

modification) và RESPMOD nghĩa là phản hồi sửa đổi(response modification). Đây là một điểm đặc biệt của tiêu đề gói gọn ICAP(ICAP header "Encapsulated"), đây là cái mà trình bày về thông điệp HTTP một cách tự nhiên(req for requests), nếu ở đây là phần chính(không có phần chính khi phần chính không tồn tại?) và độ rộng của tiêu đề HTTP(68). Chúng ta sẽ qua bộ kiểm tra để tìm kiếm trong cơ sở dữ liệu của các trang wed và tìm url(www.sex.com) trên nó. Như vậy ICAP trả lời sẽ tìm thấy nó (step 3).

```
ICAP/1.0 200 OK
Date: Mon, 10 Jan 2000 09:55:21 GMT
Server: POESIA-Monitor/1.0
Connection: close
ISTag: "W3E4R7U9-L2E4-2"
Encapsulated: res-hdr=0, res-body=(...)

HTTP/1.1 403 Forbidden
Date: Wed, 08 Nov 2000 16:02:10 GMT
Server: Apache/1.3.12 (Unix)
Last-Modified: Thu, 02 Nov 2000 13:51:37 GMT
ETag: "63600-1989-3a017169"
Content-Length: 58
Content-Type: text/html

3a
Sorry, you are not allowed to access that naughty content.
0
```

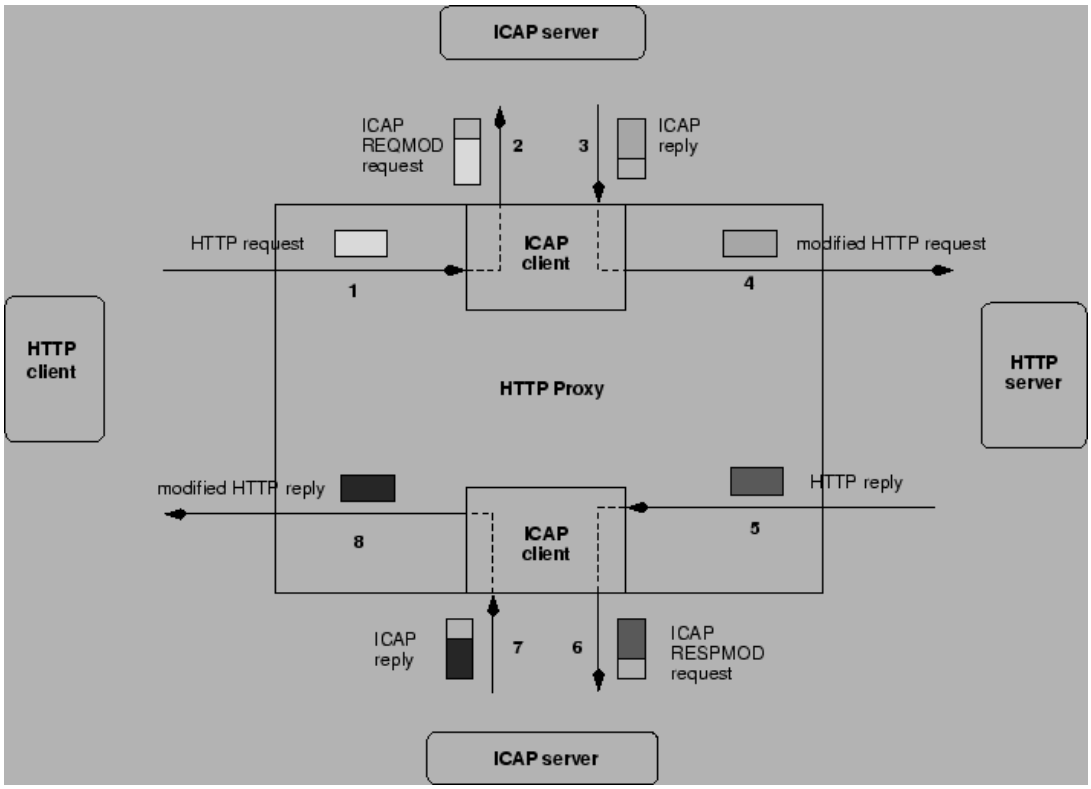
Giống như giao thức ICAP, khi proxy nhận một yêu cầu HTTP cho một HTTP yêu cầu đang được một ICAP đang thực hiện, nõ sẽ bỏ qua yêu cầu của trình duyệt web client.

```
HTTP/1.1 403 Forbidden
Date: Wed, 08 Nov 2000 16:02:10 GMT
Server: Apache/1.3.12 (Unix)
Last-Modified: Thu, 02 Nov 2000 13:51:37 GMT
ETag: "63600-1989-3a017169"
Content-Length: 58
Content-Type: text/html
Connection: close

Sorry, you are not allowed to access that naughty content.
```

Chú ý rằng ví dụ này thực hiện một quá trình vận chuyển mã khó khăn trong ICAP communication.

Ví dụ 2: Khối các nội dung xấu cùng thời gian được trả lời



Khi một web client yêu cầu nội dung của một trang web cái mà chúng ta chưa nhận dạng như là chưa được nói đến hoặc không từ một URL, chúng ta phải download nộ dung của nó về trước khi quét nó. Trong ví dụ này, yêu cầu URL là www.unknow-server.com .Một client gửi yêu cầu tới proxy giống nó!(bước 1)

```
GET /index.html HTTP/1.1
Host: www.unknow-server.com
User-Agent: Mozilla
```

Proxy bước tiếp theo sẽ tới monitor giống như ví dụ trước. Khi này thì monitor sẽ cản yêu cầu đó và sẽ trả lời lại giống như sau đây(bước 3)

```
ICAP/1.0 200 OK
Date: Mon, 10 Jan 2000  09:55:21 GMT
Server: POESIA-Monitor/1.0
Connection: close
```

```
ISTag: "W3E4R7U9-L2E4-2"
Encapsulated: res-hdr=0, null-body=(...)

GET /index.html HTTP/1.1
Host: www.unknow-server.com
User-Agent: Mozilla
```

Bộ kiểm soát sẽ không thay đổi một yêu cầu nào. Trong giao thức ICAP, ICAP server có thể nói rằng nó sẽ không thay đổi yêu cầu(hoặc phản hồi), nếu proxy hỏi rằng có tiếp tục làm tiếp không, với điều kiện là 204 "không có nội dung" phản hồi. Trong tình huống này proxy phải nhớ lại thông điệp HTTP trong bộ đệm trước khi đề nghị ICAP server cho một đáp ứng. Chúng ta tin rằng bây giờ proxy có thể giao tiếp với www.unknow-server.com và đưa một vài nội dung xấu. ICAP có thể sẽ giải quyết nó giống như sau đây:

Từ proxy tới bộ kiểm soát (bước 6)

```
RESPMOD icap://icap.example.org/satisf ICAP/1.0
Host: icap.example.org
Encapsulated: req-hdr=0, res-hdr=(...), res-body=(...)

GET /origin-resource HTTP/1.1
Host: www.origin-server.com
Accept: text/html, text/plain, image/gif
Accept-Encoding: gzip, compress

HTTP/1.1 200 OK
Date: Mon, 10 Jan 2000 09:52:22 GMT
Server: Apache/1.3.6 (Unix)
ETag: "63840-1ab7-378d415b"
Content-Type: text/html

1d
Here is some naughty content.
0
```

Từ bộ kiểm soát tới proxy(bước 7)

```
ICAP/1.0 200 OK
Date: Mon, 10 Jan 2000  09:55:21 GMT
Server: POESIA-Monitor/1.0
Connection: close
ISTag: "W3E4R7U9-L2E4-2"
Encapsulated: res-hdr=0, res-body=(...)

HTTP/1.1 403 Forbidden
Date: Wed, 08 Nov 2000 16:02:10 GMT
Server: Apache/1.3.12 (Unix)
Last-Modified: Thu, 02 Nov 2000 13:51:37 GMT
ETag: "63600-1989-3a017169"
```

```
Content-Length: 58
Content-Type: text/html

3a
Sorry, you are not allowed to access that naughty content.
0
```

Như vậy cuối cùng phản hồi tới máy chủ(server) sẽ là(bước 8)

```
HTTP/1.1 403 Forbidden
Date: Wed, 08 Nov 2000 16:02:10 GMT
Server: Apache/1.3.12 (Unix)
Last-Modified: Thu, 02 Nov 2000 13:51:37 GMT
ETag: "63600-1989-3a017169"
Content-Length: 58
Content-Type: text/html
Connection: close

Sorry, you are not allowed to access that naughty content.
```

Nếu bộ kiểm soát chấp nhận nội dung này, nó sẽ trả lời với mã 200 hoặc mã 204. Vậy nguồn gốc của nội dung sẽ được đưa tới máy khách(client).

Lọc theo chuẩn PICS

Chuẩn PICS (Platform for Internet Content Selection) được W3C(World Wide Web Consortium) đưa ra vào năm 1995, chuẩn này cung cấp tập từ điển các nhãn nhằm xác định nội dung các trang tài liệu trên Internet. Ban đầu chuẩn được đưa ra giúp người dùng dễ dàng tìm các trang web có nội dung phù hợp và tránh những trang có nội dung không mong muốn, nhằm kiểm soát việc truy cập Internet của trẻ nhỏ. Sau đó, dần trở thành chuẩn định dạng để đánh giá nội dung các tài liệu.

Chuẩn PICS là cung cấp một khung (framework) cho việc phát triển các lược đồ gán nhãn và cho phép người dùng lựa chọn linh động các tiêu chuẩn lọc theo mong muốn.

Đánh giá và gán nhãn

Chuẩn PICS đưa ra 2 loại nhãn: *self-label* và *third-label.*_Mỗi tài liệu Web được gán nhãn phù hợp với nội dung bởi người tạo tài liệu gọi là self-label. Tuy nhiên có thể tác giả không gán nhãn hoặc không gán nhãn đúng cho trang tài liệu. Do vậy giải pháp gán nhãn của bên thứ ba, third-party label,

được đưa ra để đánh giá và gán nhãn cho các tài liệu web. Và khi đó người dùng có thể lựa chọn sử dụng nhãn do tác giả hoặc của bộ gán nhãn bên thứ 3 đưa ra. Một nhãn bao gồm 3 phần: *service indentifier, label options,* và *rating. Service identifier:* khai báo URL của server được dùng để đánh giá. *Label option*: đưa ra các thuộc tính của tài liệu được đánh giá. *Rating:* Trọng số được đánh giá.Tương ứng có các cách đánh giá(thực hiện gán nhãn).

Self-rating

Người cung cấp(tạo tài liệu) tự gán nhãn cho tài liệu mà họ tạo ra…

Third-party rating

Dịch vụ thực hiện gán nhãn độc lập với phía tạo và phân phối tài liệu, và có hệ thống gán nhãn riêng. Do vậy cùng một nội dung, các hệ thống khác nhau có thể đưa ra các nhãn khác nhau.

Ease-of-use

Người dùng được phép tự đánh giá, gán nhãn các tài liệu.

Do sự đa dạng của các tiêu chuẩn đạo đức, văn hóa của người dùng nên chuẩn PICS được triển khai trên nhiều hệ thống gán nhãn để có thể đáp ứng được các nhu cầu sử dụng khác nhau trên NET. Không một hệ thống gán nhãn đơn nào phù hợp với tất cả mọi người trên cộng đồng web.

Vấn đề gán nhãn cần đảm bảo tính toàn vẹn nội dung của các tài liệu web và trong suốt đối với các browser. Các hệ thống gán nhãn dựa trên PICS cần linh động trong các tiêu chuẩn được sử dụng để đánh giá, và các tiêu chuẩn cần mở để người dùng có thể kiểm tra.

Các nhãn được thêm vào mã HTML được coi như thông tin của thẻ META, và ở phần đầu (header) của trang tài liệu.

```
<META http-equiv="PICS-Label" content='labellist'>
```

Ví dụ một trang HTML có header chứa thông tin gán nhãn PICS:

```
<head>
    <META http-equiv="PICS-Label" content='
```

```
(PICS-1.1 "http://www.gcf.org/v2.5"
   labels on "1994.11.05T08:15-0500"
       until "1995.12.31T23:59-0000"
     for "http://w3.org/PICS/Overview.html"
   ratings (suds 0.5 density 0 color/hue 1))
'>
```
</head>

Phần mềm lọc văn bản tiếng Anh (SP.01-2)

Mục tiêu

Module lọc văn bản tiếng Anh cung cấp chức năng tự động xác định một tài liệu có chứa văn hoá phẩm độc hại, hay các vấn đề chính trị, tôn giáo không. Về bản chất có thể coi bài toán này là bài toán phân lớp trang web. Với mỗi trang web đầu vào, module sẽ phân loại tài liệu vào một trong các lớp cho trước dựa vào nội dung trang web.

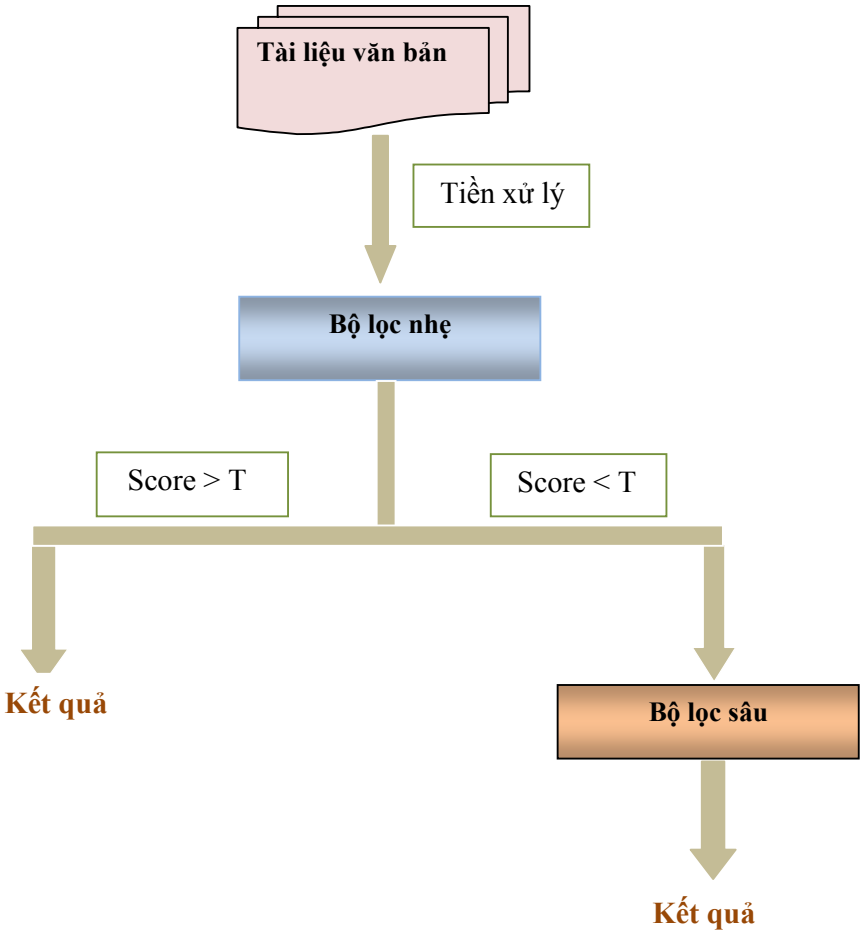
Phạm vi

Module lọc văn bản tiếng Anh được phát triển trên nền công nghệ hiện đại, cung cấp các chức năng tiện dụng cho phép ta có thể cài đặt, tích hợp vào hệ thống lớn dễ dàng.

Tổng quan hệ thống

Lược đồ áp dụng bài toán phân lớp web vào hệ thống lọc nội dung trên Internet sẽ bao gồm hai bộ lọc (*Hình 3.11*):

i. Bộ lọc nhẹ (*light fitler / light classifer)* có khả năng xử lý nhanh với số lượng dữ liệu lớn. Sử dụng phương pháp Naïve Bayes cho bộ lọc này.



Hình 3. 11. Kiến trúc bộ lọc tiếng Anh trong hệ thống lọc nội dung

ii. Bộ lọc sâu (*heavy filter / heavy classifer)* tập trung vào các trang web chưa được phân lớp bởi bộ lọc đơn giản trong quá trình học. Ta sử dụng phương pháp máy vector hỗ trợ (Support Vector Machine – SVM) cho bộ lọc này.

Với mỗi tài liệu văn bản, sau khi tiền xử lý lựa chọn thuộc tính, dữ liệu được đưa tới bộ lọc nhẹ. Nếu kết quả phân lớp trang web của bộ lọc nhẹ lớn hơn ngưỡng T thì quy trình chấm dứt và lấy kết quả quyết định từ bộ lọc nhẹ; trong trường hợp ngược lại, bộ lọc sâu được kích hoạt.

Module được phát triển bằng ngôn ngữ Java, bao gồm 4 thành phần chính:

– Bộ phân lớp Naïve Bayes
– Bộ phân lớp SVMs
– Module tiền xử lý dữ liệu đầu vào cho bộ lọc nhẹ
– Module tiền xử lý dữ liệu đầu vào cho bộ lọc sâu

Trong đó, sử dụng hai bộ phân lớp mã nguồn mở được phát triển bằng ngôn ngữ Java: bộ phân lớp Naïve Bayes được tải miễn phí từ địa chỉ sau: http://www.idsia.ch/%7Egiorgio/jncc2.html và bộ phân lớp Libsvm được tải từ địa chỉ http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Module cũng cho phép người phát triển lựa chọn các phương pháp:

- Chỉ dùng bộ lọc nhẹ
- Chỉ dùng bộ lọc sâu
- Dùng kết hợp cả hai bộ lọc

Module tiền xử lý cho bộ lọc nhẹ

Module tiền xử lý cho bộ lọc nhẹ tách văn bản thành một tập từ đơn và chuyển dữ liệu từ dạng văn bản về định dạng thuộc tính – quan hệ (Attribute-Relation File Format - ARFF). ARFF là định dạng được thiết kết cho bài toán phân lớp. Với mỗi tệp tài liệu được chuyển về theo cấu trúc sau:

```
@relation example

@attribute F1 {0, 1}
@attribute F2 {0, 1}
@attribute F3 {0, 1}
……
……
@attribute F_{n-2} {0, 1}
@attribute F_{n-2} {0, 1}
@attribute F_n {0, 1}
@attribute Class {1, 2, 3, 4}

@data
V_1, V_2, V_3, … ,V_n, Class
```

Trong đó, các data, attribute, real, numeric, …. là các từ khoá. Phần đầu của tệp ARFF bắt đầu với định nghĩa tên của tập dữ liệu: @relation <tên quan hệ> trong đó <tên quan hệ> là một chuỗi.

Các dòng phía dưới định nghĩa các thuộc tính của tập dữ liệu. Giả sử có N thuộc tính kí hiệu là F_1, F_2, …, F_N. Mỗi thuộc tính được định nghĩa trên một dòng: @attribute <tên thuộc tính> <kiểu dữ liệu> trong đó <tên thuộc tính> phải bắt đầu bằng kí tự thuộc bảng chữ cái; <kiểu dữ liệu> có thể là:

- Kiểu số (numeric) hoặc số thực (real)
- Nếu thuộc tính nhận một trong các giá trị nào đó, ví dụ thuộc tính nhận giá trị là 0 hoặc 1 thì kiểu dữ liệu là một danh sách các giá trị được liệt kê trong cặp dấu {}, mỗi giá trị cách nhau bởi dấu phẩy.

Tiếp đó là phần dữ liệu được bắt đầu bởi từ khoá @data. Mỗi ví dụ được ghi trên một dòng và được biểu diễn dưới dạng vector như sau:

V_1, V_2, V_3, … ,V_n, Class

Với V_i là giá trị của thuộc tính F_i Vi nhận giá trị 0 hoặc 1.

[2]. Vi = 0 nếu tài liệu đang xét không chứa thuộc tính Fi

[3]. Vi = 1 nếu tài liệu đang xét chứa thuộc tính Fi

Giá trị Class là nhãn lớp, nhận giá trị là các số nguyên. Vì module chỉ lọc các trang web xấu liên quan đến giới tính và chính trị nên Class chỉ gồm 4 giá trị tương ứng với 4 lớp: Dữ liệu xấu về giới tính (Sex), dữ liệu không xấu về giới tính (No_Sex), dữ liệu xấu về chính trị (Politics), dữ liệu không xấu về chính trị (No_Politics)

– Class = 1: tài liệu thuộc lớp Sex
– Class = 2: tài liệu thuộc lớp No_Sex
– Class = 3: tài liệu thuộc lớp Politics
– Class = 4: tài liệu thuộc lớp No_Politics

– Vì bộ lọc đòi hỏi tốc độ nhanh, nên ta lựa chọn tập đặc trưng là 300 thuộc tính có tần suất tài liệu cao nhất, tức là lựa chọn 300 thuộc tính mà số tài liệu chứa thuộc tính đó là lớn nhất.

– Huấn luyện trên tập dữ liệu học, kiểm tra trên tập kiểm tra thu được kết quả như sau:

Ví dụ về một tài liệu trước và sau khi tiền xử lý như sau:

Tài liệu trước khi tiền xử lý:

```
      <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
      <html xmlns="http://www.w3.org/1999/xhtml" dir="ltr" lang="en">
      <head>
      <meta http-equiv="Content-Type" content="text/html; charset=ISO-
8859-1" />
      <meta name="keywords" content=" Phone Slut By Chickle, sex stories,
porn stories, sex story, sexstoriespost, post stories, erotic stories,
erotic stories, xxx story" />
      <meta name="description" content="[Archive]  Phone Slut By Chickle
Erotic Audio Stories" />

      <title> Phone Slut By Chickle [Archive]   - Sex Stories Post
Forums</title>
      <link rel="stylesheet" type="text/css" href="../archive.css" />
      </head>
      <body>
      <div class="pagebody">
      <div id="navbar"><a href="../index.php">Sex Stories Post Forums</a>
&gt; <a href="f-10.html">Members Erotic Stories</a> &gt; <a href="f-
95.html">Erotic Audio Stories</a> &gt;  Phone Slut By Chickle</div>
      <hr />
      <div class="pda"><a href="t-12866.html@login=1" rel="nofollow">Log
in</a></div>
      <p      class="largefont">View      Full      Version      :      <a
href="../../showthread.php@t=12866">Phone Slut By Chickle</a></p>
      <hr />

      <div             class="post"><div             class="posttop"><div
class="username">Scorpio8</div><div    class="date">03-16-2008,    04:41
PM</div></div><div class="posttext">Oh the things she does when her
Master calls her up<br />
      and instructs her what to do over the phone.<br />
```

```
       <br />
       She describes it all in vivid detail.</div></div><hr />

       </div>
       </body>
</html>
```

Tài liệu sau khi tiền xử lý:

```
       @relation ContentFiltering
       @attribute 'herself' integer
       @attribute 'php' integer
       @attribute 'option' integer
       @attribute 'tightli' integer
       @attribute 'against' integer
       @attribute 'shake' integer
       @attribute 'couch' integer
       …
       @attribute 'quiz' integer
       @attribute 'safe' integer
       @attribute class {0,1}
       @data
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0, … ,0,1
```

Module tiền xử lý cho bộ lọc sâu

Module tiền xử lý tách văn bản thành các tập từ đơn (không tách các cụm từ). Sau khi tách từ và loại bỏ các từ dừng, chương trình tính trọng số từ khoá TF và chương trình đưa mỗi tài liệu về dạng vector các từ mục.

Mỗi văn bản được biểu diễn trên một dòng và dưới dạng vector như sau:

<lớp><thuộc tính>:<giá trị thuộc tính> <thuộc tính>:<giá trị thuộc tính> .... <thuộc tính>:<giá trị thuộc tính>

Trong đó:

– ***<lớp>***: biểu diễn chủ đề của văn bản, nhận giá trị là các số nguyên, có giá trị 1, 2, 3, 4 như module tiền xử lý cho bộ lọc nhẹ

- ***<thuộc tính>***: là số nguyên dương, tham chiếu đến tập thuộc tính được lựa chọn trong quá trình tiền xử lý dữ liệu. Văn bản được sắp xếp theo thứ tự tăng dần của <thuộc tính>.
- ***<giá trị thuộc tính>***: biểu diễn độ quan trọng của thuộc tính trong tài liệu. Giá trị thuộc tính của từ mục t_k trong tài liệu D_i được tính bằng số lần xuất hiện của từ mục t_k trong tài liệu D_i chia cho tổng số từ mục trong tài liệu D_i:

$$TF(t_k, D_i) = \frac{occ(t_k, D_i)}{N}$$

Trong đó N là tổng số từ mục của tài liệu D_i và $occ(t_k,D_i)$ là số lần xuất hiện của từ t_k trong văn bản D_i .

Ví dụ tài liệu sau khi tiền xử lý như sau :

Tài liệu trước khi tiền xử lý:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" dir="ltr" lang="en">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-
8859-1" />
<meta name="keywords" content=" Other Personals Links, sex stories,
porn stories, sex story, sexstoriespost, post stories, erotic stories,
erotic stories, xxx story" />
<meta name="description" content="[Archive] Other Personals Links
Personals Ladies seeking Men" />
<title> Other Personals Links [Archive] - Sex Stories Post
Forums</title>
<link rel="stylesheet" type="text/css" href="../archive.css" />
</head>
<body>
<div class="pagebody">
<div id="navbar"><a href="../index.php">Sex Stories Post Forums</a>
```

```
&gt; <a href="f-39.html">Personals Relationships Members</a> &gt; <a href="f-40.html">Personals Ladies seeking Men</a> &gt; Other Personals Links</div>
	<hr />
	<div class="pda"><a href="t-4770.html@login=1" rel="nofollow">Log in</a></div>
	<p class="largefont">View Full Version : <a href="../../showthread.php@t=4770">Other Personals Links</a></p>
	<hr />
	<div class="post"><div class="posttop"><div class="username">Tiger</div><div class="date">09-09-2004, 02:55 PM</div></div><div class="posttext">If you can't find what you are looking for in our member personals try a site from this list:<br />
	http://www.myadultfriendfinder.com/</div></div><hr /></div>
	</body>
</html>
```

Tài liệu sau khi tiền xử lý:

```
1 0:0 1:0 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0 10:1 11:0 12:0 13:0 14:0 …1008:0
```

Bộ lọc nhẹ Naïve Bayes

Naive Bayes là một thuật toán đơn giản nhưng hiệu quả với bài toán phân lớp có giám sát. Ta khai thác bộ phân lớp Naïve Bayes JNCC mã nguồn mở của Giorgio Corani và Marco Zaffalon, phiên bản 1.0 giới thiệu vào 5/2008, được tải miễn phí từ website http://www.idsia.ch/%7Egiorgio/jncc2.html. JNCC chạy từ dòng lệnh và được phát triển trên nền Java 6.0. Đầu vào của bộ phân lớp này là dữ liệu định dạng theo ARFF. Như vậy, dữ liệu văn bản ban đầu được tiền xử lý để chuyển về định dạng ARFF sau đó được đưa tới bộ lọc Naïve Bayes. Đã tiến hành tích

hợp JNCC vào module lọc văn bản tiếng Anh, kết quả thực nghiệm cho thấy tính khả quan của phương pháp.

Bộ lọc sâu SVMs

Đối với các trang web mà bộ lọc nhẹ kết quả nhỏ hơn ngưỡng T thì bộ lọc sâu được kích hoạt. Ta sử dụng bộ phân lớp SVMs để xây dựng bộ lọc sâu. Hiện nay có khá nhiều bộ phân lớp SVMs mã nguồn mở, do đó đã lựa chọn việc khai thác mã nguồn mở LIBSVM của Chin-Chung Chang và Chih-Jen Lin, được tải miễn phí từ địa chỉ http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Nhóm đề tài đã sử dụng phiên bản 2.86 được giới thiệu tháng 4/2008, phát triển trên nền Java. Đầu vào của bộ lọc sâu là dữ liệu đã qua module tiền xử lý văn bản, được biểu diễn dưới dạng vector các thuộc tính (phần 2.4.2).

Đánh giá, thử nghiệm

Dữ liệu

Tập dữ liệu được thu thập trên từ nhiều trang trên Internet. Để thu thập tập dữ liệu này đã sử dụng phương pháp loang dần các url. Ban đầu, lựa chọn tập nhân gồm một số trang được tìm từ Google (ví dụ trang http://www.sexstoriespost.com/), sau đó mở rộng dần tập url dựa trên các liên kết từ trang này tới các trang web khác. Đối với dữ liệu liên quan đến chính trị, sau khi thu thập tập dữ liệu đủ lớn, đã thực hiện rà soát thủ công để xác định chính xác trang web nào thực sự là dữ liệu xấu.

Tập dữ liệu gồm các trang web về khiêu dâm, chính trị được chia thành tập dữ liệu học và dữ liệu kiểm tra. Thống kê về tập dữ liệu này như sau:

Bảng 3. 3. Thống kê tập dữ liệu

Lớp	Dữ liệu học	Dữ liệu kiểm tra	Tổng

Lớp	Dữ liệu học	Dữ liệu kiểm tra	Tổng
Sex	4500	500	5000
No_Sex	2000	282	2282
Politics	4500	500	5000
No_Politics	2700	300	3000

Lựa chọn đặc trưng:

Việc lựa chọn các đặc trưng là rất quan trọng, có vai trò quyết định trong bài toán phân lớp. Các đặc trưng được lựa chọn càng tinh tế và càng chuyên biệt với từng lớp thì khả năng đoán nhận càng chính xác. Do tập dữ liệu học khá lớn nên việc lọc các đặc trưng thực sự có ích cho quá trình huấn luyện sẽ rất quan trọng. Trong module này, sử dụng phương pháp lọc đơn giản: Loại tất cả các đặc trưng có số lần xuất hiện nhỏ hơn một ngưỡng T cho trước khi huấn luyện mô hình. Với bước lọc này, số lượng đặc trưng sử dụng cho huấn luyện mô hình giảm đi rất nhiều. Thống kê về số đặc trưng khi lựa chọn ngưỡng T = 1000 như sau:

Bảng 3. 4. Thống kê tập đặc trưng

Số đặc trưng ban đầu	Số đặc trưng sau khi đã trích chọn
32415	1009

Kết quả thực nghiệm

Huấn luyện trên tập dữ liệu học 2 lớp Sex và No-Sex, kiểm tra trên tập kiểm tra thu được kết quả như sau:

Bảng 3. 5. Kết quả thực nghiệm với bộ lọc nhẹ

Độ chính xác
88.325%

Bảng 3. 6. Kết quả thực nghiệm với bộ lọc sâu

Độ chính xác
99.175%

Theo kiến trúc bộ lọc tiếng Anh trong hệ thống lọc nội dung, nếu kết quả của bộ lọc nhẹ nhỏ hơn ngưỡng T thì bộ lọc sâu mới được kích hoạt. Trong thực nghiệm này, đã thử lựa chọn ngưỡng T = 0.9

Nhận xét: Từ kết quả thực nghiệm cho thấy, module lọc văn bản tiếng Anh cho kết quả tốt với tập dữ liệu lớn.

Phần mềm lọc văn bản tiếng Việt (SP.01-3)

Mục tiêu của Module

Với một tài liệu (trang web) tiếng Việt module lọc văn bản tiếng Việt cung cấp chức năng tự động xác định được tài liệu đó:

- Có chứa văn hóa phẩm độc hại hay không.
- Có chứa các vấn đề chính trị, tôn giáo nhạy cảm hay không.

Về bản chất có thể quy bài toán trên về bài toán phân lớp trang web. Theo đó, với mỗi trang web đầu vào module phân lớp sẽ phân loại trang web đó vào một trong các lớp cho trước dựa vào nội dung của nó.

Phạm vi của Module

Module lọc văn bản tiếng Việt được xây dựng trên các kỹ thuật, công nghệ hiện đại hiện nay.

Module có thể dễ dàng triển khai thành ứng dụng riêng hoặc tích hợp vào các hệ thống lớn.

Tổng quan về hệ thống

Để phân loại trang web một cách tốt nhất, khuyến cáo nên sử dụng kết hợp hai bộ lọc (*hình 3.12*):

- Bộ lọc nhẹ *(light filter/ light classifier)*: Bộ lọc này có đặc điểm là xử lý nhanh, giải quyết tốt với trường hợp hợp dữ liệu lớn với thời gian và độ chính xác chấp nhận được. Do vậy đã sử dụng phương pháp Naïve Bayes để xây dựng bộ lọc này.
- Bộ lọc sâu *(heavy filter/ heavy classifier)*: Trong quá trình lọc, nếu bộ lọc nhẹ không xác định được phân loại của trang web thì bộ lọc sâu sẽ được kích hoạt. Phương pháp Entroy cực đại (Maximum Entropy-ME) được sử dụng cho bộ lọc này.

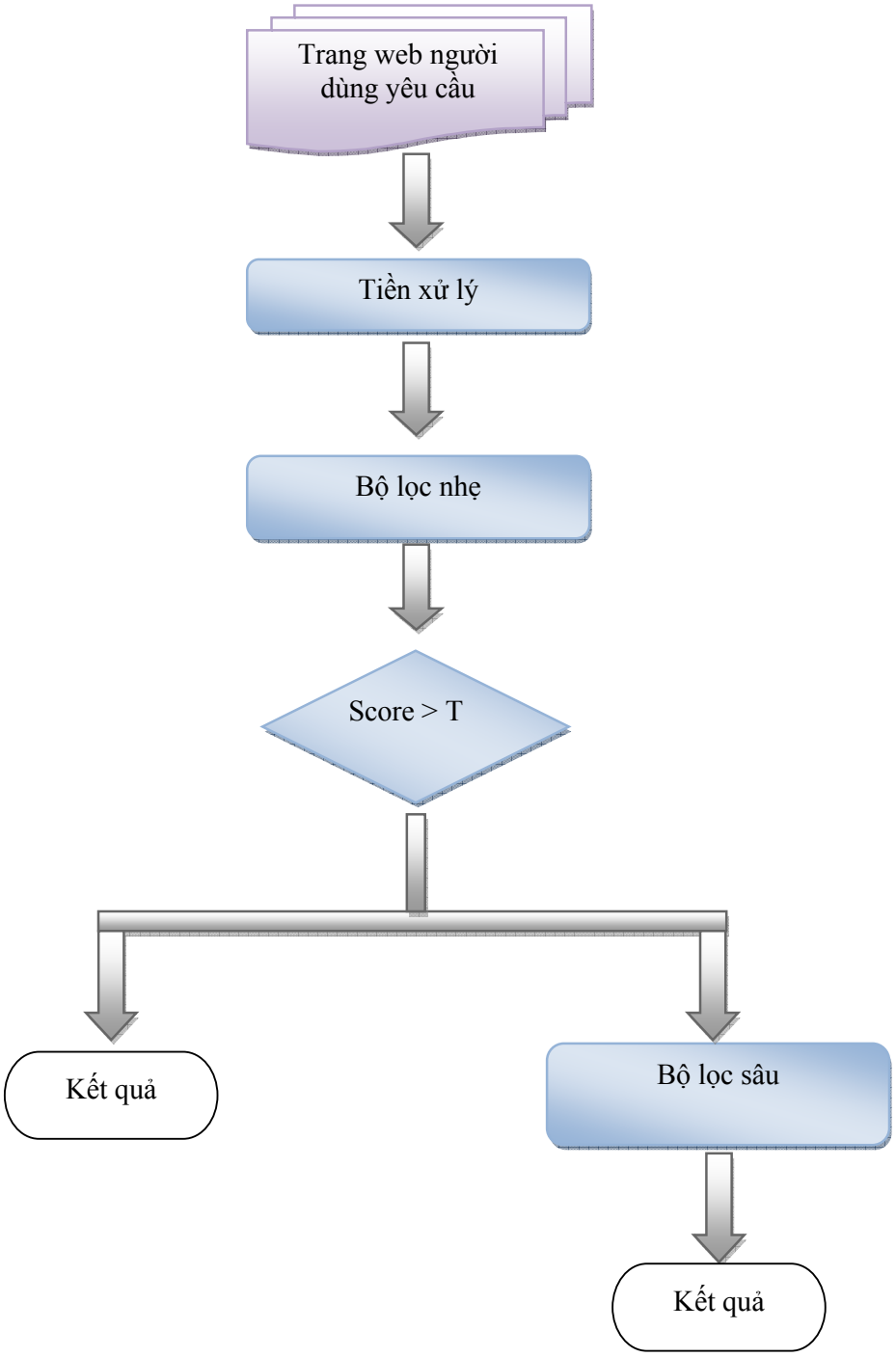
Quá trình lọc trang web được thực hiện như sau: với mỗi trang web sau khi tiền xử lý: loại bỏ các thẻ html, tách từ và lựa chọn thuộc tính, dữ liệu được đưa tới bộ lọc nhẹ. Nếu kết quả phân lớp trang web của bộ lọc nhẹ lớn hơn ngưỡng *T* thì quá trình lọc kết thúc và lấy kết quả quyết định từ bộ lọc nhẹ. Trong trường hợp ngược lại, khi kết quả phân lớp không vượt ngưỡng *T* thì bộ lọc sâu sẽ được kích hoạt.

Toàn bộ module được phát triển bằng ngôn ngữ Java bao gồm các thành phần chính:

- Module tiền xử lý dữ liệu.
- Bộ phân lớp Naïve Bayes
- Bộ phân lớp Maximum Entropy

Module được thiết kế cho phép người dùng có thể lựa chọn phương thức lọc trang web cho phù hợp:

- Chỉ lọc với bộ lọc nhẹ (*light filter)*
- Chỉ lọc với bộ lọc sâu *(heavy filter)*
- Lọc kết hợp cả hai bộ lọc



Hình 3. 11. Kiến trúc bộ lọc tiếng Việt

Module tiền xử lý cho các bộ lọc

Module tiền xử lý cho bộ lọc mang hai nhiệm vụ chính:

- Loại bỏ các thẻ HTML của trang web
- Biểu diễn trang web dưới dạng các đặc trưng - ở đây sử dụng mẫu 3-gram để biểu diễn.

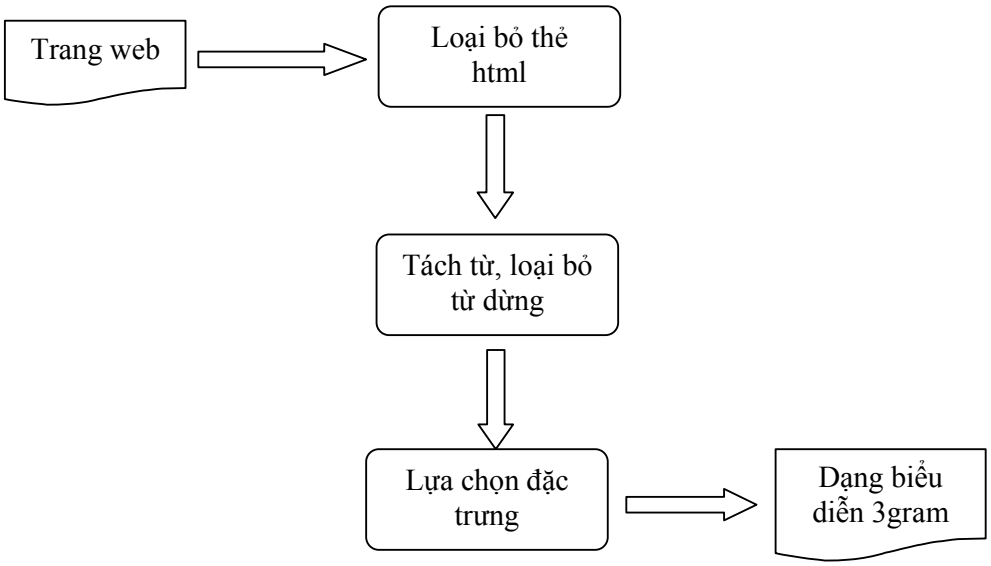
Quá trình tiền xử lý diễn ra như sau (Hình 3. 13):

(1) Loại bỏ các thẻ HTML của trang web, sau đó lưu lại ở định dạng 1 WebObject với khuôn mẫu sau:

```
<blackweb>
      <title> tiêu đề của trang web </title>
      <meta keywords=Tập các keywords></meta>
      <content> Nội dung của trang web </content>
      <outlinks>Các liên kết</outlinks>
</blackweb>
```

(2) Tách từ, loại bỏ từ dừng, loại bỏ các ký tự đặc biệt với phần tiêu đề và nội dung của trang web.

(3) Lựa chọn đặc trưng, biểu diễn phần nội dung trang web bằng các mẫu 3gram



Hình 3. 12. Quá trình tiền xử lý

Module tiền xử lý được đóng gói trong package: *jvntextfilter.perpro*

Bộ lọc nhẹ Naïve Bayes

Với bộ lọc nhẹ Naïve Bayes sử dụng mô hình đa thức *(multinomial model)* vì nó được chứng minh là tốt nhất so với các mô hình còn lại trong nhiều trường hợp của phân lớp văn bản [1,4]. Mô hình đa thức biểu diễn văn bản bằng tập các lần xuất hiện của các từ. Mô hình không quan tâm đến trật tự của từ mà chỉ quan tâm đến số lần xuất hiện của các từ trong một văn bản.

Với bộ bọc nhẹ, ta đưa ra một ngưỡng T , theo đó kết quả lọc của mỗi trang web sẽ được so sánh với T nếu vượt ngưỡng kết quả lọc sẽ được thông báo và ngược lại sẽ chuyển sang bộ lọc sâu.

Bộ lọc nhẹ được đóng gói trong package chính *jvntextfilter* trong đó lớp *lightfilter*.java sẽ thực hiện phân lớp với trang web.

Quá trình phân lớp với bộ lọc nhẹ được tiến hành như sau:

(1). Thiết lập tham số $T = \theta$ cho bộ phân lớp

(2). For mỗi lớp $d \in D$ $c \in C$ do

(3). {

(4). For mỗi tài liệu $d \in T$ do

(5). {

(6). Tính giỏ trị P(*c*|*d*)

(7). Nếu P(c|d) >= θ kết luận d thuộc lớp c, kết thúc

(8). Ngược lại để cho bộ phân lớp mạnh quyết định

(9). }

(10). }

Bộ lọc sâu Maximum Entropy

Bộ lọc sâu sẽ được kích hoạt khi kết quả của bộ lọc nhẹ không đạt ngưỡng. Nhóm đề tài sử dụng thư viện JTextPro: A Java-based Text Processing Toolkit của tác giả Phan, X.H**Error! Reference source not found.** để tiến hành huấn luyện mô hình. Sau khi có được mô hình huấn luyện, ta tiến hành thiết kế lớp *heavyfilter.java* để lọc trang web.

Cụ thể *heavyfilter.java* được thiết kế như sau:

heavyfilter.java
-classifier : Classification = null -FeatureSelection : FeatureSelection = null
+HeavyFilter(in modelDir:String) +Init(in modelDir:String) : Boolean +filter(in Content:String) : Label[]

Hình 3. 13. Bộ lọc sâu

Hàm *filter* sẽ tiến hành phân lớp trang web và trả về nhãn của phân lớp, nhãn này bao gồm 3 giá trị:

- Black: tài liệu mang nội dung xấu
- Gray: tài liệu mang trung tính, còn nhập nhằng trong phân loại
- White: tài liệu sạch, không chứa nội dung xấu, nhạy cảm.

Thử nghiệm và đánh giá

Dữ liệu

Tập dữ liệu của thực nghiệm đã được thu thu thập từ nhiều trang web trên Internet điển hình như: *http://9binh.com; http://tudodanchuvn.com ; http://truyenviet.com* ... cho tập dữ liệu mang nội dung xấu và các trang: *http://cpv.org.vn; http://tapchicongsan.org.vn* … cho tập dữ liệu mang nội dung sạch. Để thực hiện điều này đã sử dụng công cụ crawler của máy tìm kiếm tiếng việt đặt tại *http://203.113.130.205:8080/sise*

Tập dữ liệu được chia thành tập dữ liệu học và tập kiểm tra, thống kê về tập dữ liệu này như sau:

Bảng 3. 7. Thống kê tập dữ liệu

Lớp	Dữ liệu học	Dữ liệu kiểm tra	Tổng
Black	760	71	831
White	690	62	752

Việc lựa chọn số lượng dữ liệu học không lớn bởi vì:

- Tài liệu mang nội dung chính trị, tôn giáo nhạy cảm khó xác định, trong số trang web đã lấy được từ internet cóc rất nhiều tài liệu có nội dung mập mờ.
- Không phải cứ tập dữ liệu huấn luyện nhiều là tốt, nhiều trường hợp số lượng đặc trưng bị bão hòa, không làm nổi một số đặc trưng chính

  Bởi vậy nên quyết định chọn tập dữ liệu như bảng 2.

Kết quả thực nghiệm

Sau khi huấn luyện tập dữ liệu học và tiến hành kiểm tra, ta thu được kết quả như sau:

Bảng 3. 8. Kết quả thực nghiệm với bộ lọc nhẹ

Precision	Recall	F1
73.54	64.64	68.80

Bảng 3. 8. Kết quả thực nghiệm với bộ lọc sâu

Precision	Recall	F1
88.46	90.79	89.61

Đánh giá

Kết quả của module lọc nội dung tiếng Việt là khá tốt. Bộ lọc nhẹ với độ chính xác 73% đủ khả năng đảm bảo lọc nhanh các trang web với độ chính xác chấp nhận được. Bộ lọc sâu cho kết quả tốt hơn với độ chính xác vào khoảng 90% đủ khả năng xử lý các trường hợp mà bộ lọc nhẹ không phân loại được

Yêu cầu dữ liệu

Chuẩn hóa dữ liệu là một module trong hệ thống lọc nội dung hỗ trợ quản lý và đảm bảo an toàn – an ninh thông tin trên mạng Internet nên dữ liệu vào của module là dữ liệu trên môi trường mạng. Dữ liệu trên mạng rất đa dạng, các tài liệu dạng văn bản có các định dạng phổ biến: rtf, doc, pdf, html, txt. Module chuẩn hóa dữ liệu cần chuyển dữ liệu ở các định dạng khác nhau về dạng thống nhất văn bản thuần (txt).

Hiện nay, tiếng Việt có nhiều bảng mã khác nhau và yêu cầu đặt ra cần tiến hành chuẩn hóa, chuyển các tài liệu về cùng bảng mã thông nhất là Unicode dựng sẵn. Ngoài ra module cũng cần cung cấp giao diện dữ liệu vào/ra dạng xâu ký tự hoặc chuỗi bit ký tự cho thành phần chuyển mã nhằm đảm bảo tính khả mở và khả chuyển thi tích hợp vào hệ thống.

Vậy, dữ liệu vào/ra của module có thể ở dạng các file hoặc đoạn văn bản, có thể là xâu ký tự (chuỗi bit).

Yêu cầu chức năng

Module chuẩn hóa dữ liệu thực hiện chuyển các tài liệu ở các định dạng *doc, pdf, rtf* về dạng văn bản thuần (*txt*) và chuyển mã cho các văn bản ở định dạng html hoặc văn bản thuần trong các bảng mã khác nhau (như: VNI, NCR, TCVN3, VIQR, và Unicode tổ hợp) về cùng một mã thống nhất Unicode dựng sẵn.

Vậy, chuẩn hóa dữ liệu gồm có 2 thành phần chính:

[1] Chuyển các định dạng dữ liệu khác nhau (rtf, doc, pdf) về dạng văn bản thuần (txt) thông qua các bộ chuyển *parse*.

[2] Chuẩn hóa bảng mã cho các tài liệu tiếng Việt:

Yêu cầu phát triển

Module viết bằng ngôn ngữ Java (version 1.6) và được đóng gói thành file .jar (dạng file đóng gói của Java). Có thể được gọi thực thi độc lập một cách dễ dàng bằng giao diện dòng lệnh:

java –jar convert.jar << tên file cần chuẩn hóa>>

Dễ dàng tích hợp vào hệ thống khác qua các giao diện (API) như đã nêu trong yêu cầu về dữ liệu.

Yêu cầu chất lượng

Phần này liệt kê các yêu cầu về chất lượng module về độ ổn định, tính mở, độ chịu tải, khả năng tương thích... là những yếu tố có ảnh hưởng quan trọng đến kiến trúc hệ thống.

Tính tiện dụng

- Module được đóng gói thành một file duy nhất.
- Dễ dàng sử dụng với giao diện hỗ trợ nhiều kiểu dữ liệu vào/ra khác nhau: Module cung cấp các giao diện (API) khác nhau cho phép chuẩn hóa từng file hoặc tất cả các file trong thư mục được chỉ dẫn hoặc một trang web hoặc có thể chỉ một đoạn (paragraph) văn bản hoặc một chuỗi ký tự/ chuỗi bit.

Tính tương thích

- Chạy được trên nhiều hệ điều hành: Module có thể thực thi trên các môi trường khác nhau như Windows, Linux/Unix, Mac, Solaris, .. (sử dụng ngôn ngữ Java).

- Nằm trong hệ thống lọc nội dung, thuộc thành phần tiền xử lý dữ liệu, module chuẩn hóa dữ liệu cần có giao diện phù hợp với phân tích của hệ thống tổng thể và dễ dàng ghép nối với các thành phần khác trong hệ thống.

Độ bền và an toàn dữ liệu

- Dữ liệu vào từ các file: module chỉ đọc, không sửa đổi, xoá dữ liệu file khỏi hệ thống.
- Dữ liệu sau khi chuẩn hóa với định dạng đưa ra dưới dạng file cần được lưu trên file riêng không đè lên file gốc với thư mục mặc định trước được xác định trong file cấu hình.

Khả năng chịu "tải"

- Cho phép chạy đa luồng (multi-thread).

Tính mở

- Có khả năng giao tiếp dễ dàng với các module khác, và có thể sử dụng lại trong các hệ thống cần thành phần tiền xử lý dữ liệu.

Chiến lược thiết kế

Hiện nay có một số phần mềm mã nguồn mở hỗ trợ chuyển mã tiếng Việt như UnicodeConverter [3a.1], vietuni[3a.2]. Do đó việc xây dựng module được phát triển dựa trên các mã nguồn đó và phát triển các thành phần cần thiết phù hợp với hệ thống.

Việc chuyển đổi các tệp dữ liệu ở định dạng *rtf, doc, pdf, html* về dữ liệu văn bản thuần (*txt*) sử dụng các gói (package) đã được sử dụng trong máy tìm kiếm tiếng Việt (VNSen [3a.3] dựa trên mã nguồn mở Nutch [3a.4]) để chuyển đổi các định dạng file khác nhau.

Tất cả các mã nguồn trên đều được viết bằng ngôn ngữ java và đã được sử dụng cho kết quả tốt để chuẩn hóa dữ liệu trong máy tìm kiếm tiếng Việt.

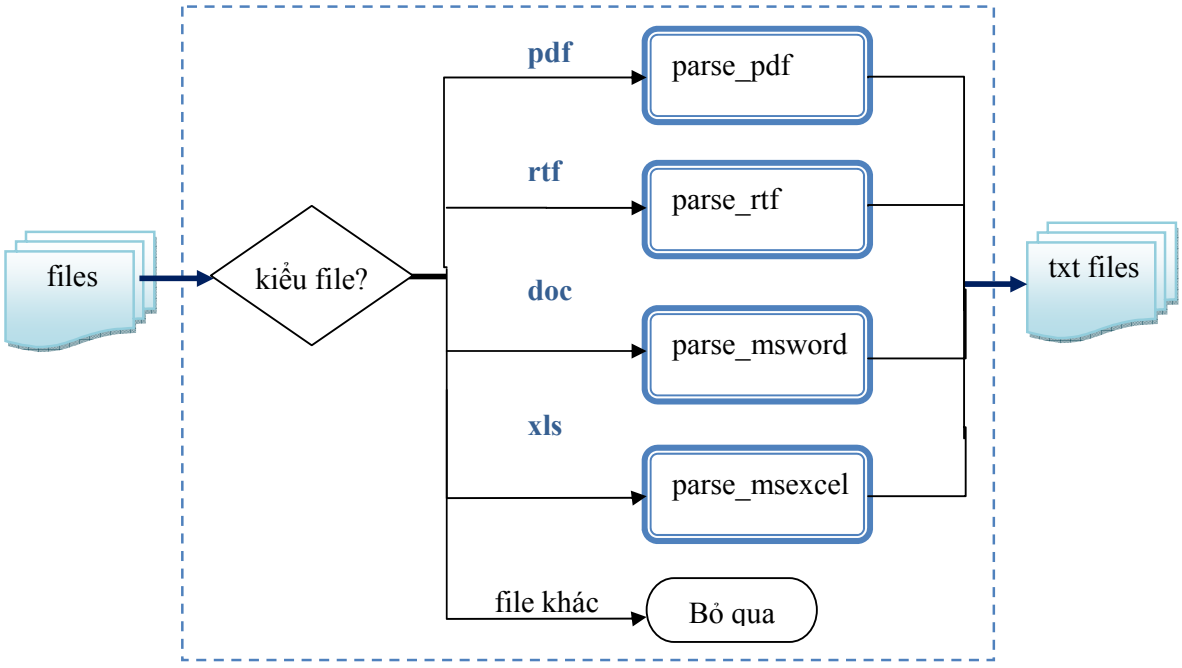
Do đó cũng rất phù hợp để sử dụng phát triển module chuẩn hóa cho hệ thống lọc.

Kiến trúc mức cao

Module bao gồm 2 thành phần chính: chuyển đổi định dạng file dữ liệu từ các dạng rtf, doc, pdf, html về văn bản thuần và chuẩn hóa bảng mã cho các tài liệu văn bản thuần tiếng Việt về dạng thống nhất Unicode dựng sẵn.

Chuyển định dạng file dữ liệu

Thông qua các bộ chuyển đổi từng định dạng file về file văn bản thuần (txt), module chuyển định dạng file dữ liệu dựa vào phần mở rộng file xác định loại file cần chuyển, sau đó đưa vào từng bộ chuyển tương ứng. Ở đây, module chỉ thực hiện chuyển các file pdf, rtf, doc, xls về txt, còn khi gặp các file định dạng khác như js, swf, bmp.. sẽ không xử lý.



Hình 3. 14. Module chuyển định dạng

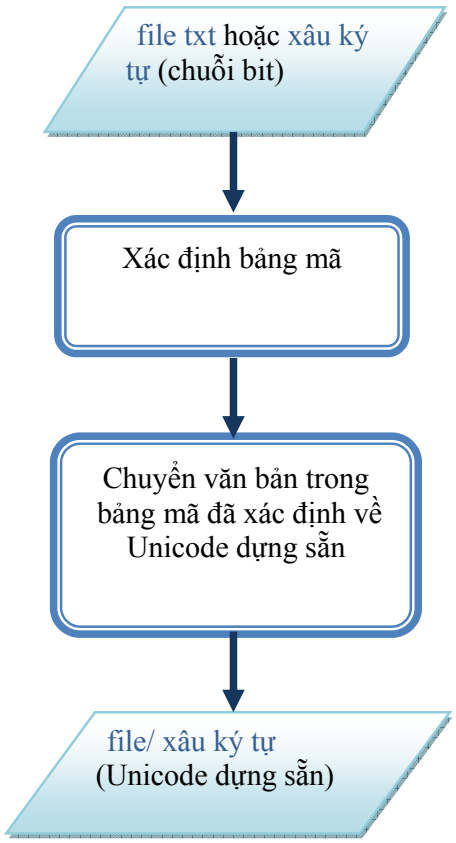
Ngoài ra, trường hợp đặc biệt với file định dạng html, do định dạng html có các thông tin về bảng mã, font chữ hỗ trợ chuyển đổi bảng mã. Nên thay vì

đưa qua module chuyển đổi định dạng file trước, file html sẽ được đưa trực tiếp qua module chuẩn hóa bảng mã và file kết quả là văn bản thuần (txt) được chuẩn hóa.

Chuẩn hóa bảng mã cho các tài liệu tiếng Việt:

Các tài liệu, văn bản tiếng Việt có thể sử dụng nhiều bảng mã khác nhau như VNI, VPS, NCR, TCVN3, VIQR, hay Unicode tổ hợp. Do đó để chuẩn hóa chuyển tất cả các văn bản về cùng một bảng mã thống nhất Unicode dựng sẵn, trước tiên cần xác định văn bản đang dùng bảng mã nào sau đó mới có thể chuyển văn bản từ bảng mã hiện tại sang bảng mã Unicode dựng sẵn.

a. **Nhận diện/ xác định bảng mã**: Xác định bảng mã được sử dụng trong từng đoạn văn bản, và chuyển tiếp cho thành phần chuyển mã phía sau. Thông thường trong văn bản chỉ sử dụng một bảng mã, tuy nhiên với tài liệu trên môi trường mạng do một văn bản có thể do nhiều người cùng soạn nên có thể xảy ra trường hợp mỗi đoạn văn bản được sử dụng một bảng mã riêng. Vì vậy module cần xác định bảng mã trên từng đoạn.
b. **Chuyển bảng mã:** Chuyển đoạn văn bản từ bảng mã tương ứng xác định ở trên sang bảng mã Unicode dựng sẵn.



Hình 3. 15. Module chuẩn hóa bảng

Thiết kế chi tiết

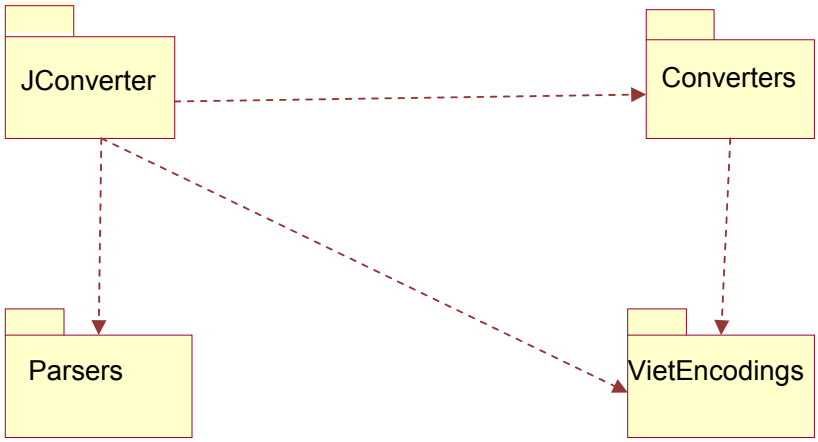
Mô hình các phân hệ bên trong (logic)

Bộ chuẩn hóa dữ liệu bao gồm các phân hệ:

1. Parsers: gồm các lớp chuyển định dạng file dữ liệu parse_pdf, parse_rtf, parse_msword, parse_msexcel
2. JConverter: Cung cấp các API chuyển bảng mã, và thực hiện xác định bảng mã dựa vào VietEncodings trước khi gọi tới Converter tương ứng.

3. Converters: Các lớp chuyển bảng mã gồm ViqrConverter, VniConverter, VpsConverter, Tcvn3Converter, NcrConverter, IscConverter.
4. VietEncodings: Chứa các thông tin đặc trưng của mỗi bảng mã.

Mối quan hệ giữa các phân hệ được đưa ra ở Hình 3. .



Hình 3. 17. Các phân hệ trong bộ chuẩn hóa dữ liệu

Mô hình vật lý các thành phần

Mô hình phân tầng

[1] JConverter

- ***Tầng giao diện***

  Gui.java

- ***Tầng xử lý***

  Gồm các lớp trong gói *jconverter.util*

  Quan trọng nhất*: jconverter.java*

- ***Tầng dữ liệu***

  VietEncoding.java

[2] Converters

- ***Tầng giao diện***

  Converter.java

- ***Tầng xử lý***

  UnicodeConverter.java

  CompositeConverter.java

  Tcvn3Converter.java

  ViqrConverter.java

  VisciiConverter.java

  VniConverter.java

  VpsConverter.java

  IscConverter.java

  NcrConverter.java

- ***Tầng dữ liệu***

  VietEncoding.java

[3] **VietEncodings**: là 1 lớp (class) dữ liệu, chứa đặc trưng của các bảng mã, có danh sách/ mảng các bit ký tự đặc trưng của từng bảng mã với khai báo **final char**[].

[4] Parsers

- ***Tầng giao diện***

  Parse.java

- ***Tầng xử lý***

  ParseImpl.java

  ParsePluginList.java

Và các gói chuyển định dạng:

parse_rtf.jar, parse_pdf.jar, parse_msword.jar, parse_msexcel.jar

Mô hình các thành phần (component)

Bộ chuẩn hóa được đóng gói thành một file duy nhất ***converter.jar***.

Trong đó bao gồm các gói thành phần (package):

- *UnicodeConverter.jar* : đóng gói các thành phần thuộc UnicodeConverter được sử dụng lại
- *Parser.jar* : đóng gói các bộ chuyển định dạng file
- *JConverter.jar*: đóng gói module xác định bảng mã và các thành phần khác của bộ chuẩn hóa.

Môi trường

a. Yêu cầu về phần cứng

- CPU: Celeron 1 GHz trở lên
- Memory: ≥ 512 MB RAM

b. Yêu cầu về phần mềm

- jdk 1.6 trở lên

Đánh giá

Dữ liệu thử nghiệm

- Sử dụng 100 tài liệu tiếng Việt trên wiki (ở bảng mã chuẩn Unicode dựng sẵn)
- Dùng công cụ Unikey chuyển 100 tài liệu trên sang các bảng mã khác: TCVN3, Vni, VIQR, VISCII, VPS, Unicode tổ hợp, NCR

Kết quả

Bảng mã	*TCVN3*	*Vni*	*VIQR*	*VISCII*	*VPS*	*Unicode tổ hợp*	*NCR*
Độ chính xác (%)	~~99~~	~~95~~	~~93~~	~~80~~	~~87~~	~~100~~	~~99~~

Yêu cầu

Yêu cầu dữ liệu

Nhận dạng ngôn ngữ là một module trong hệ thống lọc nội dung hỗ trợ quản lý và đảm bảo an toàn – an ninh thông tin trên mạng Internet nên dữ liệu vào của module là dữ liệu trên môi trường mạng. Theo thống kê hiện nay trên mạng internet có khoảng 600 ngôn ngữ, do đó module nhận dạng ngôn ngữ cần xác định được ngôn ngữ của một trang web định hướng đi vào bộ lọc tương ứng.

Để đảm bảo tính khả mở và khả chuyển khi tích hợp vào hệ thống, module nhận dạng ngôn ngữ cũng cần cung cấp giao diện vào/ra dạng xâu ký tự hoặc chuỗi bit ký tự.

Vậy, dữ liệu vào/ra của module có thể ở dạng các file hoặc đoạn văn bản hoặc cũng có thể là xâu ký tự.

Yêu cầu chức năng

Module nhận dạng ngôn ngữ cần cung cấp chức năng tự động xác định ngôn ngữ cho một văn bản mới. Về bản chất có thể coi bài toán này là bài toán phân lớp văn bản. Với mỗi trang web đầu vào, module sẽ phân loại tài liệu vào một trong các lớp cho trước dựa vào chính nội dung trang web.

Yêu cầu phát triển

Module viết bằng ngôn ngữ java (version 1.6), có thể được gọi thực thi độc lập một cách dễ dàng bằng giao diện dòng lệnh:

java -jar test.jar ten_file_can_nhan_dang_ngon_ngu

Yêu cầu chất lượng

1. ***Tính tiện dụng***: Dễ dàng sử dụng các giao diện khác nhau, hỗ trợ nhiều kiểu vào/ra khác nhau. Điều này còn có lợi trong quá trình kiểm thử và khi muốn sử dụng lại module.

2. ***Tính tương thích***: có thể chạy trên nhiều hệ điều hành, dễ dàng ghép nối với các thành phần khác trong hệ thống lọc nội dung
3. ***Độ bền và an toàn dữ liệu***: Dữ liệu vào từ các file chỉ đọc, không sửa đổi hay xoá dữ liệu file khỏi hệ thống. Đối với bộ phân lớp được sinh ra có thể dễ dàng sao lưu vì kích cỡ nhỏ.
4. ***Tính mở***: Dễ dàng giao tiếp với các thành phần khác. Đặc biệt có thể sử dụng lại trong các thành phần tiền xử lý dữ liệu.

Phạm vi

Module nhận dạng ngôn ngữ được xây dựng trên nền hệ thống hiện đại, viết trên nền Java1.6. Module này có tính độc lập nên cung cấp khả năng tích hợp rất dễ dàng. Đặc biệt do áp dụng phương pháp học tốt, nên module này cho phép mở rộng phạm vi ngôn ngữ dễ dàng. Hiện tại đã xây dựng được bộ nhận dạng cho hai ngôn ngữ tiếng Việt và tiếng Anh.

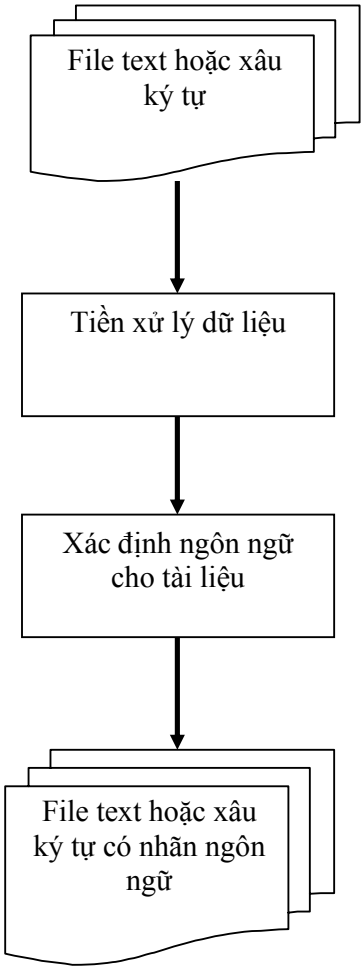
Chiến lược thiết kế

Module nhận dạng ngôn ngữ nhận là module đầu tiên trong hệ thống lọc. Kết quả của module nhận dạng là đầu vào cho module chuẩn hoá dữ liệu sau đó tiếp tục đi đến các bộ lọc tương ứng. Với vai trò đứng đầu tiên trong hệ thống lọc module nhận dạng ngôn ngữ đòi hỏi tốc độ nhanh (performance & avaiability), độ chính xác (performance)

Dung lượng bộ phân lớp được thiết kế sao cho không quá lớn, đảm bảo hiệu năng mong muốn của hệ thống. Hơn nữa, phải đảm bảo cho việc dễ dàng sao lưu, backup dữ liệu khi hệ thống xảy ra sự cố.

Trong module này, sử dụng thuật toán học Naïve Bayes – một thuật toán phân lớp rất hiệu quả, và cho độ chính xác cao 98%.

Thiết kế kiến trúc



Hình 3. 16. Module nhận dạng ngôn ngữ

Thiết kế chi tiết

Mô hình logic

Bộ chuẩn hoá dữ liệu bao gồm các thành phần

1. Train: các lớp xây dựng bộ nhận dạng ngôn ngữ

2. Test: gồm các lớp cho phép xác định ngôn ngữ của một trang web mới

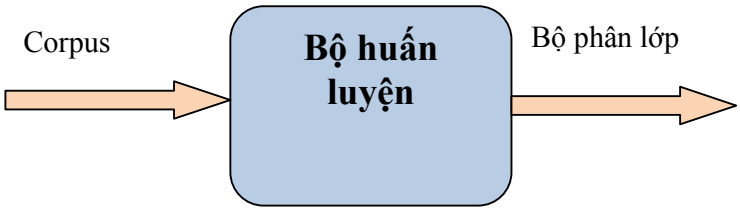
3. Bộ phân lớp: gồm dữ liệu đã được học thống kê được sử dụng để nhận dạng một file (chuỗi ký tự) mới.

4. Bộ chuẩn hoá dữ liệu: loại bỏ từ dừng, số, các đặc trưng không có nhiều ý nghĩa để nhận dạng ngôn ngữ.

Thiết kế vật lý

Module nhận dạng ngôn ngữ **LanguageID** viết trên nền Java 1.6.

Train



Các module con thuộc Bộ huấn luyện được cài đặt như sau:

a. **docThuMuc**(thumuc **fd**)

- Mô tả:
- Đầu vào:
- Đầu ra:

b. **sinhNgram**(file **f**)

- Mô tả: từ file văn bản đầu vào **f**, sinh file 3gram
- Đầu vào: file văn bản **f**
- Đầu ra: ghi các 3gram vào file ngram.txt

c. **tinhTanSuat**(file **ngram**)

- Mô tả: tính tần suất xuất hiện của từng 3gram trong file ngram.txt
- Đầu vào: file ngram.txt
- Đầu ra: bảng lưu các 3gram và số lần xuất hiện của chúng trong corpus

d. **sinhTrain**()

- Mô tả: ghi bảng tần xuất vào file train.txt
- Đầu vào: bảng tần xuất
- Đầu ra: file train.txt

Test

Mô tả: gán nhãn cho một văn bản mới

Đầu vào: Một file văn bản

Đầu ra: Nhãn ngôn ngữ có xác suất cao nhất

Module test.java thừa kế từ train.java để sử dụng 2 module chuanHoa() va sinhNgram. Nó được khai báo thêm một số các module con sau:

a. **tinhXacSuatNhan**(file **ngram**)

- Mô tả: từ file ngram.txt sinh ra của file văn bản mới, tính xác suất nhãn theo công thức Bayes.
- Đầu vào: dữ liệu từ file ngram.txt
- Đầu ra: xác suất sinh cho từng nhãn

Bộ học

Được lưu dưới dạng file .text, gồm nhiều dòng ghi. Mỗi dòng ghi có định dạng như sau:

3-gram <tần_suất_xuất_hiện>

Chuẩn hoá dữ liệu

a. **chuanHoa**(String **s**)

- Mô tả: Loại bỏ các kí tự không có ý nghĩa phân biệt đặc trưng ngôn ngữ, ví dụ dấu câu (., !, ?, /,…). Thay thế dấu cách là kí tự nối "_"
- Đầu vào: Câu vào **s**
- Đầu ra: Câu **s** đã được chuẩn hoá

Môi trường

a. Yêu cầu về phần cứng
   - CPU: Celeron 1GHz trở lên
   - Memory: lớn hơn 512MB RAM

b. Yêu cầu về phần mềm
   - jdk 1.6 trở lên

Đánh giá

a. Dữ liệu thử nghiệm

- Sử dụng các tài liệu web với độ ngắn dài khác nhau để kiểm tra khả năng nhận dạng của module.
- Sử dụng công cụ lọc thẻ web được xây dựng bởi nhóm nghiên cứu "Công nghệ tri thức và tương tác người máy"

b. Kết quả

Số ký tự	**5 ký tự**	**10 ký tự**	**20 ký tự**	**>50 ký tự**
Độ chính xác	60%	85%	92%	98%

Phần mềm lọc ảnh (SP.01-4)

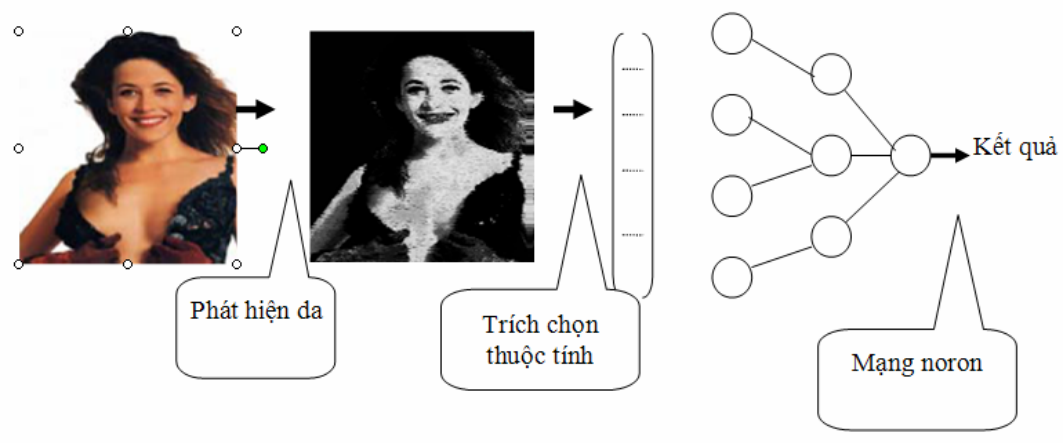
Đối với chức năng lọc hình ảnh:

Có nhiều kỹ thuật được đề xuất và được lựa chọn để giải quyết vấn đề lọc hình ảnh. Tài liệu báo cáo này sẽ thực hiện đánh giá mức độ tốt xấu, định dạng của bức ảnh (cụ thể là những bức ảnh có định dạng: JPG/JPEG, GIP, PNG, BMP) dựa trên sự kết hợp của các phương pháp và giải thuật, và được xây dựng theo 3 bước chính sau:

Bước đầu tiên: Xác định bản đồ da từ ảnh gốc. Bước này được thực hiện theo mô hình MaxEnt.

Bước thứ hai: Trích chọn các thuộc tính là ảnh bản đồ da, cụ thể là các điểm màu da, dựa trên các hình ellipse(GFE, LFE) để tạo nên vector thuộc tính của ảnh.

Bước cuối cùng: Phân loại ảnh với đầu vào là các thuộc tính đã tìm được ở bước thứ hai, sử dụng mạng noron(MLP).



Hình 3. 19. Các bước trong quá trình phát hiện nội dung ảnh

Như vậy, với những ảnh đầu vào từ trên Internet, và được thực thi theo ba bước trên, kết quả mong muốn cần thực hiện là đánh giá tốt được mức độ tốt, xấu của ảnh đó và nhận được định dạng của ảnh, để từ đó cho ra quyết định đúng đắn với những web có chứa ảnh này.

Module lọc ảnh xây dựng phải phân tích, phân loại, được nội dung của ảnh. Nội dung của module cần xây dựng được xác định theo các bước đã giới thiệu trên, kết quả cuối cùng của module là đưa ra kết quả quyết định được ảnh thuộc loại nào: tốt hay xấu? định dạng của ảnh?

MÔ TẢ HỆ THỐNG LỌC

Để có thể minh hoạ rõ nét qúa trình hoạt động của bộ ra quyết định, chúng ta sẽ xem xét scenario thể hiện quá trình hoạt động của bộ ra quyết định trong quá trình tương tác với các mô đun khác trong hệ thống lọc nội dung.

Bộ ra quyết định được cài đặt trong hệ thống phân tích nội dung Web. Quá trình hoạt động, tương tác của bộ ra quyết định với các mô đun khác trong bộ lọc Web được diễn ra như sau:

1. Trình duyệt web yêu cầu kết nối tới trang web server.

2. Dữ liệu đầu vào xuất phát từ hệ thống proxy sẽ được kiểm soát thông qua bộ kiểm soát. Dữ liệu này sẽ được gửi xử lý ở mô đun chuẩn hoá dữ liệu với nhiệm vụ chuyển tất cả các tài liệu Web/Mail về hai dạng cơ bản : plaintext và hình ảnh. Song song quá trình này, riêng với hệ thống lọc Web, bộ kiểm soát sẽ gửi URL của tài liệu tương ứng đến bộ lọc URL/PICS.
3. Sau khi chuẩn hoá dữ liệu về hai dạng cơ bản xong, dữ liệu này sẽ được xem như dữ liệu đầu vào của mô đun kiểm soát
4. Mô đun kiếm soát làm nhiệm vụ điều phối quá trình phân tích nội dung tài liệu. Trong trường hợp tài liệu này là văn bản plaintext, nó sẽ được chuyển đến mô đun xác định ngôn ngữ, nếu là hình ảnh nó sẽ chuyển trực tiếp đến mô đun lọc hình ảnh.
5. Sau khi xác định được ngôn ngữ của tài liệu, kết quả này sẽ được mô đun xác định ngôn ngữ chuyển lại cho bộ kiểm soát. Mô đun này sẽ từ đó chuyển tài liệu đó đến bộ lọc tương ứng với ngôn ngữ vừa xác định được (Việt/Anh).
6. Kết quả của các mô đun lọc văn bản plaintext tiếng Việt, tiếng Anh, lọc hình ảnh và URL+PICS sẽ được mô đun kiểm soát thu nhận và chuyển đến bộ ra quyết định.
7. Bộ ra quyết định sẽ dựa vào những chính sách lọc đã được quy định, tiến hành ra quyết định cấm, không cấm hay xét duyệt lại nội dung tài liệu. Từ đó mô đun kiểm soát có thể tương tác với hệ thống firewall/proxy để thực hiện những quyết định đưa ra.

Trong số các mô đun đề cập trên đây, các mô đun kiểm soát, ra quyết đinh, và lọc URL/PICS được tích hợp trực tiếp vào trong hạ tầng hệ thống lọc nhằm tăng hiệu quả xử lý. Các mô đun lọc tiếng Việt, tiếng Anh và hình ảnh, chuẩn hoá dữ liệu và xác đinh ngôn ngữ sẽ được tách riêng thành những tiến trình riêng và được triệu gọi từ hệ thống lọc thông qua những lời gọi hệ thống.

Đặc tả chi tiết từng thành phần trong hệ thống lọc tài liệu này sẽ được trình bày cụ thể ở phần dưới đây. Trong số những mô đun liệt kê dưới đây, riêng bộ lọc Mail sẽ không cần cài đặt bộ lọc URL và PICS.

Hạ tầng cơ bản của hệ thống lọc tài liệu sẽ được phát triển tích hợp các thành phần dưới đây:

- Đọc các tệp cấu hình hệ thống liên quan đến các bộ lọc thành phần, cập nhật cũng như tối ưu hoá các dữ liệu liên quan.
- Tiến hành khởi tạo thi hành các mô đun lọc thành phần: lọc tiếng Việt, lọc tiếng Anh và lọc ảnh, chuẩn hoá dữ liệu, xác định ngôn ngữ. Việc tương tác giữa các mô đun này thông qua những hàm của bộ kiểm soát.
- Tiến hành thiết lập kết nối đến hệ thống proxy tương ứng khi có yêu cầu từ phía người sử dụng.
- Listen đợi các kết nối đến từ phía người dùng.
- Trả lại kết quả cho phía người dùng tuỳ thuộc và quyết định của bộ kiểm soát.

YÊU CẦU ĐẶT RA

Với các kết quả đã được đánh giá từ các thành phần lọc thông tin khác nhau như lọc văn bản, lọc hình ảnh… ta cần phải đưa ra quyết định cho phép hay từ chối quyền truy cập đến một thông tin nào đó được cung cấp trên Internet. Do có nhiều các thành phần lọc tài liệu khác nhau nên việc xây dựng một bộ quyết định phải đáp ứng được một số các yêu cầu nhất định mà ta cần xem xét dưới đây.

Một bộ lọc thông tin luôn cần phải có tính đáng tin cậy, sự nhanh chóng và một cấu hình tốt. Thật dễ hiểu lý do tại sao cần có 2 yếu tố đầu trong 3 yêu cầu trên, còn yếu tố thứ 3, chúng ta có thể hiểu như sau: Những người dùng khác nhau luôn có những nhu cầu truy cập thông tin khác nhau do đó, bộ lọc thông tin không thể áp dụng một quy tắc lọc tốt cho tất cả nhưng có thể

áp dụng linh hoạt và có tính thích ứng cao. Để đơn giản hơn chúng ta sẽ dùng một phương tiện tạo quyết định sẽ được đề cập cụ thể trong kiến trúc mô đun dưới đây.

Một điểm khá quan trọng của bộ quyết định là làm sao để hạn chế tối đa trạng thái không đồng bộ của những kết quả các thành phần lọc dữ liệu khác nhau. Để giải quyết tốt vấn đề này, ta cần xây dựng một giải thuật tốt để tổng hợp, đánh giá các thang điểm đầu vào để đưa ra một quyết định phù hợp.

Giải thuật này sẽ được đề cập cụ thể dưới đây.

Ngoài những vấn đề trên, bộ quyết định còn cần phải đáp ứng được tính có thể mở rộng. Điều này có nghĩa là cần xây dựng một bộ quyết định với tính năng dễ dàng mở rộng, thực thi một thuật toán mới hỗ trợ cho 1 kiểu lọc mới mở rộng.

Với những yêu cầu đặt ra trên đây, nhóm nghiên cứu đã xây dựng một mô hình cho phép xác định giải pháp quyết định hợp lý cho toàn bộ thông tin từ các thành phần lọc khác nhau đưa ra.

MÔ ĐUN RA QUYẾT ĐỊNH

Chính sách ra quyết định

Thông thường, bộ ra quyết định sẽ phải đợi cho đến khi nhận được tất các các messages được bộ kiểm soát truyền đến từ tất các các bộ lọc trong hệ thống. Tuy nhiên, với luật như thế, bộ ra quyết định sẽ có tốc độ xử lý rất hạn chế và hiệu quả của cả hệ thống lọc nội dung sẽ bị ảnh hưởng. Chính vì thế, việc đưa ra những chính sách ra quyết định hiệu quả hơn cần phải được xem xét và sử dụng.

Chính sách đề xuất trong giải pháp ra quyết định ở đây được xây dựng dựa trên mô hình mạng Bayesian. Đối với chính sách này, dựa trên những luật lọc, bộ ra quyết định có thể trả lời mà không cần thiết phải đợi scores của tất

cả các thành phần lọc khác trong hệ thống. Chẳng hạn, bộ ra quyết định sẽ có thể trả lời ngay khi nhận được kết quả từ mô đun lọc URL và PICS.

Chính sách ra quyết định trong mô đun này cũng cần phải xử lý vấn đề liên quan đến quá trình xử lý các ảnh trong một tài liệu yêu cầu. Thông thường trong một trang Web thường có nhiều hơn một ảnh, vì thế bộ ra quyết định phải có những chính sách riêng liên quan đến tập ảnh trong một tài liệu đang được phân tích.

Dựa trên các chính xách ra quyết định, bộ ra quyết định sẽ có thể gửi trả lời lại cho bộ kiểm soát theo những thông điệp dưới đây:

1. **ACCEPT** *req-number* : Thông điệp này xác định nội dung tài liệu có định danh *req_number* đã được chấp nhận.

Chẳng hạn *ACCEPT 2345*.

2. **READY** *protocol-ID* : Đây là thông điệp của bộ ra quyết định gửi đến bộ điều khiển trong quá trình khởi tạo hoạt động của bộ ra quyết định để thông báo đã sẵn sàng nhận những yêu cầu theo giao thức đã được xác định bằng protocol_ID.

3. **REJECT** *req- number* : Đây là thông điệp xác định tài liệu yêu cầu sẽ bị hệ thống lọc nội dung cấm không cho phép truy cập.

Ví dụ : *REJECT 1234.*

4. **SUSPECT** *req- number* : Đây là thông điệp xác định tài liệu yêu cầu cần phải được kiểm duyệt lại bởi những nhà chức trách liên quan, từ đó cập nhật lại trực tiếp vào danh sách trắng (nếu cho phép truy cập) hoặc danh sách đen (nếu cấm)

Ví dụ : SUSPECT 2345.

Cơ chế hoạt động

Để giải quyết những vấn đề đặt ra trong mô đun ra quyết định, giải thuật ra quyết định sẽ bao gồm:

- Đầu vào của mô đun sẽ bao gồm các thông số sau:

  - Mã ID yêu cầu
  - Điểm đánh giá của các thành phần lọc.
  - Phạm vi đánh giá của thông tin, được xác định là bạo lực hay khiêu dâm để có những chính sách thuật giải khác nhau.
  - Nguồn thông tin được đánh giá, ví dụ như văn bản hay hình ảnh…
  - Ngôn ngữ của văn bản (nếu điểm được đánh giá từ bộ lọc văn bản)
  - Cấp độ xấu của thông tin. Cấp độ sẽ được xếp từ thấp đến cao.
- Khi điểm đánh giá của phần lọc cao hơn 80/100 hoặc cấp độ xấu của thông tin là cao thì mô đun sẽ đưa ra quyết từ chối mà không cần xem xét đầy đủ cá thông tin khác.
- Khi thời gian lọc dữ liệu vượt quá mức cho phép hoặc dữ liệu quá lớn, quyền truy cập sẽ được trả lại cho ngươi sử dụng với nguồn tài liệu sẽ được xem xét đánh giá ngoài luồng và cập nhật lại danh sách trắng, danh sách đen sau này.
- Với các giá trị trung gian khác:
  - Nếu số các điểm trung gian (có điểm lớn hơn 30/100 và nhỏ hơn 70/100) hoặc cấp độ xấu trung bình chiếm hơn một nửa số thông tin từ các bộ lọc trên cùng một ID thì quyền truy cập sẽ bị cấm.
  - Nếu các điểm trung gian đủ nhỏ hoặc độ xấu đủ thấp thì việc truy cập thông tin sẽ được cho phép.
  - Các trường hợp còn lại sẽ được đưa về quyết định mặc định của cấu hình quyết định. Quyết đinh này giống như quyết định khi thời gian phân tích đánh giá của các bộ lọc vượt quá giới hạn cho phép.

- Trên đây là những thuật toán chính giúp đánh giá các bộ thông tin và đưa ra các quyết định cho quyền truy cập của người sử dụng.

Giải thuật quyết định

Để có thể đưa ra quyết định trả về bộ kiểm soát, dựa trên các chính sách đã đề cập ở phần trên, bộ ra quyết định sẽ sử dụng giải thuật ra quyết định dựa trên các luật đã xác định trước, để xây dựng, cài đặt mô đun này.

Bộ ra quyết định sẽ tiến hành phân tích các dữ liệu đến từ các bộ lọc, xác định luật tương ứng và từ đó ra quyết định tương ứng với những luật đó. Các luật này cần phải tuân theo một số đề xuất sau:

- Ra quyết định theo từng thể loại, chẳng hạn như theo thể loại ảnh khiêu dâm, văn bản xấu tiếng Việt, văn bản xấu tiếng Anh, …
- Ra quyết định theo từng domain của tài liệu yêu cầu.
- Ra quyết định dựa vào trễ thời gian (bỏ qua đầu vào) hay theo một yêu cầu cụ thể nào.
- Ra quyết định theo cấu hình mặc định hoặc cấu hình đề nghị của người quản lý đưa ra.

3.2. Phần mềm lọc thư điện tử - MAIL GATEWAY (SP.02)

3.2.1. Giới thiệu

Gần 40 năm sau khi ra đời, thư điện tử (email) hiện tại vẫn là phương tiện giao tiếp phổ biến nhất trên. Mặc dù thế giới Internet ngày càng xuất hiện thêm nhiều công cụ hỗ trợ giao tiếp trực tuyến mới như tin nhắn tức thời, mạng xã hội ảo, email vẫn nhận được sự ưu ái từ người sử dụng. Tuy nhiên trong các cuộc thống kê, email có nội dung ngoài mong muốn của người sử dụng luôn chiếm số lượng áp đảo. Các email ngoài mong đợi của người sử dụng được gọi chung là spam.

Khái niệm thư rác(spam)

Theo Nghị định của chính phủ số 90/2008/NĐ-CP ngày 13 tháng 08 năm 2008 về chống thư rác:

Thư rác (spam) là thư điện tử, tin nhắn được gửi đến người nhận mà người nhận đó không mong muốn hoặc không có trách nhiệm phải tiếp nhận theo quy định của pháp luật.

Phân loại thư rác

Trong khuôn khổ nghiên cứu này, nhóm tác giả quy ước các dạng thư điện tử dưới đây sẽ bị coi là thư rác(spam):

- Thư điện tử với mục đích lừa đảo, quấy rối hoặc phát tán virus máy tính, phần mềm gây hại hoặc vi phạm khoản 2 Điều 12 Luật Công nghệ thông tin.
- Thư điện tử quảng cáo quảng cáo vi phạm các nguyên tắc gửi thư điện tử quảng cáo quảng cáo tại các Điều 7, Điều 9 và Điều 13 Nghị định 90/2008/NĐ-CP.
- Thư điện tử có nội dung xuyên tạc đường lối chủ trương của Đảng, tuyên truyền chống phá các hoạt động của Nhà nước CHXHCN Việt Nam

Thư rác gây phiền toái cho người dùng thông thường vì họ phải mất thời gian đọc và xóa chúng, nhưng đối với doanh nghiệp thì đó là tiền bạc. Chẳng hạn, thư rác sẽ lấy mất không gian đĩa giá trị trên các máy chủ dành cho dữ liệu và thông điệp có ích, giảm khả năng của mạng và tốc độ của các ứng dụng mạng do phải tải những e-mail vô bổ. Thư rác là một trong những vấn đề đau đầu đối với người dùng máy tính hiện nay.

Thư rác làm hao phí rất nhiều tiền bạc của các ISP cũng như người dùng. Phí tổn này gồm những khoản sau:

- Tốn nhiều băng thông hơn vì nhiều thư hơn, điều này cũng có nghĩa là chi phí cao hơn cho người dùng.

- Các thư rác có chứa virus, mã độc có thể làm sập toàn bộ hệ thống, đánh cắp những thông tin nhạy cảm, dữ liệu quan trọng.
- Nhà cung cấp dịch vụ mạng phải tốn nhiều kinh phí để lưu trữ rất nhiều thư rác.
- Chi phí mua, phát triển, bảo trì các phần mềm lọc thư.
- Chi phí cho nhân lực, vật lực để đối phó với vấn nạn thư rác.
- Thời gian kết nối lâu hơn đối với người dùng trên những mạng bị phát tán thư rác.
- Lãng phí thời gian làm việc vì nhiều nhân viên phải mất thời gian đọc và xóa những thông điệp vô bổ này.

3.2.2. Mục tiêu

Xây dựng một hệ thống Mail Gateway có khả năng kiểm soát được toàn bộ email trước khi chúng đi tới các máy chủ mail và người dùng cuối. Dựa vào các tiêu chí đánh giá, Mail Gateway sẽ phân loại các email trung chuyển qua nó. Các email vô hại sẽ được phép đi qua Mail Gateway, tới các máy chủ mail. Còn các email bị nghi ngờ là thư rác với mục đích xấu (quảng cáo, lừa đảo) hoặc phát tán virus, mã độc hoặc có nội dung không phù hợp (nội dung đồi trụy, nội dung chống phá Nhà nước) sẽ bị xử lý tùy theo mức độ vi phạm, tùy theo cấu hình của người quản trị. Cụ thể, nếu email vượt quá ngưỡng quy định sẽ bị loại bỏ ngay tại Mail Gateway, không được gửi tới máy chủ mail. Các mail có mức vi phạm nhẹ hơn sẽ được đánh dấu, cảnh báo đối với người dùng cuối. Dựa vào những cảnh báo này, từ phía mail client sẽ có mức hành xử phù hợp như chuyển các email bị nghi ngờ vào Hộp thư rác (SPAM BOX), không cho các email này đi vào INBOX (Hộp thư đến).

3.2.3. Yêu cầu đối với hệ thống

3.2.3.1. Yêu cầu chức năng

- Lọc theo địa chỉ người gửi

- Địa chỉ IP
- Một dải các địa chỉ IP
- Địa chỉ email

- Lọc theo tên miền
- Lọc các trường ở trong header của thư điện tử
- Lọc nội dung thư điện tử
  - Nội dung tiếng Việt
  - Nội dung tiếng Anh
  - Nội dung phức hợp có cả tiếng Việt và tiếng Anh
- Lọc các tệp tin đính kèm
  - Lọc theo phần mở rộng của tệp tin đính kèm theo thư điện tử. Cấm một số loại tệp tin mà phần mở rộng nằm trong danh sách nghi ngờ có chứa virus như exe, bat, reg, com, scr…
  - Phát hiện virus, mã độc nếu có trong các tệp tin đính kèm ở trong thư điện tử
  - Hỗ trợ lọc nhiều định dạng tệp tin phổ biến như doc, xls, txt, pdf…
  - Hỗ trợ quét virus các tập tin đính kèm ở dạng bị nén (tệp tin có phần mở rộng zip, rar)
- Tùy từng trường hợp cụ thể, hệ thống cần hỗ trợ cách thức hành xử thích hợp với các thư điện tử:
  - Đối với thư điện tử bình thường, không vi phạm chính sách nào sẽ được Mail Gateway chuyển giao tới các máy chủ mail, người dùng cuối.
  - Đối với thư điện tử được xác định là đã vi phạm thì hệ thống sẽ xử lý tùy theo mức độ:
    - Nếu mức độ vi phạm cao, cần thiết phải chặn lại thì Mail Gateway sẽ tiến hành loại bỏ email đó mà không

chuyển giao tới máy chủ mail, người dùng cuối phía sau. Bên gửi sẽ nhận được thông báo từ chối chuyển giao email từ Mail Gateway.

- Nếu mức độ vi phạm vừa phải, thư điện tử được xếp vào diện bị nghi ngờ và có cách thức cảnh báo cho người dùng cuối được biết.

- Chức năng lưu trữ thư điện tử:

Đối với các thư điện tử thuộc diện bị nghi ngờ hoặc bị phát hiện vi phạm các chính sách, quy định, hệ thống sẽ phải lưu trữ lại các thư điện tử đó và có hình thức thông báo tới người quản trị. Các thông tin quan trọng này sẽ phục vụ cho công tác điều tra sau này

3.2.3.2. Yêu cầu phi chức năng

Yêu cầu về giao diện

Hệ thống cần hỗ trợ giao diện quản trị dạng web-based. Bố cục giao diện hợp lý giúp cho các thao tác tra cứu, chỉnh sửa dữ liệu nhanh và thuận tiện, dễ sử dụng.

Với từng thành phần trong hệ thống, cần hỗ trợ giao diện tương tác trực tiếp từ dòng lệnh (command-line) để việc cấu hình, tối ưu hệ thống được dễ dàng hơn.

Yêu cầu về trao đổi và tích hợp

Hệ thống phải đảm bảo tính mở, có thể tích hợp các mô-đun mới vào dễ dàng, không gây ảnh hưởng tới hoạt động chung. Việc thay đổi, nâng cấp phiên bản, cập nhật cơ sở dữ liệu cũng như các phần mềm tiện ích trong toàn bộ hệ thống phải được hỗ trợ tối đa, đảm bảo tính tương thích giữa các thành phần trong hệ thống.

Yêu cầu về hiệu năng

Hệ thống phải đảm bảo khả năng phục vụ một số lượng lớn người sử dụng.

Yêu cầu về tính bảo mật

Để sử dụng các dịch vụ trong hệ thống, người sử dụng cần được xác thực và cung cấp các quyền tương ứng. Việc xác thực thông qua tên đăng nhập và mật khẩu.

Yêu cầu về tính mở

Các thành phần trong hệ thống cần được xây dựng từ mã nguồn. Cụ thể là các phần mềm tiện ích đều phải có mã nguồn kèm theo, nên hạn chế sử dụng các phần mềm đã đóng gói. Tính mở ở đây có ý nghĩa đặc biệt quan trọng. Khi đã có mã nguồn, người quản trị hoàn toàn có thể thay đổi, tối ưu hệ thống dựa vào các tham số đã được cung cấp một cách dễ dàng. Với mã nguồn của chương trình, chúng ta hoàn toàn có thể kiểm soát, can thiệp vào được mọi hoạt động của nó, đảm bảo không có những nguy cơ, những mối đe dọa về an ninh, bảo mật như ở phần mềm đã đóng gói.

3.2.4. Các phương pháp lọc SPAM

Spam gây ra rất nhiều tác hại, do vậy việc phòng chống và ngăn chặn các thư rác là một việc rất cần thiết, cấp bách. Hiện có nhiều công ty phần mềm cung cấp giải pháp chống spam, mỗi dòng sản phẩm có những tính năng và các ưu nhược điểm riêng, tuy nhiên để xây dựng một hệ thống Mail Gateway cần một giải pháp có tính tổng thể, linh hoạt có thể áp dụng linh hoạt cho những hoàn cảnh, điều kiện và môi trường khác nhau. Trên cơ sở nghiên cứu của đề tài này, nhóm nghiên cứu đề xuất một giải pháp xây dựng hệ thống Mail Gateway phòng thủ nhiều lớp. Hệ thống sẽ bao gồm nhiều tầng, nhiều lớp lọc khác nhau. Mỗi lớp sẽ đảm nhiệm lọc một hoặc một số thông tin cụ

thể trong thư điện tử khi đi qua Mail Gateway. Việc phòng thủ nhiều lớp có ý nghĩa đặc biệt quan trọng bởi vì:

- Hệ thống có thể phát hiện sớm các thư điện tử có biểu hiện đáng ngờ và có hình thức xử lý sớm. Ví dụ nếu như phát hiện địa chỉ IP của bên gửi có dấu hiệu phát tán thư rác, hệ thống có thể ngắt luôn kết nối, không cho phép khởi tạo các kết nối SMTP, tiết kiệm được thời gian, băng thông, tốc độ xử lý.
- Hệ thống được tách thành nhiều lớp phòng thủ khác nhau. Tại mỗi lớp sẽ có những phần mềm tiện ích đảm nhiệm nhiệm vụ lọc thư điện tử. Do có tính linh hoạt nên tùy vào môi trường, phạm vi bài toán cụ thể mà người quản trị sẽ quyết định là sử dụng những lớp phòng thủ và công cụ nào phù hợp, đảm bảo tính an toàn, ổn định và hiệu suất hoạt động của hệ thống. Các phần mềm tiện ích ở từng lớp phòng thủ tuy hoạt động riêng biệt nhưng có phối hợp với nhau chặt chẽ. Việc nâng cấp, thay đổi hay gỡ bỏ các thành phần này hoàn toàn dễ dàng, không gây ảnh hưởng tới hoạt động của toàn bộ hệ thống.

3.2.4.1. Phương pháp lọc theo IP (Spamhaus DNS-based Blocklists)

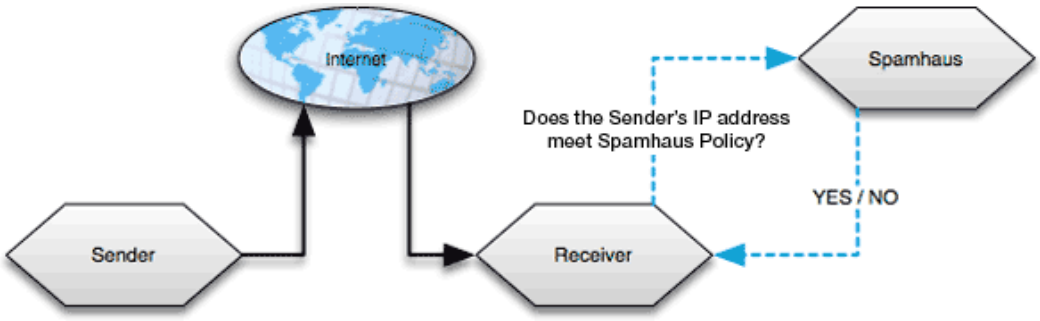
Phương pháp sử dụng DNS Blocklists sẽ chặn các email đến từ các địa chỉ nằm trong danh sách DNS Blocklist. Có hai loại danh sách DNS Blocklist thường được sử dụng, đó là:

Danh sách các tên miền gửi spam đã biết, danh sách các miền này được liệt kê và cập nhật tại địa chỉ http://spamhaus.org/sbl. Danh sách các máy chủ email cho phép hoặc bị lợi dụng thực hiện việc chuyển tiếp spam được gửi đi từ spammer.

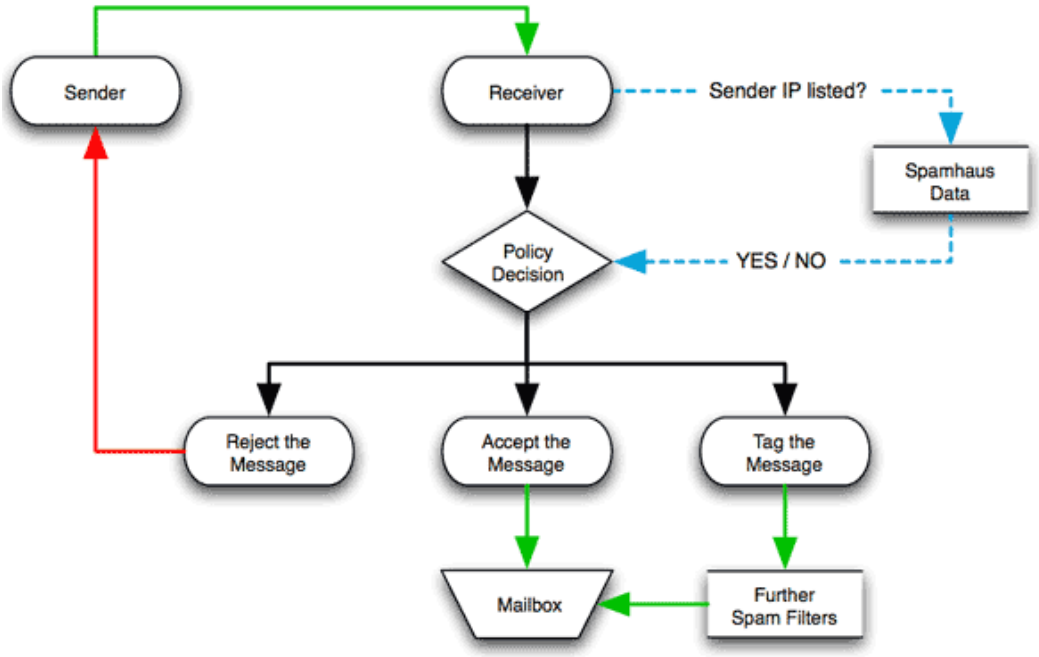
Những kẻ phát tán thư rác chuyên nghiệp thường sở hữu một dãy máy chủ chuyên để gửi thư rác, mỗi máy chủ có một địa chỉ IP riêng. Khi các địa chỉ IP này bị nghi ngờ là nguồn phát tán thư rác, chúng sẽ được báo cáo lên

các tổ chức chuyên theo dõi tình trạng phát tán thư rác trên Internet. Các tổ chức này sẽ tiến hành xác minh, thẩm định. Nếu đúng địa chỉ IP đó đã bị những kẻ phát tán thư rác lợi dụng thì nó sẽ được đưa vào danh sách DNS Blocklist. Spamhaus.org là một trong rất nhiều tổ chức cập nhật danh sách này. Những kẻ phát tán có thể vượt qua chướng ngại vật này bằng hai cách: Một là thường xuyên thay đổi địa chỉ IP. Cách làm này còn khiến những địa chỉ IP cũ được tái sử dụng, và người dùng sau không thể sử dụng được địa chỉ IP này bởi chúng đã bị khoá do gửi spam, và họ không thể dùng nó dể gửi email hợp pháp.

Các chương trình chống spam sẽ kiểm địa chỉ IP của SMTP server bên gửi trong phần header của thư điện tử đó sau đó so sánh với cơ sở dữ liệu DNS Blocklist đã biết. Nếu địa chỉ IP tìm thấy trong phần này có trong cơ sở dữ liệu về các DNS Blocklist, nó sẽ bị coi là thư rác, còn nếu không, email đó sẽ được coi là một email hợp lệ.



Hình 3. 20. Mô hình hoạt động của phương pháp lọc theo địa chỉ IP



Hình 3. 17. Cách thức ứng xử của bộ lọc

Quy trình xử lý thư điện tử khi sử dụng phương pháp lọc theo địa chỉ IP:

Bước 1: Khi thư điện tử đi tới Mail Gateway, đầu tiên bộ lọc sẽ lấy địa chỉ IP của máy chủ SMTP đã gửi thư điện tử này đến và gửi yêu cầu tới cơ sở dữ liệu DNS Blocklist của tổ chức Spamhaus.

Bước 2: Khi Spamhaus trả về kết quả cho bộ quyết định là thư điện tử đó có phải là thư rác hay không, bộ quyết định sẽ lựa chọn một trong 3 cách xử lý:

- Từ chối thư điện tử này nếu như nó được xác định là thư rác
- Chấp nhận thư điện tử này và gửi tới các máy chủ mail ở sau Mail Gateway
- Đánh dấu thư điện tử này thuộc diện bị nghi ngờ và khi thư điện tử tới người dùng cuối, nó sẽ bị xếp vào Hộp thư rác (Spam Box).

Phương pháp này có ưu điểm là các thư điện tử có thể được kiểm tra trước khi tải xuống, do đó tiết kiệm được băng thông đường truyền. Nhược

điểm của phương pháp này là không phát hiện ra được những email giả mạo địa chỉ người gửi.

3.2.4.2. Phương pháp lọc các liên kết trong nội dung thư diện tử sử dụng danh sách SURBL

Phương pháp sử dụng danh sách SURBL phát hiện thư rác dựa vào nội dung của thư điện tử. Chương trình chống thư rác sẽ phân tích nội dung của thư điện tử xem bên trong nó có chứa các liên kết đã được liệt kê trong Spam URI Realtime Blocklists (SURBL) hay không. SURBL chứa danh sách các miền và địa chỉ của những nguồn phát tán thư rác đã biết. Cơ sở dữ liệu này được cung cấp và cập nhật thường xuyên tại địa chỉ www.surbl.org.

Có nhiều danh sách SURBL khác nhau như sc.surbl.org, ws.surbl.org, ob.surbl.org, ab.surbl.org..., các danh sách này được cập nhật từ nhiều nguồn. Thông thường, người quản trị thường kết hợp các danh sách SURBL bằng cách tham chiếu tới địa chỉ multi.surbl.org. Nếu một thư điện tử sau khi kiểm tra nội dung có chứa các liên kết được chỉ ra trong danh sách SURBL thì nó sẽ được đánh dấu là thư rác, còn không nó sẽ được cho là một thư điện tử thông thường.

Phương pháp này có ưu điểm phát hiện được các thư điện tử giả mạo địa chỉ người gửi để đánh lừa các bộ lọc. Nhược điểm của nó là thư điện tử phải được tải xuống trước khi tiến hành kiểm tra, do đó sẽ chiếm băng thông đường truyền và tài nguyên của máy tính để phân tích các nội dung thư điện tử.

3.2.4.3. Phương pháp lọc theo địa chỉ người nhận(receiver*)*

Tấn công phát tán thư rác kiểu "từ điển" sử dụng các địa chỉ email và tên miền đã biết để tạo ra các địa chỉ email hợp lệ khác. Bằng kỹ thuật này những kẻ phát tán thư rác có thể gửi spam tới các địa chỉ email được sinh ra một cách ngẫu nhiên. Một số địa chỉ email trong số đó có thực, tuy nhiên một

lượng lớn trong đó là địa chỉ không tồn tại và chúng gây ra hiện tượng "lụt" (flood) ở các máy chủ mail.

Phương pháp kiểm tra người nhận sẽ ngăn chặn kiểu tấn công này bằng cách chặn lại các email gửi tới các địa chỉ không tồn tại trên Active Directory hoặc trên máy chủ mail nằm phía sau Mail Gateway. Tính năng này sẽ sử dụng Active Directory hoặc LDAP server để xác minh các địa chỉ người nhận có tồn tại hay không. Nếu số địa chỉ người nhận không tồn tại vượt quá một ngưỡng nào đó (do người quản trị thiết lập) thì email gửi tới đó sẽ bị coi là spam và chặn lại.

3.2.4.4. Giới hạn số lần nhận thư từ mỗi địa chỉ IP

Những chương trình gửi thư rác sẽ gửi thư liên tiếp tới từng hộp thư của người dùng ở máy chủ mail. Có nhiều trường hợp các chương trình này còn gửi tới những hộp thư mà địa chỉ của nó không có thật trong máy chủ mail. Ở tại Mail Gateway, người quản trị có thể cấu hình cho phép máy chủ mail sẽ chỉ nhận số lượng thư hạn chế từ mỗi địa chỉ IP trong một khoảng thời gian nhất định, ví dụ chỉ nhận 1 thư trong vòng 5 phút. Khi một địa chỉ IP gửi liên tiếp nhiều thư đến và vượt giới hạn cho phép thì nó sẽ bị từ chối cho đến hết khoảng thời gian đã định rồi mới được gửi tiếp. Thông thường chỉ có những kẻ chuyên phát tán thư rác mới gửi liên tiếp dồn dập nhiều email vào hệ thống, người dùng bình thường hiếm khi gửi đi nhiều email liên tiếp như vậy. Việc đưa ra ngưỡng giới hạn số lần nhận, giới hạn thời gian nhận là một phương pháp đơn giản để nhận diện những kẻ phát tán thư rác với người dùng chân chính. Đối với các máy chủ mail lớn, thực hiện việc chuyển nhận nhiều thư liên tục, người quản trị sẽ cấu hình Mail Gateway đưa địa chỉ của các máy chủ mail này vào Whitelist cho phép chúng được phép giao tiếp thường xuyên, liên tục, với tốc độ cao không bị giới hạn bởi thời gian, số lượng email chuyển nhận.

Phương pháp lọc này giúp các máy chủ mail hạn chế được hình thức tấn công flood email. Với mô hình mạng có Mail Gateway đứng trước các máy chủ mail, những kẻ phát tán thư rác sẽ không thể gửi dồn dập một lúc hàng nghìn email vào hệ thống. Mail Gateway được cấu hình đúng sẽ chặn đứng hoàn toàn những đợt tấn công này, không làm ảnh hưởng tới hoạt động của các máy chủ mail phía sau nó.

3.2.4.5. Giới hạn dung lượng của mỗi thư điện tử

Hệ thống cần hỗ trợ chức năng cho phép người quản trị có thể giới hạn dung lượng từng email khi đi qua Mail Gateway. Chức năng này giúp phát hiện và hạn chế các email có dung lượng quá lớn, tránh gây tình trạng tràn hộp thư của người sử dụng. Chức năng này sẽ hữu ích trong trường hợp hệ thống bị tấn công bomb mail – kẻ tấn công cố tình gửi rất nhiều thư điện tử có dung lượng lớn tới các máy chủ mail cũng như là hộp thư của người dùng. Việc phát hiện sớm các cuộc tấn công dạng này và tổ chức ngăn chặn ngay tại Mail Gateway sẽ hạn chế được tối đa ảnh hưởng hoạt động của máy chủ mail cũng như những người sử dụng trong hệ thống.

3.2.4.6. Phương pháp lọc sử dụng kiểm tra địa chỉ

Bằng cách kiểm tra địa chỉ người gửi và người nhận, phần lớn spam sẽ được phát hiện và chặn lại. Thực hiện kiểm tra địa chỉ người gửi trước khi email được tải xuống sẽ tiết kiệm được băng thông đường truyền cho toàn hệ thống.

Kỹ thuật Sender Policy Framework (SPF, www.openspf.org) được sử dụng để kiểm tra địa chỉ người gửi email. Kỹ thuật SPF cho phép chủ sở hữu của một tên miền Internet sử dụng các bản ghi DNS đặc biệt (gọi là bản ghi SPF) chỉ rõ các máy được dùng để gửi email từ miền của họ. Khi một email được gửi tới, bộ lọc SPF sẽ phân tích các thông tin trong trường “From” hoặc “Sender” để kiểm tra địa chỉ người gửi. Sau đó SPF sẽ đối chiếu địa chỉ đó

với các thông tin đã được công bố trong bản ghi SPF của miền đó xem máy gửi email có được phép gửi email hay không. Nếu email đến từ một server không có trong bản ghi SPF mà miền đó đã công bố thì email đó bị coi là giả mạo.

3.2.4.7. Phương pháp lọc theo IP, một dải IP

Phương pháp này sẽ chặn các email được gửi đến từ các địa chỉ IP biết trước. Khi một email đến, bộ lọc sẽ phân tích địa chỉ máy gửi và so sánh với danh sách địa chỉ bị chặn. Nếu email đó đến từ một máy có địa chỉ trong danh sách này thì nó sẽ bị coi là spam, ngược lại nó sẽ được coi là email hợp lệ. Tương tự, người quản trị hệ thống có thể đưa vào danh sách cấm một dải các địa chỉ IP (ví dụ: 203.162.x.x). Tất cả các email đến từ IP bắt đầu bằng 203.162 đều sẽ bị Mail Gateway chặn lại

3.2.4.8. Phương pháp lọc theo tên miền

Khi đã biết rõ một số tên miền độc hại, chuyên phát tán thư rác vào hệ thống, người quản trị có thể tạo ra một cơ sở dữ liệu bao gồm các tên miền bị cấm, không cho phép gởi email tới các máy chủ mail trong hệ thống.

3.2.4.9. Phương pháp lọc sử dụng bộ lọc Bayesian

Bộ lọc Bayesian hoạt động dựa trên định lý Bayes để tính toán xác suất xảy ra một sự kiện dựa vào những sự kiện xảy ra trước đó. Kỹ thuật tương tự như vậy được sử dụng để phân loại spam. Nếu một số phần văn bản xuất hiện thường xuyên trong các spam nhưng thường không xuất hiện trong các email thông thường, thì có thể kết luận rằng email đó là spam.

Trước khi có thể lọc email bằng bộ lọc Bayesian, người dùng cần tạo ra cơ sở dữ liệu từ khóa và dấu hiệu (như là ký hiệu $, địa chỉ IP và các miền...) sưu tầm từ các spam và các email không hợp lệ khác.

Mỗi từ hoặc mỗi dấu hiệu sẽ được cho một giá trị xác suất xuất hiện, giá trị này dựa trên việc tính toán có bao nhiêu từ thường hay sử dụng trong

spam, mà trong các email hợp lệ thường không sử dụng. Việc tính toán này được thực hiện bằng cách phân tích những email gửi đi của người dùng và phân tích các kiểu spam đã biết.

Để bộ lọc Bayesian hoạt động chính xác và có hiệu quả cao, cần phải tạo ra cơ sở dữ liệu về các email thông thường và spam phù hợp với đặc thù kinh doanh của từng công ty. Cơ sở dữ liệu này được hình thành khi bộ lọc trải qua giai đoạn "huấn luyện". Người quản trị phải cung cấp khoảng 1000 email thông thường và 1000 spam để bộ lọc phân tích tạo ra cơ sở dữ liệu cho riêng nó.

3.2.4.10. Phương pháp lọc sử dụng danh sách trắng / danh sách đen (Blacklist/Whitelist)

Việc sử dụng các danh sách blacklist, whitelist giúp cho việc lọc email hiệu quả hơn.

Blacklist(danh sách đen) là cơ sở dữ liệu các địa chỉ email và các tên miền mà Mail Gateway sẽ không bao giờ cho các email từ đó đi qua. Các email gửi tới từ các địa chỉ này sẽ bị đánh dấu là spam hoặc từ chối chuyển/nhận tùy theo yêu cầu cấu hình của người quản trị. Ví dụ: trong quá trình vận hành hệ thống, người quản trị nhận thấy một số cuộc tấn công hoặc các hành vi đáng ngờ xuất phát từ một địa chỉ IP nào đó. Khi đó, người quản trị sẽ đưa địa chỉ IP này vào danh sách đen, như vậy mỗi lần nhận được yêu cầu từ địa chỉ IP đó, hệ thống sẽ không đáp ứng lại yêu cầu, đảm bảo an toàn không để kẻ tấn công có thể gây hại tới hệ thống cũng như người sử dụng.

Whitelist(danh sách trắng) là cơ sở dữ liệu các địa chỉ email và các tên miền mà Mail Gateway cho phép các email từ trong danh sách này hoặc tới từ các tên miền này đi qua. Nếu các email được gửi đến từ những địa chỉ nằm trong danh sách này thì chúng luôn được cho qua mà không cần phải đi qua quá trình kiểm tra sàng lọc nào. Danh sách trắng được sử dụng trong trường hợp người quản trị muốn thiết lập một địa chỉ thư điện tử hoặc một danh sách

các địa chỉ thư điện đã được tin tưởng (ví dụ của những người quản lý, điều hành, những thành phần chủ chốt của công ty). Khi đó tất cả các thư điện tử của những người này gửi đi đều được mặc định cho đi qua Mail Gateway mà không bị giữ lại kiểm tra. Việc này đảm bảo tất cả các thư điện tử của những người quan trọng khi gửi đi đều tới được người nhận, không bị bỏ sót, không bị lọc nhầm các thông điệp quan trọng.

Thông thường các bộ lọc có tính năng tự học, khi một email bị đánh dấu là spam thì địa chỉ người gửi sẽ được tự động đưa vào danh sách blacklist. Ngược lại, khi một email được gửi đi từ trong công ty thì địa chỉ người nhận sẽ được tự động đưa vào danh sách whitelist.

3.2.4.11. Phương pháp lọc dựa vào các thông tin của mail header

Phương pháp này sẽ phân tích các trường trong phần header của email để đánh giá email đó là email thông thường hay là spam. Spam thường có một số đặc điểm như:

- Để trống trường From: hoặc trường To: .
- Trường From: chứa địa chỉ email không tuân theo các chuẩn RFC.
- Các URL trong phần header và nội dung của thư điện tử có chứa địa chỉ IP được mã hóa dưới dạng hệ hex/oct hoặc có sự kết hợp theo dạng username/password
- Phần tiêu đề của thư điện tử có thể chứa địa chỉ email người nhận để cá nhân hóa email đó.
- Gửi tới một số lượng rất lớn người nhận khác nhau: Các email thông thường khi gửi đi đều chỉ có một số lượng người nhận giới hạn. Một số loại thư rác thường được gửi tới một số lượng lớn người nhận, do vậy hệ thống có thể cho phép cấu hình đặt ngưỡng số lượng người nhận của email để đảm bảo các email này sẽ bị chặn lại.

- Chỉ chứa những file ảnh mà không chứa các từ để đánh lừa các bộ lọc.

Sử dụng ngôn ngữ khác với ngôn ngữ mà người nhận đang sử dụng.

Dựa vào những đặc điểm này của spam, các bộ lọc có thể lọc chặn được các email vi phạm quy định.

3.2.4.12. Phương pháp lọc theo nội dung

Người quản trị sẽ xây dựng cơ sở dữ liệu các luật lọc, đưa vào các từ khóa nhạy cảm đặc trưng thường xuất hiện trong thư rác và chỉ số tương ứng cho từ khóa đó. Khi email đi qua bộ lọc nội dung, bộ lọc sẽ tiến hành phân tích, tính tổng chỉ số của toàn bộ email đó và đưa ra quyết định là email này có vi phạm ngưỡng SPAM cho phép hay không. Nếu email bị nghi ngờ là SPAM, bộ lọc sẽ thêm một số trường vào mail header để cung cấp thêm thông tin cho người dùng cuối ví dụ: đưa ra cảnh báo đây là thư spam, nó đã vi phạm những luật lọc gì của hệ thống. Đồng thời, bộ lọc sẽ viết lại tiêu đề của email, cảnh báo người dùng đây là thư rác, khuyến cáo không nên mở ra.

3.2.4.13. Phương pháp lọc dựa vào các tệp tin đính kèm

Các hệ thống lọc email cho phép người quản trị có thể cấm một số loại tệp tin đính kèm dựa trên phần mở rộng của chúng(extension). Ví dụ các loại tệp tin đính kèm có tiềm ẩn nguy cơ gây hại tới người dùng và hệ thống như: exe, scr, com, bat

Khi các email đi qua Mail Gateway, hệ thống sẽ thực hiện việc đối chiếu với cơ sở dữ liệu các phần mở rộng bị cấm, nếu tệp tin đính kèm không hợp lệ, email đó sẽ bị loại bỏ.

3.2.4.14. Quét virus

Hệ thống lọc mail sẽ được tích hợp với chương trình AntiVirus, tất cả các email khi được trung chuyển qua Mail Gateway đều được kiểm tra xem có chứa virus hoặc mã độc gây hại tới người sử dụng hoặc hệ thống hay

không. Chương trình AntiVirus cần hỗ trợ việc nhận dạng được các loại virus phổ biến, cơ sở dữ liệu nhận dạng các mẫu virus cần được cập nhật thường xuyên liên tục để đảm bảo hệ thống luôn được bảo vệ ở mức cao nhất, phòng chống được những loại virus, mã độc mới nhất.

3.2.5. Giải pháp sử dụng QMAIL làm MAIL GATEWAY

3.2.5.1. Giới thiệu

Qmail Mail Gateway là một giải pháp toàn diện cung cấp cơ chế bảo vệ hệ thống máy chủ mail và người dùng cuối. Việc sử dụng Qmail Mail Gateway là rất linh hoạt, chúng ta có thể cài đặt hệ thống này trên một máy chủ riêng biệt mà không cần yêu cầu phải được tích hợp vào một hệ thống máy chủ mail hiện có. Giải pháp hiệu quả này giúp tăng mức độ bảo vệ hệ thống máy chủ mail trước những mối đe dọa tiềm ẩn trên Internet như virus, thư rác, mã độc...

Qmail Mail Gateway được thiết kế để hoạt động hoàn toàn tự động, vì vậy ứng dụng có thể hoạt động tốt trên các môi trường các hệ điều hành Linux cũng như kết hợp được với những ứng dụng khác được cài đặt trên các nốt trong mạng. Việc cài đặt và cấu hình Qmail Mail Gateway là dễ dàng đối với những nhà quản trị đã có kinh nghiệm sử dụng hệ điều hành Linux.

Có hai điểm tối quan trọng của mô hình Qmail Mail Gateway là:

Tính ổn định và hiệu năng cao

Qmail có khả năng chuyển nhận hàng triệu thông điệp một ngày. Tuy nhiên, nếu biến nó thành một máy chủ mail phục vụ tất cả các giao thức mail (IMAP, POP3)thì bị giới hạn trong vấn đề xác thực. Nếu các máy chủ mail thuộc một mạng dùng hoàn toàn Unix thì giới hạn này có thể khắc phục dễ dàng. Trong một hệ thống không phải tất cả các máy chủ đều dùng hệ điều hành giống nhau.. Vì thế, Qmail hoạt động như một mail gateway chỉ có trách nhiệm chuyển mail đến các máy chủ mail khác (trong giới hạn cho phép

domain) mà không phải lo vấn đề xác thực, vì vậy ngoài tính bảo mật cao thì hiệu năng được nâng cao đáng kể.

Tính bảo mật

Như đã nêu ra ở trên những khó khăn trong cơ chế quản lý xác thực của một (hoặc nhiều) mạng có đa hệ điều hành không những giảm sút tính hiệu năng mà còn ảnh hưởng lớn đến tính bảo mật. Lý do, quản lý một trung tâm dữ liệu người sử dụng (central user database) dễ dàng và ổn định hơn so với việc có nhiều nhiều cơ sở dữ liệu người sử dụng (user database). Hơn thế, cơ chế Internet <--> Mail Gateway <--> Tường lửa(Firewall) <--> Các máy chủ mail phía trong <--> Người dùng cuối chặc chẽ và an toàn hơn. Đó là chưa kể ứng dụng kiểm soát / ngăn chặn virus, trojan và cách loại mã độc mang tính phá hoại trên Mail Gateway trước khi thông điệp được chuyển vào các máy chủ mail bên trong (Microsoft Exchange hoặc IBM Lotus chẳng hạn). Các ứng dụng cho POP3 hoặc IMAP được thiết lập một cách độc lập trên các máy chủ mail ở phía sau mail gateway.

3.2.5.2. Chức năng chính của hệ thống Qmail Mail Gateway

Chống virus

Hệ thống sẽ thực hiện thao tác quén và loại bỏ các loại virus, mã độc và những chương trình có tiềm ẩn nguy cơ đe dọa tới sự an toàn của người sử dụng. Các thông tin được kiểm duyệt bao gồm tất cả các thành phần của các thông điệp được gửi tới máy chủ mail hoặc từ người dùng ra hệ thống bên ngoài: header, các tệp tin đính kèm theo thông điệp được gửi đi.

Lọc thư rác

Hệ thống sẽ thực hiện quá trình quét và phân tích các thông điệp đi qua Qmail Mail Gateway để phát hiện những thông điệp vi phạm các quy định do người quản trị đặt ra. Các thông tin mà Qmail Mail Gateway quan tâm bao gồm các trường trong header của thông điệp(người gửi, địa chỉ IP, tên miền),

tiêu đề cũng như nội dung chính của thông điệp đó và tất cả các tệp tin đính kèm đi cùng. Với mỗi vi phạm được phát hiện, hệ thống sẽ đánh chỉ số (score) tương ứng dựa trên cơ sở dữ liệu các chỉ số đã được thiết lập. Sau khi tính toán chỉ số cho từng thông điệp, Qmail Mail Gateway sẽ so sánh với ngưỡng đã được quy định để ra quyết định đối với thông điệp này:

- Nếu thông điệp là hoàn toàn hợp lệ, nó sẽ được chuyển tiếp tới máy chủ mail tương ứng.
- Nếu thông điệp bị nghi ngờ là thư rác, nó sẽ được đánh dấu là SPAM, thông điệp được chuyển tới máy chủ mail và bị di chuyển vào thư mục SPAM.
- Nếu thông điệp có chứa virus, mã độc hoặc chứa các tệp tin đính kèm không được phép trao đổi, hệ thống sẽ tự động loại bỏ thông điệp này và gửi thông báo về cho bên gửi đi.

Cảnh báo người sử dụng

Đối với các thông điệp bị nghi ngờ là thư rác, hệ thống sẽ thực hiện việc cách li (quarantine) bằng cách di chuyển các thông điệp này vào một thư mục riêng. Tuy nhiên để đảm bảo không xóa nhầm các email quan trọng, việc chuyển các email này về cho người sử dụng đầu cuối vẫn diễn ra, điểm khác biệt ở đây là các thông điệp này đã được Qmail Mail Gateway điều chỉnh lại phần tiêu đề(subject). Tiêu đề của các thông điệp này sẽ được thêm vào một dòng cảnh báo SPAM và ở mail client của người dùng cuối, các thông điệp đó sẽ được chuyển trực tiếp vào thư mục SPAM chứ không phải là INBOX.

Cách li các email nghi ngờ

Các thông điệp có chứa virus hoặc bị nghi ngờ là thư rác sẽ được chuyển vào một thư mục cách li riêng. Hệ thống cũng cho phép người quản trị có thể xem, xóa các thông điệp này hoặc tiếp tục chuyển thông điệp đó tới người dùng cuối

Lưu trữ email

Các chức năng lọc nâng cao

- Lọc theo địa chỉ IP
- Lọc theo một dải địa chỉ IP
- Lọc theo tên miền
- Lọc theo định dạng tệp tin đính kèm

Hệ thống cho phép cấu hình lọc các tệp tin đính kèm đi theo thông điệp dựa trên tên tệp tin, phần mở rộng của tập tin đó, chức năng này hỗ trợ việc phát hiện nhanh các đối tượng có khả năng bao gồm virus, mã độc gây hại cho người sử dụng.

Lọc theo nhóm người sử dụng

Hệ thống có cơ chế hỗ trợ người quản trị có thể định nghĩa cách thức ứng xử đối với những thông điệp gửi tới hoặc gửi đi từng nhóm người dùng hoặc từng người dùng cụ thể.

Quản trị linh hoạt, dễ dàng

Qmail Mail Gateway cho phép thực hiện quản trị từ xa thông qua giao diện web-based hoặc giao diện dòng lệnh (command-line) để làm việc trực tiếp với các tệp tin cấu hình của từng thành phần trong hệ thống. Tuy nhiên,

Cấu hình, tối ưu hóa hệ thống

Tùy vào đặc điểm của hệ thống như lưu lượng email chuyển/nhận hàng ngày, cơ sở hạ tầng mạng, chính sách quy định của tổ chức mà người quản trị có thể cấu hình hệ thống một cách phù hợp, linh hoạt. Các tệp tin cấu hình của Qmail được lưu tại thư mục */var/qmail/control*.

Có một số thông số mà người quản trị quan tâm lúc cấu hình, tối ưu hóa hệ thống như số lượng email xử lý cùng một lúc, kích thước tệp tin log của Qmail, số lượng các bản ghi lưu trữ trong tệp tin log, thời gian timeout khi gửi mail…

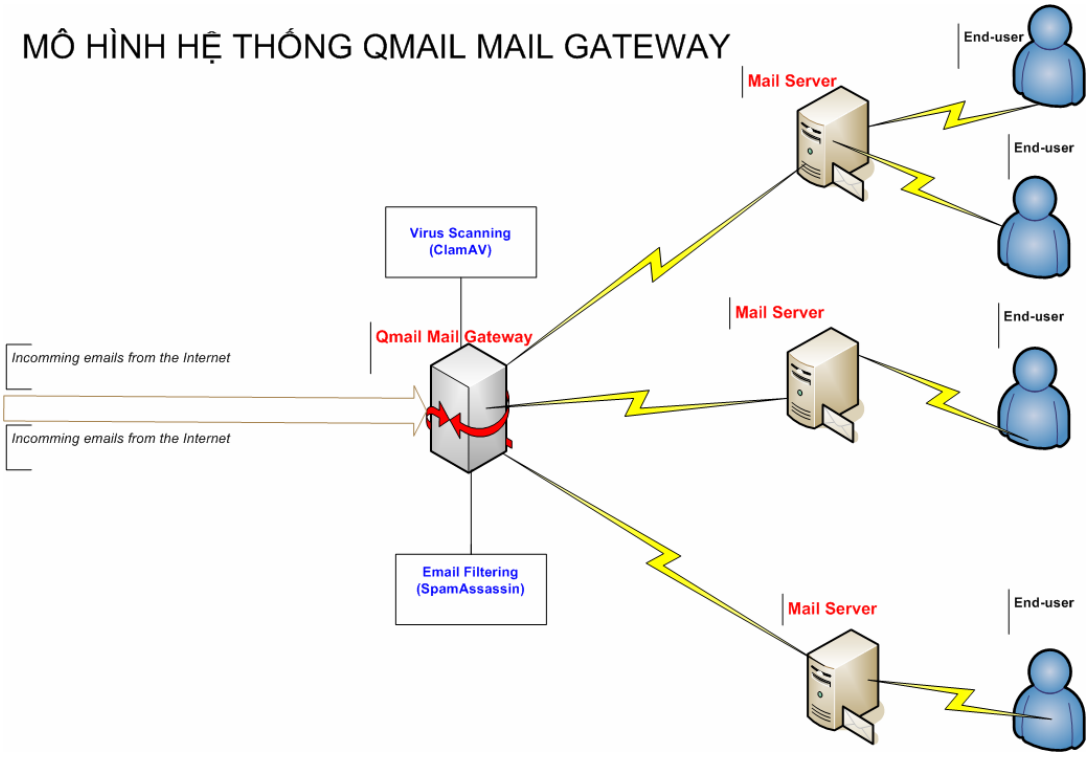
Cơ chế cập nhật

Qmail Mail Gateway hỗ trợ việc cập nhật cơ sở dữ liệu luật (rule database) và cơ sở dữ liệu chống virus (antivirus database). Người quản trị có thể cấu hình thực hiện cập nhật cơ sở dữ liệu tự động theo lịch trình hoặc có thể cập nhật bất kì thời điểm nào bằng cách sử dụng các lệnh cho trước.

Báo cáo dưới dạng biểu đồ

Hệ thống hỗ trợ khả năng xem báo cáo dưới dạng biểu đồ. Người quản trị có thể lựa chọn một khoảng thời gian cần theo dõi, hệ thống sẽ thực hiện thống kê số lượng email đã nhận về, số lượng email bị nghi ngờ là spam/virus, số lượng email bình thường, các thông số này sẽ được thể hiện trên biểu đồ.

3.2.5.3. Mô hình kiến trúc hệ thống Qmail Mail Gateway



Ở chiều nhận đến, thư điện tử được gửi từ Internet thay vì trực tiếp kết nối đến máy chủ xử lý mail của các tổ chức, doanh nghiệp sẽ được cấu hình

để đi qua trung tâm lọc thư điện tử Mail Gateway. Tại đây, việc nhận diện, phân loại các kết nối SMTP cũng như phân loại và rà quét virus trong các thư điện tử và các tệp tin đính kèm nhận được sẽ được thực hiện một cách kỹ lưỡng, nhanh chóng và cực kỳ chính xác.

Một kết nối giữa máy chủ muốn gửi mail với hệ thống Mail Gateway sẽ được thiết lập trước khi có sự chuyển/nhận thư điện thực sự. Mail Gateway sẽ từ chối tất cả các yêu cầu kết nối từ các máy chủ phát tán thư rác hoặc các máy chủ đã được nhận diện nằm trong danh sách đen(Blacklist).

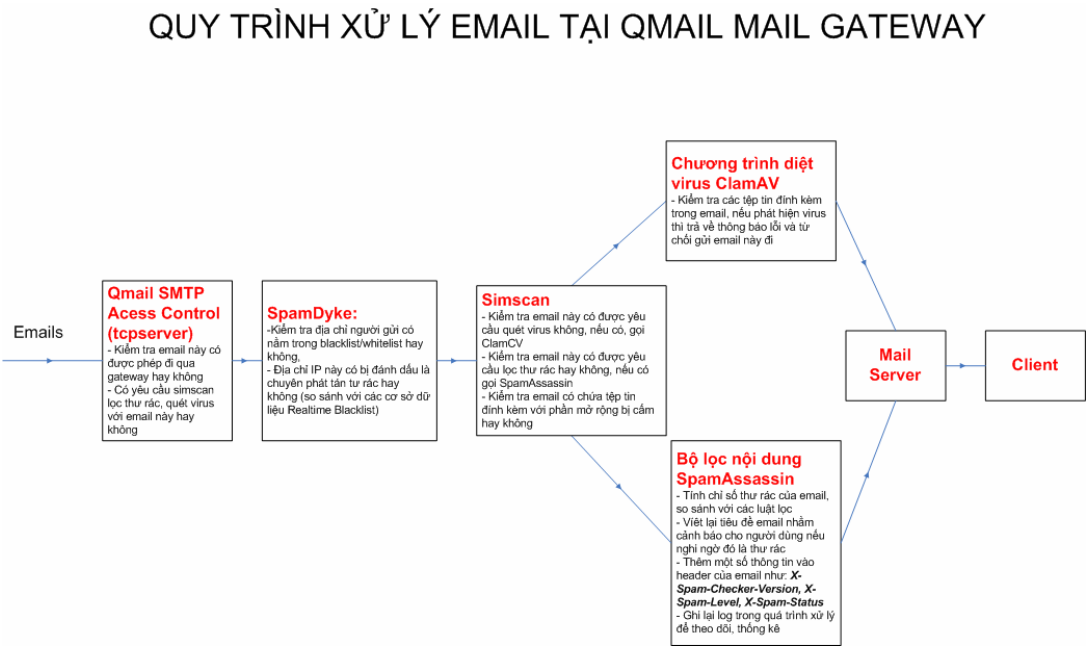
Các thư điện tử khi đi vào hệ thống Mail Gateway sau đó sẽ được kiểm tra để một lần nữa để nhận điện các thư điện tử giả mạo, thư rác, thư nhiễm virus hoặc thư mang nội dung không phù hợp.Chỉ những thư điện tử sạch mới được chuyển đến máy chủ xử lý mail của khách hàng. Những thư điện tử nếu bị nghi ngờ sẽ bị đánh dấu và có hình thức thông báo về cho người dùng cuối để đảm bảo các thư điện tử đó sẽ đi vào Hộp thư rác chứ không phải là Hộp thư đến của họ. Số email được xác định là thư rác, nhiễm virus, có nội dung không phù hợp, hoặc mang tính chất lừa đảo sẽ được cô lập có thời hạn. Và trong thời hạn quy định người dùng có thể tự mình truy nhập vào hệ thống Mail Gateway để kiểm tra, xóa bỏ hoặc lấy lại những email mình muốn đã bị cô lập trước đó. Chức năng lưu trữ đóng vai trò quan trọng cho công tác điều tra, kiểm chứng sau này.

Hệ thống cũng có khả năng gửi thông báo cho người nhận, người gửi thư điện tử khi có bất kỳ thư điện nào liên quan bị cô lập hoặc chậm trễ trong việc phân phát đến đích.

Ở chiều gửi ra, chỉ những thư điện tử gửi ra từ máy chủ mail phía sau Mail Gateway đã được cấu hình mới được phép đi qua hệ thống Mail Gateway. Chức năng này nhằm đảm bảo việc bảo vệ các máy chủ mail không bị những kẻ tấn công lợi dụng để phát tán thư rác, mã độc.

Các email gửi ra trước tiên sẽ đến hệ thống Mail Gateway để được kiểm sóat tính xác thực, đảm bảo không mang virus,…sau đó nếu thư điện tử không vi phạm các chính sách đã được đề ra chúng sẽ được nhanh chóng chuyển đến người nhận.

Trong mô hình này, Mail Gateway đóng vai trò như một bức tường lửa bảo vệ máy chủ mail trước các mối nguy, các hành động xâm nhập, tấn công trái phép từ bên ngoài, tránh các cuộc tấn công từ chối dịch vụ, loại bỏ các thư rác, thư nhiễm mã độc. Hệ thống Mail Gateway hoạt động tốt với tất cả các máy chủ mail SMTP/POP3/IMAP4/S. Việc triển khai cài đặt, cấu hình là hoàn toàn dễ dàng, không làm ảnh hưởng chung tới toàn bộ hoạt động hiện tại của hệ thống.



3.2.5.4. Các thành phần trong hệ thống Qmail Mail Gateway

1. Hệ điều hành CentOS 5.2

Trang chủ: http://centos.org

Phiên bản mới nhất: CentOS 5.3

CentOS (viết tắt của cụm từ Community ENTerprise Operating System) là môt phân phối của Linux dành cho các doanh nghiệp và các nhóm phát triển. CentOS đươc phát triển từ Red Hat Enterpires Linux (RHEL) (Theo các giấy phép GPL và một số giấy phép tương tự khác). Nó hoàn toàn miễn phí. Phiên bản hiện nay (2009) là phiên bản 5.3 (kernel 2.6.18), hỗ trợ với hai dòng chíp phổ biến là i386, x86-64.

CentOS lấy chỉ số version giống như RHEL. Ví dụ CentOS 5.1 sẽ tương ứng với RHEL5 Update 1.

Ưu điểm của hệ điều hành CentOS

- CentOS gần như giống hệt Redhat Enterprise Linux nên gói nào đã được biên dịch cho Redhat Enterprise Linux đều dùng được trên CentOS phiên bản tương ứng. Có thể nói là tương thích 100% cũng được.

Redhat Enterprise Linux được dùng một phiên bản cho các hệ thống Enterprise mà không mất tiền. Vì Redhat Enterprise Linux bắt phải mua kèm dịch vụ nên khá đắt.

- Chất lượng của Redhat Enterprise Linux được đảm bảo một cách không chính thức từ Red Hat.

- Tính bảo mật, độ ổn định cao.

2. Qmail

Qmail được viết bởi Tiến sĩ toán của trường đại học Illinois, Chicago, tiến sĩ Dan Bernstein. Qmail ra đời vào tháng Giêng năm 1996 với một phiên bản Beta 0.70 và sau đó phiên bản Gamma 0.90 được cập nhật vào tháng 8 năm 1996. Phiên bản ổn định 1.0 được ra mắt vào tháng 2 năm 1997. Phiên bản được lưu hành hiện nay là 1.03 được phát hành vào tháng 6 năm 1997.

Có rất nhiều MTA trên môi trường Unix hiện nay và mỗi khi nhắc đến MTA chúng ta phải nhắc đến Sendmail. Tuy nhiên, Dan Bernstein viết Qmail vì ông ta thấy rằng Sendmail thừa hưởng nhiều lỗi bảo mật từ các phiên bản trước đây và phần mềm này rất cồng kềnh, cho dù những năm gần đây, nhóm

Sendmail không ngừng điều chỉnh và cải tiến phần mềm này để giảm thiểu những yếu điểm. Khi viết Qmail, ngoài ưu tiên cho vấn đề bảo mật, Dan Bernstein chú trọng rất nhiều đến khả năng hoạt động và tính dễ dùng của nó. Qmail mang tính truyền thống của các hoạt trình Unix: mỗi tiểu ứng trình có khả năng đảm đương trọn vẹn một chức năng chuyên biệt và các tiểu ứng trình này có thể chuyền (pipe) sang các tiểu ứng trình khác để đáp ứng các quy trình phức tạp. Bởi thể, Qmail bao gồm nhiều gói đã biên dịch(binaries) tạo thành một dây chuyền hoạt động. Đây là một điển hình nặng tính bảo mật và tính hiệu năng trong cơ chế điều hành của một MTA.

3. QmailToaster

Trang chủ: http://Qmailtoaster.com/

QmailToaster được xây dựng như là một bộ hoàn chỉnh bao gồm các gói phần mềm đã được biên dịch sẵn. Tất cả các gói này được phân phối dưới dạng các gói RPM vì vậy việc lựa chọn các gói phù hợp cho từng hệ điều hành và kiến trúc đều rất dễ dàng. Để cài đặt Qmail-Toaster, người quản trị hệ thống chỉ cần tải về một số đoạn script đã được viết sẵn và chạy các đoạn script này, quá trình tải các gói phần mềm phù hợp và cài đặt sẽ diễn ra tự động. Người quản trị hệ thống không mất nhiều thời gian để cấu hình từng thành phần riêng lẻ, do vậy quá trình cài đặt Qmail-Toaster chỉ diễn ra trong khoảng một giờ đồng hồ đối với các máy chủ có đường truyền với tốc độ tốt. Khi quá trình cài đặt hoàn thành, chúng ta sẽ có một máy chủ mail Qmail hoàn chỉnh. Theo các con số thống kê, Qmail-Toaster hoạt động hiệu quả với số lượng lớn người sử dụng trong hệ thống.

4. Simscan

Trang chủ: http://sourceforge.net/projects/Simscan/

Simscan là một ứng dụng hỗ trợ dịch vụ Qmail smtpd trong việc quét virus, lọc thư rác và loại bỏ những email có chứa những tệp tin đính kèm độc

hại trong quá trình giao tiếp qua giao thức SMTP. Simscan là phần mềm mã nguồn mở, nhỏ gọn, hiệu quả được viết bằng ngôn ngữ C.

Các chức năng nổi bật của Simscan

* Hoạt động như là một chương trình độc lập chạy dưới quyền một người dùng thuộc nhóm khác với Qmail, cơ chế này giúp hạn chế những mối nguy về bảo mật. Trong trường hợp có sự cố xảy ra với Simscan, Qmail vẫn hoạt động bình thường.

* Hỗ trợ quét virus thông qua chương trình ClamAV. Tự động loại bỏ các email có chứa virus.

* Hỗ trợ SpamAssassin các phiên bản 3.0 và dòng phiên bản 2.6.

* Hỗ trợ DSpam

* Hỗ trợ lọc thư rác sử dụng chương trình SpamAssassin. Có thể cấu hình để loại bỏ các email mà cờ X-Spam-Flag được bật lên.

* Hỗ trợ ngưỡng loại bỏ email (hit level rejection). Các email sẽ bị loại bỏ nếu như chỉ số spam vượt ngưỡng cho phép của SpamAssassin. Các email khi được lọc qua Simscan sẽ được thay đổi phần tiêu đề tương ứng. Cụ thể là Simscan sẽ gọi tới SpamAssassin, SpamAssassin sẽ ghi thêm các thông tin vào phần tiêu đề của email đi qua nó. Một trường hợp cụ thể là Simscan sẽ ra lệnh viết lại phần tiêu đề của email nếu như nó phát hiện đây là thư rác. Việc này sẽ hỗ trợ cho quá trình phân loại thư ở phía người sử dụng.

* Hỗ trợ việc loại bỏ các tệp tin đính kèm có nguy cơ chứa virus, mã độc dựa trên phần mở rộng của tập tin đính kèm đó.

* Cho phép cài đặt cơ chế quét virus, lọc thư rác trên từng người dùng hoặc domain hoặc hệ thống cụ thể.

* Ghi lại log trong trường hợp phát hiện email có chứa virus hoặc thư rác

5. SpamAssassin

Trang chủ: http://spamassassin.apache.org/

Phiên bản mới nhất: **SpamAssassin 3.2.5**

SpamAssasin là một bộ lọc mail sử dụng nhiều cơ chế khác nhau để phát hiện thư rác bao gồm: phân tích văn bản(text analysis), phương pháp lọc Bayesian(Bayesian filtering), lọc theo DNS (DNS blocklists) hoặc dựa trên các cơ sở dữ liệu lọc cộng tác(collaborative filtering databases)

SpamAssassin là một dự án của tổ chức the Apache Software Foundation (ASF).

Một điểm rất quan trọng là SpamAssassin không phải là chương trình cho phép xóa thư rác hoặc dẫn đường cho thư rác hoặc email bình thường tới những thư mục khác nhau hoặc gửi phản hồi từ chối thư rác tới đối tượng gửi. Những chức năng trên được gọi là dẫn đường mail (mail routing) còn SpamAssassin ở đây không đóng vai trò là bộ định tuyến mail (mail router). SpamAssassin chỉ đóng vài trò là một bộ lọc hay phân loại các email mà thôi. Nó sẽ thực hiện phân tích từng thông điệp, với mỗi thông điệp sẽ được gán một chỉ số thư rác tương ứng (score). Một chương trình khác sẽ làm nhiệm vụ nhận đầu ra là chỉ số thư rác từ SpamAssassin và tiến hành việc dẫn đường cho các email này. Có nhiều chương trình dễ dàng thực hiện chức năng này sau khi đã nhận được chỉ số thư rác được tính toán bởi SpamAssassin.

Khi đã được nhận diện, email sẽ được đánh dấu là thư rác, từ đó mail user-agent ở phía người sử dụng cuối sẽ có cách hành xử phù hợp như chuyển vào thư mục SPAM, đánh dấu tiêu đề để cảnh báo với người sử dụng đó là thư rác. Xác suất phát hiện thư rác của SpamAssassin giao động từ 95% tới 100% tùy vào loại email mà hệ thống xử lý, mức độ hiệu quả của bộ lọc mà người quản trị sử dụng. Những con số thống kê cho thấy SpamAssassin để lọt 1.5% thư rác và phát hiện nhầm 0.06% số lượng email vô hại nhưng bị đánh dấu là thư rác

SpamAssassin cũng bao gồm những plugin hỗ trợ việc báo cáo, thống kê thư rác tự động theo lịch trình hoặc thao tác bằng tay tới các cơ sở dữ liệu luật lọc cộng tác (collaborative filtering databases) như Pyzor, DCC, và Vipul's Razor. Chương trình SpamAssassin cung cấp một công cụ lọc thư rác có thể chạy từ dòng lệnh có tên là spamassassin. Bên cạnh đó mô-đun Mail::SpamAssassin sẽ cho phép SpamAssassin được sử dụng như là một máy chủ SMTP hoặc POP/IMAP proxy hoặc ứng dụng trong nhiều trường hợp khác nhau cho mục đính chống thư rác.

6. ClamAV

Trang chủ: http://www.clamav.net/

Phiên bản mới nhất: **ClamAV 0.95.2**

Clam Antivirus là một phần mềm diệt virus mã nguồn mở (GPL) dành cho hệ điều hành Unix. ClamAV được thiết kế đặc biệt dành cho việc quét virus trong email tại các mail gateway. Nó cung cấp nhiều tiện ích bao gồm tính linh động, khả năng chạy dưới dạng các tiến trình đa nhiệm (scalable multi-threaded daemon), cung cấp công cụ quét virus từ dòng lệnh, hỗ trợ cơ chế cập nhật tự động thường xuyên.

Các tính năng nổi trội của ClamAV:

* Là phần mềm mã nguồn mở được phát hành tuân theo giấy phép GNU General Public License, Phiên bản 2

* Quét virus tốc độ cao

* Hỗ trợ quét virus khi truy cập vào file hoặc thư mục (chức năng này chỉ có trên các hệ điều hành Linux và FreeBSD)

* Có khả năng phát hiện hơn 618440 virus, worm và trojan, bao gồm cả virus Microsoft Office macro, mã độc và các mối nguy tiềm ẩn khác.

* Có khả năng quét các tệp tin đã bị nén, các định dạng nén hỗ trợ bao gồm:

o Zip (including SFX)

o RAR (including SFX)

o ARJ (including SFX)

o Tar

o Gzip

o Bzip2

o MS OLE2

o MS Cabinet Files (including SFX)

o MS CHM (Compiled HTML)

o MS SZDD compression format

o BinHex

o SIS (SymbianOS packages)

o AutoIt

* Hỗ trợ hầu hết các định dạng email

* Hỗ trợ các định dạng tệp tin đặc biệt khác bao gồm:

o HTML

o RTF

o PDF

o Các tệp tin được mã hóa bởi CryptFF và ScrEnc

o uuencode

o TNEF (winmail.dat)

* Có công cụ cập nhật cơ sở dữ liệu tiên tiến cho phép người quản trị có thể cấu hình hệ thống tự động cập nhật cơ sở dữ liệu virus nhiều lần trong ngày.

3.2.6. Thử nghiệm hệ thống lọc thư điện tử

Trong quá trình nghiên cứu, đã xây dựng, triển khai hệ thống Mail Gateway và tiến hành thử nghiệm theo mô hình (hình 3.28). Các trường hợp

thử nghiệm thư rác được thể hiện qua các ví dụ lọc ở phần trên. Dưới đây là thống kê một số kết quả bước đầu hoạt động của hệ thống lọc thư điện tử:

Tổng cộng các email được gửi tới Mail Gateway	2538
Số lượng thư rác	2152
Số lượng thư rác bị phát hiện	2035
Tỉ lệ thư rác bị phát hiện	94.6%
Số lượng thư rác bị bỏ sót	117
Số lượng thư rác bị nhận diện nhầm	78
Tỉ lệ thư rác bị nhận diện nhầm	3.85%

Dưới đây là các một số kết quả cụ thể đã tiến hành thử nghiệm trên hệ thống lọc thư điện tử Mail Gateway:

3.2.6.1. Chặn email từ các máy chủ định trước

1. Mô tả

Trong trường hợp khi phát hiện thông tin của chính xác các server có khả năng được sử dụng làm nơi để phát tán thư rác, người quản lý hệ thống sẽ dựa vào thông tin của các server đó(dải IP, đuôi địa chỉ email, domain server...) nhằm ngăn chặn thư rác.

Lọc sử dụng thông tin tương ứng với server sẽ chặn tất cả những email xuất phát từ server đó hoặc có liên quan.

Các thông tin này thường thể hiện ở trong mục header của thư.

2. Các trường hợp

Trường hợp 01 - Chặn email từ domain định trước

a. Input – Output dự kiến

i. Kiểm tra header nếu xuất phát từ domain google.com, chỉ số spam sẽ là 5.1. Thông số trong file .cf sẽ có dạng:

Header DOMAIN_HEADER_FILTER Received =~ /google\.com/i

Score DOMAIN_HEADER_FILTER 5.2

ii. Khi chỉ số vượt ngưỡng(5), email sẽ bị đánh dấu là spam

b. Gửi một email từ địa chỉ ****@gmail.com (tương ứng với server có domain google.com)

DOMAIN_HEADER_FILTER: (trích header sau khi đã lọc)

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.3 required=5.0 tests=DK_SIGNED,
     DOMAIN_HEADER_FILTER,HTML_MESSAGE,RDNS_NONE autolearn=no version=3.2.4
X-Spam-Report:
     * 5.2 DOMAIN_HEADER_FILTER Domain forbidden
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-pz0-f178.google.com) (209.85.222.178)
     by qmt.localhost.local with SMTP; 5 Sep 2009 03:48:21 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209.85.222.178 as permitted sender)
Received: by pzk8 with SMTP id 8so1231166pzk.22
```

c. Chỉ số nhận được là 5.3, do đó email sẽ được đánh dấu Spam.

d. Đối chiếu, trùng với output dự kiến.

e. Thử với một email từ server khác với google.com, bộ lọc bỏ qua không đánh dấu spam.

Trường hợp 02 - Chặn email đến từ một server có ip/dải ip định trước

a. Input – Output dự kiến

i. Kiểm tra header nếu xuất phát từ IP 209.85.*.*, chỉ số spam sẽ là 5.2. Thông số file .cf có dạng:

header IP_FILTER Received =~ /209\.85/

describe IP_FILTER IP forbidden

score IP_FILTER 5.3

ii. Khi chỉ số vượt ngưỡng(5), email sẽ bị đánh dấu spam

b. Gửi một email từ server IP tương ứng là 209.85.222.178 (trong trường hợp này là ip tương ứng email từ ****@gmail.com) IP_FILTER: (trích header sau khi đã lọc)

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.4 required=5.0 tests=DK_SIGNED,HTML_MESSAGE,
     IP_FILTER,RDNS_NONE autolearn=no version=3.2.4
X-Spam-Report:
     * 5.3 IP_FILTER IP forbidden
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-px0-f190.google.com) (209.85.216.190)
     by qmt.localhost.local with SMTP; 7 Sep 2009 03:04:42 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209.85.216.190
Received: by pxi28 with SMTP id 28so2215779pxi.2
```

c. Chỉ số nhận được vượt ngưỡng 5, đánh dấu spam.

d. Đối chiếu, trùng với output dự kiến.

e. Thử với email gửi từ server có địa chỉ IP nằm trong dải khác, bộ lọc bỏ qua không đánh dấu spam.

Trường hợp 03 - Chặn các email có đuôi địa chỉ thuộc diện spam

a. Input – Output dự kiến

i. Kiểm tra header trong trường From nếu là từ ****@hotmail.com, chỉ số spam sẽ là 5.2. Thông số file .cf có dạng:

header FROM_ADDR_FILTER From =~ /hotmail\.com/
describe FROM_ADDR_FILTER From address forbidden
score FROM_ADDR_FILTER 5.2

ii. Khi chỉ số vượt ngưỡng cho phép(5), email sẽ bị đánh dấu spam

b. Gửi một email từ địa chỉ ****@hotmail.com: FROM_ADDR_FILTER (trích từ header sau khi đã lọc)

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.3 required=5.0 tests=FROM_ADDR_FITER,HTML_MESSAGE,
     RDNS_NONE autolearn=no version=3.2.4
X-Spam-Report:
     * 5.2 FROM_ADDR_FITER From address forbidden
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO snt0-omc1-s37.snt0.hotmail.com) (65.55.90.48)
     by qmt.localhost.local with SMTP; 5 Sep 2009 03:57:05 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at spf-a.hotmail.com designates 65.55.90.48 as permitted sender)
Received: from SNT104-W55 ([65.55.90.7]) by snt0-omc1-s37.snt0.hotmail.com with Microsoft SMTPSVC(6.0.3790.3959);
     Fri, 4 Sep 2009 20:53:10 -0700
Message-ID: <SNT104-W554344B26F80095C3725BFABED0@phx.gbl>
Return-Path: hoahongxukhac@hotmail.com
Content-Type: multipart/alternative;
     boundary="_d0b5ec39-e2fb-4b9a-bfc4-cbf4b61c841f_"
X-Originating-IP: [58.187.51.190]
From: Theo Ghost <hoahongxukhac@hotmail.com>
To: <huy@pielab.mine.nu>
Subject: ***SPAM*** Hi
```

c. Chỉ số nhận được vượt ngưỡng 5, email bị đánh dấu spam.

d. Kiểm tra, trùng với output dự kiến.

e. Thử với email có địa chỉ có đuôi khác @hotmail.com, bộ lọc bỏ qua không đánh dấu spam.

3.2.6.2. Chặn các email lừa đảo

1. Mô tả

Email lừa đảo trực tuyến và ăn cắp danh tính đã và đang rất phổ biến. Có rất nhiều phương cách, nhưng hình thức tập trung chủ yếu vào một số kiểu lừa đảo như sở hữu thừa kế, trúng xổ số hay ăn cắp danh tính.

Trong trường hợp sở hữu thừa kế hay trúng xổ số, kẻ tấn công thường gửi email với nội dung người nhận đã trúng xổ số hoặc kêu gọi hợp tác để chia chác tài sản thừa kế. Những kiểu lừa đảo này tựu chung sẽ yêu cầu người nhận email gửi một khoản tiền nhỏ nào đấy đặt cọc, sau đó có thể rút ra được số tiền lớn hơn(thường khá lớn).

Trong trường hợp ăn cắp danh tính, người dùng nhận được một email giả mạo từ các trang mua bán hay thanh toán trực tuyến nổi tiếng như amazon, ebay, paypal,... yêu cầu đăng nhập vào trang web qua đường link trong email. Thực tế trang web mà đường link trong email dẫn tới là một

trang web giả mạo được lập ra nhằm đánh cắp thông tin tài khoản của người dùng.

Các email dạng lừa đảo sẽ được bộ lọc kiểm tra nội dung thư, sử dụng định nghĩa các từ khoá nhất định tương ứng với từng loại.

2. Các trường hợp – Tiếng Anh

Trường hợp 01 - Chặn email thuộc diện lừa đảo bằng chiêu thức trúng xổ số

a. Input – Output dự kiến

i. Các từ khoá nhạy cảm tương đương với lừa đảo trúng xổ số bao gồm “lucky number”, “deposit”, “pay out”, “prize”, … sẽ có chỉ số spam tương ứng là 1.66.

ii. Thông số file .cf có dạng

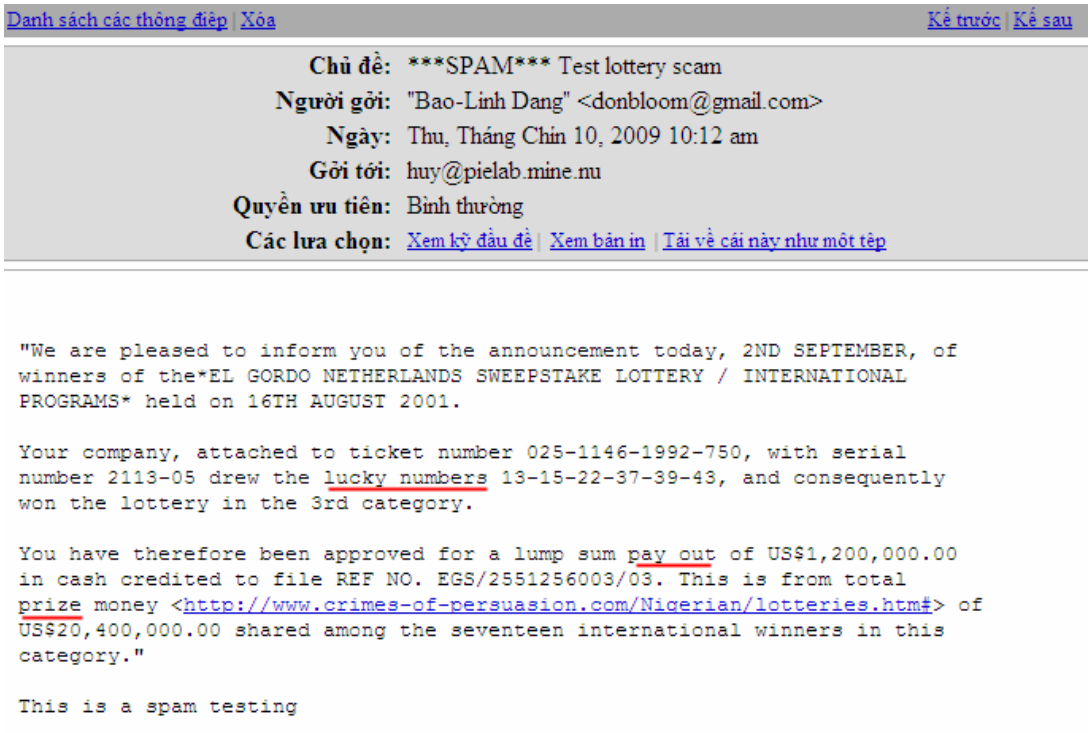
body EN_BODY_505 /lucky number/

describe EN_BODY_505 scam

score EN_BODY_505 1.66

iii. Khi email có từ 3 từ khoá nhạy cảm, chỉ số spam thấp nhất sẽ >5, vượt ngưỡng cho phép, email sẽ bị đánh dấu spam.

b. Gửi một email có nội dung chứa 3 trong số các từ khoá nhạy cảm



Danh sách các thông điệp | Xóa — Kế trước | Kế sau

Chủ đề: ***SPAM*** Test lottery scam
Người gởi: "Bao-Linh Dang" <donbloom@gmail.com>
Ngày: Thu, Tháng Chín 10, 2009 10:12 am
Gởi tới: huy@pielab.mine.nu
Quyền ưu tiên: Bình thường
Các lựa chọn: Xem kỹ đầu đề | Xem bản in | Tải về cái này như một tệp

"We are pleased to inform you of the announcement today, 2ND SEPTEMBER, of winners of the*EL GORDO NETHERLANDS SWEEPSTAKE LOTTERY / INTERNATIONAL PROGRAMS* held on 16TH AUGUST 2001.

Your company, attached to ticket number 025-1146-1992-750, with serial number 2113-05 drew the lucky numbers 13-15-22-37-39-43, and consequently won the lottery in the 3rd category.

You have therefore been approved for a lump sum pay out of US$1,200,000.00 in cash credited to file REF NO. EGS/2551256003/03. This is from total prize money <http://www.crimes-of-persuasion.com/Nigerian/lotteries.htm#> of US$20,400,000.00 shared among the seventeen international winners in this category."

This is a spam testing

Trích header sau khi qua bộ lọc:

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: ***********
X-Spam-Status: Yes, score=11.5 required=5.0 tests=ADVANCE_FEE_2,DK_SIGNED,
     EN_BODY_504,EN_BODY_505,EN_BODY_510,HTML_MESSAGE,HTML_OBFUSCATE_10_20,
     RDNS_NONE,US_DOLLARS_3 autolearn=no version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.7 EN_BODY_510 BODY: scam
     * 1.7 EN_BODY_505 BODY: scam
     * 1.7 EN_BODY_504 BODY: scam
     * 1.2 US_DOLLARS_3 BODY: Mentions millions of $ ($NN,NNN,NNN.NN)
     * 3.2 HTML_OBFUSCATE_10_20 BODY: Message is 10% to 20% HTML obfuscation
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 2.0 ADVANCE_FEE_2 Appears to be advance fee fraud (Nigerian 419)
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-pz0-f186.google.com) (209.85.222.186)
     by qmt.localhost.local with SMTP; 10 Sep 2009 03:12:40 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209.85.222.186 a
Received: by pzk16 with SMTP id 16so4626065pzk.18
     for <huy@pielab.mine.nu>; Wed, 09 Sep 2009 20:12:26 -0700 (PDT)
```

c. Chỉ số nhận được là 11.5, trong đó có 3 từ khoá nhạy cảm tương ứng kiểu lừa đảo trúng xổ số (EN_BODY). Do vượt ngưỡng là 5, email bị đánh dấu spam.

d. Đối chiếu trùng với output dự kiến.

Trường hợp 02 - Chặn email lừa đảo bằng chiêu thức tài sản thừa kế

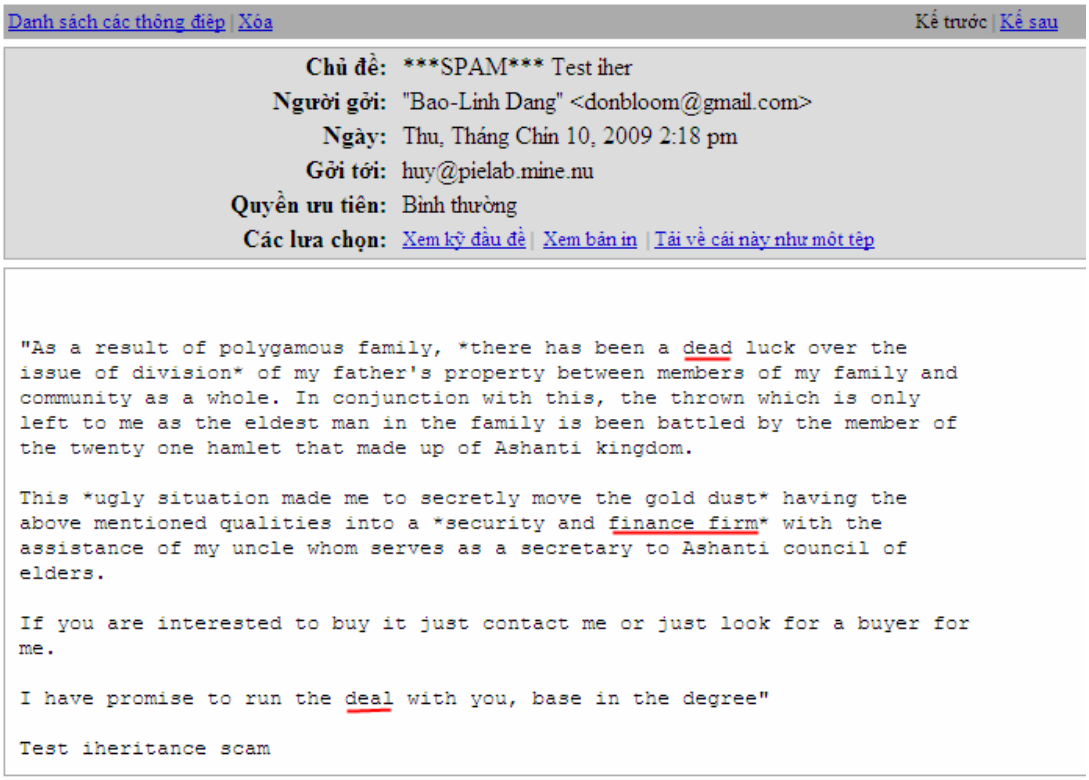
a. Thiết lập Input – Output dự kiến:

i. Các tự khoá nhạy cảm tương ứng lừa đảo thông qua tài sản thừa kế bao gồm "dead", "financial firm", "deal", "sincerity", "transaction",… sẽ có chỉ số spam tương đương 1.6.

ii. Thông số file .cf có dạng:
body EN_BODY_514 /finance firm/
describe EN_BODY_514 inheritance scam
score EN_BODY_514 1.66

iii. Khi email có nội dung chứa từ 3 từ khoá nhạy cảm hoặc nhiều hơn, chỉ số spam sẽ vượt ngưỡng(5), email bị đánh dấu spam.

b. Gửi email có nội dung chứa ít nhất 3 trong số các từ khoá nhạy cảm



Trích header sau khi qua bộ lọc:

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: ********
X-Spam-Status: Yes, score=8.3 required=5.0 tests=DK_SIGNED,EN_BODY_512,
     EN_BODY_514,EN_BODY_515,HTML_MESSAGE,HTML_OBFUSCATE_10_20,RDNS_NONE
     autolearn=no version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.7 EN_BODY_512 BODY: inheritance scam
     * 1.7 EN_BODY_514 BODY: inheritance scam
     * 1.7 EN_BODY_515 BODY: inheritance scam
     * 3.2 HTML_OBFUSCATE_10_20 BODY: Message is 10% to 20% HTML obfuscation
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-px0-f197.google.com) (209.85.216.197)
     by qmt.localhost.local with SMTP; 10 Sep 2009 07:19:09 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209
Received: by pxi35 with SMTP id 35so1614178pxi.10
     for <huy@pielab.mine.nu>; Thu, 10 Sep 2009 00:18:46 -0700 (PDT)
```

c. Chỉ số spam nhận được là 8.5, trong đó có 3 từ khoá tương ứng với kiểu lừa đảo qua thừa kế tài sản (EN_BODY). Do vượt ngưỡng 5, email bị đánh dấu spam.

d. Đối chiếu, trùng với output dự kiến.

Trường hợp 03 - Chặn các email giả mạo trang web thanh toán trực tuyến thông dụng – paypal

a. Thiết lập Input – Output dự kiến:

i. Các từ khoá nằm trong các email giả mạo Paypal sẽ có chỉ số spam tương ứng tuỳ thuộc vào mức độ liên quan.

ii. File .cf sẽ có dạng:

body EN_PPP_01 //

describe EN_PPP_01 gia mao paypal

score EN_PPP_01 1.2

iii. Khi email có chỉ số vượt ngưỡng(5), email sẽ bị đánh dấu spam

b. Gửi email có nội dung giả mạo paypal

Trích header sau khi qua bộ lọc

c. Chỉ số nhận được là , vượt ngưỡng, email sẽ bị đánh dấu spam

d. Đối chiếu, trùng với output dự kiến.

3.2.6.3. Chặn các email quảng cáo

1. Mô tả

Thư quảng cáo là một trong những hình thức thư rác phổ biến nhất hiện nay. Người dùng sẽ nhận được các quảng cáo ngẫu nhiên không mong muốn. Quảng cáo thường có nội dung thu hút người sử dụng như dược, chứng khoán, kiếm tiền, các trang web người lớn...

Có rất nhiều phương thức để chặn thư rác quảng cáo, một trong số đó là duyệt nội dung email đối chiếu với các từ khoá tương ứng để lọc.

2. Các trường hợp – Tiếng Anh

Trường hợp 01 - Chặng các email quảng cáo về thị trường chứng khoán

a. Thiết lập Input – Output dự kiến:

i. Các từ khoá liên quan đến quảng cáo thị trường chứng khoán như "1nvest", "short term target", "smart money equities"… sẽ có chỉ số spam tương ứng, thường ~1.66.

ii. File .cf có dạng:

```
body        EN_BODY_25      /smart m.ney equ.{3}es/is
describe    EN_BODY_25  Smart Money Equities
score       EN_BODY_25  1.66
```

iii. Nếu chỉ số spam của email vượt ngưỡng 5, email sẽ bị đánh dấu spam.

b. Gửi email có nội dung chứa các từ khoá liên quan tới chứng khoán

Trích header sau khi qua bộ lọc:

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.4 required=5.0 tests=DK_SIGNED,EN_BODY_24,
     EN_BODY_25,HTML_MESSAGE,RDNS_NONE,SHORT_TERM_PRICE autolearn=no version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.7 EN_BODY_25 BODY: Smart Money Equities
     * 1.9 SHORT_TERM_PRICE BODY: SHORT_TERM_PRICE
     * 1.7 EN_BODY_24 BODY: Stock ads
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-pz0-f183.google.com) (209.85.222.183)
     by qmt.localhost.local with SMTP; 10 Sep 2009 08:59:43 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209
Received: by pzk13 with SMTP id 13so513175pzk.5
     for <huy@pielab.mine.nu>; Thu, 10 Sep 2009 01:59:27 -0700 (PDT)
```

c. Chỉ số nhận được là 5.4, vượt ngưỡng nên email bị đánh dấu spam. Chú ý các từ khoá liên quan tới chứng khoán đều bị đánh dấu.

d. Đối chiếu, trùng với output dự kiến.

3. Các trường hợp – tiếng Việt

Trường hợp 01 - Chặn các email lừa đảo giới thiệu kiếm tiền qua mạng

a. Thiết lập Input – Output dự kiến:

i. Các từ khoá có liên quan đến giới thiệu kiếm tiền qua mạng như "tài khoản", "chuyển tiền", "kiếm tiền", "thành công", "sponsor", "earning",… sẽ có chỉ số spam tương ứng tuỳ vào mức độ liên quan

ii. Thông số file .cf sẽ có dạng

body	VI_BODY_LD_02	/kiếm tiền/
describe	VI_BODY_LD_02	noi dung lua dao
score	VI_BODY_LD_02	1.66

iii. Nếu chỉ số spam vượt ngưỡng(5), email sẽ bị đánh dấu spam.

b. Gửi một email có chứa nội dung giới thiệu kiếm tiền qua mạng

Danh sách các thông điệp | Xóa — Kế trước | Kế sau

Chủ đề: ***SPAM*** Test egold
Người gởi: "Bao-Linh Dang" <donbloom@gmail.com>
Ngày: Sat, Tháng Chín 12, 2009 10:03 am
Gởi tới: huy@pielab.mine.nu
Quyền ưu tiên: Bình thường
Các lựa chọn: Xem kỹ đầu đề | Xem bản in | Tải về cái này như một tệp

keim tien co that voi 500 cents
Chào bạn !

Dưới đây tôi xin giới thiệu với các bạn các trang web nước ngoài. Đầu tiên là :
http://www.500cents-500dollars.org/pages/index.php?refid=satthumang
Bạn chép link này vào thanh address và Enter.

Bạn vào mục Sign Up, nhấp vào đó và nhập địa chỉ email của bạn vào ô trống ở dưới rồi bấm Continue. Sau đó một phút, bạn mở hộp thư Inbox trong Yahoo ra, rồi bấm vào link đăng kí (URL) mà họ đã gửi cho bạn. Và bạn sẽ thấy được phần đăng kí của họ như sau :

Username: Tùy bạn chọn
Email: Địa chỉ email của bạn (có sẵn, không cần điền)
Send emails to: Bạn chọn Email Address hoặc Inbox (tốt nhất chỉ chọn Inbox để e-mail bạn khỏi phiền phức).
First name: Tên thật của bạn, ví dụ : Nam
Last name: Họ và tên đệm, ví dụ : Tran Ngoc Phuong
Address: Địa chỉ của bạn
City: Thành phố/ Tỉnh nơi bạn ở
State: Bạn chọn chữ N/A vào ô này
Zip Code: 70000 (Nếu bạn ở Tp.HCM), 10000 (Nếu bạn ở Hà Nội)
Các nơi khác thì bạn xem theo link sau :
http://danhba.vdc.com.vn/tracuu/danhba/mavungdt.asp
Xem mục Mã bưu chính. Công thức : Mã bưu chính + 000.
(Vd : Mã bưu chính Tp.HCM = 70 => Zip Code : 70+000 = 70000)
Country: Vietnam
Referred: Để nguyên
Payment method: Cách thanh toán. Tùy bạn chọn : e-gold (chọn cái này để dễ chuyển tiền về VN, còn bạn chọn cái khác thì tự tìm hiểu, tôi chỉ dùng e-gold), MoneyBookers, StormPay và điền số tài khoản e-gold vào bên dưới (nếu chưa có thì để trống, sau này điền vào sau).
Select categories of interests to you: Bạn nên đánh dấu hết vào những ô trống này vì bạn đánh dấu càng nhiều, bạn sẽ nhận được càng nhiều email.

Trích header sau khi đã qua bộ lọc:

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.6 required=5.0 tests=DK_SIGNED,FB_WORD1_END_DOLLAR,
     HTML_MESSAGE,RDNS_NONE,VI_BODY_LD_02,VI_BODY_LD_03,VI_BODY_LD_04,VN_BODY_503
     autolearn=no version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.4 VN_BODY_503 BODY: Body contains "xxx"
     * 0.0 FB_WORD1_END_DOLLAR BODY: Looks like a word ending with a $
     * 1.2 VI_BODY_LD_03 BODY: noi dung lua dao
     * 1.7 VI_BODY_LD_02 BODY: noi dung lua dao
     * 1.2 VI_BODY_LD_04 BODY: noi dung lua dao
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-px0-f199.google.com) (209.85.216.199)
     by qmt.localhost.local with SMTP; 12 Sep 2009 03:04:07 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209.85.216.199 as permitted s
Received: by pxi37 with SMTP id 37so1295230pxi.28
     for <huy@pielab.mine.nu>; Fri, 11 Sep 2009 20:03:54 -0700 (PDT)
```

c. Chỉ số spam nhận được là 5.6, vượt ngưỡng 5, email bị đánh dấu spam.

d. Kiểm tra đối chiếu, trùng với output dự kiến.

Trường hợp 02 - Chặn email quảng cáo trang web người lớn

a. Thiết lập Input – Output dự kiến:

i. Các từ khoá nhạy cảm liên quan đến trang web người lớn như "phim sex", "phim người lớn", "direct link", … sẽ có chỉ số spam là 1.66 hoặc 1.4 tuỳ vào độ liên quan.

ii. File .cf sẽ có dạng:

body	EN_BODY_522 /link trực tiếp/
describe	EN_BODY_522 quang cao trang web nguoi lon
score	EN_BODY_522 1.2

iii. Khi email có chỉ số vượt ngưỡng(5), email sẽ bị đánh dấu spam.

b. Gửi email có nội dung quảng cáo trang web người lớn chứa một số từ khoá nhạy cảm

Trích header sau khi qua bộ lọc

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: ******
X-Spam-Status: Yes, score=6.3 required=5.0 tests=DK_SIGNED,EN_BODY_519,
     EN_BODY_520,EN_BODY_521,EN_BODY_522,HTML_MESSAGE,RDNS_NONE autolearn=no
     version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.2 EN_BODY_522 BODY: quang cao trang web nguoi lon
     * 1.7 EN_BODY_519 BODY: quang cao trang web nguoi lon
     * 1.7 EN_BODY_520 BODY: quang cao trang web nguoi lon
     * 1.7 EN_BODY_521 BODY: quang cao trang web nguoi lon
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-pz0-f183.google.com) (209.85.222.183)
     by qmt.localhost.local with SMTP; 10 Sep 2009 08:40:15 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designate
Received: by pzk13 with SMTP id 13so503048pzk.5
     for <huy@pielab.mine.nu>; Thu, 10 Sep 2009 01:39:59 -0700 (PDT)
```

c. Chỉ số nhận được là 6.3, vượt ngưỡng cho phép, email bị đánh dấu spam.

d. Đối chiếu, trùng với output dự kiến.

3.2.6.4. Chặn các email có nội dung xấu

1. Mô tả

Một dạng thư rác phổ biến nữa ở Việt Nam là các loại thư có nội dung đi ngược lại chính sách chủ trương nhà nước, vi phạm các quy định nhà nước về văn hoá, xã hội.

Các thư rác này sẽ được bộ lọc duyệt qua nội dung đối chiếu với các từ khoá có liên quan.

2. Các trường hợp – tiếng Việt

Trường hợp 01 - Chặn các email có nội dung phản động

a. Thiết lập Input – Output dự kiến:

i. Các từ khoá nhạy cảm có liên quan đến nội dung phản động như "dân chủ", "biểu tình", "đa đảng", "cộng sản,… sẽ có chỉ số spam tuỳ vào mức độ liên quan.

ii. File .cf sẽ có dạng

body VI_BODY_PD_01 /dân chủ/

describe VI_BODY_PD_01 noi dung phan dong

score VI_BODY_PD_01 1.2

iii. Khi email có chỉ số spam vượt ngưỡng (5), email sẽ bị đánh dấu spam.

b. Gửi email có chứa nội dung phản động(chứa các từ khoá nhạy cảm)

Trích header sau khi qua bộ lọc

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 3.2.4 (2008-01-01) on qmt.localhost.local
X-Spam-Level: *****
X-Spam-Status: Yes, score=5.2 required=5.0 tests=DK_SIGNED,HTML_FONT_FACE_BAD,
     HTML_MESSAGE,RDNS_NONE,VI_BODY_PD_01,VI_BODY_PD_03,VI_BODY_PD_05 autolearn=no
     version=3.2.4
X-Spam-Report:
     * 0.0 DK_SIGNED Domain Keys: message has a signature
     * 1.2 VI_BODY_PD_01 BODY: noi dung phan dong
     * 1.7 VI_BODY_PD_05 BODY: noi dung phan dong
     * 1.7 VI_BODY_PD_03 BODY: noi dung phan dong
     * 0.0 HTML_MESSAGE BODY: HTML included in message
     * 0.6 HTML_FONT_FACE_BAD BODY: HTML font face is not a word
     * 0.1 RDNS_NONE Delivered to trusted network by a host with no rDNS
Received: from unknown (HELO mail-pz0-f198.google.com) (209.85.222.198)
     by qmt.localhost.local with SMTP; 12 Sep 2009 02:24:27 -0000
Received-SPF: pass (qmt.localhost.local: SPF record at _spf.google.com designates 209.85.222.198 as
Received: by pzk36 with SMTP id 36so1292599pzk.2
     for <huy@pielab.mine.nu>; Fri, 11 Sep 2009 19:24:13 -0700 (PDT)
```

c. Chỉ số nhận được là 5.2, vượt ngưỡng cho phép, email bị đánh dấu spam.

d. Đối chiếu, trùng với output dự kiến.

3.2.6.5. Chặn email có chứa virus

1. Mô tả

Ngoài các loại thư rác quảng cáo, lừa đảo,… có một số không nhỏ email có chứa virus với mục đích tấn công hay chiếm quyền điều khiển máy người dùng, ăn cắp thông tin,…

Hệ thống lọc có thể quét email để phát hiện virus đính kèm.

2. Trường hợp

Trường hợp 01 – chặn email chứa virus Worm.Kido-95

a. Thiết lập Input – Output dự kiến:

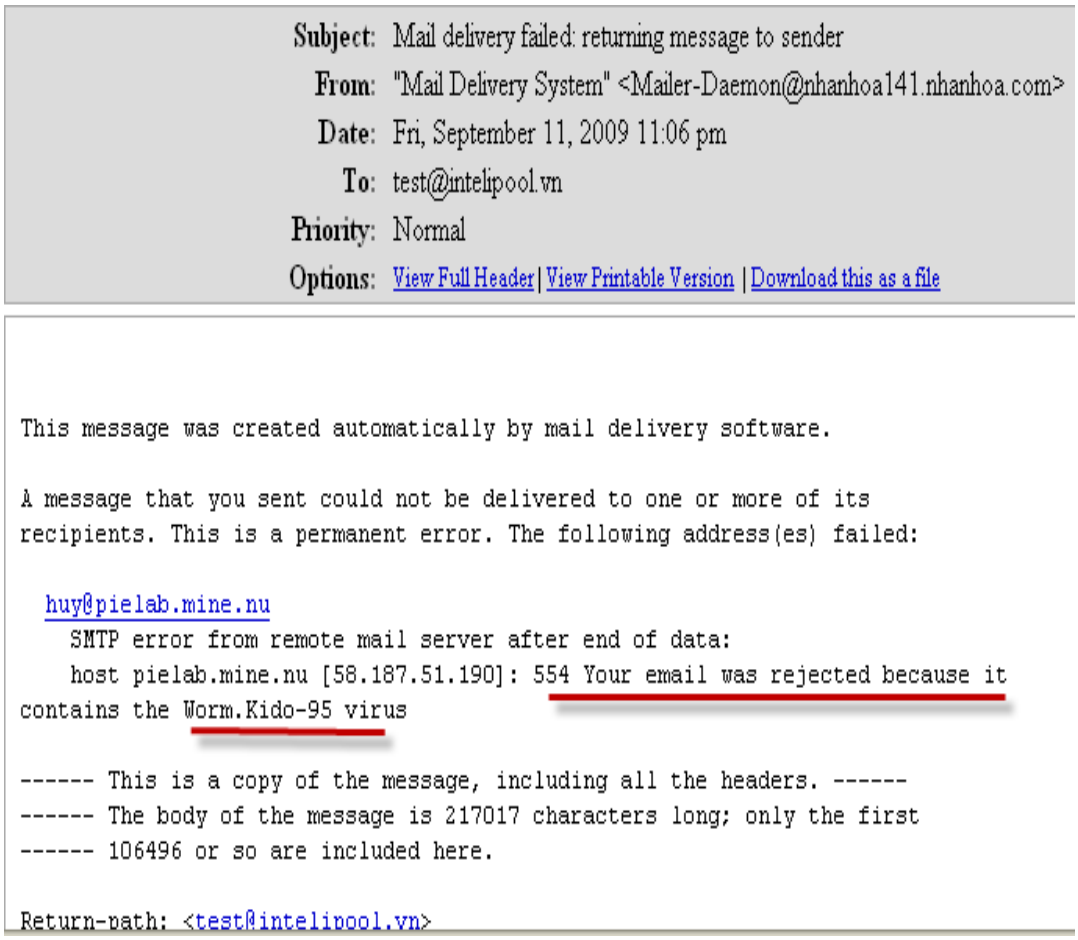
i. Khi nhận được email có chứa virus, lập tức hệ thống coi đó là spam nguy hiểm và trả email lại nơi gửi.

b. Gửi email có chứa virus:

Hệ thống phát hiện ra có virus

```
@400000004aaa79440ddc8004 /var/qmail/simscan/1252686138.21002.9244/msg.125268613
8.21002.9244: Worm.Kido-95 FOUND
@400000004aaa794412c5cb34 /var/qmail/simscan/1252686138.21002.9244/BRPHOW.DLL.zi
p: Worm.Kido-95 FOUND
```

Email gửi trả nơi xuất phát:



Subject: Mail delivery failed: returning message to sender
From: "Mail Delivery System" <Mailer-Daemon@nhanhoa141.nhanhoa.com>
Date: Fri, September 11, 2009 11:06 pm
To: test@intelipool.vn
Priority: Normal
Options: View Full Header | View Printable Version | Download this as a file

```
This message was created automatically by mail delivery software.

A message that you sent could not be delivered to one or more of its
recipients. This is a permanent error. The following address(es) failed:

  huy@pielab.mine.nu
    SMTP error from remote mail server after end of data:
    host pielab.mine.nu [58.187.51.190]: 554 Your email was rejected because it
contains the Worm.Kido-95 virus

------ This is a copy of the message, including all the headers. ------
------ The body of the message is 217017 characters long; only the first
------ 106496 or so are included here.

Return-path: <test@intelipool.vn>
```

a. Email được coi là spam bị trả lại.

b. Đối chiếu, trùng với output dự kiến

3.2.6.6. Các trường hợp hệ thống chưa ngăn chặn được

1. Mô tả

Hệ thống lọc thư rác hiện tại chưa lọc được các email có nội dung chứa hình ảnh hoặc các email chứa nội dung quảng cáo nhưng các từ khoá nhạy cảm được biến đổi khéo léo.

2. Các trường hợp

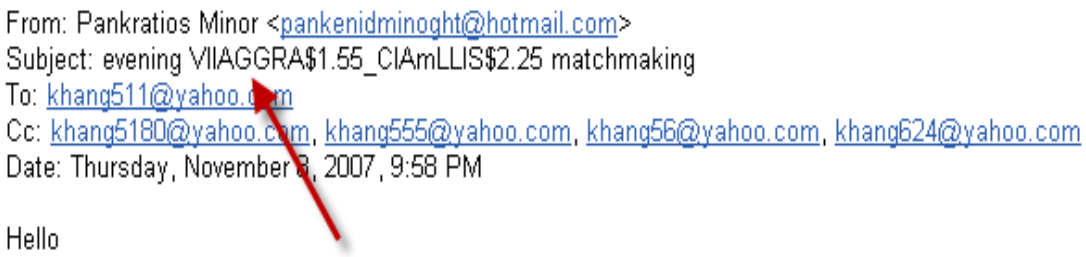
Trường hợp 01 – biến đổi từ khoá

a. Thiết lập Input – Output dự kiến:

i. Gửi email có nội dung quảng cáo nhưng với các từ khoá biến đổi

ii. Bộ lọc không phát hiện được.

b. Gửi email có nội dung quảng cáo nhưng có từ khoá biến đổi


From: Pankratios Minor <pankenidminoght@hotmail.com>
Subject: evening VIIAGGRA$1.55_ClAmLLIS$2.25 matchmaking
To: khang511@yahoo.com
Cc: khang5180@yahoo.com, khang555@yahoo.com, khang56@yahoo.com, khang624@yahoo.com
Date: Thursday, November 8, 2007, 9:58 PM

Hello

c. Hệ thống không phát hiện được spam.

d. Đối chiều, trùng output dự kiến.

Trường hợp 02 – thư nội dung nằm trong ảnh

a. Thiết lập Input – Output dự kiến:

i. Gửi email quảng cáo sử dụng ảnh

ii. Bộ lọc không phát hiện được do không lọc được nội dung ảnh

b. Gửi email quảng cáo sử dụng ảnh chứa nội dung:

c. Email được bộ lọc bỏ qua.

d. Đối chiếu, trùng output dự kiến.

3.3. Phần mềm lọc web trên máy cá nhân (SP.03)

3.3.1. Yêu cầu

3.3.1.1. Yêu cầu dữ liệu

Dữ liệu xử lý của phần mềm là các dữ liệu Internet của người sử dụng, bao gồm:

- URL/domain/IP website mà người dùng truy cập.
- Nội dung Internet dạng văn bản (text).
- Nội dung Internet dạng hình ảnh.
- Các luật lọc chặn, gồm có:
  - Các URL/domain/IP không được phép truy cập.
  - Các từ khóa lọc chặn.

3.3.1.2. Yêu cầu chức năng

Phần mềm lọc nội dung tại máy cá nhân phải đảm bảo các chức năng sau đây:

- Đối với module Client:
  - Không cho phép truy cập các URL/domain/IP không được phép dựa vào bộ luật lọc chặn.
  - Không cho phép truy cập các website có chứa những từ khóa không được phép dựa vào bộ luật lọc chặn.
  - Chặn các hình ảnh sex trên một website nào đó.
  - Cập nhật tự động bộ luật lọc chặn.
- Đối với module Server: quản lý (xem, thêm, xóa, sửa) bộ luật lọc chặn, bao gồm:
  - Các URL/domain/IP không được phép truy cập
  - Các từ khóa lọc chặn.
  - Cho phép Client tự động cập nhật bộ luật lọc chặn.

3.3.1.3. Yêu cầu phát triển

- Kích thước phần mềm nhỏ gọn.
- Không gây ảnh hưởng lớn đến toàn bộ máy tính cá nhân.
- Có khả năng mở rộng.

3.3.1.4. Yêu cầu chất lượng

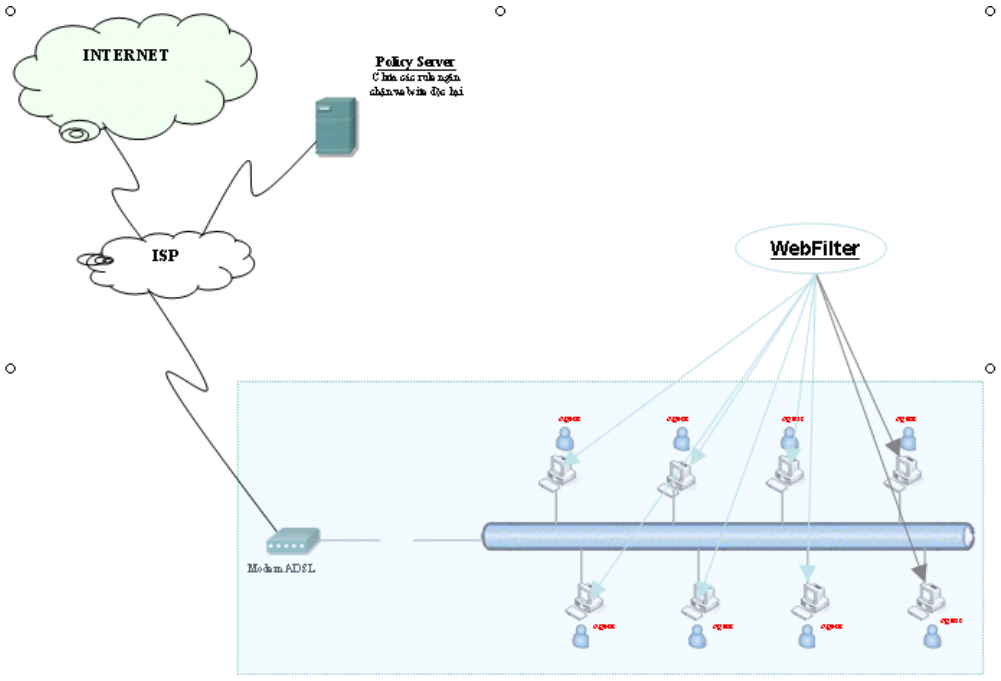
- Đối với module Client:
  - Lọc chặn chính xác.
  - Tương thích với nhiều phiên bản hệ điều hành.
  - Khả năng chịu lỗi cao.
  - Không ảnh hưởng nhiều đến hoạt động của toàn bộ máy tính.
  - Dễ sử dụng.
- Đối với module Server:

- Tính ổn định và sẵn sàng cao.
- Dễ sử dụng.

3.3.2. Chiến lược thiết kế

- Thiết kế tổng thể dựa trên yêu cầu chức năng phần mềm.
- Thiết kế chi tiết dựa trên thiết kế tổng thể và đặc tả phần mềm.
- Sử dụng công cụ thiết kế Rational Rose để thiết kế UML cho các module Client và Server.
- Đối với Module Server: chỉ cần thiết kế cho chức năng quản lý URL, chức năng quản lý DOMAIN và KEYWORD tương tự.
- Đối với Module Client: chỉ cần thiết kế chức năng lọc chặn theo URL, theo KEYWORD và lọc hình ảnh, chức năng lọc chặn theo DOMAIN tương tự lọc chặn theo URL.

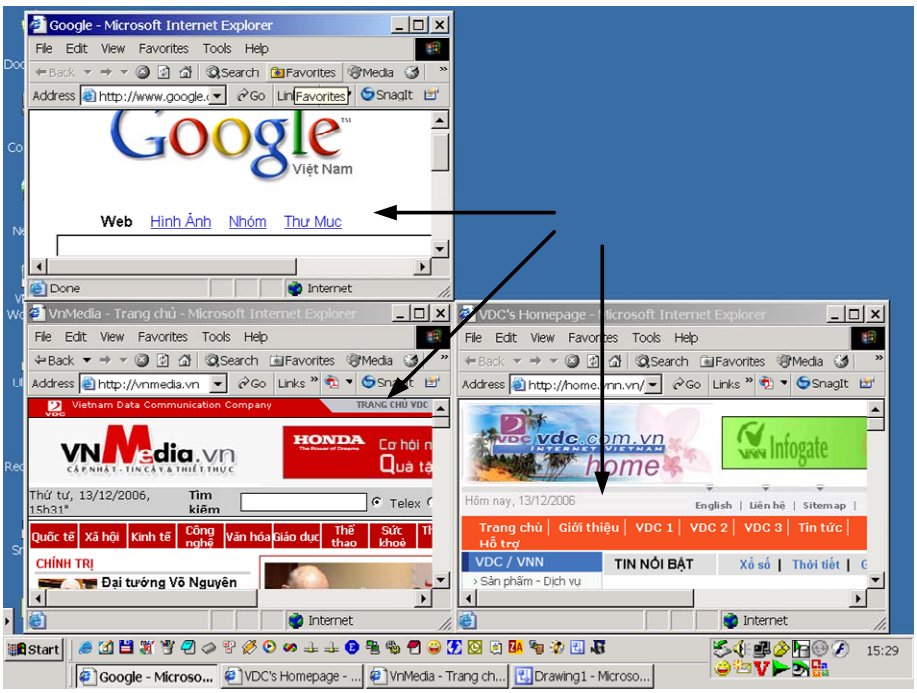
3.3.3. Mô hình kiến trúc hệ thống



Hình 3. 22. Mô hình module Client

Trong mô hình này, ta gọi WebFilter là module lọc chặn tại các máy tính cá nhân, module Server chạy trên Policy Server. Mọi máy tính truy cập Internet đều được cài đặt module WebFilter. Hình trên là mô hình một điểm truy cập Internet có sử dụng WebFilter cho toàn bộ hệ thống mạng.

WebFilter sử dụng kỹ thuật Sniffer trong phạm vi một máy tính, nghe toàn bộ nội dung các gói tin đến máy tính, tổng hợp các gói tin thành các phiên giao dịch HTTP. Sau đó sẽ phân tích ngữ nghĩa trang web, kiểm tra xem đó có phải là trang web 'đen' không, theo một bộ luật nhất định. Nếu là web đen, WebFilter sẽ tác động đến trình ứng dụng đang hiển thị web đen đó. Tác động ở đây có thể là thay đổi nội dung cửa sổ hoặc kết thúc trình duyệt, đưa ra thông báo cho người dùng…

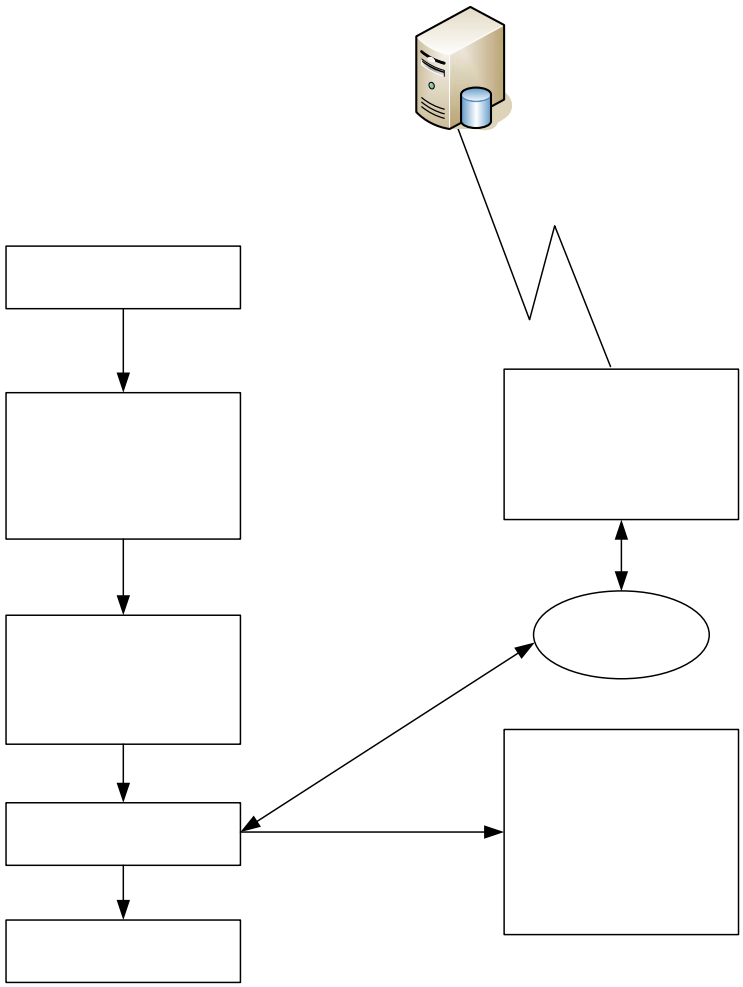


Hình 3. 218. WebFilter lắng nghe các gói tin

WebFilter có các thành phần sau:

- Driver bắt gói tin
- Phân tích nội dung gói tin.

- Tổng hợp gói tin thành các phiên giao dịch HTTP
- Phân tích nội dung trang web.
- Liệt kê danh sách tiến trình tương ứng với cổng truy cập, kết thúc tiến trình truy cập web đen.
- Liên kết server cập nhật địa chỉ lọc chặn, từ khoá lọc chặn
- Database lưu trữ luật lọc chặn tại Policy Server và Client.



Hình 3. 24. Cơ chế hoạt động WebFilter

Ở mức chi tiết hơn, quá trình hoạt động của WebFilter diễn ra như sau: Driver bắt gói tin chịu trách nhiệm thu thập tất cả các gói tin đến máy. Module phân tích gói tin sẽ phân tích cấu trúc header của gói tin theo từng

Driver bắt gói tin

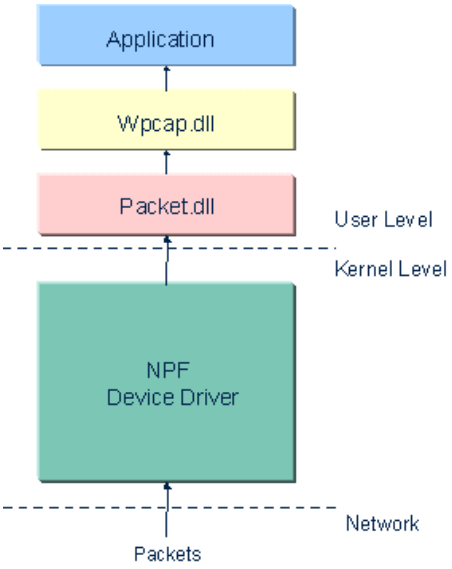
Phân tích cấu trúc

tầng mạng, bắt đầu từ tầng 2 trong mô hình OSI. Tiếp theo, chỉ chuyển các gói tin theo giao thức HTTP đến module tổng hợp gói tin thành các phiên giao dịch HTTP. Sau đó các phiên giao dịch đã được tổng hợp được đưa vào module phân tích nội dung trang web. Tại đây sẽ quyết định nội dung trang web là 'đen' hay không. Nếu đó là trang web 'đen', thông tin về cổng của phiên giao dịch tương ứng sẽ được chuyển đến module xử lý tiến trình. Tại đây sẽ liệt kê ra toàn bộ tiến trình và cổng truy cập tương ứng, để cuối cùng sẽ kết thúc tiến trình nào có truy cập trang web 'đen'.

3.3.3.1. Thành phần bắt gói tin

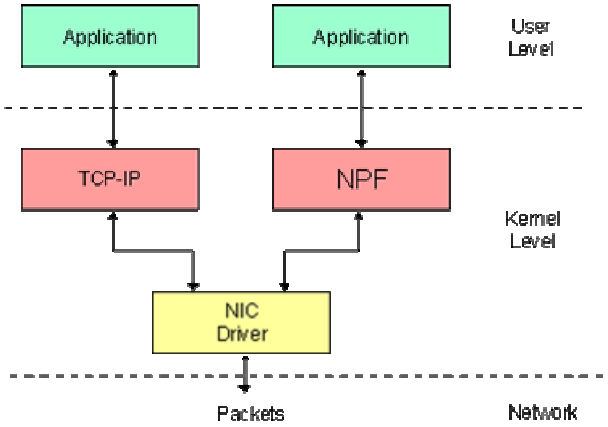
Trên hệ điều hành Windows, có thể bắt gói tin một cách đơn giản bằng cách dùng Winsock – Raw socket. Cách này rất đơn giản, tuy nhiên lại không đầy đủ. Các gói tin bị thất lạc nhiều, nhất là khi việc truy cập mạng ở tốc độ cao. Chính vì vậy mà cần có một nền tảng bắt gói tin thật tốt, ứng dụng cho nhiều việc, trong đó WebFilter là một trong các ứng dụng đó. Driver bắt gói tin trên windows phổ biến nhất hiện nay là Pcap.

Pcap bao gồm một trình driver bắt gói tin tương tự driver TCP/IP của windows.



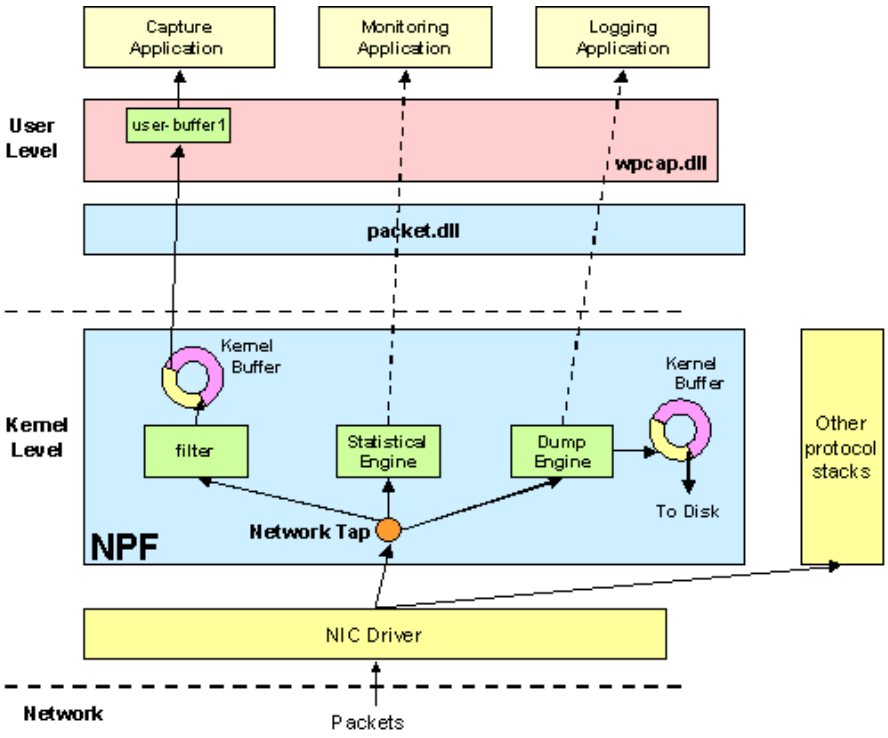
Hình 3. 25. Phân lớp bắt gói tin.

Driver bắt gói tin Pcap hoạt động song song với driver TCP/IP của Windows

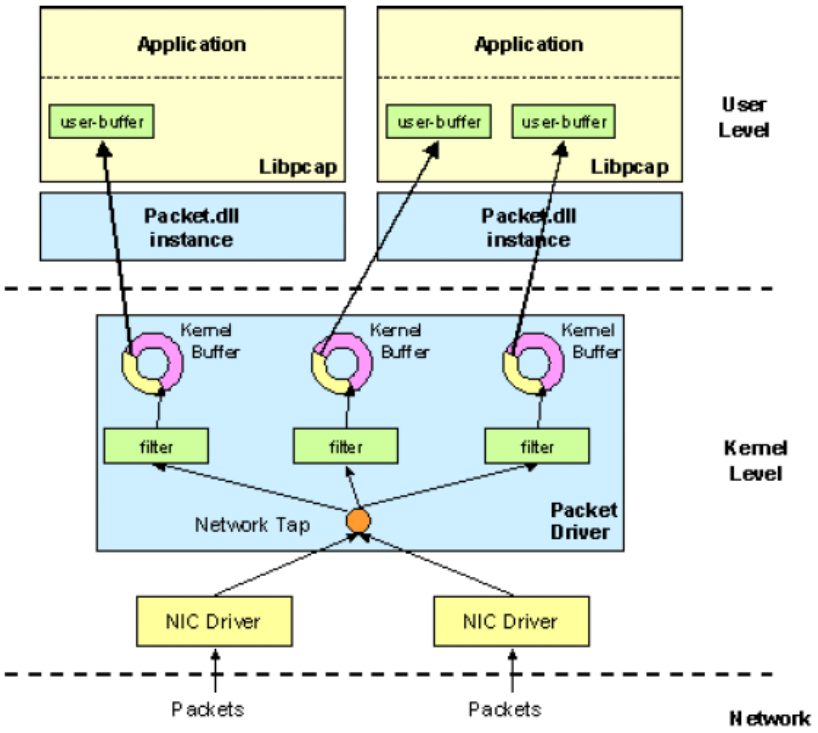


Hình 3. 26. Driver bắt gói tin hoạt động song song với TCP/IP

Ở mức chi tiết hơn, Pcap có 2 lớp KernelMode là driver NPF (npf.sys) và UserMode (wpcap.dll và packet.dll). Các ứng dụng sẽ tương tác ở UserMode. Thành phần Sniffer của WebFilter với lõi sẽ tương tác ở lớp này, sử dụng các API (giao diện lập trình) ở đây.

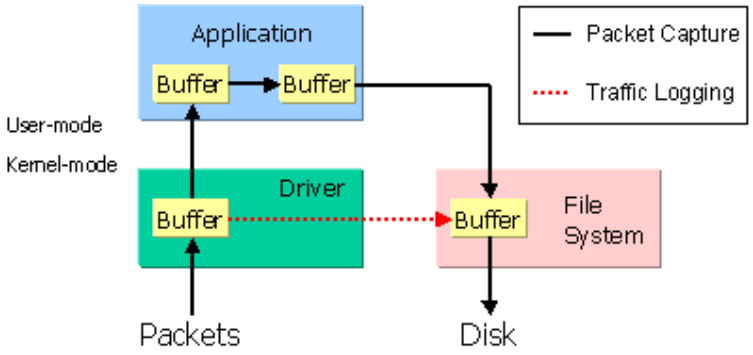


Hình 3. 27. Chu trình các gói tin đi qua driver sniffer NPF Pcap



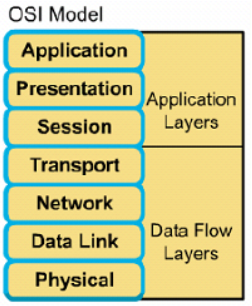
Hình 3.28. Tại một thời điểm, có nhiều NIC và nhiều ứng dụng được theo dõi

Nhìn ở góc độ bộ đệm dữ liệu (Buffer), có thể thấy chu trình dữ liệu đi như ở hình dưới. Dữ liệu gói tin đi qua driver, lên ứng dụng, và được thao tác tại ứng dụng, và có thể đưa ra hệ thống file…

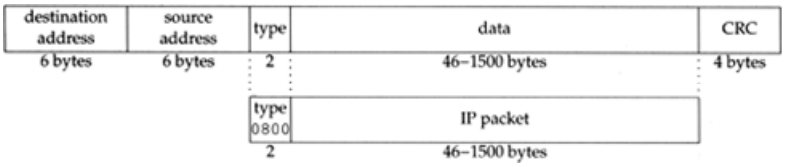


Hình 3.29. Các gói tin được lưu trong các buffer và có thể được xử lý hoặc kết xuất .

Driver bắt gói tin của Pcap được chạy song song với driver TCP/IP của windows. Không giống như các chương trình Firewall thường *chặn giữa* luồng dữ liệu vào ra. Thành phần bắt gói tin này được hoạt động ở tầng DataLink trong mô hình OSI.



Gói tin được bắt tại tầng 2 (OSI) dưới dạng Frame

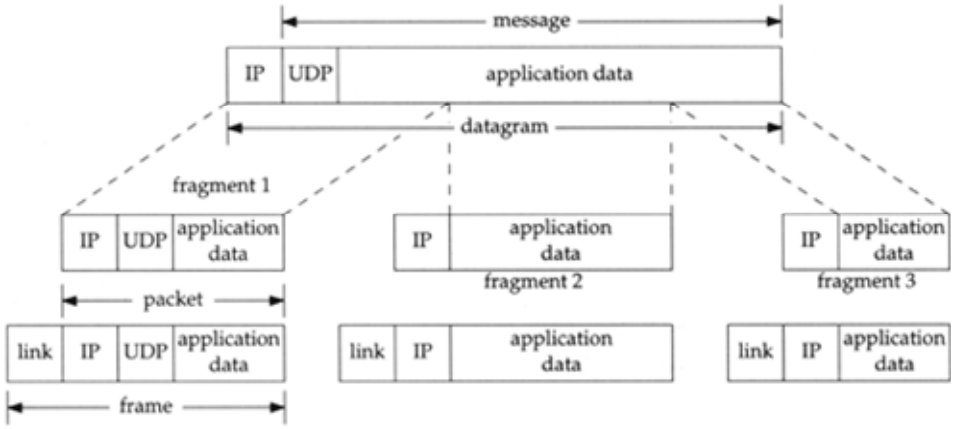


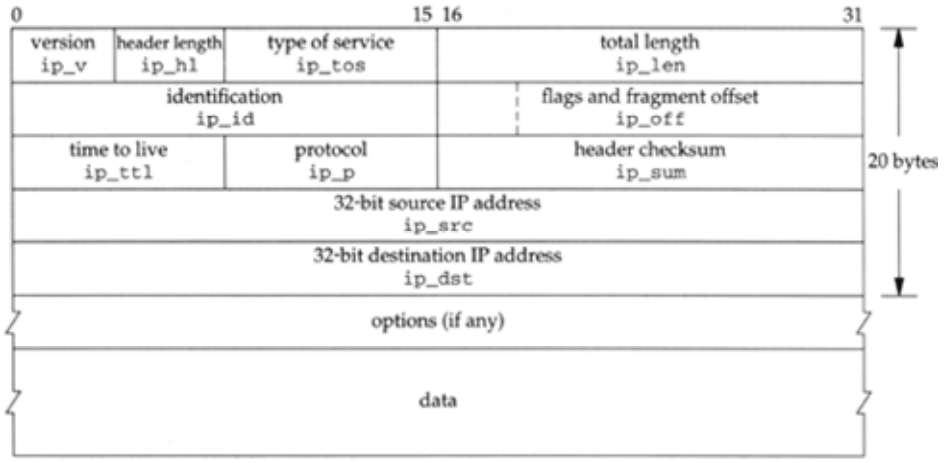
Cấu trúc dữ liệu gói tin (tầng data link)

Hình 3. 30. Các gói tin được thu thập tại tầng 2 - Datalink

Cấu trúc dữ liệu tại tầng DataLink được thiết kế trong phần mềm như sau:

```
// Ethernet Header
struct eth_header   // 14 bytes
{
      u_char dmac[6]; //destination mac address
      u_char smac[6]; //source mac address
      u_short type;                 //IP ,ARP , RARP
};
```



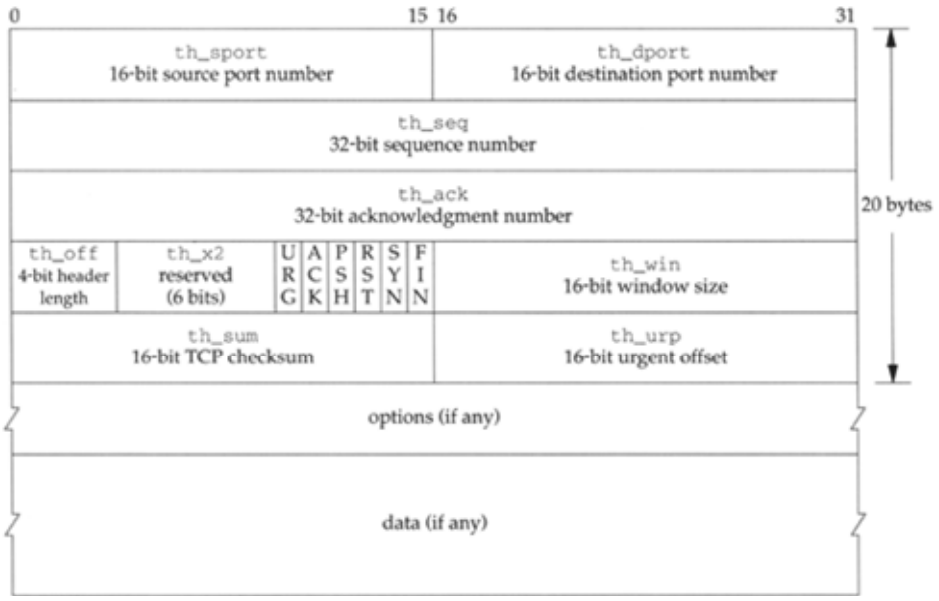


Hình 3.31. Cấu trúc dữ liệu gói tin IP (tầng Network)

Cấu trúc dữ liệu tại tầng Network như sau:

```
/* IPv4 header */
struct ip_header
{
      u_char  ver_ihl;        // Version (4 bits)
      u_char  tos;            // Type of service
      u_short tlen;           // Total length
      u_short identification; // Identification
      u_short  flags_fo;  //  Flags  (3  bits)  +  Fragment  offset  (13
bits)
      u_char  ttl;            // Time to live
```

```
      u_char  proto;          // Protocol
      u_short crc;            // Header checksum
      ip_address  saddr;      // Source address
      ip_address  daddr;      // Destination address
};
```



Hình 3.32. Cấu trúc gói tin TCP (Tầng Transport)

Cấu trúc dữ liệu tầng Transport được thiết kế trong phần mềm như sau:

```
struct tcp_header  //20 bytes : default   {
      u_short sport;      //Source port
      u_short dport;      //Destination port
      u_long seqno;       //Sequence no
      u_long ackno;       //Ack no
      u_char offset;//Higher level 4 bit indicates data offset
      u_char flag;        //Message flag
      u_short win;
      u_short checksum;
      u_short uptr;
};
```

Như vậy với thành phần Driver bắt gói tin, ta đã có thể thu thập được toàn bộ các gói tin vào ra máy tính. Công việc tiếp theo là phân tích, tổng hợp các gói tin đi và đến. Nhiệm vụ này được thành phần tiếp theo sau đây thực hiện.

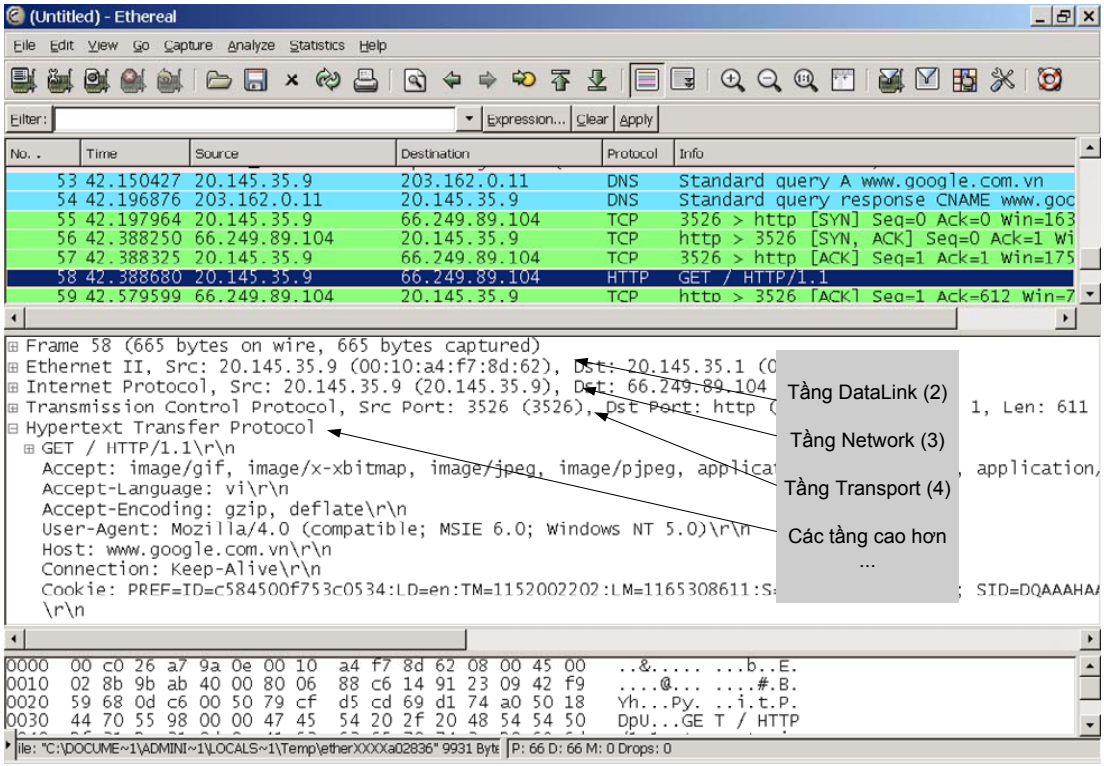
*3.3.3.*2. Thành phần phân tích nội dung gói tin

Trong một khối lượng lớn gói tin đến máy tính, nhiệm vụ của WebFilter là phải tách biệt được các gói tin tương ứng với giao dịch Web.

WebFilter bắt gói tin từ tầng DataLink, tách header để thu được gói tin thuộc tầng IP, ở đây sẽ có tham số về địa chỉ IP nguồn, đích. Tiếp theo là tầng TCP, có tham số về port, session, ack… Qua tầng TCP là dữ liệu gói tin HTTP.

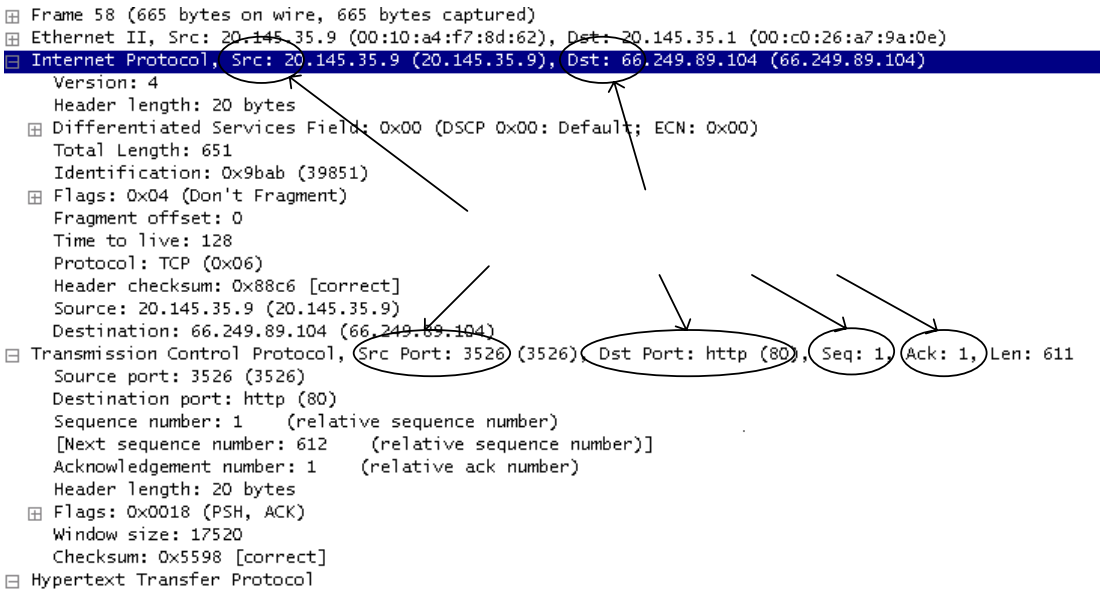
Trình tự tách gói tin HTTP như sau:

1. Quét Buffer để thu thập gói tin đến trên tất cả card mạng hiện có trên máy, mỗi card mạng tương ứng với một địa chỉ IP.
2. Phân tích header IP trong gói tin, kiểm tra địa chỉ IP trong header tương ứng, theo dõi các gói tin theo địa chỉ IP định sẵn của card mạng cần theo dõi.
3. Tương ứng với các gói tin trên, sẽ phân tích TCP header của nó. Nếu số cổng nguồn hoặc đích là 80 (tương ứng với Client gửi và nhận gói tin từ cổng 80 của server) thì tiếp tục xử lý bước sau. Các gói tin cổng khác bỏ qua.
4. Xử lý các gói tin cổng 80, đó sẽ là các gói gửi và nhận. Ở đây quan tâm chính đến các gói HTTP nhận từ server. Dữ liệu sẽ được tổng hợp từ các gói này, nội dung Web sẽ được tổng hợp tại đây.



Hình 3.33. Các gói tin thô được bắt và được phân tích theo từng tầng mạng

Cùng một thời điểm có thể có nhiều phiên giao dịch HTTP tương ứng với các đối tượng của trang web (text, ảnh…). Các phiên này được phân biệt bởi các tham số trong cấu trúc dữ liệu TCP/IP (ở trên), bao gồm: địa chỉ ip, số cổng, số ACK, SEQUENT (xem hình dưới).



Hình 3. 319. Cấu trúc dữ liệu chi tiết tầng TCP/IP của một gói tin

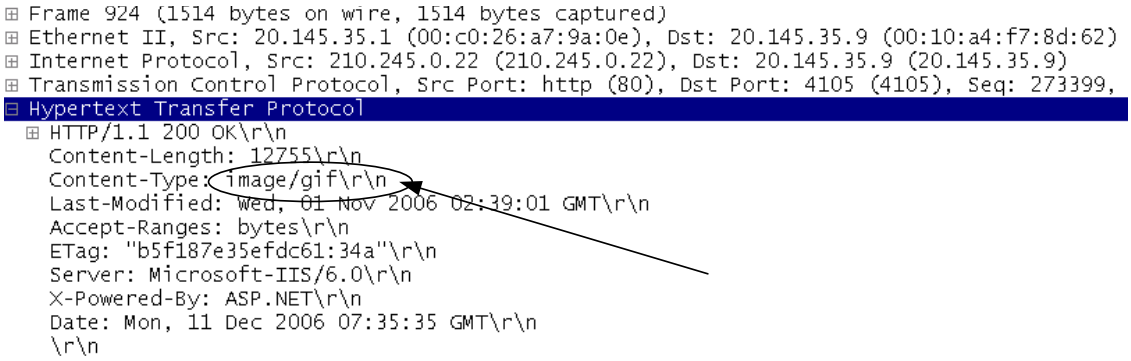
Các gói tin thông thường chỉ có kích thước nhỏ hơn 1500 bytes, trong khi đó một trang web có kích thước gấp nhiều lần như vậy. Chính vì vậy cần phải có bước tổng hợp các gói tin để thu về nguyên vẹn một phiên giao dịch, có thể đó là phiên giao dịch FTP, NetBios, SSH, YahooMessenger, HTTP... WebFilter sẽ quan tâm đến các giao dịch HTTP. Đó sẽ bao gồm các gói tin đến và đi qua cổng 80. Webfilter dựa vào việc phân tích cấu trúc gói tin mức chi tiết như trên để thực hiện việc này.

*3.3.3.*3. Thành phần tách, tổng hợp gói tin thành các phiên giao dịch HTTP

Ở phần trên, chúng ta đã phân tích quá trình xử lý các gói tin trong tầng TCP để thu được các phiên giao dịch HTTP. Khi thu được dữ liệu các phiên giao dịch HTTP, nhiệm vụ tiếp theo của WebFilter sẽ là phân tích nội dung giao dịch để biết đó là giao dịch HTTP theo dạng nào. Dạng của giao dịch HTTP ở đây có ý nghĩa là: một trang web vốn có nhiều thành phần, gồm ảnh,

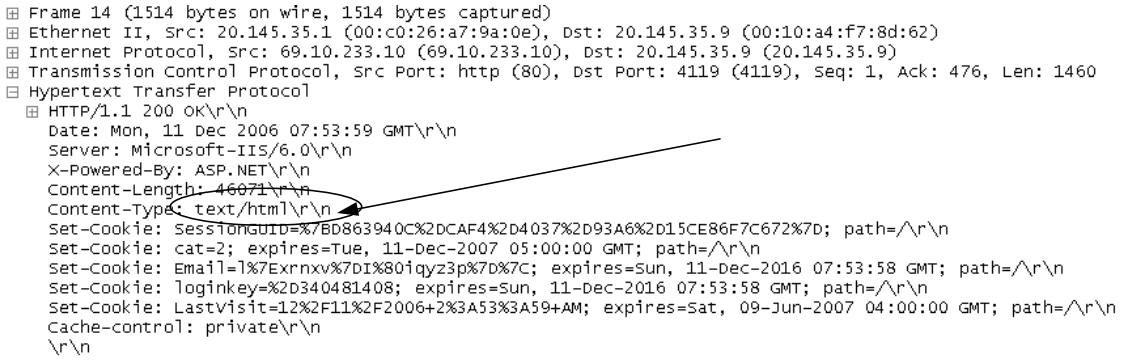
phim, flash file, javascript, text file… Do đó cần phân loại các đối tượng này, để có thể phân tích nội dung tương ứng. Trong phạm vi đề tài này các text file sẽ được quan tâm xử lý.

Dưới đây là ví dụ về nội dung HTTP session được phân tích.



Hình 3.35. Một HTTP Header cho biết đây là file ảnh

Trong hình trên ta có thể thấy một cấu trúc HTTP header, với nội dung cho biết đây là một session ảnh gif, với kích thước 12755 bytes.



Hình 3.36. HTTP Header cho biết đây là một file Text/html

Trên đây là Header của phiên giao dịch HTTP text/html, với một trang text có nội dung 46071 bytes. WebFilter sẽ nhận dạng ra các phiên giao dịch này và thực hiện bước tiếp theo đối với phiên giao dịch. Tất cả các giao dịch không phải text HTML sẽ được bỏ qua.

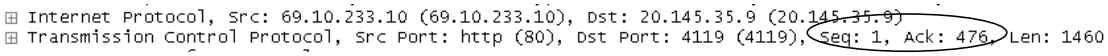
Bước tiếp theo của WebFilter đó là phải tổng hợp các gói tin Text HTML để ra một trang web hoàn chỉnh. Như trên ta thấy phiên giao dịch này cho biết text file có kích thước 46071 bytes, với kích thước tương đối của một

gói tin khoảng 1500 bytes, như vậy phiên giao dịch này sẽ có khoảng 30 gói tin.

Trọn vẹn gói tin HTTP đầu tiên của phiên giao dịch có nội dung như sau:

Hình 3.37. Nội dung gói tin đầu tiên của một phiên giao dịch web.

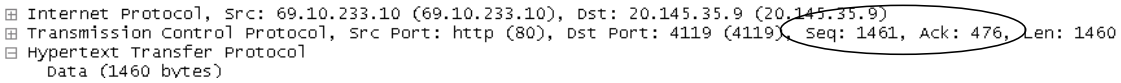
Công việc tiếp theo đó là tập hợp các gói tin kế tiếp của phiên giao dịch. Vậy làm sao để biết gói tin nào thuộc về phiên giao dịch này? Gói tin nào sẽ nối tiếp gói tin này theo đúng thứ tự của phiên? Như phần trên đã nói về vấn đề xác định phiên giao dịch, dựa trên các tham số: địa chỉ ip, số cổng, số ACK, SEQUENT.



Hình 3.38. Trong một gói tin luôn chứa tham số Seq và Ack

WebFilter sẽ kiểm tra các gói tin có các tham số như trên và Seq theo qui luật, để ghép nội dung gói tiếp theo vào phiên. Cụ thể trong ví dụ này, gói

tiếp theo như sau, tham số Ack là của gói tin trước nó, tham số Seq bằng tổng của Seq của gói tin trước và kích thước gói tin trước nó (1+1460=1461):

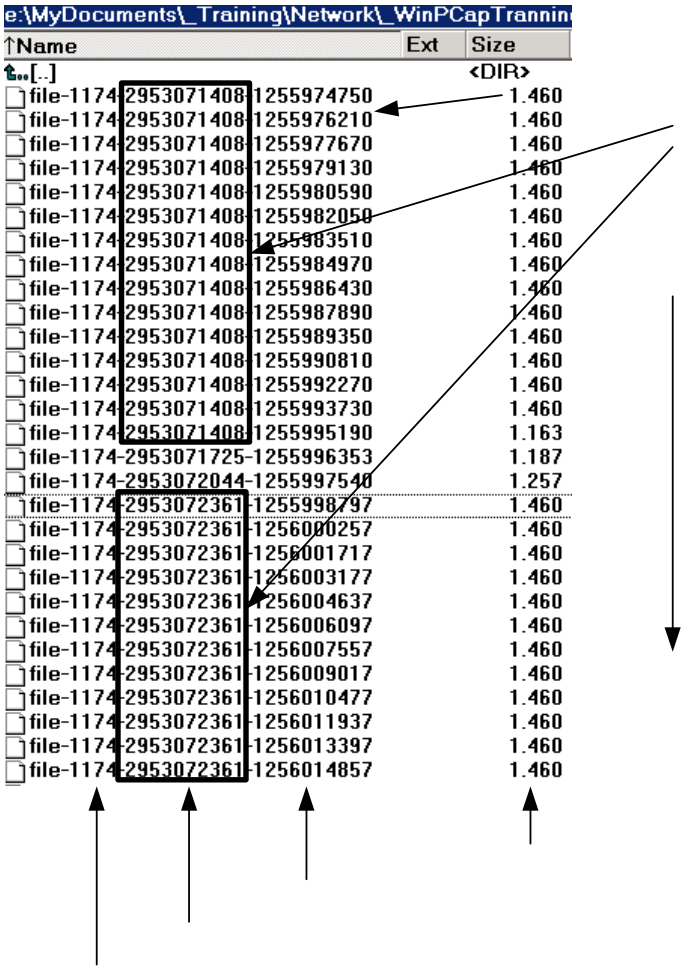


Hình 3.39. Một gói tin tiếp theo với tham số Seq và Ack

Theo cách như vậy, WebFilter có thể tổng hợp các gói tin tiếp theo của phiên giao dịch.

Còn một vấn đề đó là phiên giao dịch được kết thúc bằng tín hiệu nào? WebFilter xác định kết thúc phiên giao dịch tại thời điểm nó nhận thấy một gói tin đến tại 1 cổng và Ack đã bị thay đổi. Hoặc trong tình huống thứ hai là trường hợp dữ liệu của phiên giao dịch đã nhận được đủ: tổng số **Len** cho biết kích thước (xem trên hình) bằng kích thước phiên giao dịch, đã được xác định trong Header trên (46071 bytes).

Hình dưới đây mô tả thứ tự các gói tin của phiên giao dịch.



Hình 3.40. Tổng hợp các gói tin đơn lẻ thành phiên giao dịch HTTP.

Sau khi đã tổng hợp được hoàn chỉnh một phiên giao dịch Text, bước tiếp theo sẽ là phân tích nội dung phiên giao dịch đó, cụ thể là phân tích nội dung một text file.

*3.3.3.*4. Thành phần phân tích nội dung trang web

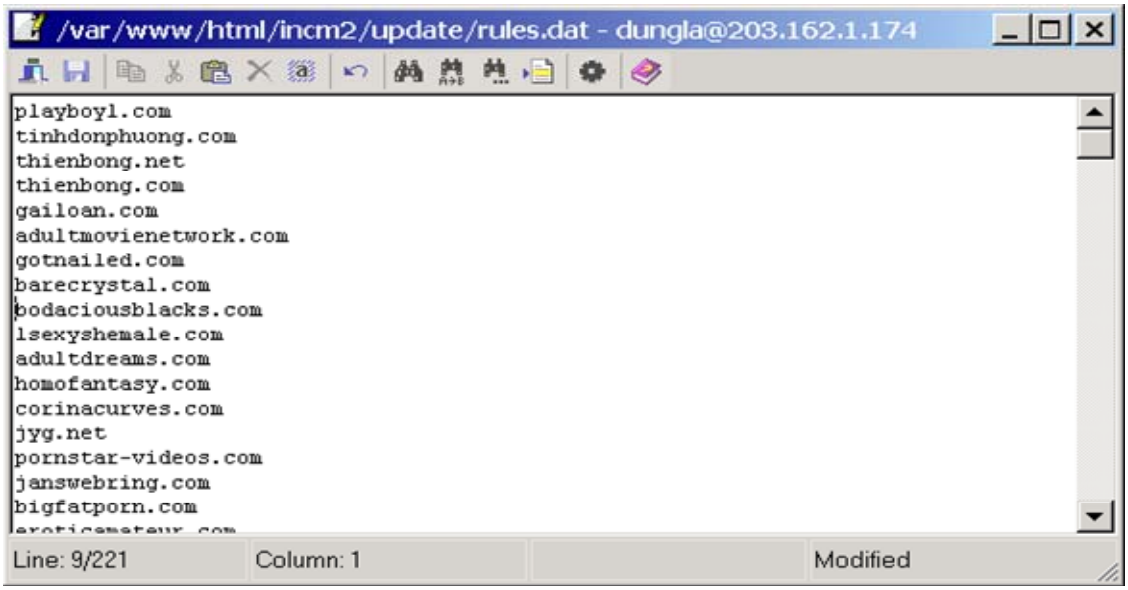
Giả sử ta thu được một nội dung trang web như sau, công việc của WebFilter lúc này là phân tích nội dung trang web.

Hình 3.41. Phân tích nội dung trang web dựa theo từ khoá

Trong phạm vi ứng dụng lọc chặn Web 'đen' của WebFilter, sẽ phân tích nội dung theo URL và từ khoá.

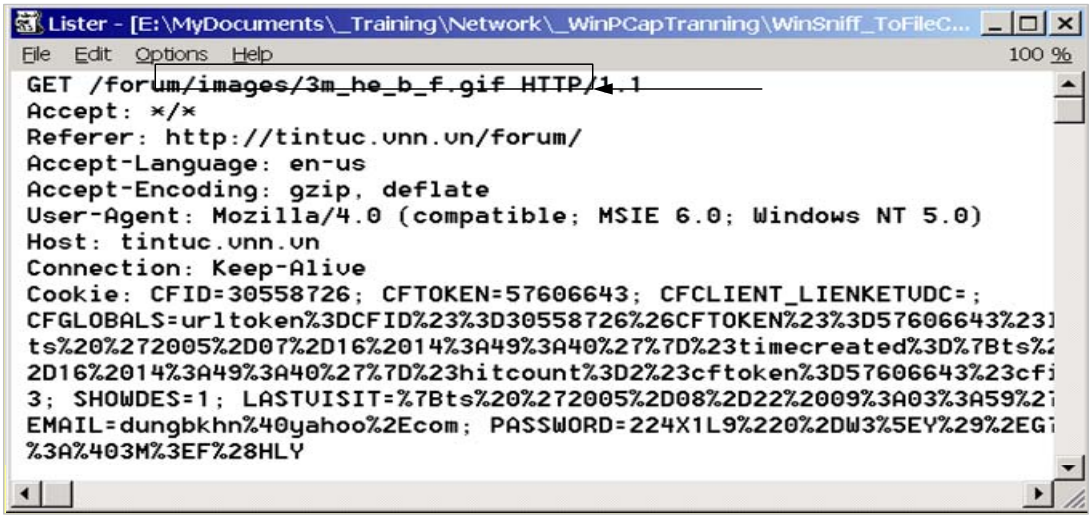
3.3.3.4.1. Lọc chặn theo URL

Công việc nhận như sau: theo dõi các gói tin header gửi và nhận là có thể phát hiện ra URL có vi phạm qui tắc hay không. Ví dụ ta có một danh sách lọc chặn như sau:



Hình 3. 42. Danh sách lọc chặn đang được áp dụng cho mạng VNN

WebFilter sẽ kiểm tra tất cả các *gói tin đi*, chẳng hạn một gói tin request như sau:



Hình 3. 43. Gói tin request đến trang tintuc.vnn.vn

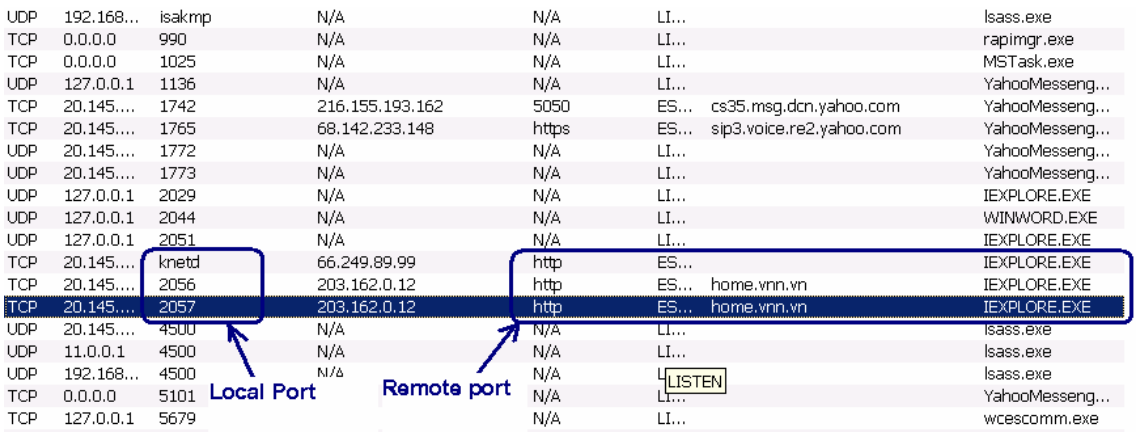
WebFilter sẽ nhận dạng ra dòng Referer trong gói tin, cho biết có tiến trình truy cập đên địa chỉ http://tintuc.vnn.vn/forum/. Nếu địa chỉ này nằm trong danh sách 'đen' ở trên, webfilter sẽ tác động vào trình duyệt đang truy cập trang web.

3.3.3.4.2. Lọc chặn theo từ khoá

WebFilter được thiết kế theo qui tắc tính điểm cho các từ khoá, ví dụ từ xxx với điểm là 10, từ sex với điểm là 20, từ nude với điểm là 30… Một trang web có 2 từ xxx, điểm sẽ là 20, có 3 từ sex thì điểm sẽ là 20x3=60 điểm… Tổng số điểm giới hạn được cho trước, giả sử là 100 - số điểm cho biết trang web này đã vi phạm nội dung. Giả sử trang web đó có ít nhất 10 từ khoá xxx, thì tổng số điểm ít nhất là 10x10=100 điểm, như vậy đó là một trang web 'đen' theo định nghĩa trên.

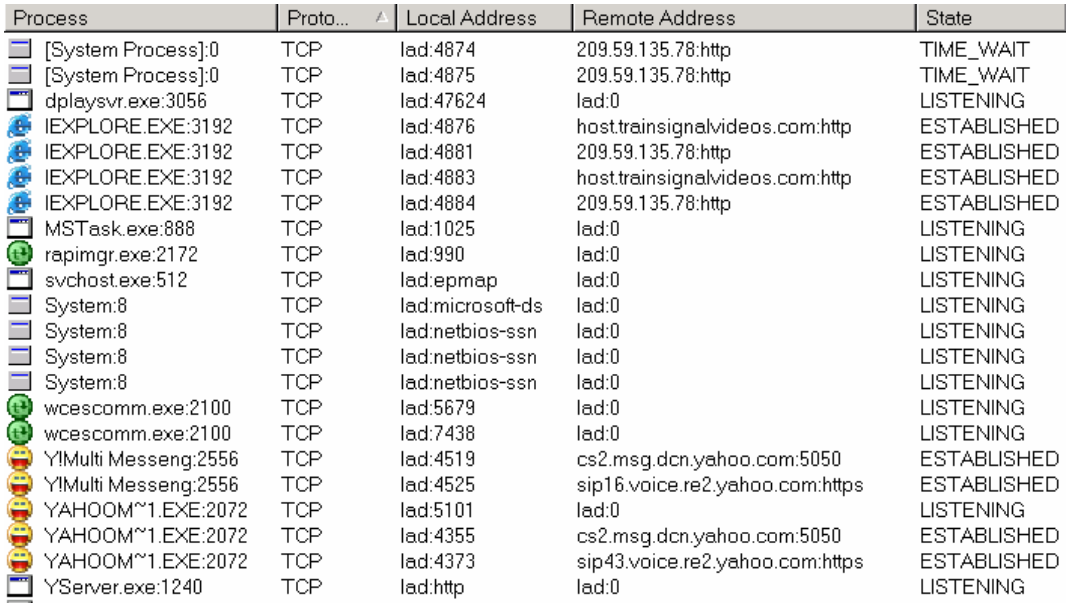
3.3.3.5. Thành phần quản lý cổng tiến trình và kết thúc tiến trình

Về cơ bản, WebFilter sẽ sử dụng các API (hàm lập trình hệ thống) để liệt kê các tiến trình và cổng truy cập tương ứng. Trong mỗi gói tin sẽ có cổng nguồn và cổng đích. Ở đây ta quan tâm đến các gói tin trên cổng 80 (remote), từ đó lấy được số cổng tương ứng trong gói tin, từ số cổng này tìm được tiến trình tương ứng. Đến đây có thể tác động vào tiến trình, có thể là kết thúc tiến trình, hoặc thay đổi giao diện của tiến trình… Nghĩa là có thể biến đổi thông tin tiến trình nhận được thành một dạng thông tin bất kỳ. Ví dụ khi phát hiện ra một tiến trình Internet Explorer đã truy cập vào một trang web có nội dung nào đó (chẳng hạn một trang web đen), ta có thể tác động vào tiến trình IE đó theo các cách như kết thúc tiến trình, hoặc thay đổi nội dung cửa sổ hiển thị…



Hình 3.44. Ánh xạ tên tiến trình và cổng tương ứng

Ứng dụng WebFilter trong trường hợp lọc web 'đen' cần biết tiến trình nào truy cập tương ứng với nội dung nào. Nhiệm vụ của module liệt kê tiến trình tương ứng với cổng TCP sẽ cho biết điều này. Tiến trình và số cổng TCP vốn không có sự liên hệ. Ta cần dùng các API (các hàm thư viện Windows) để tìm ra sự liên kết này.



Process	Proto...	Local Address	Remote Address	State
[System Process]:0	TCP	lad:4874	209.59.135.78:http	TIME_WAIT
[System Process]:0	TCP	lad:4875	209.59.135.78:http	TIME_WAIT
dplaysvr.exe:3056	TCP	lad:47624	lad:0	LISTENING
IEXPLORE.EXE:3192	TCP	lad:4876	host.trainsignalvideos.com:http	ESTABLISHED
IEXPLORE.EXE:3192	TCP	lad:4881	209.59.135.78:http	ESTABLISHED
IEXPLORE.EXE:3192	TCP	lad:4883	host.trainsignalvideos.com:http	ESTABLISHED
IEXPLORE.EXE:3192	TCP	lad:4884	209.59.135.78:http	ESTABLISHED
MSTask.exe:888	TCP	lad:1025	lad:0	LISTENING
rapimgr.exe:2172	TCP	lad:990	lad:0	LISTENING
svchost.exe:512	TCP	lad:epmap	lad:0	LISTENING
System:8	TCP	lad:microsoft-ds	lad:0	LISTENING
System:8	TCP	lad:netbios-ssn	lad:0	LISTENING
System:8	TCP	lad:netbios-ssn	lad:0	LISTENING
System:8	TCP	lad:netbios-ssn	lad:0	LISTENING
wcescomm.exe:2100	TCP	lad:5679	lad:0	LISTENING
wcescomm.exe:2100	TCP	lad:7438	lad:0	LISTENING
YIMulti Messeng:2556	TCP	lad:4519	cs2.msg.dcn.yahoo.com:5050	ESTABLISHED
YIMulti Messeng:2556	TCP	lad:4525	sip16.voice.re2.yahoo.com:https	ESTABLISHED
YAHOOM~1.EXE:2072	TCP	lad:5101	lad:0	LISTENING
YAHOOM~1.EXE:2072	TCP	lad:4355	cs2.msg.dcn.yahoo.com:5050	ESTABLISHED
YAHOOM~1.EXE:2072	TCP	lad:4373	sip43.voice.re2.yahoo.com:https	ESTABLISHED
YServer.exe:1240	TCP	lad:http	lad:0	LISTENING

Hình 3.45. Ánh xạ chi tiết Process number và số cổng

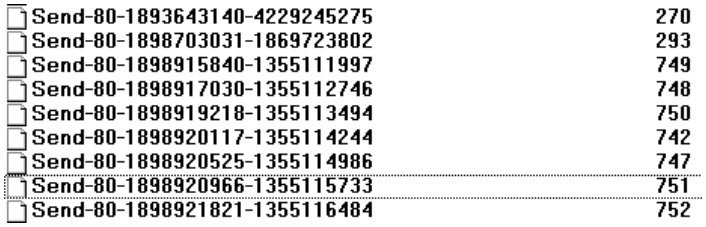
WebFilter sẽ có tác động thế nào nếu phát hiện có sự truy cập web đen? WebFilter sẽ tìm ra tiến trình nào truy cập web đen, và có tác động với tiến

trình đó. Tiến trình truy cập web đen sẽ là một trình duyệt nào đó (IE, Mozilar…). Tác động được lựa chọn ở đây đó là kết thúc tiến trình truy cập.

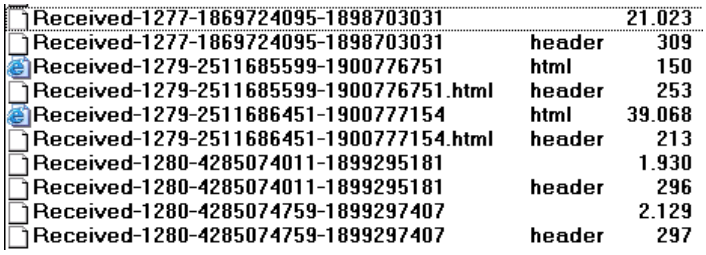
WebFilter sẽ dựa trên số cổng của phiên giao dịch HTTP có dấu hiệu truy cập web đen để tìm ra tiến trình tương ứng. Và động tác tiếp theo đó là đóng tiến trình đó lại hoặc thay đổi nộidung cửa sổ, đưa ra thông báo…

3.3.3.6. Thành phần tổng hợp ghi log truy cập

Trong quá trình Sniffer, webfilter có thể ghi log toàn bộ quá trình truy cập mạng vào và ra máy tính. Các log file có thể được ghi ra file một cách đầy đủ như sau, gồm các gói tin gửi đi và nhận lại:



Các gói tin request HTTP gửi đi



Hình 3.46. Các gói tin gửi/nhận được tổng hợp thành session và ghi ra file tương ứng

3.3.3.7. Thành phần tổng hợp trang web dưới dạng nén Gzip

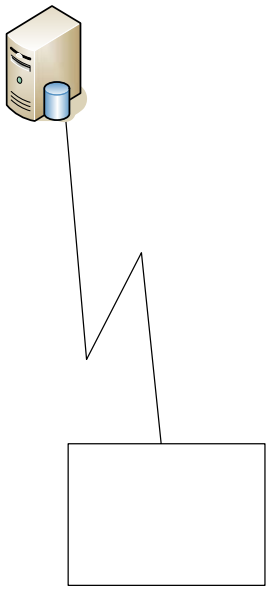
Hầu hết các WebServer đều có chế độ nén TextHTML để giảm thiểu dung lượng truy cập. Vì thực tế Text file có độ nén rất cao, dung lượng có thể giảm đi đến 10 lần, tải băng thông server vì thế cũng giảm bớt.

Thông thường, sau khi client truy cập lấy được nội dung file nén đó, trình duyệt sẽ giải mã (unzip) để hiển thị lên giao diện trình duyệt. Vì vậy WebFilter khi nghe lén các gói tin dạng HTML thì nội dung không phải là text nữa, mà mỗi gói tin là một phần của file nén Gzip. Do đó, khi nhận toàn bộ một phiên giao dịch HTML là nhận được một file nén Gzip. Công việc còn lại của WebFilter là giải nén file này ra để nhận về nguyên vẹn trang HTML (như những gì mà người sử dụng nhìn thấy) mới có thể tiếp tục công đoạn phân tích trang đó.

Việc xử lý Gzip chỉ tiến hành với Windows WebFilter, đối với Gateway WebFilter không cần bước xử lý này, bởi vì đã có module WebCache đảm nhiệm công việc này.

3.3.3.8. Thành phần liên kết server cập nhật địa chỉ lọc chặn, từ khoá lọc chặn

Trong sơ đồ ứng dụng thực tế của WebFilter, có cơ chế cho phép WebFilter cập nhật nội dung lọc chặn thường xuyên với Server quản lý của cơ quan chức năng.



Hình 3.47. Cập nhật chính sách lọc chặn từ policy server.

Nguyên tắc cập nhật như sau: theo thời gian định kỳ, WebFilter sẽ kết nối với Policy Server. WebFilter sẽ tải mã số đánh số Version luật lọc chặn trên policy server, nếu phát hiện số phiên bản luật lọc chặn đã thay đổi, WebFilter sẽ download bản luật lọc chặn mới nhất về Client (PC). Ngay sau đó WebFilter sẽ cập nhật thời gian thực cho hệ thống lọc chặn để đáp ứng ngay luật lọc chặn mới. Hiện tại, WebFilter được thiết kế có thể cập nhật một phút một. Nghĩa là nếu có luật lọc chặn mới, thì chỉ vài phút sau là hệ thống client sẽ cập nhật đầy đủ và tiến hành lọc chặn ngay sau đó.

3.3.4. Thiết kế chi tiết

3.3.4.1. MODULE SERVER

Chú ý: Chỉ cần thiết kế UML cho chức năng quản lý URL, các chức năng quản lý DOMAIN và KEYWORD tương tự như chức năng quản lý URL

Mô tả bài toán

Xây dựng phần mềm phía Server quản lý các danh sách trắng/đen cho phép:

- Thêm URL
- Sửa URL
- Xóa URL
- Tìm kiếm URL

Thiết kế

- Xác định các tác nhân, use case và mô tả chi tiết các use case

Xác định các tác nhân (actor)

Chức năng của các actor như sau:

Actor	Chức năng
Adminstrator	Quản trị hệ thống
Workstation	Cập nhật rule từ server

Xác định các use case

a. Phía Server

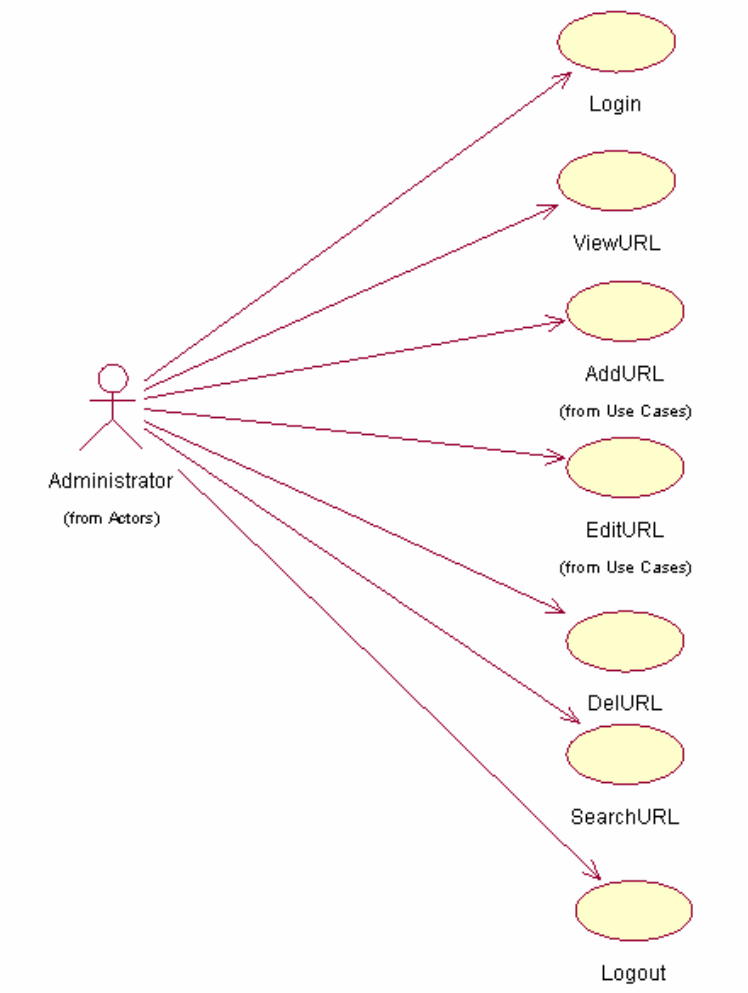
Login:	Đăng nhập vào hệ thống
ViewURL:	Hiển thị danh sách
AddURL:	Thêm URL
EditURL:	Thay đổi URL
DelURL:	Xóa URL
SearchURL:	Tìm kiếm URL
Logout:	Đăng xuất khỏi hệ thống

b. Phía Client

UpdateURL:	Máy trạm cập nhật URL từ server

Biểu đồ các use case

a. Phía Server



Hình 3.48. Biểu đồ use-case phía server

b. Phía Client

Hình 3.49. Biểu đồ use-case phía client tương tác

Mô tả các use case

a. Phía Server

Login

- Tên: Login
- Mô tả khái quát: Người quản trị đăng nhập vào hệ thống
- Luồng sự kiện

Administrator	Hệ thống
1. Nhập user/pass	2. Kiểm tra user/pass, xác định quyền hạn 2.1 Nếu đúng thì ghi lại thời điểm đăng nhập; xác định quyền hạn hiển thị form chương chính của chương trình. 2.2 Nếu sai yêu cầu nhập lại.

- Tiền điều kiện: Nhập đúng user/pass.
- Hậu điều kiện: Login vào hệ thống.

AddURL

- Tên: CreateURL
- Mô tả khái quát: Người quản trị thêm URL mới vào
- Luồng sự kiện

Administrator	Hệ thống
1. Chọn chức năng thêm URL	2. Hiển thị form cho admin nhập URL
3. Điền các thông tin và chọn submit	4. Kiểm tra thông tin nhập vào 4.1 Nếu đúng thì lưu và cơ sở dữ liệu 4.2 Nếu sai yêu cầu nhập lại
5.2 Nhập lại thông tin	

- Tiền điều kiện: Nhập URL

- Hậu điều kiện: URL mới được thêm vào

EditURL

- Tên: EditURL
- Mô tả khái quát: Người quản trị sửa URL
- Luồng sự kiện

Administrator	Hệ thống
1. Chọn chức năng thay đổi URL	2. Hiển thị form cho admin thay đổi URL
3. Thay đổi URL và chọn submit	4. Kiểm tra thông tin nhập vào 4.1 Nếu đúng thì lưu vào cơ sở dữ liệu 4.2 Nếu sai yêu cầu nhập lại
4.3 Nhập lại thông tin	

- Tiền điều kiện: Thay đổi thông tin URL
- Hậu điều kiện: URL được thay đổi

DelURL

- Tên: DelURL
- Mô tả khái quát: Người quản trị xóa URL
- Luồng sự kiện

Administrator	Hệ thống
1. Chọn chức năng xóa URL	2. Hiển thị một message cảnh báo, xác nhận việc xóa 2.1 Xác nhận xóa 2.2 Không xóa
3.1 Xóa người dùng đó 3.2 Không xóa người dùng	4.1 Gửi thông tin xóa thành công hoặc không cho admin

- Tiền điều kiện: Yêu cầu xóa URL

- Hậu điều kiện: URL bị xóa

SearchURL

- Tên: SearchURL
- Mô tả khái quát: Người quản trị tìm kiếm URL
- Luồng sự kiện

Administrator	Hệ thống
1. Nhập từ khóa cần tìm	2. Thực hiện yêu cầu và trả về kết quả (dưới dạng form)

- Tiền điều kiện: Nhập từ khóa tìm kiếm
- Hậu điều kiện: Kết quả được hiển thị

Logout

- Tên: Logout
- Mô tả khái quát: Người quản trị đăng xuất
- Luồng sự kiện

Administrator	Hệ thống
1. Người quản trị chọn logout	2. Đưa ra cảnh báo có thực sự muốn thoát khỏi chương trình hay không?
2.1 Xác nhận thoát	2.1.1 Lưu thông tin về thời gian kết thúc của người quản lý.

- Tiền điều kiện: Xác nhận logout.
- Hậu điều kiện: Logout khỏi hệ thống.

b. Phía Client

UpdateURL

- Tên: UpdateURL
- Mô tả khái quát: Máy trạm yêu cầu cập nhật danh sách URL mới
- Luồng sự kiện

Workstation	Hệ thống

1. Máy trạm yêu cầu cập nhật danh sách	2. Gửi danh sách cho máy trạm
3. Cập nhật danh sách	

- Tiền điều kiện: Yêu cầu danh sách URL mới
- Hậu điều kiện: Workstation cập nhật vào danh sách.

3.3.4.2. MODULE CLIENT

Mô tả bài toán

Bài toán đặt ra là phải xây dựng một hệ thống phần mềm lọc nội dung tại máy cá nhân. Hệ thống này có các chức năng lọc chặn web "đen" theo URL và từ khóa, quản lý danh sách lọc chặn, cập nhật danh sách lọc chặn tự động.

Nội dung của tài liệu này là thiết kế chi tiết hệ thống lọc nội dung máy cá nhân phía người dùng. Module Client bao gồm các chức năng sau:

- Lọc chặn web đen theo URL/DOMAIN
- Lọc chặn web đen theo từ khóa
- Cập nhật rule lọc chặn tự động

Thiết kế chi tiết

(Chú ý: Chỉ cần thiết kế chức năng lọc chặn theo URL, chức năng lọc chặn theo DOMAIN tương tự như lọc chặn theo URL)

a) Xác định các tác nhân (actor), ca sử dụng (use case)

Xác định các tác nhân

Actor	Chức năng
User	Người sử dụng Internet để truy cập web
Server	Server quản lý danh sách lọc chặn

Xác định các Use Case

1.Cl_FilterURL: lọc chặn theo URL

1.1.Cl_GetURL: Lấy URL mà người dùng truy cập

1.2.Cl_FindURL: Tìm trong danh sách lọc chặn xem có URL mà người dùng truy cập không

1.3.Cl_DropConnectionURL: Chặn kết nối đến website nếu URL đó không được phép

1.4.Cl_AllowConnectionURL: Cho phép kết nối tới website nếu URL đó hợp lệ

2.Cl_FilterKeyword: lọc chặn theo từ khóa

2.1.Cl_GetContent: Lấy nội dung website người dùng truy cập

2.2.Cl_FindKeyword: Tìm xem trong nội dung website có chứa keyword bị chặn hay không, nếu có thì tổng trọng số của từ khóa đó là bao nhiêu (tính theo số lần xuất hiện từ khóa)

2.3.Cl_TotalKeyword: Tính tổng trọng số các keyword bị chặn xuất hiện trong nội dung website. Kiểm tra xem tổng trọng số đó có vượt quá giới hạn cho phép không

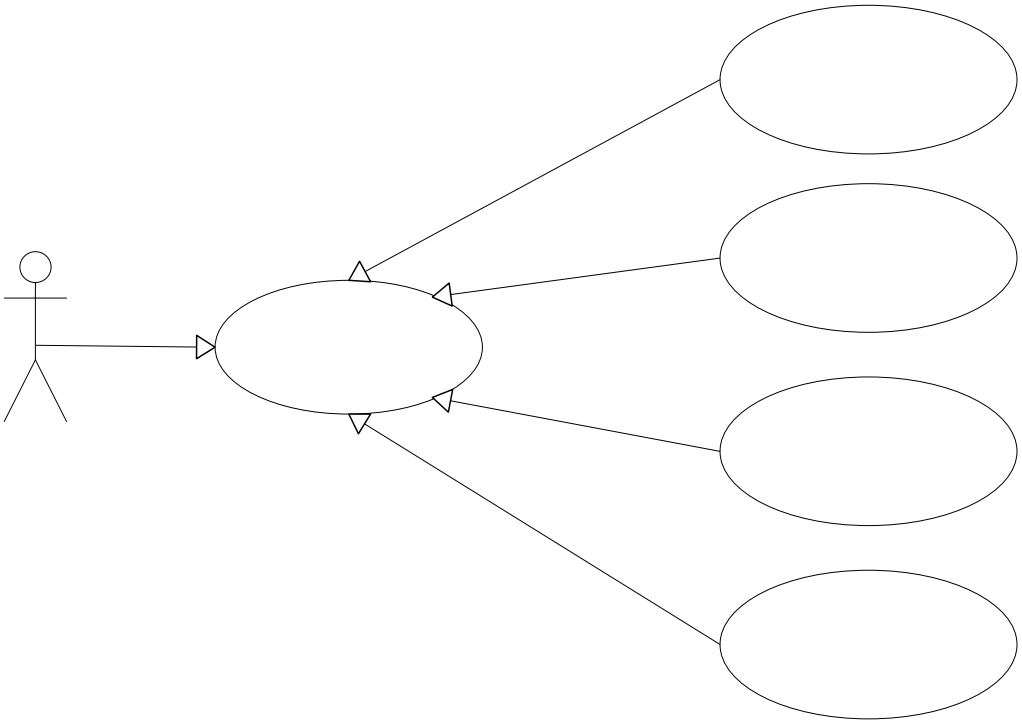
2.4.Cl_ DropConnectionKW: Chặn kết nối đến website nếu tổng trọng số các keyword bị chặn vượt quá giới hạn cho phép

2.5.Cl_ AllowConnectionKW: Cho phép kết nối tới website tổng trọng số các keyword bị chặn không vượt quá giới hạn cho phép

3.Cl_UpdateRule: cập nhật các Rule lọc chặn (cả rule cho URL và rule cho Keyword).

Biểu đồ các Use Case

Lọc chặn theo URL

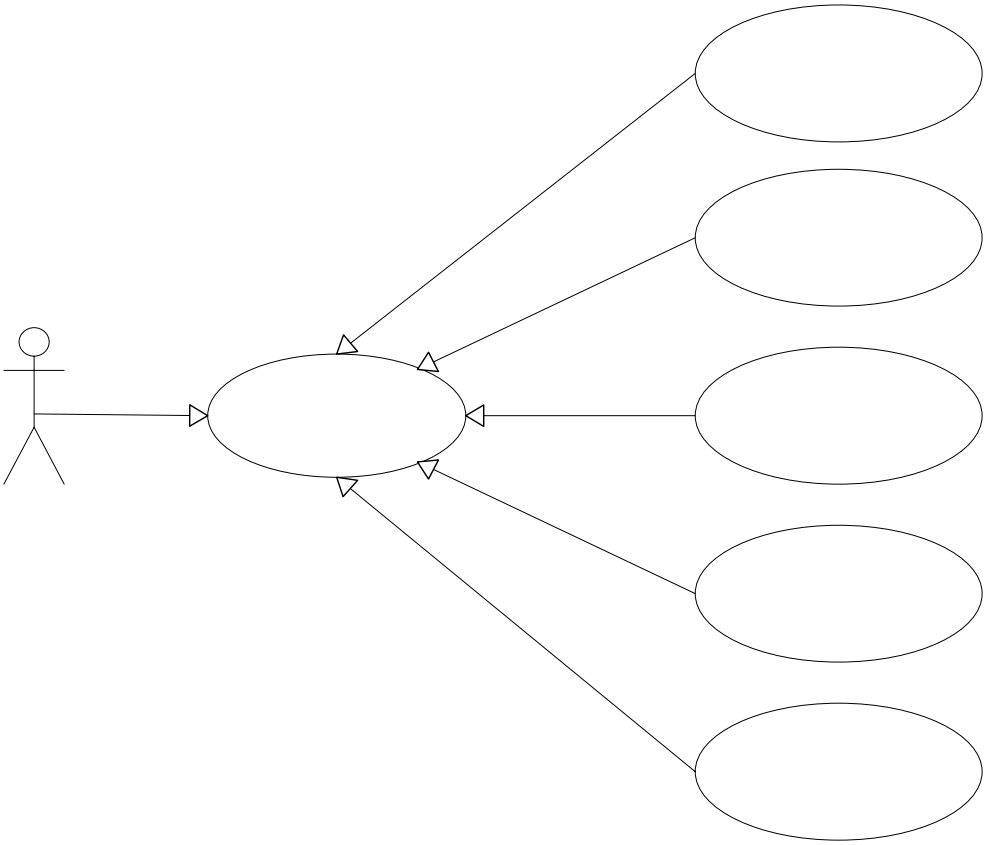


Hình 3.50. Biểu đồ Use Case gói lọc chặn theo URL

1.Cl_FilterURL

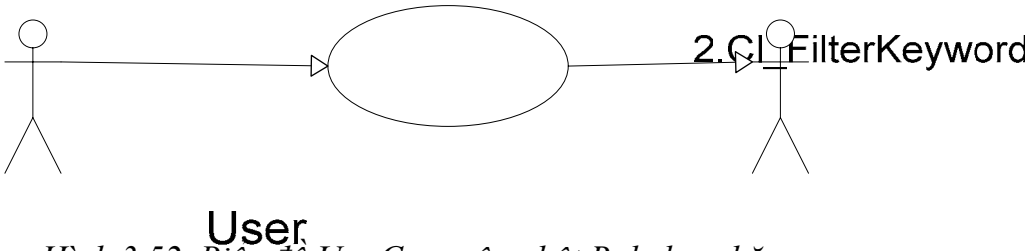
User

Lọc chặn theo Keyword



Hình 3.51. Biểu đồ Use Case gói lọc chặn theo Keyword

Cập nhật Rule lọc chặn



Hình 3.52. Biểu đồ Use Case cập nhật Rule lọc chặn

Mô tả các Use Case

Lọc chặn theo URL

Cl_GetURL

- Tên: Cl_GetURL

- Mô tả khái quát: chương trình bắt URL do người sử dụng truy cập
- Luồng sự kiện:

User	System
1. Người dùng mở một URL	2. Phần mềm sẽ bắt URL trên trình duyệt do người dùng truy cập

- Tiền điều kiện: người dùng mở một URL
- Hậu điều kiện: có được URL mà người dùng truy cập

Cl_FindURL

- Tên: Cl_ FindRule
- Mô tả khái quát: Kiểm tra xem URL mà người dùng truy cập có trong danh sách lọc chặn không
- Luồng sự kiện:

System
1. Tìm kiếm trong danh sách lọc chặn có URL mà người dùng truy cập không 2. Trả kết quả tìm kiếm (có/không)

- Tiền điều kiện: có URL mà người dùng truy cập
- Hậu điều kiện: kết quả tìm kiếm: URL người dùng truy cập có xuất hiện trong danh sách lọc chặn hay không

Cl_DropConnectionURL

- Tên: Cl_ DropConnectionURL
- Mô tả khái quát: chặn truy cập vào URL nếu URL đó không được phép
- Luồng sự kiện:

User	System
2. Hiển thị thông báo cho người dùng	1. Chặn truy cập vào URL

- Tiền điều kiện: URL người dùng truy cập không hợp lệ
- Hậu điều kiện: đưa ra thông báo URL đã bị chặn

1.4.Cl_AllowConnectionURL

- Tên: Cl_ AllowConnectionURL
- Mô tả khái quát: mở truy cập vào URL nếu URL đó được phép
- Luồng sự kiện:

User	System
2. Người dùng có thể xem nội dung URL	1. Mở truy cập vào URL

- Tiền điều kiện: URL người dùng truy cập hợp lệ.
- Hậu điều kiện: người dùng có thể xem được nội dung URL.

Lọc chặn theo Keyword

Cl_GetContent

- Tên: Cl_ GetContent
- Mô tả khái quát: lấy nội dung URL người dùng truy cập
- Luồng sự kiện:

User	System
1. Người dùng truy cập vào một URL	2. Lấy nội dung URL mà người dùng truy cập (trước khi người dùng có thể xem được nội dung)

- Tiền điều kiện: người dùng truy cập vào một URL.
- Hậu điều kiện: có được nội dung URL người dùng truy cập.

Cl_FindKeyword

- Tên: Cl_ FindKeyword

- Mô tả khái quát: tìm trong danh sách các từ khóa cần chặn, những từ nào xuất hiện trong nội dung lấy được, và những từ đó xuất hiện bao nhiêu lần, nhân với trọng số của từ đó ra tổng trọng số ứng với từng từ khóa.
- Luồng sự kiện:

System
1. Tìm kiếm trong danh sách các từ khóa cần chặn, những từ nào xuất hiện trong nội dung URL lấy được 2. Tính tổng trọng số ứng với mỗi từ khóa (trọng số nhân với số lần xuất hiện)

- Tiền điều kiện: có được nội dung URL người dùng truy cập.
- Hậu điều kiện: có tổng trọng số tương ứng của các từ khóa bị chặn xuất hiện trong nội dung URL đó.

Cl_TotalKeyword

- Tên: Cl_ TotalKeyword
- Mô tả khái quát: tính tổng trọng số của tất cả các từ khóa
- Luồng sự kiện:

System
1. Tính tổng trọng số của tất cả các từ khóa 2. Kiểm tra xem tổng trọng số có vượt quá giới hạn cho phép hay không

- Tiền điều kiện: có tổng trọng số tương ứng của các từ khóa bị chặn xuất hiện trong nội dung URL mà người dùng truy cập.
- Hậu điều kiện: website bị chặn/không.

Cl_ DropConnectionKW

- Tên: Cl_ DropConnectionKW
- Mô tả khái quát: chặn truy cập vào URL nếu tổng trọng số các từ khóa vượt quá giới hạn cho phép
- Luồng sự kiện:

User	System
2. Hiển thị thông báo cho người sử dụng	1. Chặn truy cập vào URL

- Tiền điều kiện: tổng trọng số các từ khóa vượt quá giới hạn cho phép.
- Hậu điều kiện: người dùng không thể truy cập vào URL đó.

Cl_ AllowConnectionKW

- Tên: Cl_ AllowConnectionKW
- Mô tả khái quát: mở truy cập vào URL nếu tổng trọng số các từ khóa
- Luồng sự kiện:

User	System
2. Người dùng có thể xem nội dung URL	1. Mở truy cập vào URL

- Tiền điều kiện: tổng trọng số các từ khóa không vượt quá giới hạn cho phép.
- Hậu điều kiện: người dùng có thể xem được nội dung URL.

Cập nhật Rule lọc chặn

Cl_UpdateRule

- Tên: Cl_ UpdateRule
- Mô tả khái quát: tự động tải các Rule lọc chặn (URL+Keyword) về máy cục bộ.

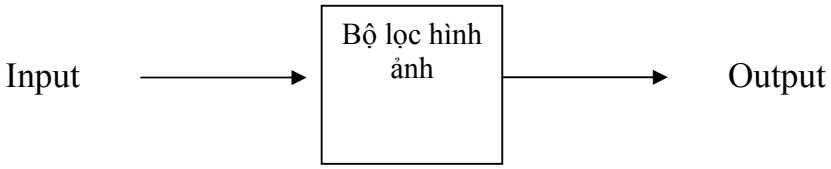
- Luồng sự kiện:

System	Server
2. Kiểm tra sự thay đổi về Rule lọc chặn định kỳ theo thời gian 3. Tự động tải Rule lọc chặn (DOMAIN+URL+Keyword)	1. Rule lọc chặn bị thay đổi

- Tiền điều kiện: thời điểm cập nhật định kỳ.
- Hậu điều kiện: Rule lọc chặn được tải về máy cục bộ (nếu có thay đổi).

3.3.4.3. Chức năng lọc hình ảnh

Module được xây dựng để nhằm đánh giá những bức ảnh xuất hiện trên internet về nội dung của nó có đảm bảo tính lành mạnh, văn hóa hay không, và theo yêu cầu đó thì module có mô hình đề xuất[2] là :



Input: Ảnh có dạng BMP, JPEG, GIF, PNG

Output:

- Độ xấu của ảnh đầu vào, đánh giá trên thang điểm 0-100 tương ứng từ tốt đến
- xấu
- Thể loại của ảnh đầu vào

Giải thuật phân tích ảnh

Căn cứ vào mô hình và lý thuyết liên quan module được thiết kế như sau:

b1: Xác định các vùng da của bức ảnh nhập vào.(sử dụng MaxEnt) Nhập ảnh vào → xác định kích cỡ ảnh,số lượng điểm ảnh →dãy các mảng chứa điểm ảnh(mảng 2 chiều về các điểm ảnh của bức ảnh)

Thực hiện quét ảnh:

Trong khi chưa quét hết các điểm ảnh(quét hết chiều dọc và chiều ngang của bức ảnh)

Xác định xác suất a[i][j] là da theo [1, 5, 6] ,nhị phân hóa các điểm ảnh tương ứng dựa trên ngưỡng,

với 0 <= i <= kích cỡ dọc của ảnh

0 <= j <= kích cỡ ngang của ảnh

Nếu mức độ da tại điểm a[i][j] = 1→ a[i][j] là điểm da

Gán các a[i][j] là màu trắng.//có thể thay bằng màu đánh dấu khác.

Còn lại thì: Gán màu của các a[i][j] là màu đen

b2: Từ các điểm màu trắng xác định trong bước 1 ta xây dựng các vùng màu trắng gọi là vùng da và tập các vùng được xây dựng gọi là bản đồ da theo[1]

Từ các ảnh được nhị phân hóa bằng cách sử dụng phân ngưỡng ở trên.

Tiếp theo các phép biến đổi hình thái *đóng/mở* được áp dụng để loại bỏ các nhiễu và nối các vùng bị dời.

Các vùng da nhỏ được coi là có tác động không đáng kể và bị loại bỏ.

Sau bước này kết quả là các vùng da xác định.

b3: Xác định đặc tính của vùng da từ bản đồ da thiết lập ở bước 2 (sử dụng GFE, LFE)

Xét duyệt những mảng chứa điểm da,

chọn lựa vùng có số lượng da lớn nhất.: max

Sử dụng GFE, LFE…tính toán tìm ra vector đặc trưng.

Kết thúc bước này, kết quả là vector đặc trưng.

b4: Từ thuộc tính xây dựng ở bước 3 ta xác định mức độ tốt, xấu của bức ảnh, và định dạng.

Trong khi lỗi còn lớn hơn ***w*** thì :

Điều chỉnh lại mẫu vào

Khi nào ổn định bằng w thì cho ra kết quả t.

Kết quả ra là ***t***.

m tính toán theo FP và TP

Nếu ***t*** → 1 hay (***t*** >=***m***) thì là ảnh khiêu dâm.

Nếu ***t*** →0 hay (***t***<***m***) thì là ảnh bình thường . Kết quả của bước này là đánh giá được mức độ tốt xấu của ảnh.

b5: In ra thông báo rằng nội dung của ảnh : Ảnh khiêu dâm? ảnh bình thường?

Hoặc gửi thông điệp, kết quả tới bộ kiểm soát và bộ ra quyết định trong hệ thống để hệ thống ra quyết định cuối cùng.

3.3.4.6. Môi trường

- Module Client:
  - Chạy trên nền tảng Windows 32 bit.
  - Định hướng cải tiến phần mềm để có thể tương thích tốt với nền tảng Windows 64 bit.
  - Nâng cấp phần mềm để có các phiên bản chạy trên hệ điều hành Linux, MAC,…
- Module Server:
  - Chạy trên hệ điều hành Linux hoặc Solaris.
  - Cơ sở dữ liệu: Postgres.
  - Ngôn ngữ lập trình: PHP.

3.4. ĐÁNH GIÁ VÀ THỬ NGHIỆM

a) Dữ liệu thử nghiệm

Dữ liệu thử nghiệm dựa trên danh sách lọc chặn như sau:

Kịch bản thử nghiệm

- Thử nghiệm lọc chặn theo DOMAIN trên các trình duyệt như Internet Explorer, Firefox, Google Chrome,…
- Lặp Thử nghiệm lọc chặn theo URL trên các trình duyệt như Internet Explorer, Firefox, Google Chrome,…
- Thử nghiệm lọc chặn theo KEYWORD trên các trình duyệt như Internet Explorer, Firefox, Google Chrome,…
- Thử nghiệm lọc chặn theo hình ảnh trên các trình duyệt như Internet Explorer, Firefox, Google Chrome,…
- Đánh giá thời gian thực thi khi một website bị chặn, và khi 1 website không bị chặn.
- Lặp lại các thử nghiệm trên với danh sách lọc chặn lớn (trên 1000 bản ghi).
- Đánh giá thời gian thực thi khi danh sách lọc chặn lớn.
- Thử nghiệm cập nhật danh sách lọc chặn mới.

Chương IV

KẾT QUẢ ĐÀO TẠO, HỢP TÁC QUỐC TẾ VÀ SẢN PHẨM BỔ SUNG

4.1. Giới thiệu

Như đã được trình bày, lọc nội dung trên Internet là một vấn đề nghiên cứu thời sự đối với các khía cạnh về quản lý Nhà nước, về giải pháp kỹ thuật và công nghệ thi hành hệ thống và nhiều khía cạnh liên quan khác. Kết quả đào tạo, hợp tác quốc tế và sản phẩm bổ sung được trình bày trong chương này liên quan trực tiếp tới khía cạnh về giải pháp kỹ thuật và công nghệ thi hành hệ thống. Một nội dung chủ yếu khi thi hành các giải pháp lọc nội dung trên Internet chính là bài toán phân lớp trang Web trong điều kiện thời gian xử lý phải nhanh và dung lượng thông tin lớn. Một số vấn đề cơ bản sau đây cần được quan tâm giải quyết:

- *Biểu diễn trang web*: Lựa chọn các phương pháp biểu diễn trang Web (tiếng Anh, tiếng Việt) làm đầu vào cho bài toán phân lớp. Các bài toán nền tảng trong xử lý ngôn ngữ tự nhiên, cụ thể là xử lý tiếng Việt như tách từ, phân cụm từ, gán nhãn từ loại... cần được quan tâm nghiên cứu. Mặt khác, định danh ngôn ngữ và tách nội dung chính của trang web phục vụ bài toán biểu diễn trang web cũng cần được tiến hành.
- *Phân lớp văn bản nội dung trang Web*: Bài toán lọc nội dung trang Web có bản chất là bài toán phân lớp trang Web vào một số lớp theo quy định của quản lý Nhà nước mà đối với một trang web theo yêu cầu người dùng hệ thống cần nhận ra nó thuộc vào lớp nào trong các lớp nói trên. Một số bài toán con được đề cập ở đây. Thứ nhất, hệ thống lớp văn bản theo quy định của quản lý Nhà nước, trong đó nhãn lớp có thể được gán theo gợi ý từ nghiên cứu của ONI trong Chương 1. Thứ hai, cần giải quyết bài toán phân lớp đa lớp đối với nội dung trang Web.

Như đã được trình bày tại Chương 2, các thuật toán phân lớp văn bản được khảo sát, đánh giá để áp dung vào hệ thống lọc nội dung.

- Các nội dung liên quan tới lọc ảnh, lọc địa chỉ, kết hợp lọc nội dung với lọc địa chỉ, lọc tại máy tính cá nhân và tích hợp hệ thống cũng là đề tài nghiên cứu khoa học và công nghệ có tiềm năng.

Trong quá trình thực hiện đề tài, các thành viên trong nhóm thực hiện đã tích hợp các nội dung nghiên cứu khoa học - công nghệ vào quá trình triển khai thực hiện đề tài để khẳng định tính tin cậy của các thành phần và toàn bộ hệ thống lọc nội dung.

4.2. Kết quả đào tạo

Trong giai đoạn 2006-2009, nội dung nghiên cứu trực tiếp và gián tiếp về lọc nội dung trên Internet đã được lấy làm chủ đề luận văn Thạc sỹ cho hàng chục học viên cao học từ nhiều cơ quan, trong đó có Trường ĐHCN, Công ty VDC, Bộ Công an và các cơ quan khác. Ở mức độ cao hơn, có ba nghiên cứu sinh đã quan tâm tới các giải pháp lọc nội dung trên Internet trong quá trình hình thành nội dung luận án Tiến sỹ. Đồng thời, ở mức độ thấp hơn, các chủ đề thuộc xử lý văn bản tiếng Việt và trích chọn thông tin trên Internet cũng là đề tài của hàng chục khóa luận tốt nghiệp đại học.

4.2.1. Luận văn Thạc sỹ

Dưới đây mô tả sơ bộ 10 luận văn Thạc sỹ có nội dung liên quan tới nội dung đề tài:

Về xử lý ngôn ngữ tự nhiên, đánh giá các yếu tố trong biểu diễn văn bản

- Lê Đắc Nhường (2009). Tối ưu hóa truy vấn trong máy tìm kiếm thực thể, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009.
- Nguyễn Thu Trang (2009). Học xếp hạng trong tính hạng đối tượng và tạo nhãn cụm tài liệu, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009.

- Trần Thị Oanh (2009). Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009.
- Nguyễn Việt Cường (2007). Tự động sinh mục lục cho văn bản, *Luận văn Thạc sỹ*, Trường ĐHCN, 2007.
- Đặng Thanh Hải (2007). The Biological Sample Classification Using Gene Expression Data, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHCN, 2007.
- Nguyễn Hoài Nam (2006). The WWW and The PageRank-Related Problems, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHKHTN, 2006.

Về phân lớp, phân cụm văn bản và thư điện tử

- Nguyễn Cẩm Tú (2009). Hidden Topic Discovery Towards Classification and Clustering in Vietnamese Documents*, Luận văn Thạc sỹ (viết bằng tiếng Anh),* Trường ĐHCN, 2008.
- Ngô Thương Huyền (2008). Phân lớp thư điện tử sử dụng máy hỗ trợ vector, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008.
- Nguyễn Thị Thu Hằng (2008). Phương pháp phân cụm tài liệu Web và áp dụng vào máy tìm kiếm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008.

Về lọc nội dung tại máy tính cá nhân

- Phạm Tiến Dũng (2009). Nghiên cứu giải pháp lọc nội dung Internet tại máy tính cá nhân và xây dựng phần mềm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009.

4.2.2. Các nghiên cứu sinh tham gia thực hiện đề tài

Các nghiên cứu sinh dưới đây đã tham gia thực hiện đề tài và các kết quả nghiên cứu trong đề tài cũng góp phần vào nội dung luận án Tiến sỹ:

- NCS. Nguyễn Cẩm Tú với đề tài *"Mô hình trích chọn thông tin trên Web dựa trên các thuật toán máy trạng thái hữu hạn"* (2006-2008). Hiện NCS

Nguyễn Cẩm Tú đang làm luận án Tiến sỹ tại ĐH Tohoku, Nhật Bản. Nghiên cứu sinh tham gia thực hiện các giải pháp xử lý tiếng Việt như biểu diễn văn bản, phân lớp văn bản, phân cụm văn bản, trích chọn nội dung, giải pháp lọc nội dung và cũng là các nội dung nghiên cứu được đề cập khi thực hiện luận án..

- NCS. Nguyễn Việt Cường với đề tài *"Mô hình tích hợp thuật toán phát hiện quan hệ ngữ nghĩa trên Web"* (2006-2007). Hiện NCS Nguyễn Việt Cường đang làm luận án Tiến sỹ tại Viện Khoa học và Công nghệ tiên tiến Nhật Bản. Nghiên cứu sinh tham gia thực hiện các giải pháp phân lớp văn bản, mô tả văn bản, trích chọn nội dung và cũng là các nội dung nghiên cứu được đề cập khi thực hiện luận án.
- NCS. Đặng Thanh Hải hoàn thành chương trình đào tạo Thạc sỹ tốt nghiệp loại xuất sắc và được Trường ĐHCN đề nghị xem xét là chuyển tiếp sinh Tiến sỹ. Hiện NCS Đặng Thanh Hải đang làm luận án Tiến sỹ tại ĐH Antwerp, Bỉ. Nghiên cứu sinh tham gia thực hiện các giải pháp phân lớp văn bản, tính hạng (từ, cụm từ quan trọng, gene...) và cũng là các nội dung nghiên cứu được đề cập khi thực hiện luận án.

4.3. Kết quả nghiên cứu công bố khoa học

Trong quá trình thực hiện đề tài, nhóm nghiên cứu đã công bố 28 công trình khoa học trong đó có 4 bài báo đăng trên tạp chí khoa học quốc tế (2 bài có chỉ số ISI), 10 bài đăng kỷ yếu hội nghị khoa học quốc tế, và 14 bài công bố tại các hội bghị khoa học quốc gia (2 bài đăng kỷ yếu).

4.3.1. Bài báo đăng tạp chí khoa học quốc tế (Hai bài tạp chí thuộc ISI)

1. Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha (2009). Classification and Contextual Match on the Web with Hidden Topics from Large Data

Collections, *The IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (accepted, 14 pages)*. ISI Journal System.

2. Cam-Tu Nguyen, Xuan-Hieu Phan, Susumu Horiguchi, Thu-Trang Nguyen, Quang-Thuy Ha (2009). Web Search Clustering and Labeling with Hidden Topics, *ACM Transactions on Asian Language Information Processing,* **8**(3), 12 (August 2009), 40 pages. DOI=10.1145/1568292.1568295. http://doi.acm.org/10.1145/1568292.1568295. ISI Journal System.

3. Ha Q. Thuy, Nguyen H. Nam, Nguyen Thu Trang (2006). Improve Performance of PageRank Computation with Connected-Component PageRank, *ICMOCCA2006*: 154-158 & *International Journal of Natural Sciences and Technology*, **1**(1): 53-60, 2006,

4. Lan N. Bui, Anh Q. Tran, Thuy Q. Ha (2006). User authentic Rating based on Email Networks, *ICMOCCA2006*: 144-148, Seoul, Korea & *International Journal of Natural Sciences and Technology*, **1**(2): 173-180, 2006,

4.3.2. Báo cáo khoa học đăng kỷ yếu Hội nghị quốc tế (có 2-3 phản biện)

1. Tran Thi Oanh, Le Anh Cuong, Ha Quang Thuy and Quynh Hoang Le (2009). An Experimental Study on Vietnamese POS tagging, *International Conference on Asian Language Processing (IALP 2009)*, Dec 7-9, 2009, Singapore (*accepted,* http://ialp2009.colips.org/index.php?id=14&pg =acceptedlist*)*

2. Vu Tran, Vinh Nguyen, Uyen Pham, Oanh Tran and Quang Thuy Ha (2009). An Experimental Study of Vietnamese Question Answering System, *International Conference on Asian Language Processing (IALP*

2009), Dec 7-9, 2009, Singapore (*accepted,* http://ialp2009.colips.org/index.php?id=14&pg =acceptedlist*)*

3. Huong-Thao Nguyen, Phuong-Thai Nguyen, Quang-Thuy Ha, and Le-Minh Nguyen (2009). Vietnam Noun Phrase Chunking based on Conditional Random Field, *The First International Conference on Knowledge and System Engineering (KSE)*: 172-178*,* Hanoi, Vietnam, 2009.

4. Dieu-Thu Le, Cam-Tu Nguyen, Quang-Thuy Ha, Xuan-Hieu Phan, and Susumu Horiguchi (2008). Matching and Ranking with Hidden Topics towards Online Contextual Advertising, *The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08):* 888-891, University of Technology, Sydney, Australia, December 9 - 12, 2008. DOI: 10.1109/WIIAT.2008.180

5. Tran Thi Oanh, Le Anh Cuong, Ha Quang Thuy (2008). Improving Vietnamese Word Segmentation by Integrating Different Knowledge Resources, *The 2008 Empirical Methods for Asian Language Workshop (EMALP 2008):* 1-12*,* Hanoi, Vietnam, Dec. 13, 2008

6. Dang Thanh Hai, Wonjun Lee, Ha Quang Thuy (2008). A pageranking based method for identifying characteristic genes of a disease, *IEEE Proceeding of International Conference on Networking, Sensing and Control, 2008. ICNSC 2008:* 1496-1499, Sanya, China, 6-8 April 2008. DOI: 10.1109/ICNSC.2008.4525457

7. Do Thi Minh Viet, Nguyen Hai Chau, Wonjun Lee, Ha Quang Thuy (2008). Using Cross-layer Heuristic and Network Coding to Improve Throughput in Multicast Wireless Mesh Networks, *Information*

Networking (ICOIN 2008), Busan, Korea, January 23-25, 2008 (*The Best paper award*). DOI: 10.1109/ICOIN.2008.4472771

8. Nguyen Viet Cuong, Nguyen Thi Thuy Linh, Ha Quang Thuy and Phan Xuan Hieu (2006). A Maximum Entropy Model for Text Classification, *The International Conference on Internet Information Retrieval 2006:* 134-139, Hankuk Aviation University, Korea, Dec. 6, 2006

9. Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen and Quang-Thuy Ha (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation, *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20):* 215-222, November 1-3, 2006, Wuhan, China.

10. Son Doan, Quang Thuy Ha, and Susumu Horiguchi (2006). A General Fuzzy-based Framework for Text Representation and its Application to Text Categorization*, Lecture Notes on Artificial Intelligence (LNAI),* **4423**: 611-620, 2006 (Springer-Verlag Berlin Heidenberg) form The *Third International Conference on Fuzzy Systems and Knowledge Discovery* - FSKD 2006. DOI: 10.1007/11881599_73

4.3.3. Báo cáo tham gia hội thảo trong nước

1. Trần Mai Vũ, Phạm Thị Thu Uyên, Hoàng Minh Hiền, Hà Quang Thụy (2008). Độ tương đồng ngữ nghĩa giữa hai câu và áp dụng vào bài toán sử dụng tóm tắt đa văn bản để đánh giá chất lượng phân cụm dữ liệu trên máy tìm kiếm VNSEN, *Hội thảo CNTT & TT (ICTFIT08):* 94-102, ĐHKHTN, ĐHQG TP Hồ Chí Minh, Thành phố Hồ Chí Minh, 14/11/2008

2. Lê Diệu Thu, Trần Thị Ngân, Nguyễn Cẩm Tú, Nguyễn Thu Trang (2008). Xây dựng Ontology hỗ trợ tìm kiếm ngữ nghĩa trong linh vực y

tế , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã đăng kỷ yếu).

3. Trần Thị Oanh, Lê Hoàng Quỳnh, Lê Anh Cường, Hà Quang Thụy (2009). Một nghiên cứu về gán nhãn từ loại tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày).
4. Trần Nam Khánh, Pham Kim Cuong Nguyễn Thu Trang, Hà Quang Thụy (2009). Finding object-oriented information in unstructured data and adapting to Vietnamese real estate domain, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày).
5. Nguyễn Tiến Thanh, Trần Nam Khánh, Nguyễn Thu Trang, Hà Quang Thụy (2009). Xếp hạng các trường đại học Việt Nam dựa trên "độ đo web" , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày).
6. Nguyễn Thị Thu Chung, Nguyễn Thu Trang, Hà Quang Thụy (2009). Xây dựng danh bạ web tiếng Việt với phân cụm phân cấp văn bản , *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI,* Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày).
7. Trần Mai Vũ, Trần Thị Oanh, Nguyễn Đức Vinh, Phạm Thị Thu Uyên, Nguyễn Đạo Thái, Hà Quang Thụy (2009). Hệ thống hỏi đáp tự động tiếng Việt sử dụng trích rút mối quan hệ ngữ nghĩa trong kho văn bản tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ*

thông tin và Truyền thông lần thứ XI, Đồng Nai, 5-6/8/2009 (đã gửi toàn văn và trình bày).

8. Phạm Thị Thu Uyên, Hoàng Minh Hiền, Trần Mai Vũ, Hà Quang Thụy (2008). Độ tương đồng câu và áp dụng vào bài toán tóm tắt đa văn bản tiếng Việt, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày).
9. Nguyễn Thị Thu Chung, Nguyễn Thu Trang, Hà Quang Thụy (2008). Đánh giá chất lượng phân cụm trên máy tìm kiếm tiếng Việt VNSEN*, Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày).
10. Nguyễn Minh Tuấn, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2008). Một mô hình Maximize Entropy phân lớp câu hỏi tiếng Việt . *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày).
11. Nguyễn Thị Thùy Linh, Nguyễn Việt Cường, Hà Quang Thụy (2008). Một mô hình phân lớp đa nhãn SVM đối với văn bản tiếng Việt*, Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày).
12. Đặng Thanh Hải, Trần Thị Oanh, Hà Quang Thụy (2007). Thuật toán Co-training phân lớp Web tiếng Việt sử dụng thông tin liên kết, FAIR 07, Nha Trang, 6-8/8/2007 (Gửi toàn văn và trình bày báo cáo).
13. Nguyễn Thị Hương Thảo, Nguyễn Thị Thùy Linh, Nguyễn Thu Trang, Hà Quang Thụy (2007). Ứng dụng thuật toán học bán giám sát SVM

phân lớp văn bản tiếng Việt, FAIR 07, Nha Trang, 6-8/8/2007 (Gửi toàn văn và trình bày báo cáo)

14. Trần Thị Oanh, Lê Anh Cường, Hà Quang Thụy (2008). Phân đoạn từ tiếng Việt sử dụng Maxent kết hợp nhiều nguồn tri thức, *Hội thảo Quốc gia Một số vấn đề chọn lọc về Công nghệ thông tin và Truyền thông lần thứ XI*, Huế, 12-13/6/2008 (đã gửi toàn văn và trình bày)

4.4. Kết quả hợp tác quốc tế

Trong quá trình thực hiện đề tài, nhóm thực hiện đề tài đã cộng tác nghiên cứu với các nhà khoa học, các nhóm nghiên cứu tại các cơ sở đào tạo nghiên cứu tiên tiến trên thế giới về các nội dung liên quan, điển hình là các hợp tác nghiên cứu với các nhà khoa học sau đây:

- GS. Susumu Horiguchi (Susumu Horiguchi Lab., Tohoku University, Japan). Đã cộng tác công bố một số công trình khoa học liên quan tới bài toán phân lớp trong lọc nội dung với kết quả nghiên cứu được thực hiện một phần tại Trường ĐHCN [PNL08, NNH08, LNH08, SQS06]. Từ tháng 10/2008 tới tháng 6/2009, GS. S. Horiguchi hướng dẫn Nghiên cứu sinh Nguyễn Cẩm Tú. Không may, GS. S. Horiguchi đã qua đời vào tháng 6/2009.
- GS. Akira Shimazu (Akira Shimazu Lab, JAIST, Japan): Nhóm đã thiết lập quan hệ với GS.Akira Shimazu về lĩnh vực Xử lý ngôn ngữ tự nhiên, có quan hệ gần gũi với lĩnh vực phân lớp văn bản và lọc nội dung trên Internet. GS. Akira Shimazu hiện đang hướng dẫn luận án Tiến sỹ cho Nghiên cứu sinh Nguyễn Việt Cường.
- PGS. Kevin Chang (the Database and Information Systems Laboratory

(DAIS)[19], Department of Computer Science at the University of Illinois at Urbana-Champaign): Nhóm thiết lập quan hệ với PGS. Kevin Chang từ tháng 7/2008 trong việc hợp tác nghiên cứu và phát triển mô hình, các phương pháp giải các bài toán thành phần trong trích chọn thông tin nội dung trang Web, xây dựng các hệ thống tìm kiếm thực thể.

4.5. Sản phẩm bổ sung: phần mềm tìm kiếm giá cả sản phẩm

4.5.1.Tính năng chính của sản phẩm

Hệ thống máy tìm kiếm giá cả sản phẩm được nghiên cứu, xây dựng và phát triển luôn tuân thủ và đảm bảo 4 tiêu chí chủ đạo :

* Tính độc đáo
* Tính chính xác
* Tính đầy đủ
* Tính hướng người dùng

Từ những tiêu chí đó, các tính năng của hệ thống được nghiên cứu và xây dựng luôn đảm bảo và tuân thủ những yêu cầu được đặt ra.

a) Tính năng tự động phát hiện các trang web bán hàng trực tuyến và sinh luật bóc tách

Thông thường trong các hệ thống tìm kiếm tương tự sử dụng một tập các trang web bán hàng trực tuyến được xác định trước bởi người quản trị hệ thống để thu thập dữ liệu, chính điều này sẽ tạo ra chi phí không nhỏ khi người quản trị muốn mở rộng tập dữ liệu này cho hệ thống.

Dựa vào phương pháp học bán giám sát áp dụng vào quá trình trích xuất thông tin tự động chúng tôi đã xây dựng cho hệ thống một mô hình phân tích và xác định các trang web bán hàng tự động. Bên cạnh đó, mô hình cũng

[19] http://dais.cs.uiuc.edu/. DAIS là phòng thí nghiệm*"Ranked number 6 in the world by US News and World Report"*. Trong phòng thí nghiệm DAIS có GS. Jiawei Han có chỉ số h-index là 51 (có 36 nhà khoa học về CNTT có h-index từ 51 trở lên, *xem trang web* http://www.cs.ucla.edu/%7Epalsberg/h-number.html)

tự động sinh ra luật bóc tách các thành phần dữ liệu tương ứng với trang web đó (Các thành phần được bóc tách ở đây là: tên sản phẩm, giá cả của sản phẩm, ảnh, các thông tin liên quan…)

b) Tính năng đảm bảo dữ liệu luôn được cập nhật nhanh nhất và mới nhất

Một vấn đề lớn đặt ra đối với các hệ thống máy tìm kiếm nói chung và hệ thống máy tìm kiếm giá cả sản phẩm VnGia nói riêng chính là việc đảm bảo cho dữ liệu trên hệ thống luôn được cập nhật nhanh nhất và thông tin giá cả luôn mới nhất. Với số lượng khổng lồ các trang web có chứa dữ liệu trên Internet đòi hỏi hệ thống phải có một cơ chế thu thập đáp ứng được yêu cầu đặt ra bao gồm các chức năng dưới đây.

- Chức năng thu thập dữ liệu song song

Chức năng thu thập dữ liệu (Crawler) đóng vai trò rất quan trọng trong các hệ thống máy tìm kiếm. Nhiệm vụ của chức năng này chính là cung cấp dữ liệu cho các máy tìm kiếm hoạt động. Chức năng này thực hiện công việc duyệt Web, đi theo các liên kết trên các trên Web để thu thập nội dung các trang Web. Đối với chức năng thu thập dữ liệu của hệ thống sản phẩm VnGia, chức năng này cũng thực hiện việc duyệt qua các liên kết có trong tập các trang web bán hàng và dựa vào các luật bóc tách đã được trích xuất ở chức năng trên để thu thập các thông tin liên quan đến sản phẩm như tên sản phẩm, giá sản phẩm, ảnh…

Tuy nhiên, khi số lượng các trang web phát sinh khổng lồ, việc thu thập một số lượng lớn các trang web cũng như giảm thiểu được các chi phí phát sinh trong quá trình thu thập dữ liệu như tiêu tốn tài nguyên bộ nhớ, tiêu tốn băng thông mạng…trở thành một bài toán phức tạp. Đã có rất nhiều các nghiên cứu được công bố nhằm giải quyết nhưng vấn đề này. Trong các nghiên cứu đó, những nghiên cứu sử dụng phương pháp song song hóa các bộ thu thập dữ liệu nhận được sự quan tâm lớn của cộng đồng, nhiều bộ thụ thập

dữ liệu trong các máy tìm kiếm nổi tiếng như Google, Nutch…đều áp dụng các nghiên cứu này cho hệ thống của mình.

Từ việc khảo sát, nghiên cứu các công trình khoa học đã công bố, chúng tôi cũng đã áp dụng để xây dựng cho chức năng thu thập dữ liệu của hệ thống máy tìm kiếm VnGia cơ chế ***song song hóa***. Cùng một thời điểm, hệ thống có thể sử dụng nhiều bộ thu thập được đặt trên nhiều nút mạng khác nhau để thu thập dữ liệu, chính nhờ cơ chế này mà hệ thống cơ sở dữ liệu của hệ thống VnGia luôn được mở rộng và cập nhật nhanh chóng.

- Chức năng cập nhật thông tin giá cả hàng ngày

Sản phẩm và giá của sản phẩm là một dạng dữ liệu biến đổi và tồn tại theo thời gian, chính điều này đòi hỏi hệ thống phải luôn cập nhật được sự biến động của thông tin này một cách nhanh chóng đáp ứng nhu cầu tìm kiếm của người sử dụng. Dựa vào cơ chế song song hóa của chức năng thu thập dữ liệu, chúng tôi cũng đã xây dựng cho hệ thống một chức năng tự động làm nhiệm vụ cập nhật thông tin giá cả cũng như xác định sự tồn tại của các sản phẩm trên trang web cửa hàng.

c) Tính năng tự động gom cụm các sản phẩm giống nhau

Một trong những vấn đề lớn đặt ra đối với các máy tìm kiếm giá cả sản phẩm chính là việc thể hiện được sự so sánh giá giữa các cửa hàng về cùng một loại sản phẩm. Nhiều người nghĩ rằng, việc gom nhóm các sản phẩm cùng một loại chỉ đơn giản là việc xác định các tên sản phẩm trùng lặp trên hệ thống. Tuy nhiên, đối với cùng một loại sản phẩm, trên từng cửa hàng khác nhau lại có các cách viết khác nhau, ví dụ: đối với sản phẩm Nokia 6085:

- Trên cửa hàng http://dienthoai.top1.vn là: "*Tên sản phẩm: Nokia 6085*"
- Trên cửa hàng http://www.ducminhmobile.net là: "*Nokia 6085 PINK + Tặng thẻ nhớ 128MB*"

- …

Để giải quyết vấn đề này các hệ thống tìm kiếm tương tự cũng như các hệ thống so sánh giá cả sản phẩm thường sử dụng các phương pháp thủ công như xây dựng trước các sản phẩm chuẩn(base product) sau đó nối thủ công các sản phẩm có cùng tên vào các sản phẩm chuẩn này hoặc cho cửa hàng tự đăng ký sản phẩm mình và liên kết với sản phẩm chuẩn. Các phương pháp này đòi hỏi chi phí về nhân lực, thời gian… rất lớn. Dựa vào các nghiên cứu về ***phân cụm phân cấp văn bản tự động*** (Hierarchy Document Clustering) trong lĩnh vực khai phá dữ liệu text và xử lý ngôn ngữ tự nhiên, chúng tôi đã xây dựng một phương pháp gom cụm phân cấp tên sản phẩm áp dụng ***mô hình học bán giám sát*** (Semi-supervised learning) kết hợp với ***độ tương đồng giữa hai sản phẩm***. Với phương pháp này, hệ thống tự động gom cụm hàng trăm nghìn sản phẩm trong thời gian ngắn và đạt được độ chính xác cao.

d) Tính năng quảng cáo ngữ cảnh

Quảng cáo trên máy tìm kiếm hiện đang là hình thức quảng cáo thu hút được nhiều sự chú ý nhất ngày nay, trong đó các quảng cáo được hiển thị bên cạnh kết quả tìm kiếm theo truy vấn của người dùng. Quảng cáo trên máy tìm kiếm là quảng cáo được thực hiện trên máy tìm kiếm, khi người dùng tìm kiếm theo một truy vấn, bên cạnh kết quả tìm kiếm, một danh sách các quảng cáo được hiển thị tương ứng với truy vấn của người dùng. Các quảng cáo được sắp xếp theo hai tiêu chí: độ phù hợp với truy vấn và số tiền người quảng cáo sẽ trả cho công ty quảng cáo cho việc hiển thị quảng cáo của họ. Quảng cáo trên máy tím kiếm là hình thức quảng cáo trực tuyến phổ biến nhất hiện nay.

Nhưng năm gần đây, bên cạnh các phương pháp quảng cáo trực tuyến thông thường xuất hiện một mô hình quảng cáo được đánh giá cao và áp dụng trên nhiều hệ thống nổi tiếng là mô hình quảng cáo ngữ cảnh.Các mô hình

quảng cáo ngữ cảnh trên thế giới được áp dụng rộng rãi và tạo ra một hiệu quả kinh tế lớn như Google Adwords của Google, quảng cáo ngữ cảnh trên Ebay….Quảng cáo theo ngữ cảnh khác với quảng cáo trên máy tìm kiếm ở chỗ danh sách quảng cáo hiển thị không phải được so khớp với nội dung của câu truy vấn mà dựa vào sự tương đồng về nội dung của trang web người sử dụng đang duyệt với các mẩu quảng cáo. Chính cơ chế này đã làm cho nội dung những mẩu quảng cáo không còn "rơi rạc" đối với nội dung trang web, điều này tạo động lực cho người sử dụng nhấn chuột vào các liên kết quảng cáo.

Từ những kết quả của công trình khoa học do các đồng nghiệp trong phòng Thí nghiệm Công nghệ tri thức công bố tại hội nghị IEEE/WIC/ACM International Conference on Web Intelligence, 2008, chúng tôi đã tiến hành cài đặt và xây dựng chức năng quảng cáo ngữ cảnh trên máy tìm kiếm giá cả sản phẩm VnGia. Với chức năng này doanh nghiệp có thể dễ dàng đăng ký quảng cáo bằng việc đưa liên kết trang web cần quảng cáo cho hệ thống, hệ thống sẽ tự động xác định nội dung, các từ khóa quan trọng của trang web và liên kết với các trang web giới thiệu sản phẩm.

e) Tính năng truy xuất và tìm kiếm nhanh

Với khả năng nổi trội là cơ chế tự động phát hiện và thu thập song song các dữ liệu sản phẩm tại các trang web bán hàng, máy tìm kiếm giá cả sản phẩm VnGia mỗi ngày phải xử lý và lưu trữ hàng nghìn sản phẩm. Chính vấn đề này đòi hỏi hệ thống phải có một cơ chế lưu trữ dữ liệu lớn và truy xuất dữ liệu nhanh phù hợp với số lượng dữ liệu. Bằng việc áp dụng một trong những kỹ thuật phổ biến là đánh chỉ mục ngược (Invert Indexing) cho quá trình lưu trữ và truy vấn dữ liệu, hệ thống VnGia có thể đáp ứng nhanh cho các quá trình truy xuất trên hàng triệu sản phẩm. Bên cạnh đó quá trình tìm kiếm (Searching) cũng được sử dụng phương pháp tìm kiếm toàn văn (Full Text

Search), với phương pháp này người sử dụng có thể dễ dàng tìm kiếm bằng các biểu thức truy vấn.

f) Tính năng cho phép hệ thống có khả năng tương tác cao với người sử dụng

Với mục đích hướng tới trở thành một môi trường liên kết giữa khách hàng và doanh nghiệp, hệ thống VnGia có nhiều thành phần chức năng cho phép người sử dụng tương tác với hệ thống như:

- Đánh giá điểm cho các sản phẩm của doanh nghiệp.
- Nêu ý kiến đánh giá về sản phẩm.
- Đánh giá điểm chất lượng dịch vụ của doanh nghiệp.
- Nêu ý kiến về chất lượng dịch của doanh nghiệp.
- …

Chính những thông tin đánh giá này của khách hàng sẽ là những đặc trưng để hệ thống xác định được độ quan trọng của sản phẩm cũng như phát triển các thành phần tư vấn trong tương lai. Bên cạnh đó hệ thống cũng tạo ra những phương thức giao tiếp giữa hệ thống với người sử dụng là các doanh nghiệp như:

- Đăng ký thông tin về cửa hàng, cũng như đăng ký về trang web của cửa hàng để hệ thống có thể tiến hành thu thập.
- Đăng ký quảng cáo trực tuyến đối với trang web hay sản phẩm của doanh nghiệp.
- …

g) Tính năng lọc và tính hạng các sản phẩm

Vấn đề tính hạng các đối tượng kết quả trả về trong máy tìm kiếm trở thành công việc hết sức cấp thiết khi xây dựng các công cụ tìm kiếm thông

tin. Đối với các máy tìm kiếm trang web, khi người dùng nhập vào một nhóm từ khóa tìm kiếm (hoặc một câu hỏi), máy tìm kiếm sẽ thực hiện nhiệm vụ tìm kiếm và trả lại một số trang Web theo yêu cầu người dùng. Nhưng số các trang Web liên quan đến từ khóa tìm kiếm có thể lên tới hàng vạn trang, trong khi người dùng chỉ quan tâm đến một số ít trang trong đó, vậy việc tìm ra các trang đáp ứng nhiều nhất yêu cầu người dùng để đưa lên đầu là cần thiết. Đó chính là công việc tính hạng trang Web của máy tìm kiếm - sắp xếp các trang Web theo độ phù hợp với câu hỏi. Khi hiển thị kết quả tìm kiếm, tập các trang Web kết quả được xuất hiện theo thứ tự giảm dần về độ quan trọng.

Tương tự như tìm kiếm các trang web, tìm kiếm các sản phẩm cũng đòi hỏi việc tính hạng các sản phẩm sao cho phù hợp với câu truy vấn của người sử dụng nhất. Từ những giải thuật đã được công bố trước đó, chúng tôi cũng đã nghiên cứu và xây dựng một công thức tính toán phù hợp với đặc trưng về ***độ quan trọng*** của sản phẩm. Công thức này là thể hiện được mối quan hệ giữa các đặc trưng là các thông tin do người dùng đánh giá về sản phẩm như: số lượng ý kiến đánh giá của người sử dụng về sản phẩm(comment), số lượt xem sản phẩm của người sử dụng(click), giá trị do người sử dụng đánh giá đối với sản phẩm (rate), số lượng bài báo giới thiệu về sản phẩm (review) và độ tương đồng về nội dung giữa tên sản phẩm và câu truy vấn. Hệ thống sẽ sử dụng công thức để xây dựng cho từng sản phẩm một trọng số thể hiện "độ quan trọng" của sản phẩm đó. Dựa vào "độ quan trọng" này hệ thống sẽ sắp xếp các kết quả trả về phù hợp hơn với câu truy vấn.

Bên cạnh việc sắp xếp dựa vào độ quan trọng, hệ thống cũng cho phép người sử dụng có thể sắp xếp các sản phẩm theo từng đặc trưng cụ thể như:

- Số lượng ý kiến đánh giá của người sử dụng về sản phẩm
- Số lượt xem sản phẩm của người sử dụng
- Giá trị do người sử dụng đánh giá đối với sản phẩm

- Số lượng bài báo giới thiệu về sản phẩm
- Thông tin về giá của sản phẩm

h) Tính năng tìm kiếm mọi lúc mọi nơi

Trong đời sống hàng ngày, mua hàng là một công việc xảy ra thường xuyên, không phải người sử dụng nào, lúc mua hàng cũng có sẵn máy tính để tìm kiếm các thông tin liên quan đến sản phẩm cũng như nắm được mặt bằng giá tại các cửa hàng khác nhau.

Chính vì vậy, ngoài phiên bản hoạt động trên nền web, hệ thống VnGia còn có một phiên bản có thể hoạt động trên các trình duyệt của thiết bị di động như điện thoại di động, pda, palm… Phiên bản này được thiết kế nhỏ gọn, hoạt động nhanh giảm thiểu tối đa chi phí truy cập cho người sử dụng. Với tính năng này, người sử dụng tại bất cứ đâu cũng có thể thuận tiện trong quá trình mua hàng cũng như tìm kiếm thông tin về hàng hóa.

4.5.2. Mức độ hoàn thiện

Sản phẩm được công bố tại địa chỉ:

- **http://www.vngia.com** (Phiên bản Web)
- **http://wap.vngia.com** (Phiên bản Wap cho thiết bị di động)

4.5.3. Một số kết quả của sản phẩm

a. Thời gian hệ thống bắt đầu hoạt động : 16/08/2009

b. Số lượng trang web được nhận dạng là trang web bán hàng

   Sau thời gian bắt đầu đi vào hoạt động đến nay số lượng trang web được nhận dạng là web bán hàng và bóc tách được các mẫu trích xuất là hơn 1300 trang.

c. Số lượng sản phẩm thu thập

Số lượng sản phẩm thu thập được qua quá trình trích xuất thông tin sản phẩm từ hơn 1300 trang web là 125000 sản phẩm khác loại thuộc nhiều mục khác nhau. Bên cạnh đó thu thập hơn 17000 bài viết liên quan đến các sản phẩm chuẩn.

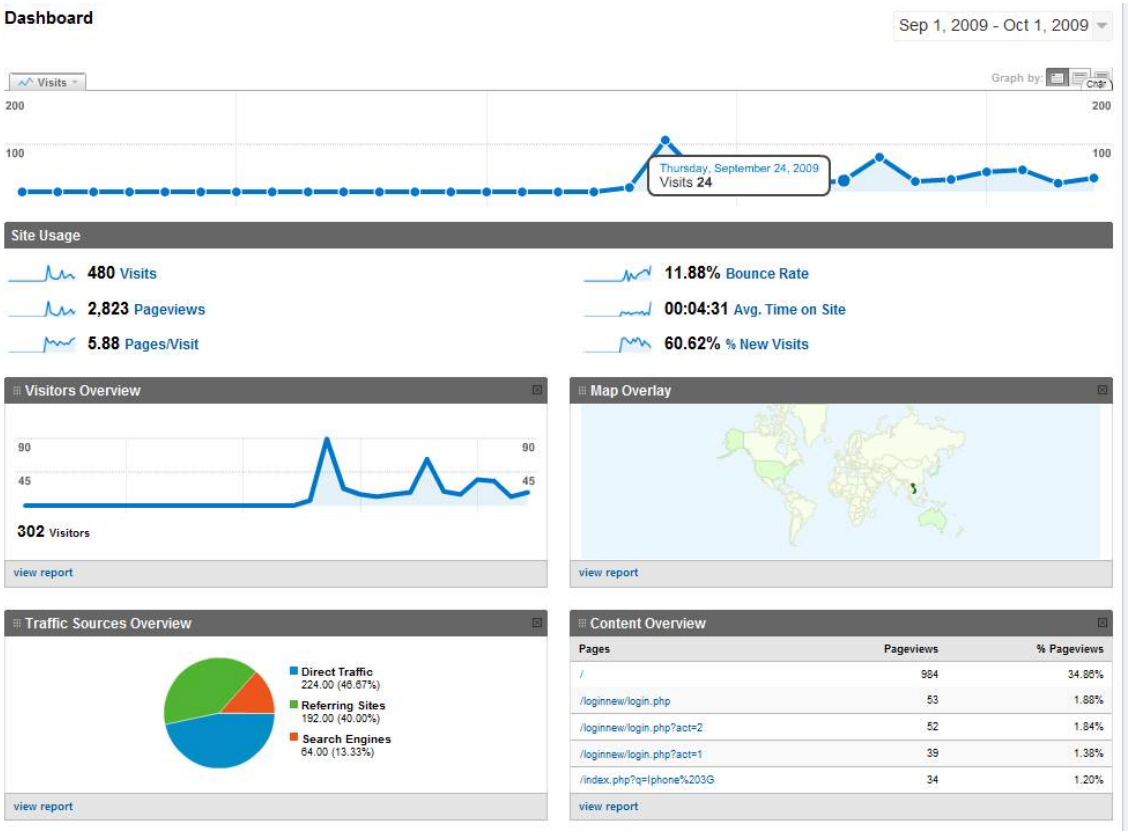
d. Kết quả đạt được của quá trình phân cụm phân cấp sản phẩm

- Sinh ra hơn 10000 sản phẩm chuẩn từ 125000 sản phẩm thu thập được.
- Tạo liên kết giữa gần 70000 sản phẩm với hơn 10000 sản phẩm chuẩn.
- Thời gian thực hiện việc phân cụm: 1h.

e. Tốc độ đáp ứng truy vấn hiện nay

Hiện nay sản phẩm được công bố tới người sử dụng ở địa chỉ http://www.vngia.com, với tốc độ đáp ứng truy vấn hiện nay là khá tốt với tốc độ đáp ứng trung bình là 2-4 giây.

Bảng 4. 1. Thống kê số lượng người truy cập (từ 18/09/2009 đến 01/10/2009)



KẾT LUẬN VÀ KIẾN NGHỊ

...

TÀI LIỆU THAM KHẢO

Tiếng Việt

[HV04] Khoa Nhà nước và Pháp luật (2004) Lý luận chung về Nhà nước và Pháp luật (tập 1), Học viện Chính trị Quốc gia Hồ Chí Minh, Nhà xuất bản lý luận chính trị, 2004.

[VN55-01] Nghị định số 55/2001/NĐ-CP ngày 23 tháng 8 năm 2001 của Chính phủ *Về quản lý, cung cấp và sử dụng dịch vụ Internet.*

[VN56-07] Quyết định số 56/2007/QĐ-TTg ngày 03 tháng 5 năm 2007 về việc phê duyệt Chương trình phát triển công nghiệp nội dung số Việt Nam đến năm 2010.

Tiếng Anh

[Ayr01] Lori Bowen Ayre (2001). Internet Filtering Options Analysis: An Interim Report, *The InFoPeople Project*, May, 2001, U.S. Institute of Museum and Library Services.

[Bau01] Patrick Baudisch (2001). *Dynamic Information Filtering*. GMD – Forschungszentrum Informationstechnik GmbH, Schloß Birlinghoven, D-53754 Sankt Augustin, Germany ISSN 1435-2699, 2001.

[BHK98] J. Breese, D. Heckerman, and C. Kadie (1998). *Empirical analysis of predictive algorithms for collaborative filtering*. In Proceedings of the 14th Conference on Uncertainty in Articial Intelligence (UAI'98), 43–52, July 1998.

[China02] People's Republic of China: *State Control of the Internet in China* (Amnesty International, November 2002)

[EC00] Commission of The European Communities (2000). *Creating a Safer Information Society by Improving the Security of Information Infrastructures and Combating Computer-related Crime*. COM(2000) 890 final.

[EC03] European Commission DG Information Society Luxembourg (2003). *The Evaluation of the Safer Internet Action Plan 1999-2002*. July 2003.

[EC04] Commission of The European Communities (2004). *Safer Internet plus (2005-2008): Commission Staff working Paper*. Ex Ante Evaluation. Dec. 2004.

[EC04b] Commission of The European Communities (2004). *Mobile Broadband Services (Text with EEA relevance)*. COM(2004) 447, Brussels, 30 June 2004.

[EC04c] Commission of The European Communities (2004). On unsolicited commercial communications or 'spam'. *COM(2004) 28 final*, Brussels, 22.01.2004

[EC04d] Commission of The European Communities (2004). *Safer use of the Internet and new online technologies. European Parliament legislative resolution on the proposal for a decision of the European Parliament and of*

the Council establishing a multiannual Community programme on promoting safer use of the Internet and new online technologies (COM(2004)0091 – C5-0132/2004 – 2004/0023(COD)).

[EC06] European Commission (2006). *Creating a Safer Information Society by Improving the Security of Information Infrastructures and Combating Computer-related Crime*, February 2006.

[EC06a] Commission of The European Communities (2006). *On Fighting spam, spyware and malicious software*. COM(2006) 688 final, Brussels, 15.11.2006

[EC06b] Commission of The European Communities (2006). *The Review of the EU Regulatory Framework for electronic communications networks and services*. SEC(2006) 817, Brussels, 28 June 2006.

[EC06c] Commission of The European Communities (2006). Communication on the implementation of the multiannual Community Programme on promoting safer use of the Internet and new online technologies (Safer Internet plus), COM(2006) 661 final, Brussels, June 2006.

[EC06d] Commission of The European Communities (2006). Final valuation of the implementation of the multiannual Community action plan on promoting safer use of the Internet by combating illegal and harmful content on global networks. COM(2006) 663 final, Brussels, 6 November 2006.

[EC06e] Commission of The European Communities (2006). *Staying safe online: EU programme leads the way*. IP/06/1512. Brussels, 6 November 2006.

[EC07] European Commission (2007). Safer Internet for children, qualitative study in 29 European countries (Summary report), May 2007.

[EC08] European Communities (2008). Making the Internet: a safer space, Information Society and Media, February 2008.

[GNO92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry (1992). *Using collaborative filtering to weave an information tapestry*. In Communications of the ACM, December 1992.

[GRT01] Paul Greenfield, Peter Rickwood, Huu Cuong Tran. *Effectiveness of Internet Filtering Software Products*. Prepared for NetAlert and the Australian Broadcasting Authority, CRISO, 9-2001.

[HTTP1] http://www.infopeople.org/resources/filtering: (1) History of Internet Filters and the Library; (2) How Filters Work; (3) Library Filtering - Best Practices và (4) Library Filtering - Resources.

[HTTP2] http://www.webir.org/: Một tập hợp các tài liệu liên quan đến lĩnh vực thu thập và trích chọn thông tin.

[HUN97] Hunter, C.D (1997). *Internet Filter Effectiveness: Testing Over and Underinclusive Blocking Decisions of Four Popular Filters*. www.copacommission.org/papers/filter_effect.pdf

[KP97] Fredrik Kilander, Jacob Palme (1997). *Intelligent Information Filtering*, Department of Computer and System Sciences, Stockholm University, 1997.

[Lanq01] Carsten Lanquillon (2001). Enhanci Text Classification to Improve Information Filtering, *PhD Thesis*, University of Madgdeburg, Germany.

[NDL97] Emmanuel Nauer, Jacques Ducloy, Jean-Charles Lamirel (1997). *Using of multiple data source for information filtering: first approaches in the MedExplore project*, 15th Filtering and Collaborative Filtering, Budapest, 10-12 November 1997.

[Nic97] M. Nichols (1997). *Implicit rating and filtering*. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, 31–36, November 1997.

[OMA03] Toshio Oka, H. Morikawa, T. Aoyama, Vineyard (2003). *A Collaborative Filtering Service Platform in Distributed Environment*, Symposium on Applications and the Internet Workshops, Japan, 2003.

[ONI04Arap] The OpenNet Initiative. *Internet Filtering in Saudi Arabia in 2004: A country study*. November 2004 (http://www.opennetinitiative.net/studies/saudi)

[ONI05Iran] The OpenNet Initiative. *Internet Filtering in Iran in 2004-2005: A Country Study* (http://www.opennetinitiative.net/studies/iran)

[ONI05Sin] The OpenNet Initiative (2005). *Internet Filtering in Singapore in 2004-2005: A Country Study*. August 2005 (http://www.opennetinitiative.net/studies/singapore)

[ONI05Tun] The OpenNet Initiative. *Internet Filtering in Tunisia in 2005: A Country Study*. November 2005 (http://www.opennetinitiative.net/studies/tunisia)

[ONI05UAE] The OpenNet Initiative (2005). *Internet Filtering in the United Arab Emirates in 2004-2005: A Country Study*. (http://www.opennetinitiative.net/studies/uae/)

[ONI06] The OpenNet Initiative (2006). *ONI Internet Watch 001*.

[ONI06Vie] The OpenNet Initiative (2006). *Internet Filtering in Vietnam in 2005-2006: A Country Study*. August, 2006 (http://www.opennetinitiative.net/studies/vietnam/ and http://www.opennet.net/vietnam)

[ONI07] The OpenNet Initiative (2007). *A Summary of the OpenNet Initiative's First Global Study of Internet Filtering*. May 18, 2007.

[ONIChina] The OpenNet Initiative (2005). *Internet Filtering in China in 2004-2005: A Country Study*. April 14, 2005 (http://www.opennetinitiative.net/studies/china/).

[POES04] POESIA (Public Open-source Environment for a Safer Internet Access), *EC-IAP 2117/27572 Project*

[QD09] Xiaoguang Qi and Brian D. Davison (2009). Web Page Classification: Features and Algorithms, *ACM Comput. Surv.*, **41** (2): 1-31.

[RIS94] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl (1994). *Grouplens: An open architecture for collaborative filtering of netnews*. In

Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'94), October 1994.

[Sa08] Sarah Houghton-Jan (2008). Internet Filtering Software Tests: CyberPatrol, FilterGate, & WebSense, *San José Public Library*, February 4, 2008

[Safer] *Safer Internet Action Plan: Intermediate Evaluation*. Business Development Research Consultants Kingsbourne House, 229 – 231 High Holborn, London WC1V 7DA (BDRC 20174/draftmainreport/final)

[SKK01] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl (2001). *Item-based collaborative filtering recommendation algorithms*. In Proceedings of the 10th International World Wide Web Conference(WWW10), 285–295, May 2001.

[Smi05] Marcia S. Smith (2005). Internet: Status Report on Legislative Attempts to Protect Children from Unsuitable Material on the Web, *Order Code RS21328* (Updated December 16, 2005).

[Sten04] Bjorn Dietmar Rafael Stenger (2004). Model-Based Hand Tracking Using A Hierarchical Bayesian Filter, *PhD Thesis*, St. John's College, USA, 2004

[Tho96] Thornton, A. (1996) *Does Internet Create Democracy?* M.A. Journalism Thesis, 1996, http://www.wr.com.au/democracy/index.html. Accessed 2001-10-01.

[US02] Lauren Weinstein (2002). *Filters, Laws Won't Clean Up Net.* http://www.wired.com/politics/law/news/2002/12/56855.

[US03] *State Internet Filter Participation Policy (Draft version)*, State of Arkansas, Office of Information Technology. www.dis.state.ar.us/poli_stan_bestpract/pdf/PS61_Filter.pdf

[US04] *Safeguarding Privacy in the fight Against Terrorism*. Technology and Privacy Advisory Committee, Department of Defense of US, Report, March 2004.

[US07] *Children and the Internet: Laws Relating to Filtering, Blocking and Usage Policies in Schools and Libraries*. http://www.ncsl.org/programs/lis/cip/filterlaws.htm (Versions: January 20, 2006 and April 23, 2007)

[Zhan05] Yi Zhang (2005). Bayesian Graphical Models for Adaptive Filtering, *PhD Thesis*, Carnegie Mellon University, USA.

[ZP07] Jonathan L. Zittrain and John G. Palfrey, Jr. (2007) *Access Denied: The Practice and Policy of Global Internet Filtering*. Oxford Internet Institute, Research Report No. 14, June 2007 (Chapter 2 of the book *Internet Filtering: The Politics and Mechanisms of Control*, MIT Press, 2007).

[4]. 2.1Aas, K. and L. Eikvil (1999, June). Text categorisation: A survey. Technical report, Norwegian Computing Center, P.B. 114 Blindern, N-0314, Oslo, Norway. Technical Report 941.

[5]. 2.2Amitay, E. (1998). Using common hypertext links to identify the best phrasal description of target web documents. In Proceedings of the SIGIR'98 Post-

Conference Workshop on Hypertext Information Retrieval for the Web, Melbourne, Australia.

[6]. 2.3Amitay, E., D. Carmel, A. Darlow, R. Lempel, and A. Soffer (2003). The connectivity sonar: Detecting site functionality by structural patterns. In HYPERTEXT '03: Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, New York, NY, pp. 38–47. ACM Press.

[7]. 2.4Angelova, R. and S. Siersdorfer (2006). A neighborhood-based approach for clustering of linked document collections. In CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, New York, NY, pp. 778–779. ACM Press.

[8]. 2.5Angelova, R. and G. Weikum (2006). Graph-based text classification: Learn from your neighbors. In SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 485–492. ACM Press.

[9]. 2.6Attardi, G., A. Gulli, and F. Sebastiani (1999). Automatic web page categorization by link and context analysis. In C. Hutchison and G. Lanzarone (Eds.), Proceedings of THAI'99, First European Symposium on Telematics, Hypermedia and Artificial Intelligence, Varese, IT, pp. 105–119.

[10]. 2.7Beeferman, D. and A. Berger (2000). Agglomerative clustering of a search engine query log. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp. 407–415. ACM Press.

[11]. 2.8Braschler, M. & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval?. *Information Retrieval*, *7*, pp. 291-316.

[12]. 2.9Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen and Quang-Thuy Ha (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. In *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, November 1-3, 2006, Wuhan, China, 215-222.

[13]. 2.10Cancedda, N., Gaussier, E., Goutte, C. & Renders, J.M. (2003). Word sequence kernels. *Journal of Machine Learning Research*, *3*, pp. 1059-1082.

[14]. 2.11Calado, P., M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Goncalves (2003). Combining link-based and content-based methods for web document classification. In CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management, New York, NY, pp. 394–401. ACM Press.

[15]. 2.12Castillo, C., D. Donato, A. Gionis, V. Murdock, and F. Silvestri (2007). Know your neighbors: Web spam detection using the web topology. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY. ACM Press. In press.

[16]. 2.13Chakrabarti, S. (2000, January). Data mining for hypertext: a tutorial survey. SIGKDD Explorations Newsletter 1(2), 1–11.

[17]. 2.14Chakrabarti, S. (2003). Mining the Web: Discovering Knowledge from Hypertext Data. San Francisco, CA: Morgan Kaufmann.

[18]. 2.15Chakrabarti, S., B. E. Dom, and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In SIGMOD '98: Proceedings of the ACM SIGMOD International Conference on Management of Data, New York, NY, pp. 307–318. ACM Press.

[19]. 2.16Chakrabarti, S., M. M. Joshi, K. Punera, and D. M. Pennock (2002). The structure of broad topics on the web. In WWW '02: Proceedings of the 11th International Conference on World Wide Web, New York, NY, pp. 251–262. ACM Press.

[20]. 2.17Chakrabarti, S., M. van den Berg, and B. Dom (1999, May). Focused crawling: A new approach to topic-specific Web resource discovery. In WWW '99: Proceeding of the 8th International Conference on World Wide Web, New York, NY, pp. 1623–1640. Elsevier.

[21]. 2.18Chekuri, C., M. Goldwasser, P. Raghavan, and E. Upfal (1997, April). Web search using automated classification. In Proceedings of the Sixth International World Wide Web Conference, Santa Clara, CA. Poster POS725.

[22]. 2.19Chen, H. and S. Dumais (2000). Bringing order to the Web: Automatically categorizing search results. In CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, pp. 145–152. ACM Press.

[23]. 2.20Chen, Z., O. Wu, M. Zhu, and W. Hu (2006). A novel web page filtering system by combining texts and images. In WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, pp. 732–735. IEEE Computer Society.

[24]. 2.21Chen, C., Lee, H. & Hwang, C. (2005). A Hierarchical Neural Network Document Classifier with Linguistic Feature Selection. *Applied Intelligence*, *23*, pp. 277-294.

[25]. 2.22Chesley, P., B. Vincent, L. Xu, and R. K. Srihari (2006, March). Using verbs and adjectives to automatically classify blog sentiment. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, pp. 27–29. AAAI Press. Technical Report SS-06-03.

[26]. 2.23Choi, B. and Z. Yao (2005). Web page classification. In W. Chu and T. Y. Lin (Eds.), Foundations and Advances in Data Mining, Volume 180 of Studies in Fuzziness and Soft Computing, pp. 221–274. Berlin: Springer-Verlag.

[27]. 2.24Cohen, W. W. (2002). Improving a page classifier with anchor extraction and link analysis. In S. Becker, S. Thrun, and K. Obermayer (Eds.), Advances in Neural Information Processing Systems, Volume 15, pp. 1481–1488. Cambridge, MA: MIT Press.

[28]. 2.25Cohn, D. and T. Hofmann (2001). The missing link — a probabilistic model of document content and hypertext connectivity. In Advances in Neural Information Processing Systems (NIPS) 13. Corporation, N. C. (2007). The dmoz open Directory Project (ODP). http://www.dmoz.com/.

[29]. 2.26Dang Thanh Hai, Nguyen Huong Giang, Ha Quang Thuy (2005). Naive Bayes text classification algorithm and problem of specifying clasifying threshold in search engine. *Journal of Computer Science & Cybernetics* **21**(2), 152-161.

[30]. 2.27Davison, B. D. (2000, July). Topical locality in the Web. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, pp. 272–279. ACM Press.

[31]. 2.28Davison, B. D. (2004, November). The potential of the metasearch engine. In Proceedings of the Annual Meeting of the American Society for Information Science and Technology, Volume 41, Providence, RI, pp. 393–402. American Society for Information Science & Technology.

[32]. 2.29Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407.

[33]. 2.30Dietterich, T. G. and G. Bakiri (1995). Solving multiclass learning problems via errorcorrecting output codes. Journal of Articial Intelligence Research 2, 263–286.

[34]. 2.31Dumais, S. & Chen, H. (2000). Hierarchical classification of Web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. pp. 256-263). : ACM Press, Athens, Greece

[35]. 2.32Drost, I., S. Bickel, and T. Scheffer (2005, March). Discovering communities in linked data by multi-view clustering. In From Data and Information Analysis to Knowledge Engineering: Proceedings of 29th Annual Conference of the German Classification Society, Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, pp. 342–349. Springer.

[36]. 2.33Elgersma, E. and M. de Rijke (2006, March). Learning to recognize blogs: A preliminary exploration. In EACL 2006 Workshop: New Text - Wikis and blogs and other dynamic text sources.

[37]. 2.34Ester, M., H.-P. Kriegel, and M. Schubert (2002). Web site mining: A new way to spot competitors, customers and suppliers in the World Wide Web. In KDD ’02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp. 249–258. ACM Press.

[38]. 2.35Fisher, M. J. and R. M. Everson (2003, April). When are links useful? Experiments in text classification. In Advances in Information Retrieval. 25th European Conference on IR Research, pp. 41–56. Springer.

[39]. 2.36Fitzpatrick, L. and M. Dent (1997, July). Automatic feedback using past queries: Social searching? In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 306–313. ACM Press.

[40]. 2.37Furnkranz, J. (1999). Exploiting structural information for text classification on the WWW. In D. J. Hand, J. N. Kok, and M. R. Berthold (Eds.), Proceedings of the 3rd Symposium on Intelligent Data Analysis (IDA-99), Volume 1642 of LNCS, Amsterdam, Netherlands, pp. 487–497. Springer-Verlag.

[41]. 2.38Furnkranz, J. (2001). Hyperlink ensembles: A case study in hypertext classification. Journal of Information Fusion 1, 299–312.

[42]. 2.39Furnkranz, J. (2005). Web mining. In O. Maimon and L. Rokach (Eds.), The Data Miningand Knowledge Discovery Handbook, pp. 899–920. Berlin: Springer.

[43]. 2.40Gao, S., Wu, W., Lee, C. & Chua, T. (2003). A maximal figure-of-merit learning approach to text categorization. In *Proceedings of SIGIR-03 26th ACM International Conference on Research and Development in Information Retrieval* (pp. pp. 174-181). ACM Press New York US

[44]. 2.41Gabrilovich, E. and S. Markovitch (2004). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the 21st International Conference on Machine learning, New York, NY, pp. 41. ACM Press.

[45]. 2.42Gabrilovich, E. and S. Markovitch (2005, July). Feature generation for text categorization using world knowledge. In Proceedings of the 19th International Joint Conference for Artificial Intelligence (IJCAI), pp. 1048–1053.

[46]. 2.43Gabrilovich, E. and S. Markovitch (2006, July). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In Proceedings of the 21st National Conference on Artificial Intelligence, Menlo Park, CA, pp. 1301–1306. AAAI Press.

[47]. 2.44Getoor, L. and C. Diehl (2005, December). Link mining: A survey. SIGKDD Explorations Newsletter Special Issue on Link Mining 7(2).

[48]. 2.45Ghani, R. (2001). Combining labeled and unlabeled data for text classification with a large number of categories. In First IEEE International Conference on Data Mining (ICDM), Los Alamitos, CA, pp. 597. IEEE Computer Society.

[49]. 2.46Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. In ICML '02: Proceedings of the 19th International Conference on Machine Learning, San Francisco, CA, pp. 187–194. Morgan Kaufmann.

[50]. 2.47Ghani, R., S. Slattery, and Y. Yang (2001). Hypertext categorization using hyperlink patterns and meta data. In ICML '01: Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, pp. 178–185. Morgan Kaufmann.

[51]. 2.48Glance, N. S. (2000, July). Community search assistant. In Artificial Intelligence for Web Search, pp. 29–34. AAAI Press. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

[52]. 2.49Glover, E. J., K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake (2002). Using web structure for classifying and describing web pages. In Proceedings of the 11th International Conference on World Wide Web, New York, NY, pp. 562–569. ACM Press.

[53]. 2.50Golub, K. and A. Ardo (2005, September). Importance of HTML structural elements and metadata in automated subject classification. In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Volume 3652 of LNCS, Berlin, pp. 368–378. Springer.

[54]. 2.51Govert, N., M. Lalmas, and N. Fuhr (1999). A probabilistic description-oriented approach for categorizing web documents. In CIKM '99: Proceedings of the 8th International Conference on Information and Knowledge Management, New York, NY, pp. 475–482. ACM Press.

[55]. 2.52Greevy, E. & Smeaton, A.F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. pp. 468-469). : ACM Press, Sheffield, United Kingdom

[56]. 2.53Haykin, S. (1998). Neural networks: a comprehensive foundation. : Prentice Hall PT. Ho, T. K. (1999). Fast Identification of Stop Words for Font Learning and Keyword Spotting. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition* (pp. pp. 333-336). : IEEE Computer Society

[57]. 2.54Hammami, M., Y. Chahir, and L. Chen (2003). Webguard: Web based adult content detection and filtering system. In WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, Washington, DC, pp. 574. IEEE Computer Society.

[58]. 2.55Haveliwala, T. H. (2002, May). Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference, New York, NY, pp. 517–526. ACM Press.

[59]. 2.56Hermjakob, U. (2001, July). Parsing and question classification for question answering. In Proceedings of the ACL Workshop on Open-Domain Question Answering, pp. 1–6.
[60]. 2.57Hofmann, T. (1999a). Probabilistic latent semantic analysis. In Proc. of Uncertainty in Artificial Intelligence, UAI'99, Stockholm.
[61]. 2.58Hofmann, T. (1999b). Probabilistic latent semantic indexing. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, pp. 50–57. ACM Press.
[62]. 2.59Huang, C.-C., S.-L. Chuang, and L.-F. Chien (2004a). Liveclassifier: Creating hierarchical text classifiers through web corpora. In WWW '04: Proceedings of the 13th International Conference on World Wide Web, New York, NY, pp. 184–192. ACM Press.
[63]. 2.60Huang, C.-C., S.-L. Chuang, and L.-F. Chien (2004b). Using a web-based categorization approach to generate thematic metadata from texts. ACM Transactions on Asian Language Information Processing (TALIP) 3(3), 190–212.
[64]. 2.61Hull, D. A. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*, pp. 70-84.
[65]. 2.62Hunnisett, D. S. & Teahan, W.J. (2004). Context-based methods for text categorisation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. pp. 578-579). ACM Press, Sheffield, United Kingdom
[66]. 2.63Jensen, D., J. Neville, and B. Gallagher (2004). Why collective inference improves relational classification. In KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp. 593–598. ACM Press.
[67]. 2.64Joachims, T. (2002). Optimizing search engines using clickthrough data. In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, pp. 133–142. ACM Press.
[68]. 2.65Joachims, T., D. Freitag, and T. Mitchell (1997, August). WebWatcher: A tour guide for the World Wide Web. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, pp. 770–775. Morgan Kaufmann.
[69]. 2.66Kaki, M. (2005). Findex: Search result categories help users when document ranking fails. In CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, pp. 131–140. ACM Press.
[70]. 2.67Kan, M.-Y. (2004). Web page classification without the web page. In WWW Alt. '04: Proceedings of the 13th International World Wide Web Conference Alternate Track Papers & Posters, New York, NY, pp. 262–263. ACM Press.
[71]. 2.68Kan, M.-Y. and H. O. N. Thi (2005). Fast webpage classification using URL features. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05), New York, NY, pp. 325–326. ACM Press.
[72]. 2.69Kiritchenko, S. (2005). Hierarchical Text Categorization and Its Application to Bioinformatics. Ph. D. thesis, University of Ottawa.
[73]. 2.70Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience.

[74]. 2.71Kurland, O. and L. Lee (2005, July). PageRank without hyperlinks: Structural re-ranking using links induced by language models. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 306–313. ACM Press.

[75]. 2.72Kurland, O. and L. Lee (2006, July). Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 83–90. ACM Press.

[76]. 2.73Kwok, C. C. T., O. Etzioni, and D. S. Weld (2001). Scaling question answering to the web. In WWW '01: Proceedings of the 10th International Conference on World Wide Web, New York, NY, pp. 150–161. ACM Press.

[77]. 2.74Kwon, O.-W. and J.-H. Lee (2000). Web page classification based on k-nearest neighbor approach. In IRAL '00: Proceedings of the 5th International Workshop on Information etrieval with Asian languages, New York, NY, pp. 9–15. ACM Press.

[78]. 2.75Kwon, O.-W. and J.-H. Lee (2003, January). Text categorization based on k-nearest neighbor approach for web site classification. Information Processing and Management 29(1), 25–44.

[79]. 2.76Lindemann, C. and L. Littig (2006). Coarse-grained classification of web sites by their structural properties. In WIDM '06: Proceedings of the 8th ACM International Workshop on Web Information and Data Management, New York, NY, pp. 35–42. ACM Press.

[80]. 2.77Liu, T.-Y., Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma (2005). Support vector machines classification with a very large-scale taxonomy. SIGKDD Explorations Newsletter 7(1), 36–43.

[81]. 2.78Liu, T.-Y., Y. Yang, H. Wan, Q. Zhou, B. Gao, H.-J. Zeng, Z. Chen, and W.-Y. Ma (2005, May). An experimental study on large-scale web categorization. In WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, New York, NY, pp. 1106–1107. ACM Press.

[82]. 2.79Liu, H. & Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. : Kluwer Academic Publisher.

[83]. 2.80Lu, Q. and L. Getoor (2003, August). Link-based classification. In Proceedings of the 20th International Conference on Machine Learning (ICML), Menlo Park, CA. AAAI Press.

[84]. 2.81Luxenburger, J. and G. Weikum (2004, November). Query-log based authority analysis for web information search. In Proceedings of 5th International Conference on Web Information Systems Engineering (WISE), Volume 3306 of LNCS, Berlin, pp. 90–101. Springer.

[85]. 2.82Macskassy, S. A. and F. Provost (2007, May). Classification in networked data: A toolkit and a univariate case study. Journal of Machine Learning Research 8, 935–983.

[86]. 2.83Menczer, F. (2005, May/June). Mapping the semantics of web text and links. IEEE Internet Computing 9(3), 27–36.

[87]. 2.84Mihalcea, R. and H. Liu (2006, March). A corpus-based approach to finding happiness. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, pp. 139–144. AAAI Press. Technical Report SS-06-03.

[88]. 2.85Mishne, G. (2005, August). Experiments with mood classification in blog posts. In Workshop on Stylistic Analysis of Text for Information Access.
[89]. 2.86Mishne, G. and M. de Rijke (2006, March). Capturing global mood levels using blog posts. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, pp. 145–152. AAAI Press. Technical Report SS-06-03.
[90]. 2.87Mitchell, T. M. (1997). Machine Learning. New York: McGraw-Hill.
[91]. 2.88Mladenic, D. (1998). Turning Yahoo into an automatic web-page classifier. In Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 473–474.
[92]. 2.89Mladenic, D. (1999, Jul/Aug). Text-learning and related intelligent agents: A survey. IEEE Intelligent Systems and Their Applications 14(4), 44–54.
[93]. 2.90Mladenic, D. & Grobelnik, M. (2003). Feature selection on hierarchy of web documents. *Decision Support Systems*, *35*, pp. 45-87.
[94]. 2.91Nanno, T., T. Fujiki, Y. Suzuki, and M. Okumura (2004). Automatically collecting, monitoring, and mining Japanese weblogs. In WWW Alt. '04: Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters, New York, NY, pp. 320–321. ACM Press.
[95]. 2.92Nie, L., B. D. Davison, and X. Qi (2006, August). Topical link analysis for web search. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, New York, NY, pp. 91–98. ACM Press.
[96]. 2.93Nigam, K., McCallum, A. K., Thrun, S. & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, *39*, pp. 103-134.
[97]. 2.94Pant, G. & Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems*, *23*, pp. 430-462.
[98]. 2.95NIST (2007). Text REtrieval Conference (TREC) home page. http://trec.nist.gov/.
[99]. 2.96Nowson, S. (2006, June). The Language of Weblogs: A study of genre and individual differences. Ph. D. thesis, University of Edinburgh, College of Science and Engineering.
[100]. 2.97Oh, H.-J., S. H. Myaeng, and M.-H. Lee (2000). A practical hypertext catergorization method using links and incrementally available class information. In SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 264–271. ACM Press.
[101]. 2.98Page, L., S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the Web. Unpublished draft, Stanford University.
[102]. 2.99Pant, G. & Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems*, *23*, pp. 430-462.
[103]. 2.100 Pham Thi Thanh Nam, Bui Quang Minh, Ha Quang Thuy (2004). *A solution to find similary documnets in the Vietnamese search engine VietSeek*. Journal of Informatics & Cybernetics, **20** (4), 293-304.
[104]. 2.101Park, S.-B. and B.-T. Zhang (2003, April). Large scale unstructured document classification using unlabeled data and syntactic information. In Advances in

Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference (PAKDD), Volume 2637 of LNCS, Berlin, pp. 88–99. Springer.
[105]. 2.102Peng, X. and B. Choi (2002). Automatic web page classification in a dynamic and hierarchical way. In Proceedings of the IEEE International Conference on Data Mining (ICDM'02), Washington, DC, pp. 386–393. IEEE Computer Society.
[106]. 2.103Pierre, J. M. (2001, February). On the automated classification of web sites. Linkoping Electronic Articles in Computer and Information Science 6. http://www.ep.liu.se/ea/cis/2001/001/.
[107]. 2.104Qi, X. and B. D. Davison (2006). Knowing a web page by the company it keeps. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), New York, NY, pp. 228–237. ACM Press.
[108]. 2.105Qu, H., A. L. Pietra, and S. Poon (2006, March). Automated blog classification: Challenges and pitfalls. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, pp. 184–186. AAAI Press. Technical Report SS-06-03.
[109]. 2.106R Du, W Susilo, F Safaei and P Boustead (University of Wollongong), Protecting an MPLS-based Programmable Virtual Network using Distributed Firewall.[http://atnac2003.atcrc.com/POSTERS/Du.pdf]
[110]. 2.107Riboni, D. (2002). Feature selection for web page classification. In Proceedings of the Workshop on Web Content Mapping: A Challenge to ICT (EURASIA-ICT).
[111]. 2.108Richardson, M., A. Prakash, and E. Brill (2006). Beyond pagerank: machine learning for static ranking. In WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, pp. 707–715. ACM Press.
[112]. 2.109Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. pp. 130-136). : ACM Press, Seattle, Washington, United States
[113]. 2.110Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (pp. pp. 41-46). : T.J. Watson Research Centre, Seattle, Washington
[114]. 2.111Ruiz, M. E. & Srinivasan, P. (2002). Hierarchical Text Categorization Using Neural Networks. *Information Retrieval*, *5*, pp. 87-118.
[115]. 2.112Rosenfeld, A., R. Hummel, and S. Zucker (1976). Scene labeling by relaxation operations. IEEE Transactions on Systems, Man and Cybernetics 6, 420–433.
[116]. 2.113Rich Caruana, Alexandru Niculescu-Mizil, *An Empirical Comparison of Supervised Learning Algorithms,* ACM International Conference Proceeding Series; Vol. 148.
[117]. 2.114Sebastiani, F. (1999). A tutorial on automated text categorisation. In Proceedings of 1st Argentinean Symposium on Artificial Intelligence (ASAI-99), pp. 7–35.
[118]. 2.115Sebastiani, F. (2002, March). Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47.
[119]. 2.116Sebastiani, F. (2004). Text classification for Web filtering. Report at POESIA Workshop, Present and Future of Open-Source Content-Based Web filtering" Pisa, IT — 21-22 January, 2004

[120]. 2.117Sen, P. and L. Getoor (2007). Link-based classification. Technical Report CS-TR-4858, University of Maryland.
[121]. 2.118Shanks, V. and H. E. Williams (2001, November). Fast categorisation of large document collections. In Proceedings of Eighth International Symposium on String Processing and Information Retrieval (SPIRE), pp. 194–204.
[122]. 2.119Shen, D., Z. Chen, Q. Yang, H.-J. Zeng, B. Zhang, Y. Lu, and W.-Y. Ma (2004). Web-page classification through summarization. In SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 242–249. ACM Press.
[123]. 2.120Shen, D., J.-T. Sun, Q. Yang, and Z. Chen (2006). A comparison of implicit and explicit links for web page classification. In Proceedings of the 15th International Conference on World Wide Web, New York, NY, pp. 643–650. ACM Press.
[124]. 2.121Slattery, S. and T. M. Mitchell (2000, June). Discovering test set regularities in relational domains. In Proceedings of the 17th International Conference on Machine Learning (ICML), pp. 895–902. Morgan Kaufmann.
[125]. 2.122Sun, A. and E.-P. Lim (2001, November). Hierarchical text classification and evaluation. In Proceedings of IEEE International Conference on Data Mining (ICDM),Washington, DC, pp. 521–528. IEEE Computer Society.
[126]. 2.123Sun, A., E.-P. Lim, and W.-K. Ng (2002). Web classification using support vector machine. In WIDM '02: Proceedings of the 4th International Workshop on Web Information and Data Management, New York, NY, pp. 96–99. ACM Press.
[127]. 2.124Tian, Y., T. Huang, W. Gao, J. Cheng, and P. Kang (2003). Two-phase web site classification based on hidden markov tree models. In WI '03: Proceedings of the IEEE/WIC International Conference on Web Intelligence, Washington, DC, USA, pp. 227. IEEE Computer Society.
[128]. 2.125T.H.**Ng**, W.B.**Goh**, and K.L. **Low**, "Feature selection, perceptron learning and a usability case study for text categorization", Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval
[129]. 2.126Thuy Q. Ha, Nam H. Nguyen, and Trang T. Nguyen (2006). Improve Performance of PageRank Computation with Connected-Component PageRank. In *Proceeding of the International Conference on Mobile Computing, Communications and Applications 2006* (ICMOCCA2006), 16-17 August, 2006, Seoul, Korea, 154-158.
[130]. 2.127Tresch, M. & Luniewski, A. (1995). Extensible classifier for semi-structured documents. In *Proceedings of the fourth international conference on Information and knowledge management* (pp. pp. 226-233). ACM Press, Baltimore, Maryland, United States
[131]. 2.128Utard, H. and J. F¨urnkranz (2005, October). Link-local features for hypertext classification. In Semantics, Web and Mining: Joint International Workshops, EWMF/KDO, Volume 4289 of LNCS, Berlin, pp. 51–64. Springer.
[132]. 2.129Wang, G. & Lochovsky, F.H. (2004). Feature selection with conditional mutual information maximin in text categorization. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. pp. 342-349). ACM Press, Washington, D.C., USA
[133]. 2.130Wibowo,W. and H. E.Williams (2002a). Simple and accurate feature selection for hierarchical categorisation. In DocEng '02: Proceedings of the 2002

ACM Symposium on Document Engineering, New York, NY, pp. 111–118. ACM Press.
[134]. 2.131Wibowo, W. and H. E. Williams (2002b). Strategies for minimising errors in hierarchical web categorisation. In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02), New York, NY, pp. 525–531. ACM Press.
[135]. 2.132Wolpert, D. (1992). Stacked generalization. Neural Networks 5, 241–259.
[136]. 2.133Xue, G.-R., Y. Yu, D. Shen, Q. Yang, H.-J. Zeng, and Z. Chen (2006). Reinforcing web-object categorization through interrelationships. Data Mining and Knowledge Discovery 12(2-3), 229–248.
[137]. 2.134Xu, J. & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, *16*, pp. 61-81.
[138]. 2.135Yahoo!, Inc. (2007). Yahoo! http://www.yahoo.com/.
[139]. 2.136Yang, H. and T.-S. Chua (2004a). Effectiveness of web page classification on finding list answers. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, pp. 522–523. ACM Press.
[140]. 2.137Yang, H. and T.-S. Chua (2004b). Web-based list question answering. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, pp. 1277. Association for Computational Linguistics.
[141]. 2.138Yang, Y. and J. O. Pedersen (1997). A comparative study on feature selection in text categorization. In ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, pp. 412–420. Morgan Kaufmann.
[142]. 2.139Yang, Y., S. Slattery, and R. Ghani (2002). A study of approaches to hypertext categorization. Journal of Intelligent Information Systems 18(2-3), 219–241.
[143]. 2.140Yang, Y., J. Zhang, and B. Kisiel (2003). A scalability analysis of classifiers in text categorization. In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, pp. 96–103. ACM Press.
[144]. 2.141Yi, J. & Sundaresan, N. (2000). A classifier for semi-structured documents. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. pp. 340-344). : ACM Press, Boston, Massachusetts, United States
[145]. 2.142Yu, H., Han, J. & Chang, K.C. (2002). PEBL: positive example based learning for Web page classification using SVM. In *Proceedings of the eighth ACM SIGKDD*
[146]. 2.143Yu, H., J. Han, and K. C.-C. Chang (2004). PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering 16 (1), 70–81.
[147]. 2.144Zelikovitz, S. and H. Hirsh (2001). Using LSI for text classification in the presence of background text. In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM), New York, NY, pp. 113–118. ACM Press.
[148]. 2.145Zhang, D. and W. S. Lee (2003). Question classification using support vector machines. In Proceedings of the 26th Annual International ACM SIGIR

Conference on Research and Development in Informaion Retrieval, New York, NY, pp. 26–32. ACM Press.

[149]. 2.146Zhang, D., Chen, X. & Lee, W.S. (2005b). Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. pp. 266-273). ACM Press, Salvador, Brazil

[150]. 2.147Zhang, Y., Zincir-Heywood, N. & Milios, E. (2005c). Narrative text classification for automatic key phrase extraction in web document corpora. In *Proceedings of the 7th annual ACM international workshop on Web information and data management* (pp. pp. 51-58). : ACM Press, Bremen, Germany

[151]. 2.148Zhang, L., Zhu, J. & Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, *3*, pp. 243-269.

[152]. 2.149Zheng, Z. & Webb, G.I. (2000). Lazy Learning of Bayesian Rules. *Machine Learning*, *41*, pp. 53-84.

[153]. 2.150Yang, Y., and Pedersen, J.O., 1997. A comparative study on feature selection in text categorization, In *Proc. of Int'l Conf. on Machine Learning* (ICML-97), pp. 412-420.

[154]. 2.151Pui Y. Lee, Siu C. Hui, and Alvis Cheuk M. Fong. Neural Networks for Web Content Filtering. Intelligent Systems, 17(5):48-57, Sep/Oct, **2002.**

[155]. 2.152Lan N. Bui, Anh Q. Tran, Thuy Q. Ha (2006). User authentic Rating based on Email Networks. In *Proceeding of the International Conference on Mobile Computing, Communications and Applications 2006* (ICMOCCA2006*)*, 16-17 August, 2006, Seoul, Korea, 144-148.

[156]. 2.153Nguyen Thi Huong Thao, Nguyen Thi Thuy Linh, Nguyen Thu Trang, Wonjun Lee, Ha Quang Thuy (2007). A Semi-Supervised SVM solution for Vietnamese web pages classification. The IEEE 22nd International Conference on Advanced Information Networking and Applications (AINA 2008), March 25 - March 28, 2008, GinoWan, Okinawa, Japan (*submitted, reviewing*).

[157]. 2.154Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen and Quang-Thuy Ha (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. In *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, November 1-3, 2006, Wuhan, China, 215-222

[158]. 2.155Dang Thanh Hai, Wonjun Lee, Ha Quang Thuy (2007). A pageranking based method for identifying characteristic genes of a disease. 2nd International Symposium on Agents and Multi-agent Systems – Technologies and Applications (KES-AMSTA-2008), 26 Mar. – 28 Mar. 2008, Incheon, Korea (*submitted, reviewing*).

[159]. 2.156Nguyen Viet Cuong, Nguyen Thi Thuy Linh Ha Quang Thuy and Phan Xuan Hieu (2006). A Maximum Entropy Model for Text Classification. In *The International Conference on Internet Information Retrieval 2006*, Hankuk Aviation University, December 6, 2006, Goyang-si, Korea, 134-139.

[160]. 2.157Nguyễn Việt Cường, Nguyễn Thị Thùy Linh, Phan Xuân Hiếu, Hà Quang Thụy (2005). Bài toán lọc và phân lớp nội dung Web tiếng Việt với hướng tiếp cận Entropy cực đại. *Kỷ yếu Hội thảo quốc gia lần thứ VIII "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, Hải Phòng, 24-26/8/2005, 174-189

[161]. 2.158Son Doan, Quang Thuy Ha, and Susumu Horiguchi (2006). A General Fuzzy-based Framework for Text Representation and its Application to Text Categorization. In *Third International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2006), Lecture Notes on Artificial Intelligence (LNAI), **4223**, 611-620 (Springer-Verlag Berlin Heidenberg, 2006).
[162]. [1] 2b.1Nguyễn Thu Trang, *LinkSpam với đồ thị Web và hạng trang Web,* khóa luận bậc Đại học, Ngành CNTT, Trường ĐH Công Nghệ, 2006.
[163]. [2] 2b.2 Thư viện Dsig http://www.w3.org/PICS/refcode/DSig/
[164]. [3] 2b.3Website: http://www.w3.org/TR/REC-PICSRules; http://www.w3.org/TR/REC-PICS-labels; http://www.w3.org/PICS/
[165]. 2c.1A. Blum and T. M. Mitchell (1998). Combining labeled and unlabeled data with co_training. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Madison, Wisconsin, USA, 1998, ACM, 92-100.
[166]. 2c.2A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
[167]. 2c.3Bennett, K., & Demiriz, A. (1999). Semi-supervised support vector machines.
[168]. Advances in Neural Information Processing Systems, 11, 368–374.
[169]. 2c.4Collins M. and Singer, Y. (1999) Unsupervised Models for Named Entity Classification. Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.
[170]. 2c.5Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT 2005).
[171]. 2c.6David Yarowsky (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics:* 189-196, 26-30 June 1995, MIT, Cambridge, Massachusetts, USA.
[172]. 2c.7Demirez, A., & Bennett, K. (2000). Optimization approaches to semisupervised learning. In M. Ferris, O. Mangasarian and J. Pang (Eds.), *Applications and algorithms of complementarity*. Boston: Kluwer Academic Publishers.
[173]. 2c.8Dang Thanh Hai, Nguyen Huong Giang, Ha Quang Thuy, *Naive Bayes text classification algorithm and problem of specifying clasifying threshold in search engine*, Journal of Computer Science And Cybernetics, **21**(2), 2005, 152-161 *(in Vietnamese)*.
[174]. 2c.9Fung, G., &Mangasarian, O. (1999). *Semi-supervised support vector machines for unlabeled data classification* (Technical Report 99-05). Data Mining Institute, University of Wisconsin Madison.
[175]. 2c.10Goldman, S., & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. *Proc. 17th International Conf. on Machine Learning* (pp. 327–334). Morgan Kaufmann, San Francisco, CA.
[176]. 2c.11Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proc. 16th International Conf. on Machine Learning* (pp. 200–209). Morgan Kaufmann, San Francisco, CA.

[177]. 2c.12Jones, R. (2005). *Learning to extract entities from labeled and unlabeled text* (Technical Report CMU-LTI-05-191). Carnegie Mellon University. Doctoral Dissertation.
[178]. 2c.13Maeireizo, B., Litman, D., & Hwa, R. (2004). Co-training for predicting emotions with spoken dialogue data. *The Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
[179]. 2c.14Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.
[180]. 2c.15Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability
[181]. of co-training. Ninth International Conference on Information and Knowledge
[182]. *Management* (pp. 86–93)
[183]. 2c.16Riloff, E., Wiebe, J., &Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*.
[184]. 2c.17Riloff, E. and Jones, R. (1999) Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)* , 1999, pp. 474-479.
[185]. 2c.18Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised selftraining of object detection models. *Seventh IEEE Workshop on Applications of Computer Vision*.
[186]. 2c.19Vapnik, V. (1998). *Statistical learning theory*. Springer.
[187]. 2c.20Xu, L., & Schuurmans, D. (2005). Unsupervised and semi-supervised multi-class support vector machines. *AAAI-05, The Twentieth National Conference on Artificial Intelligence*.
[188]. 2c.21Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189–196).
[189]. 2c.22Zhou, Y., & Goldman, S. (2004). Democratic co-learing. Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004).
[190]. [1] 2d.1Vezhnevets V., Sazonov V., Andreeva A., *A* Survey on Pixel-Based Skin Color *Detection Techniques*. Proc. Graphicon-2003, pp. 85-92, Moscow, Russia, September 2003
[191]. [2] 2d..2M.J. Jones, J.M. Rehg, .Statistical color models with application to skin detection., *Computer Vision and Pattern Recognition*, 274.280, 1999
[192]. [3] 2d..3Jean-Christophe Terrillon and M. N. Shirazi and H. Fukamachi and S. Akamatsu, Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images., *Fourth International Conference On Automatic Face and gesture Recognition*, 54-61,2000.
[193]. [4] 2d..4PEER, P., KOVAC, J., and SOLINA, F. 2003. Human skin colour clustering for face detection. In *submitted to EUROCON 2003 – International Conference on Computer as a Tool*.
[194]. [5] 2d..5L.M. Bergasa and M.Mazo. and A. Gardel and M.A. Sotelo and L. Boquete, .Unsupervised and adapative Gaussian skin-color model., *Image and Vision Computing*, 2000,18,987-1003.

[195]. [6] 2d..6A. Bosson, G.C. Cawley, Y. Chian, R. Harvey, .Non-retrieval: blocking pornographic images., *Proc. Intl Conf. on the Challenge of Image and Video Retrieval, Lecture Notes in Computer Science*, Springer-Verlag, London, 2383:50.60, 2002.

[196]. [7] 2d..7B. Jedynak, H. Zheng, M. Daoudi, .Statistical Models for Skin Detection., *IEEE Workshop on Statistical Analysis in Computer Vision*, in conjunction with CVPR 2003 Madison, Wisconsin, June 16.22, 2003.

[197]. [8] 2d..8B. Jedynak, H. Zheng, M. Daoudi, .Blocking Adult Images Based on Statistical Skin Detection., *IEEE Workshop on Statistical Analysis in Computer Vision*, in conjunction with CVPR 2003 Madison, Wisconsin, June 16.22, 2003.

[198]. [9] 2d..9J.Z. Wang, J. Li, G. Wiederhold, O. Firschein, .System for Screening Objectionable Images., *Computer Communications* (21)15:1355.1360, 1998.

[199]. [10] 2d.10M.M. Fleck, D.A. Forsyth, C. Bregler: .Finding naked people., *Proc. European Conf. on Computer Vision*, B. Buxton, R. Cipolla, Springer-Verlag, Berlin, Germany, 2:593.602, 1996.

[200]. [11] 2d.11Lee J. Y., and Yoo, S. I. 2002. An elliptical boundary model for skin color detection. In *Proc. of the 2002 International Conference on Imaging Science, Systems, and Technology*.

[201]. [12] 2d.12L. Sigal, S. Sclaroff and V. Athitsos, .Skin Color-Based Video Segmentation under Time-Varying Illumination . *IEEE Trans. on PAMI, 26(7):862-877, July, 2004*

[202]. 2.177[Lanq01] Carsten Lanquillon (2001). Enhancing Text Classification to Improve Information Filtering, *PhD thesis*, University of Madgdeburg, Germany, 2001

[203]. 2.178[Zhan05] Yi Zhang (2005). Bayesian Graphical Models for Adaptive Filtering, *PhD thesis*, Carnegie Mellon University, 2005.

[204]. 2.179[QD07] Xiaoguang Qi and Brian D. Davison (2007). Web Page Classification: Features and Algorithms, *ACM Comput. Surv.*, 41 (2): 1-31, 2009.

[205]. 2.180Michael J. Jones, James M. Rehg (2002). Statistical Color Models with Application to Skin Detection, *International Journal of Computer Vision* 46(1): 81-96 (2002).

[206]. 2.181Huicheng Zheng, Mohamed Daoudi and Bruno Jedynak (2004). Blocking adult images based on statistical skin detection, *Electronic Letters on Computer Vision and Image Analysis*, 4 (2): 1–14, 2004.

[207]. 2.182Michael J. Jones, James M. Rehg (1999). Statistical Color Models with Application to Skin Detection, *CVPR 1999*: 1274-1280

[208]. 2.183Vladimir Vezhnevets, Vassili Sazonov, Alla Andreeva (2003). A Survey on Pixel-Based Skin Color Detection Techniques, *Proc. Graphicon-2003*: 85–92

[209]. 2.184Bruno Jedynak, Huicheng Zheng, Mohamed Daoudi, Didier Barret (2002). Maximum Entropy Models for Skin Detection, *ICVGIP 2002*

[210]. 2.185Bruno Jedynak, Huicheng Zheng, Mohamed Daoudi (2003). Maximum Entropy Models for Skin Detection, *EMMCVPR 2003*: 180-193

[211]. 2.186D.A. Forsyth and M.M. Fleck (1999). Automatic Detection of Human Nudes**,** *Int. Jour. of Computer Vision*, 32**(1)**: 63-77, Aug., 1999.

[212]. 2.210G. Gomez, E. Morales (2002). Automatic feature construction and a simple rule induction algorithm for skin detection, Proceedings of Workshop on Machine Learning in Computer Vision, 2002, pp. 31-38.

[213]. 2.212 BRAND, J., AND MASON, J. (2000). A comparative assessment of three approaches to pixellevel human skin-detection, *Proc. of the International Conference on Pattern Recognition*, vol. 1, 1056–1059.

[214]. 2.211Brown, M.S., McNitt-Gray, M.F., Goldin, J.G. et al. 2001. Patientspecific models for lung nodule detection and surveillance in CT images. *IEEE Trans. Med. Imaging*., 20(12):1242–50.

[1]. 3.1Burt D (2000) Dangerous Access, 2000 Edition: Uncovering Internet Pornography in America's Libraries, Family Research Council, Washington DC, http://www.copacommission.org/papers/bi063.pdf

[2]. 3.2Calandra T (2001) CBS MarketWatch, The Exchange, 17 (1) http://www.esignal.com/exchange/v17_07/stockwatch.asp, last accessed June 1st 2002

[3]. 3.2Censorware (1999), Censored internet access in Utah public schools and libraries. The Censorware Project. http://www.bpl.burnaby.bc.ca/weblog.htm last accessed July 26th 2002

[4]. 3.4Censorware. (1998a). Cyber Patrol and Deja News. The Censorware Project. http://www.censorware.net/reports/cyberpatrol.html, last accessed July 12th 2002

[5]. 3.5 Censorware. (1998b). The X-Stop files: Deja voodo. The Censorware Project. http://censorware.net/reports/xstop/, last accessed July 24th 2002

[6]. 3.6.Center for Media Education. (1997). Youth access to alcohol and tobacco web marketing: The Filtering and rating debate, March 1997, http://www.cme.org/children/marketing/execsum.html, last accessed July 25th 2002

[7]. 3.7Donert K (2002), European Citizenship, a new educational focus: some opportunities for ODL, Proc. European Distance Education Network (EDEN) Conference, 110-115.

[8]. 3.8dotSafe Project (2002), dotSafe Project Report, http://www.eun.org/eun.org2/eun/index_dotsafe.cfm, last accessed July 29th 2002

[9]. 3.9European Commission (2001), EU schools embrace the Internet, but discrepancies remain, http://dbs.cordis.lu/cgi-bin/srchidadb Oct 12 2001, last accessed May 18th 2002

[10]. 3.10European Commission (2002a), eEurope Action Plan, http://europa.eu.int/information_society/eeurope, last accessed June 1st 2002

[11]. 3.11European Commission (2002b), Education and Culture: Cultural Activities, http://europa.eu.int/comm/culture/eac/index_en.html, accessed June 1st 2002

[12]. 3.12European Commission (2002c), eEurope 2002, European youth into the digital age, http://europa.eu.int/information_society/eeurope/news_library/documents/ index_en.hm, accessed June 8th 2002

[13]. 3.13European Commission (2002d), Benchmarking Survey: eEurope 2002, http://europa.eu.int/comm/public_opinion/, accessed June 28th 2002

[14]. 3.14European Schoolnet (2002), dotSafe Project,

[15]. 3.15http://www.eun.org/eun.org2/eun/en/index_eun.html, last accessed July 27th 2002

[16]. 3.16Gill L (2002), Experts: Filtering Software a Failure, News Factor Network, http://www.newsfactor.com/perl/story/16980.html#story-start, March 27th 2002, date accessed July 16th 2002

[17]. 3.17Greenfield P, McCrea P and Ran S (1999), NOIE Filter Evaluation Report: CSIRO, December 1999, http://www.noie.gov.au/publications/NOIE/consumer/CSIROfinalreport.html, last accessed March 6 2002

[18]. 3.18Greenfield P, Rickwood P and Tran HC (2001), Effectiveness of Internet Filtering Software Products, CSIRO, September 2001

[19]. 3.19Haddon L (2001), Issues in Managing Children's Internet Use: a Report for the European Commission, http://www.saferinternet.org/downloads/Haddon.doc, last accessed July 25th 2002

[20]. 3.20Hunter CD (1999), Internet filter effectiveness: Testing over and underinclusive blocking decisions of four popular filters, Social Science Computer Review, 18 (2).

[21]. 3.21ICRAsafe (2002), ICRAfilter Project, http://www.icra.org/, last accessed July 26th 2002.

[22]. 3.22IDATE (1999), Prepact: Review of EU Third party filtering and rating software and services (Lot 3), Final Report, Vol. 1, Dec 1999, http://www.idate.org, last accessed July 24th 2002

[23]. 3.23Internet Watch Foundation (1998), Internet Watch Foundation Rating and Filtering Internet Content, A United Kingdom Perspective, March 1998, http://www.iwf.org.uk/label/rating_r.htm, last accessed July 26th 2002

[24]. 3.24Jackson TO (2000), Concepts Document: Benchmarking of Filtering Software and Services, Oct 2000 http://e-filter.jrc.it, date accessed 19 July 2002

[25]. 3.25Jackson TO, Riva M and Puglisi F (2001), Benchmarking of filtering software and services – An analysis framework: Definition of the Evaluation Criteria, Issue 1 Draft 2, Joint Research Centre of the European Commission, ISPRA, Italy

[26]. 3.26Kerr D (2000a) INCORE (Internet COntent Rating for Europe): Final Report Executive Summary, http://www.incore.org/paper/paper.htm, last accessed May 3rd 2002

[27]. 3.27 Kerr D (2000b), INCORE (INternet COntent Rating for Europe) Project, Final Report, April 2000

[28]. 3.28McCullagh D (1999), Looking for something to blame, Wired News, http://www.wired.com/, April 23rd 1999, last accessed February 21st 2002

[29]. 3.29Montgomery KC (2002), Digital Kids: The New On-Line Children's Consumer Culture, http://www.cme.org/publications/Singer35.pdf, last accessed July 27th 2002

[30]. 3.30Montgomery KC and Pasnik S (1996), Web of Deception: Threats to Children from Online Marketing, http://www.cme.org/children/marketing/deception.pdf, last accessed July 2nd 2002

[31]. 3.31NCTE (2002a), Young children find pornography on the Net, http://www.ncte.ie/ICTAdviceSupport/TheInternet/InternetSafety/Risks/, last accessed June 21st 2002

[32]. 3.32NetProtect (2002), NetProtect Project: a European Prototype of Internet Access Filtering, http://www.net-protect.org, last accessed July 26th 2002

[33]. 3.33NUA (2001), eLearning to thrive in Europe, http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905357158&rel=true September 4th 2001, accessed June 3rd 2002

[34]. 3.34Parajon M (2002), Safer use of interactive technologies Progress and prospects, Workshop Luxembourg, 11-12 June 2001, http://www.saferinternet.org/downloads/ Speech.doc, last accessed July 29th 2002

[35]. 3.35PC Magazine (2002), Benchmarking tests: Internet Filtering, http://www.pcmag.com/article2/, last accessed July 19th 2002

[36]. 3.36RSACi (1999), Recreational Software Advisory Council: About RSACi, http://www.rsac.org/, last accessed July 15th 2002

[37]. 3.37Sims M (1998), Why censorware can't work. The Censorware Project, http://www.censorware.org/essays/whycant_2_ms.html, 6th April 1998, last accessed July 12th 2002

[38]. 3.38Turner K and Kendall M (2000), Public use of the Internet at Chester Library, UK, Information Research, 5 (3), April 2000, http://informationr.net/ir/5-3/paper75.html, last accessed July 19th 2002.

[39]. 3.39Wallace J and Mangan M (1996), Sex, Laws, and Cyberspace. New York, NY: Henry Holt & Company

[40]. 3.40Wilkins J (1997), Protecting our children from Internet smut: moral duty or moral panic? The Humanist, 17th September 1997, 57, (5), 4

[41]. 3.41 Chakrabarti, S., B. E. Dom, and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In SIGMOD '98: Proceedings of the ACM SIGMOD International Conference on Management of Data, New York, NY, pp. 307–318. ACM Press.

[42]. 3.42Cancedda, N., Gaussier, E., Goutte, C. & Renders, J.M. (2003). Word sequence kernels. *Journal of Machine Learning Research*, *3*, pp. 1059-1082.

[43]. 3.43Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen and Quang-Thuy Ha (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. In *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, November 1-3, 2006, Wuhan, China, 215-222.

[44]. (3.45)Poesia Software Architecture Definition Document, 2002

PHỤ LỤC

1. Bìa và Lời giới thiệu của 10 luận văn Thạc sỹ

- Phạm Tiến Dũng (2009). Nghiên cứu giải pháp lọc nội dung Internet tại máy tính cá nhân và xây dựng phần mềm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009
- Lê Đắc Nhường (2009). Tối ưu hóa truy vấn trong máy tìm kiếm thực thể, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009
- Nguyễn Thu Trang (2009). Học xếp hạng trong tính hạng đối tượng và tạo nhãn cụm tài liệu, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009
- Trần Thị Oanh (2009). Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt, *Luận văn Thạc sỹ*, Trường ĐHCN, 2009
- Nguyễn Cẩm Tú (2009). Hidden Topic Discovery Towards Classification and Clustering in Vietnamese Documents*, Luận văn Thạc sỹ (viết bằng tiếng Anh),* Trường ĐHCN, 2008
- Ngô Thương Huyền (2008). Phân lớp thư điện tử sử dụng máy hỗ trợ vector, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008
- Nguyễn Thị Thu Hằng (2008). Phương pháp phân cụm tài liệu Web và áp dụng vào máy tìm kiếm, *Luận văn Thạc sỹ*, Trường ĐHCN, 2008
- Nguyễn Việt Cường (2007). Tự động sinh mục lục cho văn bản, *Luận văn Thạc sỹ*, Trường ĐHCN, 2007
- Đặng Thanh Hải (2007). The Biological Sample Classification Using Gene Expression Data, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHCN, 2007
- Nguyễn Hoài Nam (2006). The WWW and The PageRank-Related Problems, *Luận văn Thạc sỹ (viết bằng tiếng Anh)*, Trường ĐHKHTN, 2006

2. Hai công trình khoa học công bố quốc tế điển hình

- Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha (2009). Classification and Contextual Match on the Web with Hidden Topics from Large Data Collections, *The IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (accepted, 14 pages)*. ISI Journal System
- Cam-Tu Nguyen, Xuan-Hieu Phan, Susumu Horiguchi, Thu-Trang Nguyen, Quang-Thuy Ha (2009). Web Search Clustering and Labeling with Hidden Topics, *ACM Transactions on Asian Language Information Processing,* **8**(3), 12 (August 2009), 40 pages. DOI=10.1145/1568292.1568295. http://doi.acm.org/10.1145/1568292.1568295. ISI Journal System