Mục lục

DANH MỤC BẢNG:

DANH MỤC HÌNH:

DANH MỤC TỪ VIẾT TẮT:

ADSL	: Asymmetric Digital Subscriber Line
CMAE	: Content Management in Adversarial Environments
COSIM	: Cosine Simarility
DNS	: Domain Name Service
DWK	: Depraved Web Killer
FTP	: File Transfer Protocol
HTTP	: Hypertext Transfer Protocol
IP	: Internet Protocol (nghi thức mạng)
IR	: Information Retrieve
ISP	: Internet Service Provider
SIM	: Simarility
Stopword	: danh mục các từ không ảnh hưởng đến nội dung văn bản
TCP	: Transmission Control Protocol
URL	: Uniform Resource Locator

Lời nói đầu

Xuất phát từ một hiện trạng sử dụng Internet ngày càng phổ biến, các dịch vụ truy cập internet phát triển mạnh mẽ. Cùng với yêu cầu quản lý chặt chẽ việc sử dụng dich vụ web của nhiều người và tránh sử dụng những trang web "đen" làm băng hoại đạo đức xã hội nhất là đối với giới trẻ.

Với những yêu cầu bức thiết đó, người nghiên cứu tiến hành thực hiện đề tài này, trên cơ sở công nghệ phân loại văn bản, đề tài này muốn đạt tới một giải pháp lọc những trang web "đen" hay những trang web "cấm" một cách hiệu quả.

Tuy nhiên, một vấn mà đề tài này chưa đạt tới là lọc những hình ảnh trong một trang web, đây là vấn đề cần được đầu tư thêm nhằm hoàn thiện mục tiêu chính đó là công việc lọc web.

Qua đây tôi xin chân thành cảm ơn *Tiến Sĩ Nguyễn Văn Hiệp* – Giảng viên Trường Đại Học Bách Khoa Tp. HCM đã tận tình hướng dẫn, truyền đạt kiến thức đến tôi, cám ơn các *Anh/Chị* phòng *Sau đại học Trường Đại Học Công Nghệ Thông Tin (thuộc Đại Học Quốc Gia TP. HCM)*, cám ơn các bạn bè đã giúp đỡ và động viên tôi hoàn thành luận văn này.

Người viết

Dương Đình Dũng

Chương 1: TỔNG QUAN

1.1. Giới thiệu:

Đề tài được chọn dựa trên hiện trạng sử dụng mạng Internet hiện nay, tại gia đình cũng như tại các dịch vụ. Môi trường Internet phát triển mạnh mẽ nhưng nó cũng tiềm ẩn những mối nguy hiểm trong đó, như những hình ảnh khiêu dâm, bạo lực và nhiều vấn đề không lành mạnh khác,… sẽ có tác động xấu đến người dùng internet nhất là giới trẻ - những người chưa ý thức đầy đủ về những nguy hại đó. Vì thế, vấn đề lọc web được nhiều người đầu tư với nhiều cách làm khác nhau, nhưng cùng hướng đến một mục tiêu là ngăn chặn những trang web độc hại.

1.2. Mục tiêu nghiên cứu:

Vấn đề nghiên cứu chính trong đề tài này là ngăn chặn các trang web và kiến thức sử dụng để xây dựng một bộ lóc web là công nghệ tri thức, cụ thể là phân loại văn bản (Text Classification). Đề tài cần phải đạt được những nội dung sau:

- ❒ Nghiên cứu những điểm mạnh của kỹ thuật phân loại văn bản và kiến thức cần thiết cho kỹ thuật này nhằm áp dụng nó tốt nhất vào đề tài nghiên cứu.
- ❒ Xác định những điểm bất cập từ những chương trình ứng dụng lọc web hiện có và những điểm mạnh yếu của những phương pháp xây dựng bộ lọc web.
- ❒ Đề xuất phương pháp lọc web và xây dựng mô hình, cài đặt một bộ lọc web dùng kỹ thuật phân loại văn bản. Đặt thuật toán lọc web đã cài đặt lên proxy để thử nghiệm và đánh giá thuật toán.
- ❒ Chỉ ra những cải tiến cần thiết cho đề tài.

1.3. Phạm vi nghiên cứu:

1.3.1. Tổng quan chung về vấn đề:

Do tính chất của mạng internet là truyền tải thông tin rộng rãi và đa dạng, nên việc khai thác nó cũng gặp nhiều rắc rối. Vấn đề không phải là dùng nó khó khăn mà giới hạn việc sử dụng nó cho từng đối tượng, dùng đúng nguồn thông tin mạng và nhất là những vấn đề có liên quan đến đạo đức, tư tưởng, thuần phong mỹ tục của người dùng và vấn đề an ninh của một quốc gia.

Ngay từ khi Internet và web phát triển vấn đề bảo mật cũng như loại bỏ những trang web chứa nội dung không lành mạnh không cho người dùng truy cập đến đã được đặt ra. Qua một thời gian dài phát triển, theo sự phát triển của công nghệ và trí thức của con người mỗi thời kỳ gắn liền với một kỹ thuật ngăn chặn trang web.

Các chương trình lọc web thường dùng các công nghệ như: danh sách trắng và danh sách đen (blacklist và whitelist), ngăn từ khóa (keyword), chặn địa chỉ liên kết (URL hay IP), kiểm soát nội dung (web content), v.v.

Đặc điểm chung: các phần mềm này có thể cài trên proxy server hay client hoặc cả hai. Một số chương trình chỉ chạy trên máy đơn, giúp phụ huynh kiểm soát việc dùng mạng Internet của con em mình tốt hơn. Đối với những cơ quan, tổ chức, doanh nghiệp thì việc dùng chương trình lọc web kiểm soát ngay đầu vào để quản lý toàn bộ các truy cập từ bên trong hay các hành vi đột nhập từ bên ngoài là vô cùng quan trọng. Với một lưu lượng lớn những yêu cầu đi qua làm cho việc xử lý ở đầu kiểm soát trở nên bận rộn hơn và luôn đòi hỏi tốc độ xử lý phải cao hơn để đáp ứng nhanh chóng các yêu cầu, nếu không bộ lọc vô tình tạo nên điểm thắt cổ chai là chậm hệ thống mạng.

1.3.2. Giới hạn vấn đề:

Trong môi trường truyền tải thông tin qua mạng Internet người ta sử dụng rất nhiều ứng dụng đồ họa và multimedia vào trong công cụ truyền tải thông tin (phần lớn là các trang web). Do đó, những nội dung được người khai thác tiếp thu rất đa dạng. Trong đề tài này, người viết chọn một phần nội dung trong một trang web để nghiên cứu đó là văn bản bên thể hiện trong một trang web.

Văn bản là những từ ngữ diễn tả cho những nội dung được con người dùng trong truyền thông. Văn bản dựa trên nền tảng chính là ngôn ngữ mà cộng đồng con người sử dụng để trao đổi với nhau. Việc phát tán trang web cũng vậy, người ta dùng nhiều ngôn ngữ trên thế giới để thể hiện nội dung. Với đề tài này, người viết chọn ngôn ngữ tiếng Anh để nghiên cứu vì đây là ngôn ngữ được sử dụng nhiều trên thế giới.

Trong đề tài này người viết dùng kỹ thuật lọc web bằng công nghệ khai mỏ văn bản (text mining) cụ thể là phân lớp văn bản (text classification). Với phương pháp lọc web này: thực hiện khám phá nội dung và phân tích URL với công nghệ khám phá văn bản. Nhưng nội dung chủ đạo ứng dụng trong luận văn này là kỹ thuật so sánh hai văn bản.

1.4. Ý nghĩa khoa học:

Xét về mặt khoa học, đề tài này là một bổ sung ý tưởng ứng dụng công nghệ tri thức vào lĩnh vực nghiên cứu an toàn mạng.

Xét về mặt kỹ thuật, là một đề tài hướng ứng dụng đến mục tiêu tự động hóa nhằm làm giảm bớt sự can thiệp của con người trong quá trình vận hành hệ thống lọc web.

Xét về tính xã hội, qua đề tài này tôi hy vọng góp một phần vào công việc bảo vệ giá trị đạo đức xã hội, thuần phong mỹ tục và tư tưởng.

1.5. Lý do chọn đề tài và phát biểu bài toán:

Ngày 14 tháng 7 năm 2005, chính phủ Ban hành “Thông tư liên tịch về quản lý đại lý Internet” số 02/2005/TTLT-BCVT-VHTT-CA-KHĐT giữa bốn bộ: Bưu chính Viễn thông; Văn hóa thông tin; Công An; Kế hoạch và Đầu tư. Bắt đầu có hiệu lực vào đầu tháng 8-2005.

Liệu các dịch vụ cho thuê Internet công cộng có thực hiện nghiêm chỉnh thông tư này? Liệu có ngăn chặn được các trang web “đen”? Làm thế nào để quản lý dịch

vụ Internet có hiệu quả? Đó là những câu hỏi đặt ra cho những người làm công nghệ thông tin.

Xây dựng một bộ lọc Web nhằm phục vụ cho vấn đề an toàn trong việc truy cập mạng Internet là một yêu cầu có thật. Nhiều nhà sản xuất phần mềm đã tung ra thị trường một số chương trình lọc web phục vụ cho máy cá nhân hoặc các Firewall của các ISP và cũng có nhiều công nghệ xây dựng chương trình lọc web. Tất cả đều có chung một thực trạng là làm chậm đường truyền do sử dụng các phép kiểm tra và so sánh liên tục, một yếu điểm khác là không tự động cập nhật các hành vi sử dụng web của người dùng.

Chính vì những lý do và điều kiện tự nhiên đó người viết chọn đề tài **"Ứng dụng phân loại văn bản xây dựng bộ lọc Web"** để xây dựng một bộ lọc cơ động và hiệu quả. Với yêu cầu khắc phục những hạn chế của những chương trình cùng loại như sau:

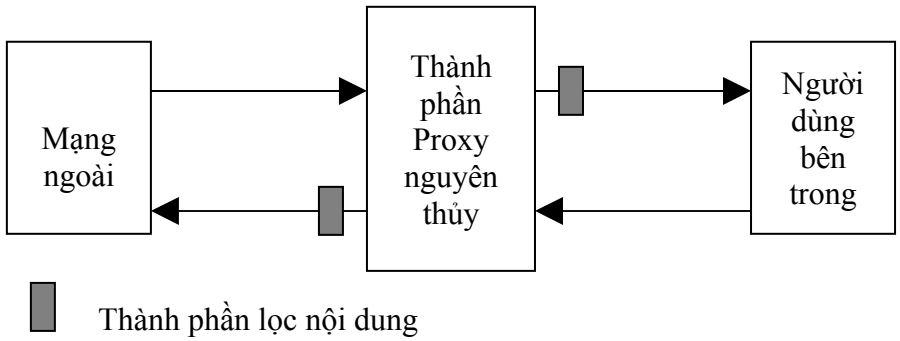
- ❒ Tốc độ làm việc: nhanh, ít làm nghẽn lưu thông mạng.
- ❒ Hiệu quả: ngăn chặn được những trang web có hại
- ❒ Đơn giản: hạn chế sự can thiệp của con người bằng cách tự động cập nhật.
- ❒ Cơ động trong vận hành: làm việc được với nhiều cơ sở dữ liệu huấn luyện cho bộ lọc khác nhau theo từng lĩnh vực.

Vấn đề lọc web bằng phương pháp phân loại văn bản này đã được nhóm nghiên cứu ở *Centre for Communication Security School of Information Technology and Computer Science University of Wollongong, Australia*. Nhóm này gồm các tác giả Rongbo Du, Reihaneh Safavi-Naini và Willy Susilo đã đưa ra đề tài *"Web Filtering Using Text Classification"* [9] và đề tài này được hỗ trợ bởi *Smart Internet Technology Cooperative Research Centre, Australia* vào tháng 11 năm 2003.

Dựa vào bài báo đã trình bày ở trên tôi nghiên cứu có cải tiến và viết thành luận văn của mình. Những cải tiến trong luận văn này tôi hy vọng sẽ cải thiện tốc độ làm việc, độ chính xác và hiệu quả thực thi.

* Phát biểu bài toán: Trong một mạng máy tính được kết nối với môi trường Internet bên ngoài, cho phép người dùng truy cập đến những trang web tại các máy chủ trên mạng. Việc kiểm soát lưu thông mạng là nhằm kiểm soát quá trình sử dụng mạng của người dùng. Trong bài toán này tôi dùng phương pháp phân loại văn bản xây dựng một bộ lọc web đặt tại ngõ ra vào mạng nhằm kiểm soát và ngăn không cho truy cập những trang web xấu!

* Sơ đồ bài toán như sau:



Hình 1.1. Vị trí của bộ lọc nội dung trên proxy của mạng

Hình 1.1 ở trên cho thấy vị trí của bộ lọc đối với việc kiểm soát truy cập của một proxy đang hoạt động trên mạng. Sự lọc web xảy ra sau khi một truy cập được cho phép. Một thành phần lọc web dựa trên văn bản xử lý nội dung trong một định dạng văn bản và vì thế các phương thức lọc như là phương thức so từ (keyword-matching) hay các phương thức tinh vi khác liên quan đến phân tích văn bản có thể được sử dụng.

1.6. Phân tích hiện trạng

1.6.1. Những phần mềm cùng lĩnh vực trên thế giới:

Hiện nay, lĩnh vực này được nhiều nhà sản xuất phần mềm quan tâm, vì những yêu cầu ngăn chặn các trang web xấu ngày càng nhiều. Nhà sản xuất luôn đổi mới

công nghệ nhằm đem đến hiệu quả cao nhất trong việc ngăn chặn trang web xấu và giữ an toàn cho mạng.

Vấn đề lọc Web đã được nhiều nhà phát triển phần mềm trên thế giới xây dựng trên những giải pháp lọc web. Một số phần mềm điển hình như:

- SurfControl – Enterprise Threat Protection: đây là phần mềm của hãng SurfControl, phần mềm này thiết kế theo cách tiếp cận lọc web và ngăn chặn từ proxy qua URL và từ khóa, có khoảng 20 loại kiểu ngăn chặn.
- Internet Filter - Web Filters: do hãng - iPrism internet filters & web filters phát triển, là phần mềm thực hiện giám sát và ngăn chặn. Phần mềm này được quảng cáo dùng kỹ thuật lọc web động kiểm soát nội dung trang web ở ngõ vào. Tuy nhiên, theo hướng dẫn quản trị của nhà sản xuất thì phần mềm này cũng có bóng dáng của kỹ thuật dùng phương pháp lọc chặn từ khóa.
- Internet Security Systems - Proventia Web Filter: ngăn chặn những trang web không mong muốn nhiều hơn các sản phẩm lọc nội dung khác. Là phần mềm được cài đặt ở dạng client/server, nhà sản xuất cung cấp sẵn một danh sách nhiều địa chỉ IP và URL liên quan đến những trang web xấu. Ngoài ra nó cũng cho người quản trị được phép bổ sung thêm những địa chỉ (IP hay URL) mới xuất hiện vào cở sở dữ liệu hiện có của nó.
- ISA Server Web Filter: đây là dịch vụ tích hợp trong sản phẩm ISA (Internet Security and Accelerators) của hãng Microsoft. Phần mềm ISA chạy ở dạng tường lửa vừa kiểm soát truy cập mạng, vừa làm ủy nhiệm (proxy). Cũng giống như các phần mềm trên ISA kèm dịch vụ lọc web với kỹ thuật phân tích liên kết, ngăn địa chỉ (IP và URL), kiểm soát nội dung bằng từ khóa. Các yêu cầu lọc web được người quản trị thiết lập thông qua các luật lọc web.

1.6.2. Tình hình triển khai bộ lọc web ở Việt nam:

Trong những năm gần đây dịch vụ internet phát triển mạnh mẽ do chính sách cung cấp thông tin của nhà nước cũng như công nghệ ADSL ra đời, cùng với sự cạnh tranh của các nhà cung cấp dịch vụ internet. Nên vấn đề tiếp cận với internet của người dân dễ dàng hơn. Từ công sở, trường học, các phòng net công cộng, quán café internet, đến gia đình người dùng truy cập nhiều hơn và việc kiểm soát truy cập cũng như quy định sử dụng mạng cũng được đề cập đến. Tuy nhiên, ở từng cấp độ sử dụng vẫn còn có điều gì đó chưa được chú trọng lắm. Điển hình nhất là những dịch vụ hay quán café Internet, tại những nơi này đa số chỉ dùng bộ định tuyến ADSL để kết nối mà không dùng một bộ lọc nào ngăn chặn, có lẽ do tốc độ truy cập của khách hàng bị ảnh hưởng (?).

Phần lớn các dịch vụ cho thuê Internet ở thành phố Hồ Chí Minh nói riêng và cả nước nói chung hiếm khi sử dụng các bộ lọc web, chủ yếu dùng phần mềm ISA với chức năng Proxy server là chính.

Do những đặc điểm như trên, lĩnh vực phát triển bộ lọc web cũng không phát triển mạnh mẽ, phần mềm điển hình nhất trong lĩnh vực này của Việt nam là DWK (Depraved Web Killer tác giả Vũ Lương Bằng Công ty Điện Thoại Đông Thành phố thực hiện). Phần mềm này vừa được cập nhật phiên bản 2.4. Với các chức năng mới như: Kiểm tra nội dung của trang Web. Gửi tập tin báo cáo tới địa chỉ mail được chỉ định. Ngăn chặn các trang Web cài phần mềm gián điệp, quay số quốc tế, quảng cáo. Tự động cập nhật danh sách các từ khóa, trang web cấm từ mạng, thông báo khi có phiên bản DWK mới. Tuy nhiên, DWK 2.4 chưa thuyết phục được người dùng mạng, do tính hiệu quả, tốc độ làm việc, cũng như chủ định của con người!

Thực hiện chủ trương của Bộ Bưu chính viễn thông về quản lý đại lý Internet, Công ty Điện toán và Truyền số liệu VDC đã phối hợp với các bưu điện tỉnh/thành phố và các công ty viễn thông triển khai cài đặt phần mềm quản lý Internet công cộng cho các đại lý thuộc mạng VNN. INCMwin là một giải pháp phần mềm quản lý đại lý Internet chạy trên nền hệ điều hành Windows. Một số tính năng chính của

INCMwin gồm: ngăn chặn các trang web có nội dung xấu, tự động cập nhật danh sách các web đen cần phải chặn, trợ giúp các đại lý Internet trong việc ghi thông tin khách hàng sử dụng dịch vụ, tính tiền sử dụng dịch vụ ... Tuy nhiên, sau một thời gian ngắn triển khai thì phần mềm này bị phản ứng của các phòng net rất mạnh mẽ, có nơi đã gỡ bỏ với cùng lý do như những phần mềm lọc web kể trên.

1.7. Sơ lược về khai mỏ văn bản (text mining):

Khai mỏ văn bản (Text Mining) là sự khám phá những thông tin mới hay trước đó không biết đến bằng máy tính thông qua việc trích xuất thông tin tự động từ những nguồn tài liệu khác nhau. Một yếu tố then chốt là sự liên kết với nhau của thông tin được trích xuất theo dạng sự kiện mới hay giả thuyết mới để phát hiện ra nhiều hơn bằng phương thức thử nghiệm thông thường.

Khai mỏ văn bản có sự khác biệt với những gì ta biết trong bộ tìm kiếm trên web. Trong đó người dùng đơn thuần tìm được những gì đã biết trước và được đưa ra bởi những người khác. Vấn đề khai mỏ văn bản là sự tách bạch rõ ràng giữa những tư liệu hiện có không liên quan cho sự cần thiết của người dùng với mục đích tìm ra những thông tin liên quan.

Trong khai mỏ văn bản, mục tiêu là khám phá thông tin chưa biết trước đây, những điều mà người ta chưa rõ lúc ấy nên cũng không thể viết ra.

Khai mỏ văn bản là một sự biến thể trong lĩnh vực khai mỏ dữ liệu, mà khai mỏ dữ liệu nhằm tìm ra những mô hình quan tâm từ những cơ sở dữ liệu lớn. Một ví dụ điển hình trong khai mỏ dữ liệu là việc dùng các hóa đơn mua hàng của khách hàng để dự đoán những sản phẩm nào đặt gần nhau trên kệ hàng. Ví dụ: nếu khách hàng mua một đèn flashlight, người bán luôn muốn bán thêm viên pin cho đèn flashlight vì thế chúng phải được đặt gần nhau. Một ứng dụng có liên quan là việc phát hiện sự gian lận trong sử dụng thẻ tín dụng. Các nhà phân tích tìm kiếm trên một lượng lớn thẻ tín dụng thu thập để tìm ra sự chênh lệch từ những thẻ tiêu xài bình thường.

Khác biệt cơ bản giữa khai mỏ dữ liệu thông thường và khai mỏ văn bản là trong khai mỏ văn bản những mẫu được trích xuất từ văn bản ngôn ngữ tự nhiên

thay cho việc dùng những cơ sở dữ liệu có cấu trúc của những sự kiện. Cơ sở dữ liệu được thiết kế từ những chương trình để xử lý một cách tự động; Văn bản được viết bởi con người để đọc.

Tuy nhiên, có một lĩnh vực được gọi là xử lý ngôn ngữ bằng máy tính (còn gọi là xử lý ngôn ngữ tự nhiên) mà quá trình này tạo ra hàng loạt xử lý trong khi thực hiện những tác vụ nhỏ trong phân tích văn bản. Chẳng hạn, người ta viết ra một chương trình trích xuất những cụm từ, từ những bài báo hay quyển sách tương đối dễ dàng, khi hiển thị cho người đọc bản tóm lược nội dung văn bản.

Có những chương trình cho độ chính xác chấp nhận được, điển hình như chương trình rút trích thông tin cá nhân từ văn bản phi cấu trúc trả về thông tin các cấu trúc, ví dụ: chương trình đọc tập tin văn bản mô tả thông tin cá nhân trả về họ tên, địa chỉ, kỹ năng nghề nghiệp của người đó. Với độ chính xác có thể lên đến 80%.

Một vấn đề thiết thực hứa hẹn nhất là việc ứng dụng khai mỏ văn bản vào lĩnh vực sinh học. Một ví dụ dễ nhận ra nhất là trong công trình của Don Swanson đưa ra giả thiết nguyên nhân của căn bệnh hiếm có bằng cách tìm kiếm trong các mối liên kết gián tiếp trong các tập con của những tư liệu về sinh học.

Một ví dụ khác, ứng dụng khai mỏ văn bản để giải quyết câu hỏi lớn trong các bộ gen là những protein nào tương tác với những protein khác.

Những giới hạn cơ bản của khai mỏ văn bản là:

- ❒ Không thể viết ra những chương trình dịch văn bản trọn vẹn trong một thời gian dài.
- ❒ Thông tin người ta cần thiết thường không được ghi lại trong những mẫu nguyên bản.

1.8. Sơ lược về phân loại văn bản (text classification):

Qua nhiều năm lượng tài liệu số hóa phát triển vươn tới một kích thước khổng lồ. Như một tất yếu, khả năng sắp xếp và phân loại tài liệu một cách tự động trở thành vấn đề quan trọng hàng đầu.

Có hai vấn đề khác nhau trong phân loại văn bản: gom cụm văn bản (text clustering) và phân loại văn bản (text categorization). Vấn đề gom cụm liên quan đến việc tìm kiếm một nhóm cấu trúc tiềm ẩn trong tập các tài liệu. Trong khi đó việc phân loại còn có tên gọi khác là phân loại văn bản (Text Classification) có thể được xem như tác vụ của việc cấu trúc kho chứa tài liệu theo nhóm cấu trúc.

Phân loại tài liệu xuất hiện trong nhiều ứng dụng: như lọc e-mail, định hướng mail, lọc thư rác (spam), giám sát tin, chỉ mục tự động các bài báo khoa học, … Phân loại văn bản tự động rất hấp dẫn vì việc tổ chức văn bản thủ công có thể chiếm chi phí quá đắt.

Tiếp cận vượt trội trong phân loại văn bản được dựa trên những kỹ thuật máy học. Ta có thể nhận ra ba giai đoạn khác nhau trong việc thiết kế hệ thống phân loại văn bản: biểu diễn tài liệu, xây dựng bộ phân loại, lượng giá bộ phân loại.

1.9. Nội dung đề tài:

Trong luận văn này cần đạt đến: một tập tài liệu nghiên cứu lý thuyết và thiết kế chương trình.

Một bản minh họa (demo) cài đặt cho thuật toán và thể nghiệm trên mạng.

Cấu trúc dự kiến của luận văn như sau:

Chương 1: Tổng quan

Chương 2: Cơ sở lý thuyết

Chương 3: Nghiên cứu vấn đề

Chương 4: Xây dựng ứng dụng – thử nghiệm – đánh giá

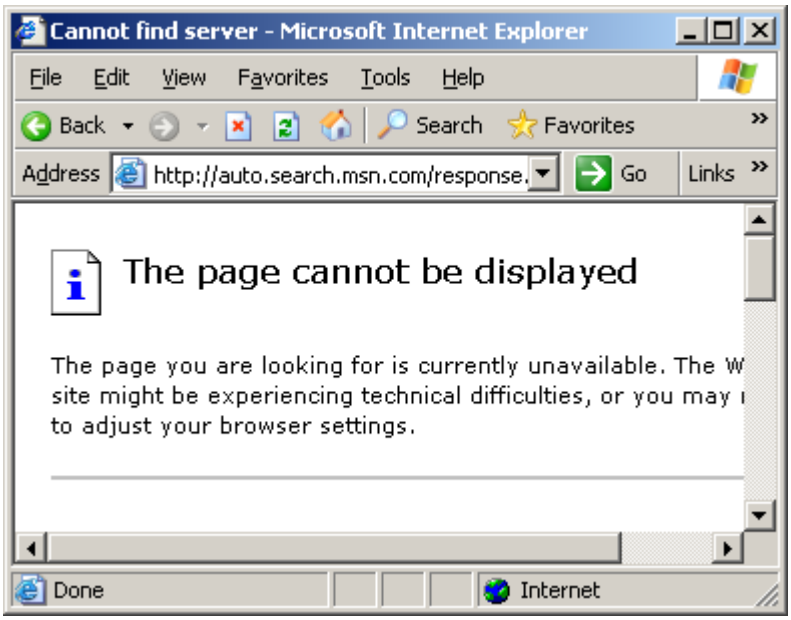
Chương 5: Kết luận và hướng phát triển

Chương 2: CƠ SỞ LÝ THUYẾT

2.1. Khái niệm bộ lọc web:

2.1.1. Bộ lọc web (Web filter):

Bộ lọc web là phần mềm có chức năng lọc vài loại nội dung hiển thị trên một trình duyệt web hay khóa vài vị trí web (web site) mà người dùng cố gắng truy cập đến. Bộ lọc kiểm tra nội dung của một trang web hay một địa chỉ web dựa vào một tập các luật và thay thế bất kỳ nội dung không mong muốn với một trang web thay thế, thường trang này nội dung có dòng "Access Denied" (cấm truy xuất).



Hình 2.1. Màn hình trình duyệt cấm truy cập

Nhà quản trị hệ thống là những người thường nắm quyền kiểm soát và cấu hình loại nội dung đi qua bộ lọc. Các bộ lọc web thường được sử dụng trong các trường học, thư viện, tiệm café internet, các dịch vụ internet công cộng, ngay cả tại nhà cũng có thể áp dụng giữ an toàn cho trẻ con tránh những nội dung không lành mạnh từ Internet.

2.1.2. Tại sao cần thiết có một bộ lọc nội dung?

Đây là bộ phận quan trọng để kiểm soát nội dung trên mạng và quy định cách dùng tài nguyên trên mạng như thế nào.

Người dùng thường có khuynh hướng làm việc riêng tư hay những điều phi pháp trong hàng giờ làm việc trong hay ngoài giờ làm việc tại cơ quan. Điều này sẽ tiêu phí hàng giờ làm việc có giá trị và có thể có khả năng làm cho hệ thống mạng rơi vào trạng thái trì trệ do mạng đang bị lạm dụng để tải xuống những tài liệu phi pháp.

2.2. Lý thuyết dùng trong nghiên cứu

2.2.1. Khai mỏ dữ liệu:

Là thao tác rút trích thông tin hữu ích, chưa biết tiềm ẩn trong một khối dữ liệu lớn. Thông tin rút trích được còn gọi là tri thức từ trong khối dữ liệu, nó có thể giải thích dữ liệu trên tập dữ liệu đó từ đó cung cấp thông tin hỗ trợ ra quyết định, dự báo hay khái quát dữ liệu. Khai mỏ dữ liệu được sử dụng rộng rãi trong các ngành: phân tích thị trường, quản lý phân tích rủi ro, quản lý và phân tích sai hỏng, khai thác web, khai thác văn bản,…

2.2.2. Khai mỏ văn bản:

2.2.2.1. Khái niệm

Là tác vụ khai thác thông tin từ nhiều tập tin văn bản, nguồn tài liệu này có thể từ bài báo, bài viết nghiên cứu, sách, thư viện điện tử, thư điện tử, trang web,… công dụng của khai thác văn bản phục vụ cho việc rút trích thông tin, gom nhóm văn bản, phân loại văn bản, …

Tìm các thông tin hữu ích trên một tập các văn bản. Khai mỏ văn bản để tìm thông tin (IR) cũng giống như truy vấn CSDL.

CSDL văn bản là tập hợp các văn bản từ nhiều nguồn khác nhau như: bài báo, bài nghiên cứu, sách, thư viện điện tử, thư điện tử, trang web,… các nguồn tư liệu này ngày càng nhiều. Con người không thể đọc hay tiếp nhận tất cả các thông tin có trong đó.

Rút trích thông tin (IR): như nói trên số lượng văn bản càng nhiều nhưng thông tin phân tán trong nhiều tài liệu khác nhau.

- Bài toán rút trích thông tin: xác định tài liệu nào nào có chứa thông tin mà người đọc mong muốn thông qua một từ khóa nào đó.
- Như vậy công việc rút trích thông tin giống như truy vấn CSDL.

Độ chính xác:

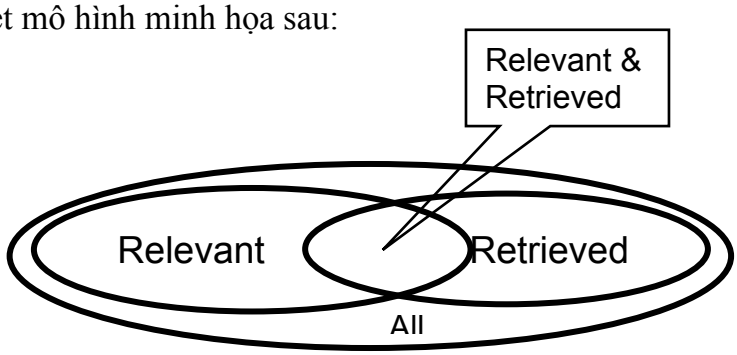
- Độ chính xác (Precision): là thành phần phần trăm của các tài liệu được tìm ra trên thực tế có liên quan đến truy vấn. (Câu trả lời chính xác)

$$precision = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{retrieved\}|}$$

- Độ tìm được (Recall): thành phần phần trăm của những tài liệu có liên quan đến câu truy vấn.

$$recall = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{relevant\}|}$$

- Xét mô hình minh họa sau:



Hình 2.2: Minh họa mối quan hệ trong rút trích thông tin

- Phần Relevant (liên quan) là tập các thông tin tìm được có liên quan đến lĩnh vựa cần quan tâm.
- Phần Retrieved (tìm được) là tập các thông tin tìm được.
- Phần giao giữa hai phần trên: càng lớn càng tốt, cho ta biết mức độ tìm được và thông tin liên quan nhiều hay ít. Phần này luôn ở tử số của 2 công thức trên.

2.2.2.2. Một số loại khai mỏ

- ***Phân tích kết hợp dựa trên từ khóa:*** Một tài liệu có thể xem như một chuỗi ký tự và có thể xác định bằng tập các từ khóa. Việc phân tích các tài liệu dựa trên từ khóa để tìm ra một kết luận về tài liệu đó.
- ***Phân tích tài liệu tự động:*** Giống như một người trợ lý, hỗ trợ đắc lực trong việc phân loại tài liệu bằng cách "đọc" tất cả các nguồn tài liệu đến và xếp nó theo từng loại một cách tự động.
- ***Đo độ tương đồng giữa các tài liệu:*** Đo độ tương đồng là việc xem xét tài liệu đó xem nó có thuộc về một dòng văn học nào hay thuộc về một tác giả nào đó. Hoặc cũng có thể dùng để xếp loại văn bản thuộc về lĩnh vực nào.
- ***Phân tích trình tự:*** Đoán sự kiện, dự báo xu hướng: Như đã nói bên trên, văn bản là một chuỗi các ký tự diễn đạt một ý. Nhiều tài liệu gởi đến, có nhiều cấp độ diễn đạt về một vấn đề. Từ các vấn đề này hệ thống có thể đưa ra dự đoán về các diễn biến của hiện tượng hay những điều sẽ xảy ra tiếp theo.
- ***Xác định các hiện tượng không bình thường:*** Hiện tượng không bình thường là một văn bản đến có sự khác biệt hay "cá tính" quá khác so với cùng loại nó đến trước đó để cho một kết luận về sự bất thường của loạt văn bản.

2.2.3. Phân loại văn bản

2.2.3.1. Định nghĩa phân loại văn bản

Phân loại văn bản là tác vụ khởi gán một hay nhiều loại văn bản ngôn ngữ tự nhiên được xác định trước vào các nhóm tài liệu nào đó đã định nghĩa trước. Chẳng hạn như, một hệ thống email nhận thông điệp đến thì thông điệp đó có thể được gán là "junk" hay "non-junk", có lẽ để thuận lợi cho việc ra quyết định nên hay không

nên xóa thông điệp đó một cách tự động. Một ví dụ khác, xếp loại các bài báo vào các thể loại như: quốc tế, thể thao, thương mại, …

Trích xuất thông tin (IR) là tác vụ tìm kiếm những đoạn liên quan của các tài liệu mà những tài liệu đó dựa trên vài thông tin cần thiết hay khái niệm được định nghĩa trước. Lấy một ví dụ, một hệ thống IR có thể nhận diện những đoạn văn bản trình bày trong đề mục báo, công ty, địa điểm, và chi tiết lương từ những mẫu quảng cáo nghề nghiệp.

Phân loại văn bản và trích xuất thông tin là hai dạng của xử lý văn bản nông cạn, nhưng có sự tương tác hay hợp lực giữa những kỹ thuật chiếm một phần khiêm tốn. Nhiều hệ thống IR phát triển một vài kỹ thuật phân loại văn bản để đảm bảo rằng những văn bản đó được xử lý để đưa đến một tập văn bản chứa dữ liệu mong muốn.

Nghĩ theo một cách khác, ta có thể mường tượng rằng một hệ thống phân loại văn bản mà hệ thống đó gán những loại dựa trên những phân đoạn được trích xuất suốt trong bước khởi gán IR. Phát biểu theo toán học, phân loại văn bản là tác vụ xấp xỉ hàm đích (target function) chưa biết rõ. Ψ: D x C Φ {T, F}, trong đó ý nghĩa của hàm Ψ: D x C Φ {T, F} được gọi là bộ phân lớp, mà Ψ và Φ "càng trùng khớp càng tốt". Trong đó:

C = {c_1, c_2, … c_m} là một tập gồm các loại được định nghĩa trước (có **m** thể loại cần phân biệt)

D là một lĩnh vực của những tài liệu.

2.2.3.2. Đặc điểm của phương pháp gán nhãn:

– Ta giả sử rằng những loại chỉ là các nhãn ký hiệu, ý nghĩa của nhãn ký hiệu chỉ là các ký hiệu không mang ngữ nghĩa giải thích để giúp xây dựng những bộ phân lớp, tóm lại văn bản nằm trong nhãn không có nghĩa.

– Sự quy kết các văn bản vào trong các loại phải được thực hiện dựa trên cơ sở nội dung văn bản chứ không dựa trên thông tin về dữ liệu mà thông tin này có thể có sẵn từ một nguồn dữ liệu bên ngoài.

– Cho rằng, nội dung của văn bản là một khái niệm mang tính chủ quan, điều này nói lên rằng thành viên của một văn bản trong một thể loại không thể được quyết định chắc chắn.

– Phụ thuộc vào ứng dụng, sự phân lớp có thể:

- ❒ Nhãn đơn: Mỗi văn bản có thể được gán chính xác vào một loại, trong trường hợp số loại m = 2 (tập C) ta có nhãn nhị phân.
- ❒ Đa nhãn: Mỗi tài liệu có thể được gán vào một số loại nhãn.

– Yêu cầu đặt ra cho bộ phân lớp văn bản:

- ❒ Sự phân lớp cứng (Hard Classification): Đối với dạng này hỗ trợ một giá trị trong tập {T, F} nhờ nó chỉ ra cho biết văn bản là thành viên hay không phải thành viên của d_j và c_i. Phương pháp này hữu dụng cho bộ phân lớp độc lập.
- ❒ Sự phân lớp mềm (Soft Classification) hỗ trợ giá trị trong [0,1] nó cho biết cấp độ tin cậy của hệ thống vào thành viên của d_j và c_i. Cách này hữu ích cho phương pháp phân loại tương tác.

2.2.4. Một số phương pháp phân loại văn bản:

Phân loại văn bản là một tác vụ học có giám sát, tác vụ đó gán những nhãn định trước cho các tài liệu mới, từ đó đem so sánh với tập huấn luyện gồm các tài liệu được gán nhãn. Các hệ thống phân loại văn bản tự động truyền thống đơn thuần là áp dụng cho các văn bản đơn giản vì thế đem ứng dụng cho trang web với các siêu liên kết phải được xem xét cẩn thận. Một số phương pháp tiếp cận chính:

- ❒ Bộ phân loại Naïve Bayes (NB) được dùng rộng rãi, bởi tính đơn giản và hiệu quả tính toán. NB sử dụng quan hệ tần suất của từ trong tài liệu như những từ có thể và sử dụng các từ có thể này để gán một loại đối với một loại tài liệu.
- ❒ K-Nearest Neighbor (KNN) là cách tiếp cận thống kê, đây là phương pháp phân loại văn bản chính xác nhất. Đưa ra một tài liệu, KNN chọn k

tài liệu tương tự từ tập huấn luyện và sử dụng những loại của các tài liệu này để phát hiện ra loại của những tài liệu đang được phân lớp. Tài liệu được biểu diễn bởi những vector của những từ và sự giống nhau giữa hai tài liệu được đo bằng cách sử dụng khoảng cách Ơ-lic (Euclidean) hay những hàm khác giữa hai vector này.

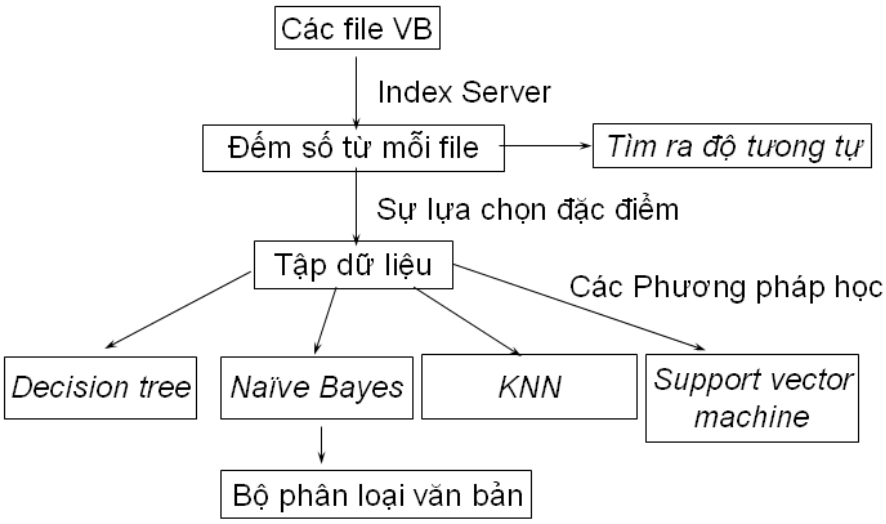
- Cây quyết định: Là phương pháp máy học tự động quy nạp các cây phân lớp dựa trên dữ liệu huấn luyện. Mỗi nút bên trong của cây quyết định được kết hợp với một kiểm thử trên một thuộc tính và nhánh ra ngoài của một nút tương ứng với kết quả kiểm thử. Một lá được kết hợp với một loại. Sự phân loại một tài liệu bắt đầu từ nút gốc và sau đó thăm các nút bên trong đến khi một nút lá được tìm đến. Tại mỗi nút, kiểm thử kết hợp với nút đã được thực thi để xác định nút tiếp theo, loại của tài liệu là loại của nút lá cuối cùng.
- Support vector machines (SVM): Sử dụng một bề mặt quyết định để chia những điểm dữ liệu vào trong các lớp. SVM cũng được ứng dụng để phân loại văn bản. Với mẫu đơn giản nhất của nó, các tài liệu huấn luyện được dùng như những vector, và thuật toán phát hiện siêu bề mặt (hyperplanes) phân chia thành những lớp khác nhau của các tài liệu huấn luyện. Những tài liệu kiểm tra được phân lớp dựa theo những vị trí của chúng đối với siêu bề mặt.
- Phân loại dữ liệu văn bản liên kết (hypertext data): Sử dụng một tập dữ liệu phổ biến để so sánh sự hiệu quả của hai thuật toán NB và KNN cho việc phân lớp trang web. Chúng xem xét cẩn thận tính hữu ích của các siêu liên kết, nội dung của những tài liệu liên kết và dữ liệu trong sự phân lớp và tìm ra dữ liệu biến đổi có thể gia tăng độ chính xác của sự phân lớp bởi một hệ số lớn.
- Phân loại văn bản cho việc lọc web: Ứng dụng mạng nơ-ron nhân tạo để lọc những trang có nội dung khiêu dâm. Chúng dùng một sự thu thập

những trang có nội dung khiêu dâm và không khiêu dâm để huấn luyện mạng nơ-ron nhân tạo, là mạng có thể quyết định một trang có mang nội dung khiêu dâm hay không? Phương pháp này đòi hỏi công suất tính toán mạnh do đó nó không thể trở thành ứng dụng thời gian thật.

2.2.5. Tiếp cận chuẩn trong phân loại văn bản:

1. Loại bỏ những từ có âm tắc (**stop word**) và những từ đánh dấu (**marking**).
2. Những từ còn lại là tất cả những thuộc tính.
3. Một tài liệu trở thành một vector ***<Từ, tần suất>.***
4. Huấn luyện một bộ phân loại luận lý *(boolean classifier)* cho mỗi lớp văn bản.
5. Đánh giá các kết quả dựa trên một mẫu văn bản chưa biết trước.

2.2.6. Quá trình phân loại văn bản



Hình 2.3. Quy trình phân loại văn bản

2.2.7. So sánh hai văn bản

2.2.7.1. Khái niệm:

So sánh hai văn bản là một phần trong khai mỏ văn bản, nó có nhiệm vụ phân tích văn bản để tìm ra mối liên quan của văn bản đó với lĩnh vực nào.

Ứng dụng chính của so sánh văn bản là dùng cho việc phân loại tài liệu, tìm sự giống nhau của hai tài liệu.

2.2.7.2. Một số phương pháp so sánh văn bản:

Năm cách đo lường phổ biến trong xác định tính tương tự của tài liệu: [14]

i) Cosine similarity dùng số lần lặp lại của từ xuất hiện trong bài viết để gán điểm. Ví dụ: từ "think (v)" xuất hiện 56 lần trong bài nói về tác phẩm Hamlet thì điểm (trọng số) của nó là 56. Nhưng trong bài về Othello, từ "think (v)" xuất hiện 86 lần. Một hình thái từ được gán điểm là 0 trong bất kỳ bài viết nào thì nó bị xem là không xuất hiện.

Để tính giá trị tương tự bằng cosine cho hai văn bản người ta thường dùng cách tính tích vô hướng của vector dựa trên những phép chia tỉ lệ của hai vector tần suất đặc trưng cho hai tài liệu.

$$\text{cosine similarity} = \frac{W_1 \bullet W_2}{|W_1| * |W_2|}$$

Trong đó W_1 là vector tần suất của bài viết đầu và W_2 là vector tần suất của bài viết thứ hai. Dấu "." đặc trưng cho toán tử nhân trong tích vô hướng. "|W1|" hay "|W2|" cho biết chiều dài khoảng cách Ơ-lic của vector W1 hay W2, cả hai đều được lấy căn bậc hai.

Xét trên phương diện hình học, giá trị tương tự cosine là giá trị lượng giác cosine của góc giữa hai vector tài liệu trong không gian đa chiều. Góc phân biệt càng nhỏ thì hai văn bản càng gần nhau hơn trong không gian hình thái từ.

Giá trị tương tự cosine được sử dụng rộng rãi hơn và thường cho biết sự tương tự của hai văn bản rõ ràng hơn những phương pháp đo đạt khác.

ii) Hệ số tương tự cosine nhị phân (binary cosine similarity coefficient) được tính toán chính xác theo cùng cách với sự tương tự cosine thông thường ngoại trừ các hình thái từ trong văn bản được đánh điểm là 1 khi nó xuất hiện trong văn bản và là 0 khi nó không xuất hiện. Tính sự tương tự cosine nhị phân như sau:

$$\text{Binary cosine similarity} = \frac{|W_1 \cap W_2|}{|W_1| * |W_2|}$$

iii) Hệ số Dice nhị phân (binary Dice coefficient) bằng giá trị cosine nhị phân khi những vector tần suất cho hai văn bản đang được so sánh chứa chính xác cùng số lượng những mẫu khác không (non-zero entries). Cosine nhị phân bị đưa vào thế bất lợi là bé hơn hệ số Dice khi số lượng các vector tần suất khác 0 trong hai văn bản rất khác biệt nhau. Giá trị Dice bằng 0.0 cho biết hai tài liệu hoàn toàn khác nhau, ngược lại khi trị giá Dice bằng 1 cho biết hai văn bản giống nhau.

$$\text{Dice coefficient} = \frac{2 * |W_1 \cap W_2|}{|W_1| + |W_2|}$$

iv) Hệ số Jaccard nhị phân (binary Jaccard coefficient) gán những giá trị tương tự cho những trường hợp low-overlap thấp hơn hệ số Dice. Số lượng hình thái từ bị chia sẻ nhỏ hơn, giá trị của hệ số Jaccard có liên quan đến hệ số Dice.

$$\text{Jaccard coefficient} = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

v) Hệ số overlap nhị phân (binary overlap coefficient) là 1.0 tất cả hình thái từ với giá trị điểm số bằng 1 thì sự giống nhau trong cả hai văn bản được so sánh.

$$\text{Overlap coefficient} = \frac{|W_1 \cap W_2|}{\min(|W_1|, |W_2|)}$$

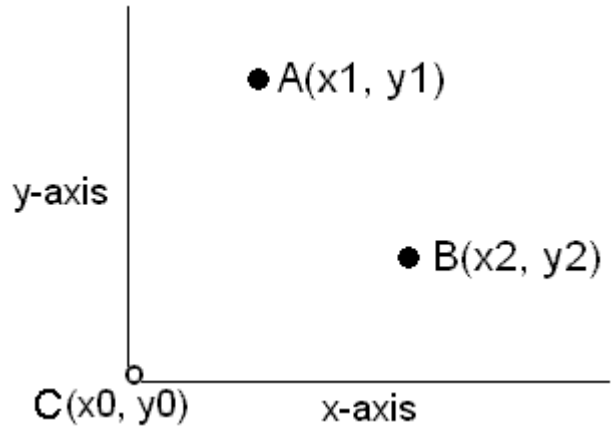
Trong năm phương pháp kể trên thì phương pháp i) thường được dùng nhất, vì tính dễ hiểu, dễ áp dụng cùng với sự chính xác về toán học của nó. Trong luận văn này, người viết áp dụng phương pháp so sánh theo phương pháp i) để so sánh hai văn bản.

2.2.7.3. Lý thuyết so sánh hai văn bản bằng công thức cosine

2.2.7.3.1. Không gian và các phép tính cơ sở

2.2.7.3.1.1. Điểm trong không gian tọa độ [3]

Điểm của một điểm có thể tìm thấy trong hệ trục của nó, ta đặt một điểm trong hệ trục phẳng x-y xem như là điểm C với hệ trục (x_0, y_0). Ta có thể nói đến điểm này là C(x_0, y_0). Trừ trường hợp $x_0 = 0$ và $y_0 = 0$. Tương tự, ta có thể liên hệ hai điểm A, B bất kỳ, trong mặt phẳng tọa độ A(x1, y1) và B(x2, y2). Biểu diễn trong hình vẽ 2.4 dưới đây:



Hình 2.4. Biểu diễn những điểm A, B, C trong một mặt phẳng hai chiều. (Bản quyền hình Dr. E. Gracia [3])

2.2.7.3.1.2. Phép tích vô hướng

Nếu ta đem nhân hai tọa độ A và B sau đó cộng hai tích số lại với nhau ta nhận được một phép tích trừu tượng A•B, và cũng được biết đến như phép tích nội và tích vô hướng. Vì thế phép tích A•B được đưa ra bởi công thức:

Công thức 1: $A \bullet B = x_1 * x_2 + y_1 * y_2$ (1)

Nếu các điểm A và B được xác định trong không gian ba chiều thì ta có tọa độ của chúng trong hệ trục ba chiều là A(x_1, y_1, z_1) và B(x_2, y_2, z_2). Phép tích A•B được biểu diễn trong công thức sau:

Công thức 2: $A \bullet B = x_1 * x_2 + y_1 * y_2 + z_1 * z_2$ (2)

Với một không gian có số chiều là n, ta chỉ cần thêm vào phép tích của một chiều như công thức 1 và 2.

2.2.7.3.1.3. Tính khoảng cách từ gốc tọa độ C đến các điểm A, B.

Để xác định một đường thẳng ta cần ít nhất hai điểm. Vì thế nếu ta vẽ một đường thẳng nối từ C đến A hay B, ta có thể tính khoảng cách d giữa hai điểm (thường gọi là khoảng cách Ơ-lit [Euclidean Distance]), khoảng cách này được tính qua 4 bước sau. Cho hai điểm bất kỳ xác định một đường thẳng:

- ❒ Đưa ra hiệu của các điểm đối với gốc tọa độ
- ❒ Bình phương tất cả các hiệu
- ❒ Cộng tất cả các hiệu đã được bình phương
- ❒ Lấy căn bậc hai của kết quả cuối cùng

Vì ta đã định nghĩa $x_0 = 0$, $y_0 = 0$, nên khi tiến hành tính khoảng cách từ C đến A ta áp dụng công thức khoảng cách Ơ-lit như sau:

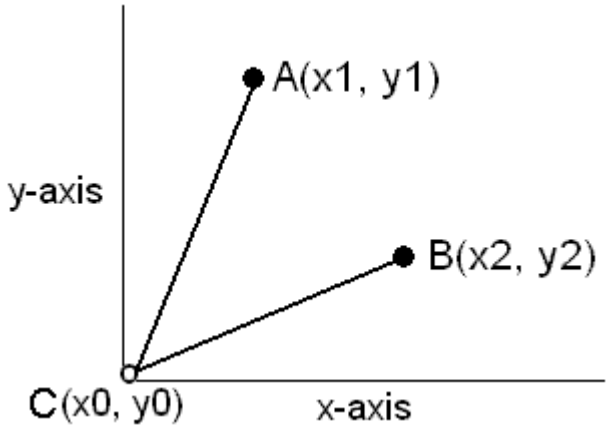
Công thức 3a:

$$d_{AC} = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} = \sqrt{(x_1^2 + y_1^2)}$$

Tương tự cho khoảng cách từ B đến C

Công thức 3b:

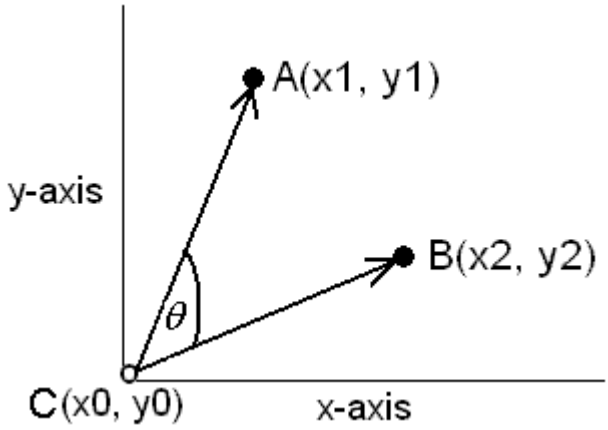
$$d_{BC} = \sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2} = \sqrt{(x_2^2 + y_2^2)}$$



Hình 2.5. Các đường thẳng biểu diễn khoảng cách Ơ-lic (Euclidean Distances) giữa các điểm A và B với điểm C. (Bản quyền hình Dr. E. Gracia [3]).

2.2.7.3.1.4. Biểu diễn dạng vector

Những đường thẳng trong hình 2.5 có thể được thay thế bằng các vector (thêm mũi tên). Một vector là một đại lượng có hướng và độ lớn xác định. Đầu và góc của mũi tên nhận biết hướng của vector đó, trong khi đó độ lớn của nó thường xác định bằng khoảng cách Ơ-lic (Euclidean Distance). Trong ví dụ đưa ra ở trên ta có $x_0 = 0$ và $y_0 = 0$, ta có thể đơn giản hóa và biểu diễn độ lớn của vector A và vector B bằng ký hiệu $d_{AC} = |A|$ và $d_{BC} = |B|$. Ký hiệu gạch sổ (bao lấy A hoặc B) là diễn đạt cho trị tuyệt đối của độ lớn. Điều này được minh họa trong hình 2.6.



Hình 2.6. Các vector A và B. (Bản quyền hình Dr. E. Gracia [3])

2.2.7.3.1.5. Biểu diễn COSIM:

Để chuẩn hoá phép tích A•B chúng ta phân tích nó thông qua khái niệm khoảng cách Ơ-lic (tức là A•B/(|A||B|)). Tỉ lệ này được xác định bằng cosine của góc giữa hai vector, với giá trị trong khoảng 0 đến 1 (xem hình 2.6 góc θ).

Trong các ứng dụng rút trích thông tin tỉ lệ này được tính toán để làm chuẩn hóa độ dài của các tài liệu vì tài liệu dài có nguy cơ làm cho tần suất từ lớn.

Hãy trở lại với phép tích vô hướng (Dot Product) được chuẩn hóa (hay cosine của góc). Tỉ lệ này cũng được sử dụng như thước đo độ tương tự của bất kỳ những

vector tương ứng với các tài liệu, truy vấn, hay sự kết hợp của chúng với nhau. Biểu thức độ tương tự cosine thường ký hiệu là Sim(A, B) hay COSIM.

$$Sim(A,B) = \cos ine\theta = \frac{A \bullet B}{|A||B|} = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

Sim(A,B) sự tương đồng của hai vector A, B

Khi góc của hai vector nhỏ lại thì cosine của góc tiến về 1, nghĩa là hai vector càng gần nhau hơn và cũng có nghĩa là sự tương tự của bất kỳ những gì được đặc trưng bởi hai vector tăng lên (có nghĩa là càng giống nhau).

Điều này là một sự thuận lợi trong việc xếp hạng tài liệu, nghĩa là đo lường sự gần nhau của các vector của chúng như thế nào đối với vector truy vấn. Ví dụ: đặt điểm A(x_1, y_1) đặc trưng cho một truy vấn. Các điểm B(x_2, y_2), D(x_3, y_3), E(x_4, y_4), F(x_5, y_5), … đại diện cho các tài liệu. Ta có thể tính được cosine của góc giữa A (truy vấn) với mỗi tài liệu và sắp xếp các kết quả (độ tương tự cosine) theo thứ tự giảm dần. Cách giải quyết này có thể được mở rộng đến toàn bộ tập huấn luyện.

Để làm được điều này ta cần xây dựng một không gian từ. Không gian từ được định nghĩa bởi một danh sách có trật tự các từ. Những từ này được trích xuất từ bộ sưu tập tài liệu để dùng cho truy vấn. Hệ trục tọa độ của các điểm biểu diễn các tài liệu và truy vấn định nghĩa bằng cách dùng lược đồ trọng số.

Nếu như trọng số thật sự là phép đếm từ (w = tf) thì tọa độ điểm được đưa ra bởi tần suất từ; tuy nhiên ta không phải xác định trọng số từ trong cách này.

Công thức tính COSIM thực tế là:

$$Sim(Q,D_i) = \frac{\sum_i W_{Q,j} W_{i,j}}{\sqrt{\sum_j W_{Q,j}^2}\sqrt{\sum_i W_{i,j}^2}}$$

Công thức sự tương tự cosine (cosine của góc) giữa truy vấn và tài liệu.

Trong công thức trên dấu sigma có nghĩa là lấy tổng của…, Q là truy vấn, D là tài liệu liên quan đến Q và w là trọng số.

2.2.7.3.2. Cơ sở lý thuyết nghiên cứu vector từ (Term Vector)

2.2.7.3.2.1. Mô hình không gian vector của Salton [2]

Những hệ thống rút trích thông tin (IR) gán trọng số cho các từ bằng cách:

- Thông tin cục bộ từ những tài liệu riêng lẻ.
- Thông tin toàn cục từ bộ sưu tập các tài liệu.

Thêm vào đó, hệ thống cũng gán thông số cho các liên kết (link) sử dụng thông tin biểu đồ web để thống kê chính xác mức độ liên quan giữa hai tài liệu.

Trong những nghiên cứu về rút trích thông tin, lược đồ trọng số cổ điển là mô hình không gian Vector của Salton, thường được hiểu như "mô hình vector từ" (term vector model). Lược đồ trọng số này được định bởi công thức:

Công thức 1: Trọng số từ

$$w_i = tf_i * \log\left(\frac{D}{df_i}\right)$$

Trong đó:

- tf_i = tần suất từ (số lượng từ đếm được) hay số lần lặp lại của từ i trong một tài liệu.
- df_i = tần suất tài liệu hay số lượng tài liệu có chứa từ i
- D = Số lượng tài liệu trong kho lưu trữ tài liệu (tập huấn luyện).

Nhiều mô hình thực hiện trích xuất vector từ trong các tài liệu hay truy vấn xuất phát từ công thức 1.

2.2.7.3.2.2. Trọng số cục bộ

Trong công thức 1 cho ta thấy rằng w_i và tf_i tỉ lệ thuận. Điều này tạo nên điểm yếu của mô hình đối với sự lạm dụng những từ lặp đi lặp lại. Vì:

- Đối với những tài liệu có độ dài tương đương, cùng với những ví dụ về từ được truy vấn thì được quan tâm trong quá trình rút trích.

- ❒ Đối với những tài liệu có độ dài khác nhau, những tài liệu dài được ưa chuộng trong suốt quá trình rút trích vì những tài liệu có vẻ phù hợp hơn chứa nhiều thể hiện hơn những từ được truy vấn.

2.2.7.3.2.3. Trọng số toàn cục

Trong công thức 1, giá trị $\log(D/df_i)$ từ được biết đến như *tần suất tài liệu nghịch đảo (*inverse document frequency), IDF_i. Đây là một phép đo lường độ hỗn tạp thông tin kết hợp với một từ trong một tập tài liệu. Trong nhiều năm qua có nhiều đề xuất sửa chữa công thức 1. Nói chung , biểu thức tf*idf nói lên cơ sở của mô hình hay nguồn gốc của công thức 1.

Công thức 1 cho thấy w_i và df_i tỉ lệ nghịch (w_i giảm thì df_i tăng). Xét một ví dụ: nếu một kho có 1000 tài liệu nhưng chỉ có 10 tài liệu có chứa từ “pet”, chỉ số IDF cho từ “pet” IDF = log(1000/10) = 2. Tuy nhiên nếu chỉ có một tài liệu chứa từ “pet” thì chỉ số IDF = log(1000/1) = 3.

Theo cách đó, những từ xuất hiện quá nhiều trong các tài liệu (chẳng hạn stopwords hay những từ lặp lại nhiều lần) sẽ nhận được trọng số thấp, trong khi đó thì những từ không phổ biến xuất hiện trong một vài tài liệu thì nhận được trọng số cao. Điều này cho ta thấy rằng nếu có quá nhiều từ phổ biến (như "a", "the", "of", etc) là không hữu ích cho việc nhận ra một tài liệu liên quan từ một tài liệu không liên quan bởi sự chi phối của nó đối với nội dung văn bản không nhiều, nhưng có ảnh hưởng lớn đến các giá trị trong quá trình tính toán.

2.2.7.3.2.4. Mật độ từ khóa có giá trị

Trong công thức 1 hiển nhiên cho thấy trọng số từ khóa bị ảnh hưởng bởi hai yếu tố:

- ❒ Số lượng từ cục bộ
- ❒ Độ hỗn tạp của các tài liệu trong cở sở dữ liệu.

Bởi thế, khái niệm trọng số từ đó được nhắc đến nhiều và có thể được ước lượng với thuật ngữ "giá trị mật độ từ khóa" ("keyword density values") ít tạo ra sự sai biệt của công thức.

Mật độ từ khóa được tính bằng công thức:

Công thức 2:

$$KD_i = \frac{tf_i}{L_i}$$

Trong công thức tf_i là số lần một từ i xuất hiện trong một tài liệu, L_i tổng số các từ có trong một tài liệu. Mật độ từ khóa thật ra chỉ là tỉ lệ từ cục bộ. Tỉ lệ này cho biết sự tập trung của những từ trong một tài liệu. Thế nên, mật độ từ khóa của một tài liệu 500 từ mà trong đó từ "pet" được lặp lại 5 lần thì $KD_{pet} = 5/500 = 0.01$ hay 1%. Cũng lưu ý rằng tỉ lệ này không dùng để giải thích cho quan hệ vị trí và quan hệ phân tán của các từ trong văn bản. Những phần tử này ảnh hưởng đến mối quan hệ của tài liệu và ngữ nghĩa của chủ đề.

2.2.7.3.2.5. Mật độ từ khóa không thích hợp

Công thức 2 không nói lên được trọng số ngữ nghĩa của từ trong quan hệ với những từ khác bên trong một tài liệu hay bên trong một tập các tài liệu.

Theo công thức 2, một từ k_1 nằm trong hai tài liệu riêng biệt nhau bằng nhau về số lần lặp sẽ có cùng mật độ từ (không quan tâm đến nội dung tài liệu hay tính tự nhiên của cơ sở dữ liệu. Tuy nhiên, nếu ta giả sử rằng mật độ từ có giá trị là hay có thể được đưa ra những trọng số từ khóa, thì ta sẽ:

- ❒ Không cần xem xét đến độ hỗn tạp của thông tin mà những từ truy vấn lấy ra được.
- ❒ Gán trọng số từ mà không cần để ý đến mối quan hệ từ.
- ❒ Gán trọng số mà không cần xem xét đến sự tự nhiên của cở sở dữ liệu được truy vấn.

Các ý 1 - 3 nói lên mâu thuẫn trong mô hình Salton. Theo công thức 1, những trọng số từ không là những tỉ lệ từ cục bộ tách rời khỏi cở sở dữ liệu được truy vấn. Một từ k_1 thường có số lần lặp trong hai tài liệu cùng độ dài bằng nhau (không chú ý đến nội dung) được đặt trọng số khác nhau trong cùng cơ sở dữ liệu được truy vấn hay trong những cơ sở dữ liệu khác.

2.2.7.3.3. Tiếp cận theo đại số tuyến tính[4]:

Đại số tuyến tính cung cấp chín phương pháp tính toán nhanh gọn cho tất cả các phép tính này. Một trong số đó là cần thiết cho việc tính toán phép đo lường sự tương tự cosine như tỉ lệ của hai phép tích:

i) Phép tích vô hướng truy vấn và các tài liệu

ii) Phép tích tiêu điểm Frobenius (Frobenius Norm).

Tiêu điểm Frobenius của một ma trận, cũng được biết đến như tiêu điểm Ơ-lit (Euclidean Norm), được định nghĩa là căn bậc hai của tổng các bình phương của từng phần tử trong nó. Bản chất vấn đề như sau: lấy ra một ma trận, bình phương tất cả các phần tử của nó, cộng chúng lại với nhau sau cùng lấy căn bậc hai cho ra kết quả. Con số được tính toán gọi là tiêu điểm Frobenius của ma trận.

Vì các dòng và các cột của ma trận là các ma trận một dòng và các ma trận một cột và chúng thể hiện dưới dạng các vector, tiêu điểm Frobenius của riêng nó bằng với chiều dài của các vector đó.

Như được đề cập trước đó, cái hay của đại số tuyến tính là nó cung cấp một phương pháp tính toán nhanh gọn cho các phép tính trên. Về cơ bản, phép tích vô hướng những vector truy vấn và tài liệu và những chiều dài của chúng sau đó đưa ra những tỉ lệ của chúng. Nếu một mô hình vector định nghĩa các phần tử của A như phép tích của cục bộ, toàn cục và những trọng số được chuẩn hóa – thay vì chỉ là những phép đếm từ - người ta có thể sử dụng cách tiếp cận này.

** Hạn chế của mô hình*

Mô hình đếm từ có những hạn chế sau:

- ❒ Dễ bị ảnh hưởng đối với những từ lặp lại.
- ❒ Xu hướng sử dụng những tài liệu lớn vì những tài liệu này chứa nhiều từ và rất thường được lặp lại và những ma trận "từ – tài liệu" của chúng có nhiều mục từ. Như vậy, những tài liệu dài có điểm cao hơn, đơn giản vì chúng dài hơn chứ không phải vì chúng có mức độ liên quan lớn hơn.

Ta có thể thực hiện tốt hơn bằng cách nhân giá trị tf với nhiều lần giá trị IDF, đó là việc xem xét thông tin cục bộ và thông tin toàn cục. Trong cách này ta cũng chú ý đến độ hỗn tạp của thông tin mà chính nó có ảnh hưởng mạnh mẽ đến từ được truy vấn. Dù vậy, trong các phép tính toán trên ta chỉ cần thay thế $w_i = tf_i$ bằng các giá trị $w_i = tf_i$*IDF và gán cho ma trận "từ-tài liệu" bằng những giá trị này.

2.2.8. Ứng dụng bộ phân loại văn bản vào việc lọc Web

Bộ phân loại văn bản được ứng dụng vào một số lĩnh vực sau:

- ❒ Sắp xếp tài liệu theo từng loại: theo chủ đề, theo cùng nội dung, từng lĩnh vực,…
- ❒ Tinh chế tài liệu theo các loại tài liệu được định nghĩa trước
- ❒ Kiểm soát các hoạt động phân lớp, để đưa ra quyết định chọn loại văn bản nào tương ứng với ngữ cảnh của văn bản đưa vào.
- ❒ Phát hiện ra tác giả của văn bản theo dòng văn được định trước.
- ❒ Phân loại hình ảnh thông qua việc phân tích đầu đề nguyên bản.
- ❒ Nhận dạng thể loại văn bản.

Trong vấn đề lọc web, ứng dụng phân loại văn bản được dùng để thiết kế một số bộ lọc web cho các hệ thống lọc. Tùy vào phương pháp phân loại văn bản đem áp dụng vào bộ lọc web mà ta có một hệ thống lọc web tương ứng. Tuy nhiên với kỹ thuật này ngày càng được các nhà sản xuất phần mềm đầu tư nghiên cứu và triển khai trong các ứng dụng internet như: Internet Filering, Spam-Filter, Web Filter…

* Giải pháp lọc web:

Lọc web được dùng để chống lại những truy cập đến những tài liệu bất hợp pháp hay không thích hợp trên Internet. Bộ lọc web theo nội dung yêu cầu tất cả các lưu thông mạng được định tuyến thông qua máy chủ ủy thác (proxy server) hay một máy chủ đóng vai trò quan sát tất cả những lưu thông đang kết nối Internet. Sau đó nó tiến hành xử lý để khóa truy cập đến (từ) những web site cụ thể hay những trang web có URL nằm trong bộ danh sách kiểm soát các URL cấm và/hay theo Blacklist/Whitelist do người quản trị định nghĩa.

Khi có sự trùng hợp của một URL với một phần tử trong blacklists, bộ lọc web theo nội dung quét qua văn bản của những trang web yêu cầu. Nếu hệ thống phát hiện ra một chuỗi các từ hay cụm từ có phong cách dùng đáng nghi ngờ thì trang đó sẽ bị khóa. Tuy nhiên, việc khóa một trang là không tùy tiện. Cách xử lý thông minh trong một hệ thống lọc web là việc đặt tập luật tùy biến áp dụng cho những văn bản không mong muốn tiềm tàng, để chắc rằng những thông tin chấp nhận luôn có thể truy cập.

Chương 3: NGHIÊN CỨU VẤN ĐỀ

3.1. Một số tiếp cận vấn đề lọc web:

3.1.1. Danh sách đen và danh sách trắng (Blacklist và Whitelist):

Đây là cách được nhiều nhà cung cấp giải pháp sử dụng, vì nó đơn giản, dễ quản lý và trong chừng mực nào đó kỹ thuật này cũng cho ra một hiệu quả tương đối có thể chấp nhận được.

Có hai danh sách riêng biệt các web site phải bị ngăn chặn hay cho phép truy cập. Blacklist thường được tạo ra thủ công bằng cách khảo sát các web site để đưa ra quyết định một trang web có thể bị xem như một thành viên của lớp "cấm" hay không, chẳng hạn như bạo lực, khiêu dâm,… các trang cũng có thể đưa vào blacklist một cách tự động nếu trong tên miền của nó có chứa các từ như "sex", "xxx",… Trong khi đó, với Whitelist chứa một danh sách trang web có thể chấp nhận cho truy cập.

Vấn đề chính với cả 2 danh sách này là các trang web mới luôn xuất hiện gây khó khăn cho việc cập nhật 2 danh sách này. Và giai đoạn cập nhật chủ yếu là bằng thủ công. Nhà quản trị phải sưu tầm những trang web cấm để bổ sung vào tập danh sách đen. Thao tác gần như thừa đối với việc phải cập nhật danh sách trắng (cho phép dùng!).

3.1.2. Chặn từ khóa (keyword blocking):

Với cách tiếp cận này một danh sách các từ khóa (keyword) được hình thành để nhận ra các trang web bị lọc. Ta biết rằng một trang web cấm chứa nhiều từ khóa bất hợp lệ, đây là cơ sở chính để nhận ra trang web bị cấm. Một vấn đề quan trọng trong phương pháp lọc này là ngữ nghĩa của từ khóa theo ngữ cảnh. Điều này cũng dễ dẫn đến sự nhầm lẫn của hệ thống khi đưa ra nhận định về một trang web có được thể hiện hay không. Ví dụ: một website chuyên nghiên cứu về bệnh ung thư có thể bị khóa với lý do: bài viết về bệnh ung thư vú, chúng ta cũng dễ thấy là chữ "vú" (breast, trong lớp khiêu dâm) xuất hiện nhiều lần như vậy là hệ thống vô tình

khóa trang này lại! Vấn đề thứ hai mà hệ thống chịu thua đó là các từ cố ý hay vô ý đánh vần sai, chẳng hạn như: có một site chứa nhiều điều ác ý thì ngôn từ được dùng trong trang web của nó bị thay đổi, ví dụ như chữ "pornographic" bị thay thành "pornogaphic" để đánh lừa hệ thống lọc (tuy nhiên người đọc vẫn có thể hiểu: "sai chính tả thôi!"). Sự thay đổi thế này dù nhỏ nhưng nó ảnh hưởng rất lớn đến hệ thống.

3.1.3. Hệ thống đánh giá (Rating systems):

Một hệ thống đánh giá điển hình là PICS (Platform for Internet Content Selection) có thể thực hiện đánh giá các Web site. Có 2 cách tiếp cận theo dạng đánh giá các site:

- ❒ Tự đánh giá (Self-rating): Cách này những trang Web được phát hành tự phát sinh thông tin phân loại của riêng chúng.
- ❒ Thành phần thứ ba đánh giá (Third-party rating): có sự phụ thuộc vào thành phần thứ ba độc lập dùng để ước lượng các web site và công bố kết quả.

Các thông tin này có thể dùng cho các mục đích lọc web. Phương pháp này vướng phải một vấn đề là nó không mang tính bắt buộc và không có sẵn. Hơn nữa vì khả năng có thể tự đánh giá, kết quả đánh giá thường không đủ tin tưởng và chính xác.

Tóm lại, phần lớn các phần mềm lọc web hiện nay dùng kỹ thuật danh sách trắng và danh sách đen, một số dùng đến phân loại từ khóa hay đánh giá. Đa số các phần mềm này chạy máy đơn, một số làm như bộ cắm thêm (plug-in) chạy dưới một browser.

Hiệu suất của một hệ thống lọc có thể được đo lường bằng đơn vị tỉ lệ khóa (bloking rate), đơn vị này nói lên phần trăm bị ngăn chặn chính xác. Và overblocking rate là tỉ lệ phần trăm những trang web hợp pháp bị khóa.

3.1.4. Lọc các yêu cầu Domain Name System (DNS)

3.1.4.1. Khái niệm:

– Sử dụng mẫu DNS giả mạo cho các hostname của trang (site) bị cấm. Như vậy mỗi URL sẽ có một entry nhân tạo được ISP tạo ra tại bộ lọc.

– Khi người dùng địa chỉ DNS cố gắng phân giải hostname về một địa chỉ IP thì bộ xử lý phân giải thì sẽ trả về giá trị mà ISP đã chọn.

3.1.4.2. Kết quả của lọc qua DNS:

Những Website bị lọc sẽ hoàn toàn không thể truy cập được đến tất cả các cấu hình sử dụng bộ lọc nameserver cho bộ phân giải tên. Vì tất cả các bộ lọc nameserver sẽ trả về thông tin bất hợp lệ khi yêu cầu phân giải một hostname của website bị lọc. Như vậy không thể truy cập đến tài liệu trên của máy chủ chứa Website. Nhưng các Website không bị lọc sẽ cho phép truy cập miễn là chúng nó có một hostname khác từ các website bị lọc. Vì tên của chúng không được hỗ trợ thông tin bất hợp lệ bởi bộ lọc nameserver nên dữ liệu đúng sẽ trả về cho bất cứ người dùng nào yêu cầu phân giải tên và website hiển nhiên là có thể truy cập vào được.

3.1.4.3. Những ưu điểm:

- ❒ Sử dụng đa nghi thức (multi-protocol): http, ftp, gropher và bất kỳ nghi thức nào khác dựa trên hệ thống tên.
- ❒ Có thể ngăn chặn những cổng phi tiêu chuẩn (non-standard ports): Ngăn chặn các website trên cổng phi tiêu chuẩn. Website bị chặn không nhất thiết phải ở cổng mặc nhiên của TCP là 80. không có sự quá tải đáng kể bị sinh ra bởi cơ chế này. Vì lưu thông mạng không thể chạy vào các site bị lọc trong vị trí đầu tiên. Và cũng không cần thiết xem xét từng gói một trên luồng mạng.
- ❒ Không bị ảnh hưởng bởi việc thay đổi IP: Khi thay đổi IP của một website không ảnh hưởng đến phương pháp lọc này, đây là phương pháp lọc hoàn toàn độc lập với địa chỉ IP.

- ❒ Cơ động: Vì kỹ thuật này sử dụng cơ sở hạ tầng có trước, và tạo ra sự ảnh hưởng về xử lý bé nhất, gần như nó sẽ theo tỷ lệ nhận biết dịch vụ tên miền đã cài đặt. Về mặt duy trì, bất kỳ một tổ chức muốn duy trì nameserver đang tồn tại có thể giữ cơ chế này. Với một ISP thường có những công cụ tự động để đơn giản hóa việc quản trị.

3.1.4.4. Những nhược điểm:

- ❒ Không hiệu quả đối với các URL có chứa địa chỉ IP:
  - ✓ Phần lớn những địa chỉ của một website ở dạng DNS (www.hcm.edu.vn/index.htm), tuy nhiên cũng có những địa chỉ được chỉ định bằng một địa chỉ IP thay vì là dạng DNS (http://203.168.0.23/index.htm).
  - ✓ Trong trường hợp này nó được truy cập đến bằng địa chỉ IP mà không phải dùng địa chỉ DNS của nó.
- ❒ Toàn bộ web server bị chặn hoàn toàn:
  - ✓ Kỹ thuật không cho phép việc khóa có chọn lựa các trang còn lại trên một webserver. Vì thế, nếu một trang bị cấm là www.exp.com/bad.htm thì có thể tất cả các truy cập không thể truy xuất đến www.exp.com dù nó không trong danh sách bị khóa.
- ❒ Ảnh hưởng đến các subdomain
  - ✓ Xét về kỹ thuật, một tên miền đơn như example.com trong URL http://www.example.com được dùng truy cập đến web server. Cùng một thời điểm, domain name có thể phục vụ như một domain cấp trên của các cổng khác như host1.example.com. Trong trường hợp này, những địa chỉ DNS dạng www.example.com có thể bị phân giải sai. Ngoài ra, nó cũng làm cho bộ phân giải tên miền bị sai đối với các miền con. Và nó còn ảnh hưởng đến các dịch vụ chạy trên mạng như e-mail.

❒ Phạm vi bị giới hạn với người dùng DNS Server

✓ Kỹ thuật này có thể bị người dùng đánh lừa bằng cách đi vòng, khi họ thực hiện đổi DNS của máy họ hướng đến một DNS không bị lọc.

✓ Thêm vào đó phần lớn người dùng của một ISP thường có quyền kiểm soát nameserver của họ, đó là nguyên nhân giúp họ có thể vượt qua được sự kiểm soát của các ISP.

3.1.5. Bộ lọc qua URL:

3.1.5.1. Tổng quan:

– Đây là kỹ thuật lọc bằng cách quan sát lưu thông web (HTTP) bằng cách theo dõi URL và các host field bên trong các yêu cầu HTTP để nhận ra đích đến của yêu cầu. Host field được dùng riêng biệt bởi các máy chủ web hosting để nhận ra tài nguyên nào được trả về.

– Lọc web qua URL thường được xếp vào loại chủ đề rộng lớn về "Content Management". Các kỹ thuật lọc qua URL ra đời từ 2 kiểu lọc "pass-by" và "pass-through".

❒ Lọc theo "pass-by":

Một sản phẩm lọc web theo cách "pass-by" xử lý trên đường mạng mà không cần phải trực tiếp trong đường nối giữa người dùng và internet. Yêu cầu ban đầu được chuyển đến máy chủ web đầu cuối. Nếu yêu cầu bị cho là không thích hợp thì bộ lọc sẽ ngăn chặn những trang gốc từ bất cứ yêu cầu truy cập nào. Kỹ thuật này cho phép thiết bị lọc không bao gồm bộ định hướng yêu cầu. Nếu thiết bị lọc bị hỏng, lưu thông mạng vẫn tiếp tục hoạt động một cách bình thường.

❒ Lọc theo "pass-through":

Kỹ thuật lọc "pass-through" gồm việc sử dụng một thiết bị trên đường của tất cả yêu cầu của người dùng. Vì thế lưu thông mạng đi qua bộ lọc "pass-through" là thiết bị lọc thực sự. Thường bộ lọc này nằm trong các kiểu firewall, router, application switch, proxy server, cache server.

3.1.5.2. Tùy chọn bộ lọc URL:

Một số sản phẩm khác và những kiểu sản phẩm có khả năng thi hành phương thức lọc qua URL. Một vài sản phẩm được thiết kế một cách đặc biệt với mục đích duy nhất là biểu diễn Quản Trị nội dung (Content Management), mà phương thức lọc URL là một thành phần trong đó. Những phần mềm như thế thường trở thành những phần mềm tích hợp với dịch vụ hỗ trợ một danh sách các website mà những website này được xác định bởi hãng sản xuất điều này không thích hợp cho nhiều môi trường, hay sẽ bị một vài cách tấn công để đánh thủng. Bản quyền cho những sản phẩm này thường dựa trên cơ sở từng người dùng, và có chi phí cho nhà sản xuất lập "danh sách các site xấu" (Bad Site List). Một vài sản phẩm thuộc về những mẫu thiết kế riêng biệt sau [12, tr.12]:

Sản Phẩm	Hãng (Công ty)
Smartfilter	Secure Computing
Web Filter	SurfControl
Web Security	Symantec
bt-WebFilter	Burst Technology
CyBlock Web Filter	Wavecrest Computing

Bảng 3.1: Một số sản phẩm lọc web theo phương thức URL.

Những sản phẩm này cho phép người dùng chỉ định các URL bằng cách thêm hay bớt các URL khỏi "danh sách các site xấu" (Bad Site List) mặc dù các website nguyên thủy trong danh sách không thể bị loại bỏ.

3.1.5.3. Ưu điểm:

- Thiết bị được cài đặt sẵn: Phụ thuộc vào thiết bị đã cài đặt: kỹ thuật này không yêu cầu thêm phần cứng. Một ISP có lẽ có sẵn phần cứng thích hợp đủ cho bộ lọc qua URL (thường tích hợp trong các router chạy các phần mềm lọc qua URL).
- Những Website ảo không bị ảnh hưởng: Kỹ thuật này không ảnh hưởng đến các máy chủ web ảo khi chúng cùng dùng một IP như những website

hạn chế. Một website bị chặn và website không bị chặn có thể chia sẻ cùng một địa chỉ IP.

- ❒ Không ảnh hưởng đối với việc thay đổi IP: Trong phần lớn tình huống, sự thay đổi IP của website bị hạn chế sẽ không ảnh hưởng đến phương pháp này. Vì phương pháp lọc này không phụ thuộc vào địa chỉ IP. Chủ sở hữu những trang web có thể đòi bất cứ IP nào họ muốn, nhưng người dùng đứng sau bộ lọc không thể truy cập được.
- ❒ Những trang cụ thể có thể bị chặn: Kỹ thuật này cho phép tuyển chọn những trang riêng lẻ để ngăn chặn từ web server. Tuy nhiên tính năng này phụ thuộc vào năng lực của sản phẩm bộ lọc được lựa chọn. Nó có thể cấm một số trang con của một website nhưng có thể truy cập được trang chủ và những trang khác.
- ❒ Có hiệu quả đối với những URL có chứa địa chỉ IP
  - ✓ Phần lớn các URL bao gồm tên miền (DNS) của máy chủ web. Tuy nhiên cũng có số số website còn mang địa chỉ IP.
  - ✓ Bộ lọc qua URL vẫn có thể cấm các truy xuất đến những trang có IP trong địa chỉ URL của yêu cầu gởi lên.
- ❒ Không hạn chế đến người dùng DNS server:
  - ✓ Không giống như phương pháp lọc qua DNS, thao tác này có thể bỏ qua việc người dùng thay đổi thiết lập DNS của máy tính của họ hướng về một server không qua bộ lọc.
  - ✓ Phương pháp này sẽ làm việc miễn sao sự kết nối của họ đến Internet có thể xác định rõ là phải qua hệ thống lọc.
- ❒ Sử dụng cho nhiều nghi thức: Sản phẩm Content Management có thể được dùng trong bộ lọc các truy xuất thông qua HTTP, FTP, Gopher và bất kỳ nghi thức nào. Có khả năng hỗ trợ các nghi thức phi HTTP là những sản phẩm và cấu hình phụ thuộc.

3.1.5.4. Hạn chế:

- ❒ Thường không thể ngăn chặn các cổng phi tiêu chuẩn:
  - ✓ Những Web server làm việc với cổng tiêu chuẩn rất tốt.
  - ✓ Website trên các cổng phi tiêu chuẩn thì khó khăn cho việc ngăn cấm vì chúng yêu cầu một cấp độ cao hơn trong bộ lọc.
  - ✓ Một giải pháp lọc qua URL có thể là kỹ thuật có khả năng cần thiết cho những kết nối HTTP trên các cổng phi tiêu chuẩn
- ❒ Không làm việc với các lưu thông bị mã hóa

– Vì HTTP yêu cầu sử dụng SSL/TLS bị mã hóa. Phương pháp lọc theo URL không thể đọc các hostfield. Cho nên, bộ lọc không có hiệu quả phát hiện một tài nguyên nào trên một địa chỉ IP mà yêu cầu thực sự định hướng vào.

- ❒ Vấn đề chi phí cao:

– Một vài nhà bán lẻ các sản phẩm thiết kế đặc biệt cho Content Management và lọc qua URL có thể đẩy giá sản phẩm lên quá cao, nhất là đối với các nhà cung cấp dịch vụ Internet. Bởi vì nhiều nhà phân phối lẻ trói buộc số lượng người dùng bị lọc vào giá cả sản phẩm. Vì các ISP có một lượng lớn người dùng, giá cả phần mềm và bất kỳ bản quyền định kỳ có tính chất bắt buộc trở thành đắt đỏ.

– Nếu chúng chạy ở chế độ "pass-though" chúng cũng sẽ chuyển sang làm gia tăng hiểm họa tiềm tàng trên mạng và vì thế lưu thông trên mạng sẽ chậm.

– Nếu chúng chạy ở chế độ "pass-by" kết nối của chúng vào mạng sẽ làm tăng gấp đôi tất cả các lưu thông trên các đoạn mạng, đây là yếu tố tiềm tàng làm cho tốc độ mạng giảm xuống.

Nói chung, các server cần có bộ lọc để thực hiện loại bỏ một số trang web không tốt, nhưng nó có thể làm cho hệ thống chậm lại.

3.1.6. Lọc IP:

Đây là kỹ thuật ngăn chặn trực tiếp trên đường mạng bằng các địa chỉ IP của một website. Kỹ thuật này có thể là thiết thực trong bối cảnh các website thường bị truy cập thông qua địa chỉ IP hay nó có thể truy cập thông qua IP thay cho tên DSN. Đa số trường hợp, không được khuyến dùng do 3 sự kém cỏi sau:

- ❒ Ngăn chặn truy cập đến một IP cũng sẽ ngăn chặn lưu thông mạng đến những site có host ảo trên cùng IP bất chấp nó có nội dung liên quan đến vấn đề cấm hay không.
- ❒ Ngăn chặn truy cập đến một IP cũng sẽ ngăn chặn lưu thông mạng đến mỗi thành viên của cổng thông tin nằm trên IP đó. Nó sẽ ngăn chặn một thành phần của Website không phải là một phần hay một tập các trang con.
- ❒ Đó là sự thay đổi thường xuyên của các website bị lọc ngay khi chủ nhân website phát hiện ra bị lọc. Hành động này dựa trên DNS để cho phép người dùng vẫn còn truy cập đến trang web.

Một đánh giá của dự án NetProtect đánh giá 50 phần mềm lọc web thương mại bằng cách dùng 2,794 URL với nội dung khiêu dâm và 1655 URL với nội dung bình thường. Kết quả nhận được như sau [12, tr.12]:

Phần mềm lọc	Tỉ lệ khóa đúng	Efectiveness Rate
BizGuard	55 %	10 %
Cyber Patrol	52 %	2 %
CYBER sitter	46 %	3 %
Cyber Snoop	65 %	23 %
Norton InternetSecurity	45 %	6 %
SurfMonkey	65 %	11 %
X-Stop	65 %	4 %

Bảng 3.2. Kết quả đánh giá của NetProject [11].

3.2. Xây dựng giả thiết

3.2.1. Đề xuất cho một phương pháp lọc Web:

Với phương pháp blacklist và whitelist sẽ khó khăn cho việc phát sinh và duy trì, còn với việc lọc web dựa trên sự so sánh keyword của Naïve có thể dễ dàng lừa bịp bằng cách cố ý đánh vần sai những keyword và công nghệ để vượt qua vấn đề này dẫn đến kết quả năng suất tính toán cao và gia tăng số lượng mẫu tích cực sai. Cuối dùng là các hệ thống phân loại (rating systems) không cung cấp thông tin đáng tin cậy.

Đề xuất phương pháp lọc web dựa trên phân loại văn bản (text classification). Sử dụng mẫu những trang web cấm để lấy đặc điểm lớp của những trang web bị chặn. Một trang web "gần giống" hay "giống" với thành viên của lớp đó sẽ bị chặn và những trang còn lại "không giống" sẽ cho qua.

Việc áp dụng thuật toán phân loại văn bản một số điểm phải được tính đến. Trước tiên, sự phân loại cho việc lọc web một lớp được phân loại, trong đó kết quả của sự phân lớp là một trong hai: *cho phép* hay *ngăn chặn một trang*. Sự phân lớp sẽ nhận ra nếu một trang thuộc về một lớp cấm, ví dụ như trang đó có phải là trang khiêu dâm hay không? Phần lớn những hệ thống phân loại văn bản truyền thống được xây dựng trên hai lớp: tích cực (positive) và tiêu cực (negative): lớp tích cực gồm những văn bản có cùng đặc điểm nổi trội, trong khi đó những văn bản không cùng đặc điểm được liệt vào lớp tiêu cực. Trong việc phân loại các trang web, rất khó cung cấp một mẫu điển hình của lớp tiêu cực vì có rất nhiều tài liệu trong lớp này.

Với phương pháp đề xuất mới này, chỉ dùng một tập những tài liệu huấn luyện tích cực vì thế loại bỏ đi vấn đề thiết lập và duy trì một tập tài liệu "tiêu cực" hoàn thiện và cân đối. Hơn nữa, trong những phương pháp phân loại văn bản truyền thống, các văn bản cần phân loại được xem xét sự độc lập, vì thế sự phân loại của một tài liệu không hỗ trợ thông tin hữu ích về sự phân lớp của những tài liệu khác. Trong việc lọc web, trang web có thể được tìm đến thông qua siêu liên kết

(hyperlink) trong văn bản cũng hỗ trợ thông tin hữu ích cho sự phân loại văn bản. Trường hợp nội dung trang web không thể đưa ra một sự phân loại rõ ràng của văn bản, sử dụng các siêu liên kết để tìm các trang được xem như tương tự với sự nghiên cứu, có thể rất hữu ích.

3.2.2. Thuật toán:

Mỗi tập tin tài liệu được thể hiện như một vec-tơ tần suất từ, độ dài của vectơ sẽ là N và vì chỉ có những tần suất N những từ phổ biến nhất được giữ lại. Sự tương đồng giữa hai tài liệu được đo bằng thuật ngữ "cosine" của "góc" giữa hai vectơ, hai tài liệu có độ tương đồng lớn thì có số đo góc giữa vectơ nhỏ vì thế giá trị cosine của nó lớn, ngược lại với hai tài liệu độ tương đồng nhỏ thì góc giữa hai vec-tơ tài liệu lớn, do đó cosine của nó nhỏ. Đo cosine của hai vec-tơ bằng công thức:

$$\cos(X,Y) = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$$

Trong đó X và Y là hai vector của hai tài liệu.

Tập huấn luyện T_s gồm các trang web mẫu mang nội dung bị cấm. Để phân loại một trang mà độ tương đồng của nó đối với tập huấn luyện T_s được lượng giá nếu nó vượt trên ngưỡng, thì nó sẽ được xem xét và đưa vào lớp cấm. Để xác định ngưỡng, người ta xây dựng một tập dữ liệu T_s' mà nó bao gồm những mẫu với nội dung bị cấm và những mẫu trang web có nội dung cho phép. Sau đó để đưa ra một phạm vi của các ứng viên ngưỡng, chúng ta dùng mỗi ứng viên để phân loại thành viên của T_s' và chọn ra một ứng viên τ mà phần lớn các thành viên của T_s' phân loại đúng. Để tính toán độ tương đồng (similarity coefficient) của một trang P vào một lớp được định nghĩa bởi T_s, sự tương tự của trang P với mỗi tài liệu huấn luyện trong T_s được tìm ra và sau đó tính trung bình cộng của n% các giá trị tương tự cao nhất được dùng như hệ số tương tự của trang P đối với T_s. Ở đây n là một con số phụ thuộc vào số lượng nhóm con trong T_s. Ví dụ: loại từ "sex" có thể chứa đựng hai nhóm con "erotic stories" – chuyện khiêu dâm, và "ponorgraphic galleries" – hình khiêu gợi. Vì một tài liệu thuộc về một nhóm con có thể không cần thiết phải

tương tự với một nhóm con khác. Để các tài liệu thuộc về một thành phần của nhóm con, các giá trị trung bình phải trên 50% sẽ cho kết quả trong một hệ số tương tự cao hơn loại từ "sex" với tất cả trung bình cộng của tất cả những giá trị tương tự.

Nếu hệ số tương tự của trang P đối với tập T_s bé hơn ngưỡng, thì các mối liên kết (hyper-link) bên trong trang P được xem xét đến và tính ra hệ số tương tự của các trang mà liên kết (link) đó chỉ đến. Nếu trong phần lớn các trường hợp các mối liên kết này chỉ đến một trang tương tự với loại cấm thì trang P cũng được phân loại vào trong lớp cấm.

3.2.3. Tóm lược các bước của thuật toán:

Bước 1: Đưa vào tập T_s gồm những tài liệu huấn luyện trong đó mỗi tài liệu thuộc về một lớp cấm. Tập này được chọn trong giai đoạn khởi tạo và được cập nhật thường xuyên.

Với một tài liệu $T \in T_s$ một vector hỗ trợ v_T của quan hệ các tần suất từ sẽ được xây dựng dựa vào những bước sau:

a. Loại bỏ những từ phổ biến như "the", "and", "for",… (bởi vì những từ loại này xuất hiện trong tất cả các tài liệu và sẽ không ảnh hưởng đến quá trình phân loại tài liệu).

b. Bỏ qua những từ có tần suất thấp (những từ này không góp phần gì trong việc nhận diện loại tài liệu)

c. Bỏ đi những từ loại ngắn hơn hai ký tự (những từ loại này như "a", "as", "to", "of", "in", … cùng với các ký hiệu như "@", "?", "#", "&", …)

Ngoài ra, trong bước 1 này còn có một vấn đề cần thực hiện đó là rút gọn các từ (Stemming), nhằm làm giảm bớt số chiều của vector.

Sau cùng, tạo ra vector tần suất của văn bản gồm từ và tần suất từ. Vector này đại diện cho văn bản (trang web) đưa vào cho hệ thống xét duyệt.

Bước 2. Tìm ra ngưỡng τ dùng cho việc quyết định một tài liệu thuộc về lớp cấm. Sử dụng những ngưỡng cao hơn sẽ dẫn đến tình trạng những trang mà chính nó thuộc về lớp bị cấm sẽ bỏ sót, và sử dụng ngưỡng thấp hơn sẽ đưa đến kết quả những trang nó không thuộc về lớp bị cấm sẽ bị khóa sai (trang này đúng ra là cho phép đi qua nhưng lại cấm!). Sau khi tìm ra được tất cả vector của văn bản huấn luyện, ta dùng đến một tập mẫu thử T_s' khác, tập này bao gồm các mẫu thử bên trong và bên ngoài lớp bị cấm và tính ra hệ số tương tự của mỗi phần tử trong T_s' đối với T_s (xem bước kế tiếp để biết cách tính hệ số tương tự ở bước 3). Sử dụng một dãy các giá trị ngưỡng ứng viên trong đoạn 0 đến 1 và dùng mỗi giá trị ngưỡng ứng viên đó để phân loại các thành viên của T_s'. Đặt $u_{\tau i}$ đại diện cho phần trăm của những văn bản trong T_s' mà những văn bản này được phân loại chính xác bằng ngưỡng τ_i. Ta chọn ngưỡng $\tau = \tau_j$ sao cho có u_τ cao nhất làm ngưỡng cho hệ thống.

Bước 3. Xét một trang P, hệ thống sẽ tìm ra hệ số tương tự σ_P như sau:

a. Tìm sự tương tự của P so với mỗi thành viên trong tập huấn luyện. Nghĩa là tìm $\cos(v_P, v_X)$ với $\forall X \in T$.

b. Xây dựng tập S, là tập chứa n% những giá trị tương tự cao nhất.

c. Lớp hệ số lớp của P, σ_P là giá trị trung bình của n% các giá trị tương tự cao nhất. Tính theo công thức:

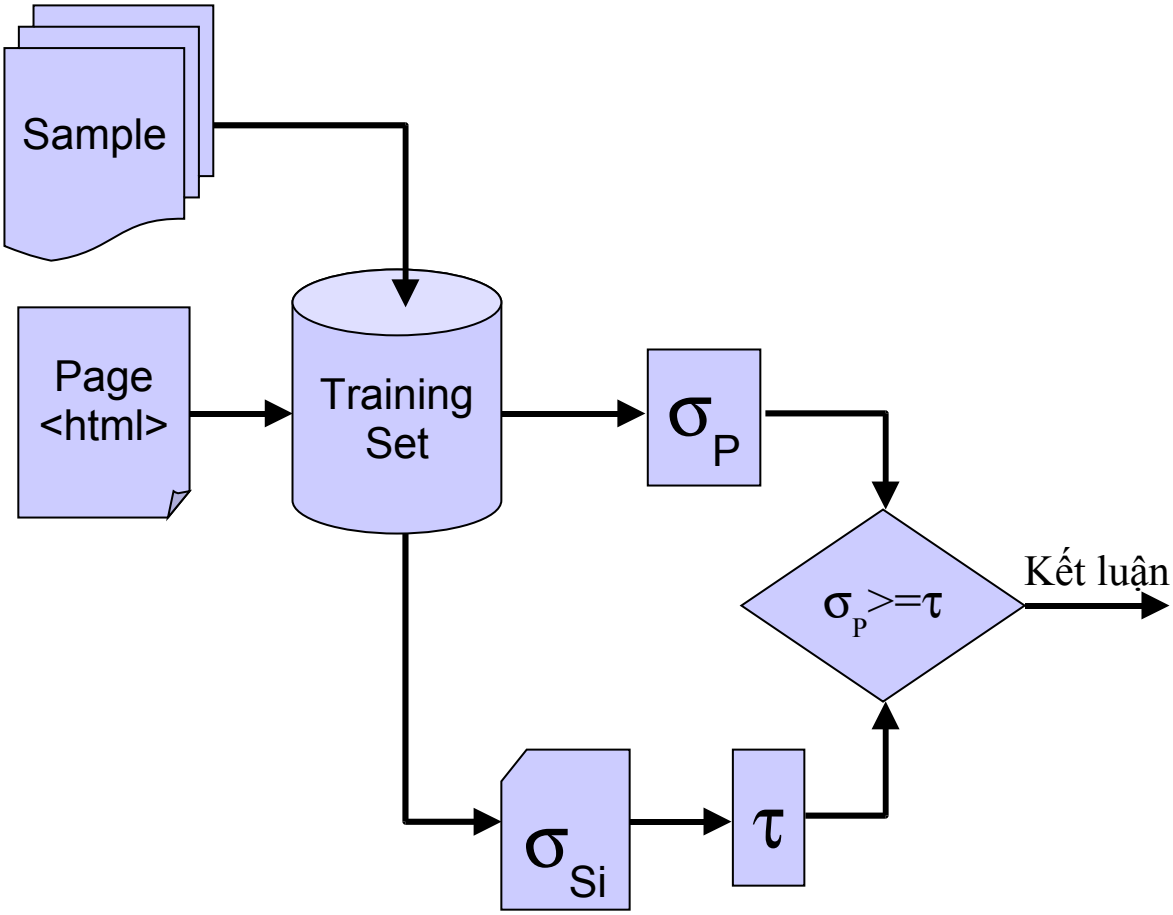
$$\sigma_P = \frac{\sum_{v \in S} v}{|T_s| \times n\%} = \frac{\sum_{v \in S} v}{|T_s| \times n\%}$$

Hệ thống sẽ so sánh σ_P với ngưỡng đã lựa chọn τ:

- Nếu $\sigma_P \geq \tau$ thì trang P sẽ bị khoá.
- Ngược lại, hệ thống sẽ xem xét r liên kết $l_1, l_2, \ldots, l_r$ một cách ngẫu nhiên có trong trang. Với mỗi l_i hệ thống sẽ tính hệ số tương tự của trang P_{li} là liên kết l_i trỏ tới. Nếu phần lớn hệ số lớp trên ngưỡng τ thì trang bị khóa, ngược lại cho phép.

3.2.4. Mô hình thuật toán:

Sơ đồ sau đây minh họa cho mô hình thuật toán.



Hình 3.3. Mô hình thuật toán phân loại văn bản

3.3. Lựa chọn phương pháp nghiên cứu:

Trong đề tài này áp dụng phương pháp nghiên cứu trí tuệ nhân tạo, chủ yếu dùng hình thức máy học để giải quyết vấn đề. Kỹ thuật "máy học" dùng trong nghiên cứu này là khai khoáng dữ liệu với nguồn dữ liệu là các văn bản (text mining, text classification).

3.3.1. Máy học là gì?

Máy học (Machine learning) nhằm giải quyết vấn đề xây dựng chương trình máy tính với mục đích làm cải tiến hiệu suất của chúng ở một số công việc thông

qua kinh nghiệm. Những thuật toán máy học đã chứng minh về giá trị hiện thực lớn lao trong nhiều lĩnh vực ứng dụng. Không ngạc nhiên gì nữa, lĩnh vực kỹ sư phần mềm đã cho ra một miền đầy hứa hẹn ở đó nhiều sự phát triển phần mềm và những tác vụ bảo quản có thể được công thức hóa như những vấn đề học và được tiếp cận dưới dạng thuật toán học.

3.3.2. Những thuận lợi của cách tiếp cận theo dạng máy học có giám sát:

Cơ chế nỗ lực tiến tới việc hình thành không chỉ là một bộ phân loại đơn thuần mà còn hướng đến một bộ phân loại tự động, tức là hệ thống luôn nạp kiến thức chứ không cố định (cơ chế học), nếu như tập các thể loại được cập nhật hay hệ thống hướng đến một lĩnh vực khác, thì tất cả những gì cần thiết chỉ là một tập những tài liệu được phân loại khác.

Cần thiết chuyên biệt hóa lĩnh vực cho việc gán nhãn không cần tri thức chuyên biệt cho cơ chế, điều này thật là thuận lợi, bởi vì nó cho thấy đặc điểm một khái niệm mở rộng dễ dàng hơn là tập trung.

Đôi khi cũng có sẵn những tài liệu được phân loại trước. Ngày nay, hiệu quả có thể đạt được bằng những bộ phân loại này tạo ra sự cạnh tranh giữa bộ phân loại thủ công (con người thực hiện) và bộ phân loại trí tuệ (con người với máy tính).

3.3.3. Đặc điểm bên trong cách tiếp cận theo dạng máy học có giám sát:

– Thường, một tài liệu được biểu diễn như một vector có trọng lượng thưa, trong đó độ dài của vector là số từ xuất hiện ít nhất trong một văn bản huấn luyện.

– Trọng lượng có thể là dạng nhị phân, nó cho biết một từ có thể xuất hiện hay không xuất hiện trong một văn bản, hay không phải dạng nhị phân: cho biết có bao nhiêu từ góp phần tạo nên ngữ nghĩa của văn bản. Trong trường hợp cuối cùng này, những hàm trọng lượng được sử dụng như hàm tìm kiếm văn bản.

– Sự phân loại là một tác vụ chủ quan, cả con người hay phân loại tự động trên máy đều dễ xảy ra lỗi, hiệu quả của một bộ phân loại (con người hay tự động) được đo lường đặc trưng bởi sự so sánh giữa sự quyết định của nó với sự quyết định của

con người mã hóa trong một tập "kiểm thử" gồm các tài liệu được phân loại trước dưới tập C.

– Phép đo lường hiệu quả "cổ điển" là $F_1 = \frac{2\pi\rho}{\pi + \rho}$ trong đó ngữ nghĩa của π (presision) và ρ (recall):

- precission (π): cho biết có bao nhiêu văn bản đã thấy bên trong một loại đích thực bên trong nó?
- recall (ρ): cho biết có bao nhiêu văn bản thực sự đang ở trong một thể loại đã thấy (hiểu theo cách thông thường).

3.3.4. Xây dựng bộ phân loại văn bản (Text Classifier)

3.3.4.1. Cách xây dựng một bộ phân loại văn bản:

Một bộ phân loại cho tập C, có thể được tạo ra theo 2 cách:

– Tạo thủ công: trang bị tri thức cho cho bộ phân loại, điều này có nghĩa là xây dựng một tập có dạng:

if ((wheat & farm) or

(corn & farm) or

(wheat & commodity) or

(wheat & tonnes) or

(wheat & winter & ¬ soft)) then Grain else ¬ Grain

– Tự động: bằng cách dùng công nghệ máy học có giám sát, từ một tập huấn luyện của các tài liệu được phân loại trước vào một lớp C.

3.3.4.2. Phân tích, những thuận lợi

Vấn đề lọc web có thể được xem như sự kết hợp của hai ứng dụng riêng biệt của phân loại văn bản:

– Phân loại trang web (Web Classification) sắp xếp các website hay tổ chức tìm kiếm các kết quả dưới những thư mục phân cấp.

- ❒ Phân loại trang web là một trường hợp đặc biệt của phân loại văn bản bởi vì sự hiện diện của các siêu liên kết. Sự cấu thành một nguồn tài nguyên phong phú về thông tin này, cũng như chúng có thể được hiểu như những sự trình bày thích hợp trang được liên kết đến trang đang liên kết.
- ❒ Nghiên cứu về sự phân loại Web cho thấy rằng: hiệu quả có thể được cải thiện bằng cách dùng heuristic "trang web đặc trưng" chẳng hạn như dùng các thể loại của các hyper-neighbour như những đặc trưng, và sử dụng "việc gán nhãn nới lỏng" kỹ thuật lặp nếu điều đó không được biết trong việc cung cấp trước.
- ❒ Các trang Web khó phân tích hơn tập tin văn bản chuẩn. Xét điều này để thấy rằng nó quan trọng để dùng đến những kỹ thuật phức tạp nhằm mục đích đạt tới những cấp độ hợp lý của hiệu quả.

– Bộ lọc (filtering): nghĩa là phân loại mỗi dòng văn bản đến vào trong $User_i$ hay không cho $User_i$ dựa trên sự liên quan đến/sự thích hợp của văn bản đến user.

- ❒ Có 2 hướng tiếp cận riêng biệt ứng dụng phân loại văn bản trong bộ lọc:
  - ✓ Tích cực sai và tiêu cực sai thường có sự quan trọng khác, điều này phải được tính đến bằng cách dùng tiện ích lý thuyết đo lường tính hiệu quả.
  - ✓ Tập huấn luyện thường đạt được sự tăng trưởng, mặc dù có sự tương tác với người dùng. Điều này nói lên rằng kỹ thuật đó phát sinh một bộ phân loại gia tăng phải được dùng đến.
- ❒ Hiện tại, các cấp độ có hiệu lực có thể tương thích với một trong những ứng dụng phân loại văn bản đã được đạt đến trong việc lọc thông qua việc dùng đến kỹ thuật "maximin-margin online learners".

– Ngoài ra còn có ứng dụng dựa trên đặc điểm của một loại ứng dụng thứ ba của phân loại văn bản, đó là phát hiện nội dung không thích hợp (Detecting unsuitable content)

- ✓ Đây là một trường hợp của quản lý nội dung trong môi trường đối lập "Content Management in Adversarial Environments-CMAE", điều này có quan hệ đến trường hợp những ứng dụng quản trị nội dung xác định nội dung hiện diện trong văn bản thuộc một tác giả nào.
- ✓ Vấn đề ở CMAE là không có kỹ thuật nào dành cho CMAE có thể được nhìn nhận chính xác và hoàn thiện, vì "động vật ăn thịt luôn thích nghi với con mồi của nó".
- ✓ Lọc những nội dung khiêu dâm và lọc thư rác (spam) là hai thể hiện của CMAE.

Chương 4: XÂY DỰNG ỨNG DỤNG, THỬ NGHIỆM, ĐÁNH GIÁ

4.1. Tổ chức dữ liệu:

4.1.1. Cấu trúc dữ liệu theo thuật toán chuẩn:

Trong thuật toán gốc, tác giả đã sử dụng tập huấn luyện T_s, đây là một tập gồm các tập tin văn bản (dạng trang web) được sưu tập làm trang mẫu tích cực (lớp cấm). Trong quá trình vận hành, các trang web cấm được bổ sung thêm vào.

Tổ chức tập huấn luyện T_s này trên lý thuyết sẽ có các phân loại để sự so sánh tìm ra đúng loại của một trang web đưa vào để xét nhanh chóng. Một trang web có thể thuộc về một phân loại chính và nhiều phân loại phụ khác. Nhưng điều quan trọng trong việc ứng dụng của nó là đưa ra quyết định trang web đưa vào có được phép đi qua hay không? Nếu trang web đưa vào giống hay gần giống theo một mức cho phép (ngưỡng) thì trang web này bị cấm: không được phép đi qua và bổ sung trang đó vào tập huấn luyện. Như vậy tập huấn luyện luôn gia tăng về số lượng tập tin.

* Xử lý một tập tin Web trong tập T_s huấn luyện hay trong P cho việc so sánh:

– Lọc bỏ những từ “Stoplist”: and, or, the, to, … (có hơn 200 từ loại này).Vì các từ này không làm ảnh hưởng nhiều đến nội dung trang web (hay văn bản). Ta cần những từ nêu lên được đặc tính của trang web hay văn bản. Xử lý này gồm các thao tác đọc nội dung tập tin, so sánh với những từ có trong “Stoplist” thì loại nó ra. Tuy nhiên với trang Web ta phải làm thêm một thao tác nữa đó là xét đến các thẻ định dạng HTML. Trong đó, đặc biệt lưu ý đến các thẻ liên hệ đến nội dung làm đại diện cho liên kết, ví dụ như: “The GNU General Public Licence” được viết trong thẻ HTML như sau: The GNU GNU General Public License.. Ở đây có hai vấn đề cần quan tâm: URL chứa bên trong và nội dung làm đại diện.

– Thống kê từ: là công việc đếm số lượng các từ còn lại và chọn ra danh mục các từ lặp lại nhiều lần nhất. Các từ này sẽ được lưu lại vào một tập tin khác (tạm thời trong quá trình xử lý). Sau khi làm hết các công việc trên, tiến hành tính độ tương tự của hai trang. Lưu ý: xem xét các từ trong các liên kết (đại diện cho các kết nối đến những trang web khác) xem nó như văn bản bình thường, mỗi từ đại diện đều được thống kê.

* Nhận xét: Dễ nhận thấy rằng, quá trình xử lý đến bước này để chuẩn bị cho việc so sánh một trang P với các trang đang lưu trong tập huấn luyện T_s thì thời gian tốn kém rất lớn để thực hiện các tao tác mô tả trong hai bước trên. Nếu ta tổ chức kho lưu trữ như trên thì mỗi lần vào thuật toán đều phải tính lại không chỉ cho trang P cần xét mà còn cho tất cả các trang mẫu trong tập huấn luyện. Do đó, ta phải tổ chức lại cấu trúc lưu trữ sao cho ít tốn kém nhất về thời gian tính toán cũng như không gian lưu trữ.

* Đề xuất cấu trúc lưu trữ để tăng tốc độ xử lý:

– Lưu trữ tập huấn luyện T_s: nếu lưu trữ trang web gốc thì quá trình xử lý trên không mang lại hiệu quả. Vì tốn kém thêm thời gian cho những xử lý không cần thiết. Ở đây, chúng ta có thể nhận thấy cách lưu trữ gồm các từ và tần suất từ là tối ưu nhất vì khi đọc lại một trang trong tập huấn luyện.

– Tổ chức thông tin ngưỡng: sử dụng tập tin văn bản ghi lại thông tin ngưỡng. Vì thế ta chỉ tính ngưỡng giới hạn cho lần đầu tiên khởi động hệ thống. Các thông tin này được giữ lại cho lần tiếp sau, khi có một mẫu mới được thêm vào thì ngưỡng được tính lại, tuy nhiên do có các thông tin trước đây nên ta tận dụng được các bước tính toán trước, không cần qua một lần thống kê toàn bộ trên tập huấn luyện T_s.

– Xử lý: theo cách tổ chức như trên, khi đưa một trang P vào để xét thì quá trình thống kê từ chỉ xảy ra đối với trang P mà không cần xử lý cho các trang trong T_s. Vì vậy vòng lặp để xét trong trang P với tập T_s như sau:

Đưa vào trang P

Xử lý biến đổi trang P thành dạng Vector

Xây dựng mảng C gồm k phần tử (k = số lượng tài liệu có trong T_s)

While T_s not Empty

{

Đọc một trang P_i trong T_s.

C[i] = So sánh (P_i, P) bằng công thức cosine

}

Chọn ra n% giá trị cao nhất trong C tạo ra C'

Dùng C' để tính hệ số tương tự của trang P (σ_P)

So sánh σ_P với ngưỡng τ để đi đến quyết định (cấm hay cho phép)

Trong đề xuất trên, không đề cập đến vấn đề xét các link, vì đây là điểm yếu sẽ làm cho quá trình chậm đi, theo đề xuất của thuật toán nguyên thủy của bài báo thì tác giả có đề cập đến các trang liên kết trong trang hiện hành trỏ đến. Như vậy quá trình xét sẽ liên quan đến đệ quy và độ sâu của đệ quy rất lớn, do những trang web thường có nhiều liên kết (vô cấp) và các trang liên tục kết nối với những trang khác nữa nên quá trình đệ quy sẽ làm chậm lưu thông mạng và máy gởi yêu cầu phải đợi lâu hơn. Thống kê từ trong các liên kết xem như các từ bình thường phục vụ cho việc so sánh toàn trang. Dù cho trang chính này có thỏa yêu cầu cho người dùng mở ra xem, nhưng khi mở một liên kết trong trang đó thì yêu cầu mới được gởi đến hệ thống và quá trình xét từng trang cho hệ thống diễn ra… như vậy tránh được quá trình đệ quy.

Kết quả sau khi xét trang P sau khi so sánh với ngưỡng τ:

– Nếu $\sigma_P >= \tau$ thì:

- ❒ Bổ sung trang P vào tập huấn luyện và tính lại ngưỡng giới hạn (dựa vào thông tin đã lưu giữ trước cùng với thông tin mới bổ sung).

- ❒ Trả về trang báo lỗi mặc định: "Trang cấm truy xuất". Các báo lỗi này ta vẫn thường thấy trên mạng khi tường lửa của các ISP ngăn chặn hay cấm một số trang nào đó.

– Ngược lại khi σ_P không vượt ngưỡng cho phép thì trang đó đi qua.

4.1.2. Cấu trúc dữ liệu đề xuất cho lập trình:

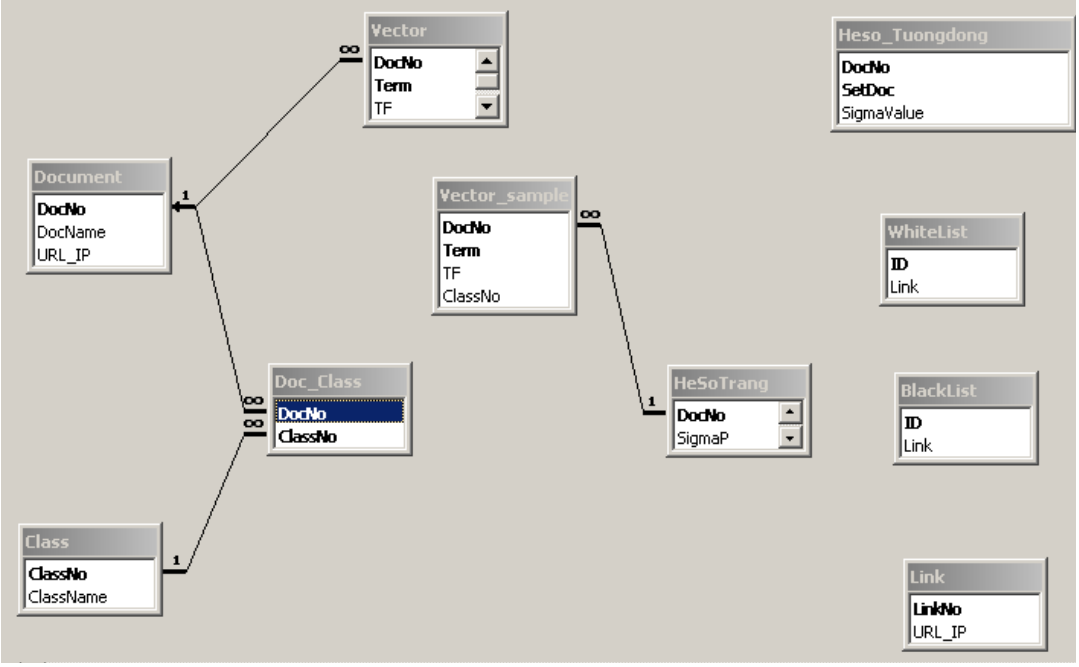
4.1.2.1. Xây dựng cấu trúc dữ liệu:

Ngôn ngữ được chọn để xây dựng ứng dụng cho thử nghiệm này là JAVA, đây là ngôn ngữ lập trình mới, đang phát triển mạnh mẽ. Phần cấu trúc dữ liệu cho lập trình và chương trình có thể chạy là một hệ quản trị cơ sở dữ liệu. Thay vì lưu trữ tập huấn luyện trong mỗi tập tin thì cấu trúc đó được sửa lại như sau:

– Table: mô tả các đặc điểm của vector tài liệu, mô tả các tập tin liên quan đến vector đó, ghi nhận liên kết (URL hay IP) của vector đó trỏ đến, ghi nhận số lần máy bên trong gởi yêu cầu lên proxy không được phục vụ.

– Query: dùng cho việc thống kê trên cơ sở các vector (từ, tần suất từ) để phục vụ cho việc tính ngưỡng giới hạn, quản trị.

Ngoài ra, cơ sở dữ liệu này còn dùng cho việc thiết kế công cụ quản trị hệ thống thông qua giao diện web.



Hình 4.1. Quan hệ giữa các bảng trong cơ sở dữ liệu tập huấn luyện và tập thử (sample)

Vì muốn đơn giản và nhanh chóng trong việc lập trình cũng như thử nghiệm, tác giả chọn hình thức tổ chức này, nhằm tận dụng những điểm mạnh của hệ quản trị cơ sở dữ liệu, cũng như quản lý tập trung mà không làm mất đi ý nghĩa lưu trữ tập huấn luyện và cải thiện tốc độ tính toán như trình bày ở mục trước.

Như vậy quá trình xét một trang P chỉ làm một số việc như sau:

– Vector hóa trang P (loại bỏ từ stoplist, tính Stemming, tính tần suất từ) trang P trở về dạng <từ> <tần suất từ> đây là dạng vector tài liệu dùng cho việc so sánh hai trang.

– Tiến hành so sánh P với từng vector X_i nằm trong tập huấn luyện (trong cơ sở dữ liệu) thông qua công thức tính độ tương tự hai văn bản cosine. Đây là quá trình sẽ chiếm nhiều thời gian xử lý do số lượng phần tử có trong tập huấn luyện lớn (và ngày càng lớn, do có sự bổ sung tự động).

– Đưa ra quyết định: kết quả cuối cùng sau khi qua xử lý chính đưa đến một quyết định: nếu cho đi qua thì trả trang web theo URL mà người dùng yêu cầu để

truy cập đến thông tin. Nếu không đủ điều kiện đi qua, thì báo lỗi và lấy hết cấu trúc của trang P đó ghi vào cơ sở dữ liệu của tập huấn luyện và đánh dấu theo chỉ mục hiện có trong table đó.

4.1.3. Chuẩn bị dữ liệu:

4.1.3.1. Một số khái niệm:

– Stopword: là các từ thường không ảnh hưởng nhiều đến nội dung của văn bản, thường là các giới từ.

– Stemming: là quá trình tìm từ gốc, quá trình này nhằm làm giảm đi số chiều trong vector văn bản. Những từ biến đổi trong các thì hay các thể sẽ được rút ngắn lại thành dạng nguyên mẫu của nó. Ví dụ: động từ go có các biến thể như sau: goes, went, gone, going. Xử lý stemming sẽ làm quá trình tính toán trên vector nhanh hơn và giảm không gian lưu trữ.

– Thống kê tính tần suất từ: đây là giai đoạn hình thành vector tần suất, quá trình thực hiện đếm các từ trong văn bản, những loại từ trùng nhau được ghi nhận số lần lặp lại trong một biến đếm đặc trưng cho từ đó.

– Tính cosine: tính hệ số tương tự của một trang X so với một trang Y. Hệ số này là cos của hai vector V_x và V_y. $Cos(V_x, V_y)$ càng lớn thì góc giữa V_x và V_y càng nhỏ → hai văn bản càng giống nhau và ngược lại. Tính cosine của hai vector n chiều dựa vào tọa độ của các chiều không gian của nó.

– Hệ số tương tự của trang: là hệ số của P so với tập huấn luyện, tính hệ số tương tự nhằm lượng giá cho trang P, giá trị lượng giá này được dùng cho việc so sánh với ngưỡng để biết trang P có vượt qua giới hạn cho phép không?

– Ngưỡng: là giới hạn mà hệ thống dùng trong việc so sánh một trang web P khi đi qua nó, trang P được tính hệ số tương tự làm giá trị so với ngưỡng.

– Xử lý: thêm trang web mới vào kho lưu trữ web mẫu, tính lại ngưỡng giới hạn cho hệ thống.

4.1.3.2. Qui trình thực hiện:

4.1.3.2.1. Lập chỉ mục

Đây là công đoạn chuẩn bị tài liệu để sử dụng cho hệ thống rút trích thông tin. Điều này cho biết việc chuẩn bị thu thập tài liệu thô thành một sự biểu diễn những tài liệu sao cho truy cập một cách dễ dàng. Sự biến đổi từ một tài liệu văn bản thành một sự biểu diễn của văn bản được biết đến như chỉ mục của những tài liệu. Thay đổi hình thức một tài liệu về dạng được chỉ mục có những yêu cầu sau:

- ❒ Một thư viện hay một tập các biểu thức đầy đủ (regular expressions)
- ❒ Phân tích từ loại
- ❒ Một thư viện của stop words (stop list)
- ❒ Những bộ lọc hỗn hợp khác

Bình thường quá trình này diễn ra theo năm bước:

- ❒ Loại bỏ đánh dấu và định dạng
- ❒ Thẻ hóa (tokenization)
- ❒ Tách lọc (filtration)
- ❒ Stemming
- ❒ Gán trọng số.

Nếu không yêu cầu loại bỏ đánh dấu và tính trọng số thì sự chuyển đổi này chỉ bao gồm các bước thể hóa, tách lọc, và stemming. Kiểu chỉ mục này được tìm thấy thường xuyên trong cơ sở dữ liệu, tại đó đơn thuần là sắp xếp các tập tin văn bản và dữ liệu thô. Tuy nhiên, trên trang Web năm bước trên được sử dụng cẩn thận vì các tài liệu được tạo trong những định dạng khác nhau và những điểm liên quan được cần đến.

4.1.3.2.2. Tuyến tính hóa tài liệu (Document Linearization)

Tuyến tính hóa tài liệu là quá trình xử lý văn bản nhằm làm giảm bớt phân loại từ. Quá trình này trải qua hai bước như sau:

i) Loại bỏ định dạng và đánh dấu (Markup and Format Removal) trong suốt giai đoạn này, tất cả các thẻ đánh dấu và những định dạng đặc biệt bị loại bỏ khỏi tài liệu. Thế nên, với một tài liệu HTML tất cả những thẻ và văn bản bên trong bị loại bỏ. Thông thường điều này có thể bao gồm tất cả những phần tử thuộc tính, kịch bản (script), những dòng ghi chú và văn bản đặt vào trong đấy.

ii) Thẻ hóa (Tokenization): Suốt giai đoạn này, tất cả những văn bản còn lại được phân tích từ loại, chữ thường và chấm câu bị loại bỏ.

Nói tóm lại, sau khi tuyến tính hóa tài liệu ta được:

- ❒ Những luồng văn bản phù hợp nên được miêu tả thành những luồng từ chặt chẽ.
- ❒ Luồng văn bản này phải đạt được những ngữ nghĩa, chủ đề, đề tài, đề tài con … trong tài liệu.
- ❒ Vị trí của những từ trong luồng văn bản được xác định bởi các dòng đánh dấu như thế nào (chẳng hạn các thẻ HTML) được công khai trong mã nguồn.

Những điều này gạch nối sự nhận thức về mối liên quan của con người (nghĩa là thông tin và ngữ nghĩa của nó được hiển thị trước người sử dụng) và sự nhận thức về mối liên quan của máy là hai vấn đề khác nhau.

4.1.3.2.3. Tách lọc (Filtration)

Tách lọc được biết đến như một quá trình của sự quyết định những từ nào nên được sử dụng để biểu diễn cho các tài liệu vì thế nó có thể được sử dụng cho:

- ❒ Mô tả nội dung của văn bản
- ❒ Có sự phân biệt tài liệu từ những tài liệu khác trong bộ sưu tập.

Những từ được sử dụng thường xuyên không thể được dùng cho mục đích này vì hai lý do. Đầu tiên, số lượng tài liệu liên quan đến một truy vấn như quan hệ tỉ lệ đối với bộ sưu tập. Một từ sẽ có hiệu quả trong việc tách những tài liệu có liên quan

ra khỏi những tài liệu không liên quan thì có thể là một từ xuất hiện trong một số ít tài liệu. Điều này có nghĩa là những từ có tần suất cao thì sự phân biệt thấp. Lý do thứ hai là những từ xuất hiện trong nhiều ngữ cảnh không xác định một đề tài hay một đề tài phụ của một tài liệu. Tuy nhiên, việc loại bỏ stopword khỏi văn bản là việc làm mất thời gian. Một sự tiếp cận có hiệu quả bao gồm việc rút ra tất cả những từ xuất hiện thường xuyên trong quá trình thu thập tài liệu không cải thiện được sự rút trích những tư liệu liên quan.

Điều này được hoàn thành với một thư viện stopword (một stop-list các từ bị loại bỏ). Những danh sách này có thể được tạo ra bằng một trong hai cách: dựa vào đặc điểm chung (áp dụng đối với tất cả các bộ dữ liệu) hay tính riêng biệt (tạo ra từ bộ dữ liệu chỉ định). Giá trị ngưỡng số xuất hiện của từ cho biết những từ bị loại bỏ trong bộ sưu tập phụ thuộc vào sự bổ sung riêng lẻ. Chẳng hạn, một số hệ thống rút trích thông tin những từ xuất hiện hơn 5% đối với bộ sưu tập bị loại bỏ, ngược lại, những từ không nằm trong stop-list nhưng nó lại xuất hiện lớn hơn 50% đối với bộ sưu tập thì bị xem là từ “tiêu cực” thì nó cũng bị loại bỏ để tránh những rắc rối về trọng số.

4.1.3.2.4. Stemming (gốc từ)

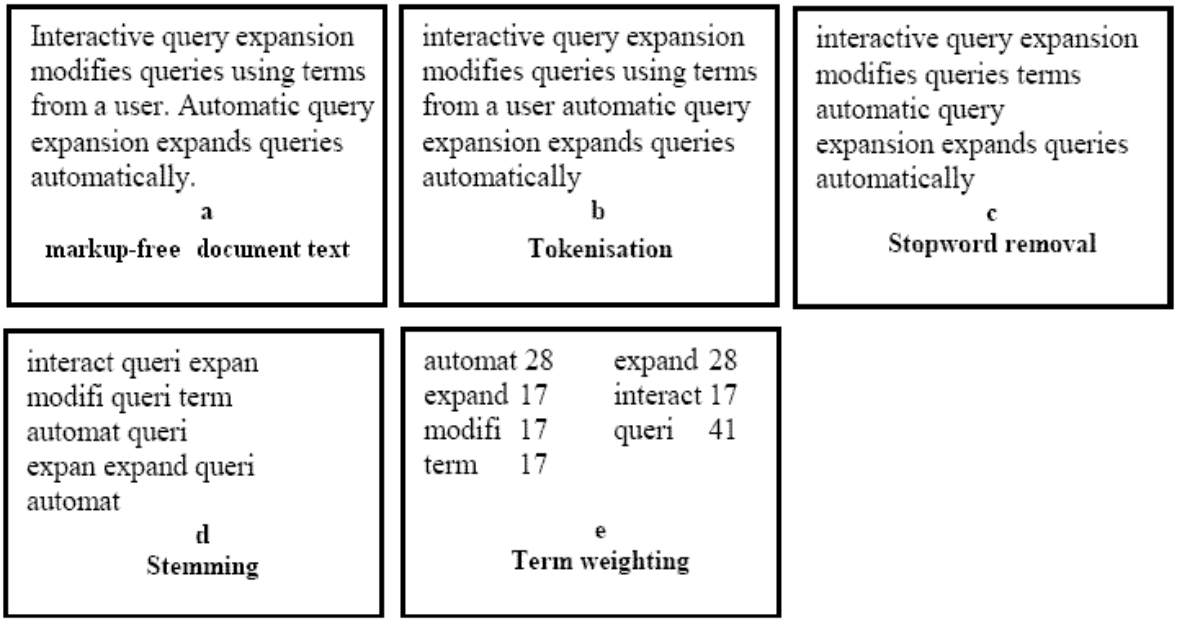
Stemming là quá trình liên quan đến việc xử lý giảm đi số từ đối với gốc từ hay cội nguồn khác nhau của chúng. Do vậy, những từ "computer", "computing", "compute" được giảm lại thành từ "compute" và "walks", "walking" và "walker" được giảm lại thành "walk". Không phải tất cả những hệ thống cùng sử dụng một bộ xác định gốc từ. Đối với tiếng Anh, bộ xác định gốc từ phổ biến là thuật toán xác định gốc từ của Martin Porter (Martin Porter's Stemming Algorithm). Một người đi đầu trong kỹ thuật rút trích thông tin – Giáo sư C. J. (Keith) van Rijsbergen – đã phát triển nó trong những đề án nghiên cứu của ông ta trong vài thập niên gần đây.

Gốc từ cũng làm giảm đi kích thước của một tập tin được chuyển hóa.

4.1.3.2.5. Gán trọng số (Weighting):

Gán trọng số (Weighting) là bước cuối cùng trong phần lớn các ứng dụng rút trích thông tin. Những từ được gán trọng số theo một mô hình trọng số được đưa ra mà trong mô hình đó có thể bao gồm trọng số cục bộ, trọng số toàn cục hay cả hai. Nếu trọng số cục bộ được dùng đến, thì trọng số từ được biểu diễn một cách bình thường như tần suất từ (tf). Nếu trọng số từ toàn cục được dùng đến, trọng số của một từ được đưa ra bởi các giá trị IDF. Nhưng trong phần lớn trường hợp (có cả trường hợp cơ bản) lược đồ trọng số là sự phối hợp giữa trọng số cục bộ và trọng số toàn cục được dùng đến (trọng số của một từ = tf*IDF). Điều này thường được xem như trọng số tf*IDF.

Sau đây là minh họa năm bước trong quá trình tạo chỉ mục tài liệu, được biểu thị bằng các hình:

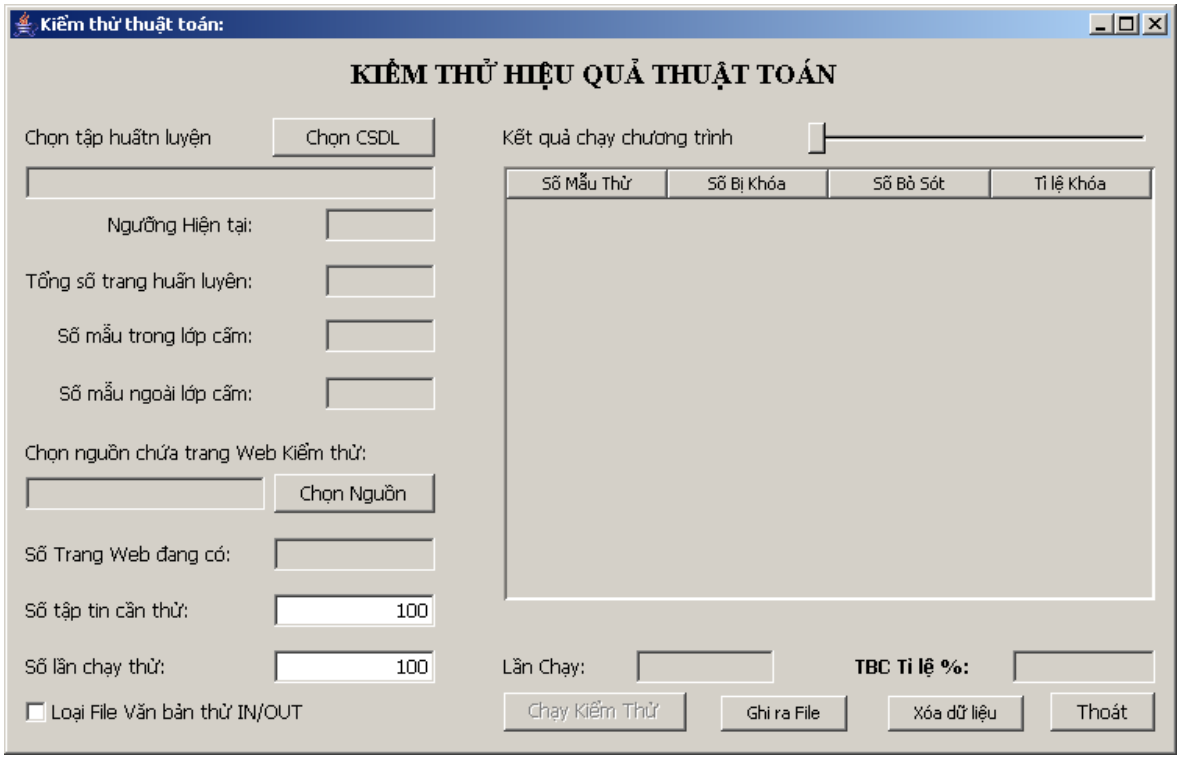


Hình 4.2. Tuyến tính hóa tài liệu bao gồm loại bỏ đánh dấu (a) và thẻ hóa (b). Tokenization is followed by stopwords tách lọc (c), stemming (d) gán trọng số (e). (Bản quyền Dr Edel Gracia [13])

4.2. Mô hình thử nghiệm:

4.2.1. Thử nghiệm theo ứng dụng: (Kiểm tra hoạt động của thuật toán)

4.2.1.1. Giao diện kiểm thử hiệu quả bộ lọc:



Hình 4.3. Màn hình ứng dụng thử nghiệm trên giao diện

4.2.1.2. Tổ chức dữ liệu và vận hành:

– Tập tích cực (tập huấn luyện):

Gồm các trang cấm, dùng để so sánh. Tập này được cập nhật thường xuyên. Dữ liệu có thể lưu ở dạng trang Web trong một thư mục hay có thể tổ chức thành các tập tin chứa vector đặc trưng (từ, tần suất từ) giúp cho việc so sánh nhanh hơn và không cần khởi động tập tích cực mỗi khi khởi động chương trình. Các trang này chỉ giữ các từ thuộc lĩnh vực thử nghiệm (từ đặc trưng của lĩnh vực lọc web).

Tập mẫu thử (sample): gồm các các trang web bên trong và bên ngoài lớp cấm. Các trang web nằm trong lớp cấm được người quản trị sưu tầm và phân loại chính xác theo chủ đề mà chương trình làm việc. Tổ chức bên trong của tập mẫu thử này giống như tổ chức tập huấn luyện, các tập tin được tổ chức thành các vector.

Điểm khác biệt duy nhất trong tập thử này là có thêm thuộc tính phân biệt (dùng để cho biết vector bên trong hay bên ngoài lớp cấm).

– *URL hay IP cấm: (Black list, IP)*

Chứa các URL hay IP cấm không cho truy xuất. Yếu tố này được dùng để tăng tốc cho hệ thống. Việc đầu tiên chương trình thực hiện là so các URL hay IP được gởi lên có tồn tại trong URL hay IP cấm không, nếu có thì cấm ngay không cần xét đến thuật toán. Nếu không chuyển sang xét thuật toán. Nếu khi xét bằng thuật toán kết quả cho là tích cực (tức là trang bị cấm) thì vector trang cấm đó được bổ sung vào tập tích cực và URL hay IP đó được đưa vào URL hay IP cấm (tự cập nhật Blacklist và IP).

* Một cơ sở dữ liệu được dùng kèm theo (file Access, text) ghi nhận các từ và tần suất từ của các tập tin trong tập tích cực. Với một trang cấm mới: các từ và tần suất từ chưa có trong trong CSDL sẽ được bổ sung vào.

– *Kết quả chạy chương trình:*

Cho hiển thị các thông tin trong quá trình kiểm tra.

– *Mô tả hoạt động:*

Đây là chương trình chạy dưới dạng ứng dụng, các bước thực hiện như sau:

– Khởi động chương trình: đọc thư mục chứa các tập huấn luyện liệt kê và danh sách "tập huấn luyện" L_h.

– Khi người dùng chọn một tập huấn luyện trong danh sách L_h thì thông tin đang có trong L_h sẽ xuất hiện trong mục "Thông tin hệ thống". Trong đó một thông tin quan trọng và phải tính toán để có được là ngưỡng hệ thống τ.

– Kiểm thử với một thư mục chứa các trang web bằng cách chọn thư mục nguồn hay gõ đường dẫn vào khung nhập chữ bên dưới (tùy vào thể loại cấm hay không cấm). Trong các thư mục chứa nguồn phải bảo đảm có đủ số lượng trang để hệ thống phát sinh ngẫu nhiên theo số tài liệu phát sinh định trước. Chọn số lượng tài liệu phát sinh bằng cách gõ vào khung chữ bên cạnh.

– "Chạy chương trình" là nút lệnh thực hiện chương trình. Kết quả trả về (cho một loại tài liệu thử, ví dụ: trang web cấm):

- ❒ Số lượng tài liệu chặn được, số lượng tài liệu cho qua (bỏ sót)
- ❒ Danh sách hệ số tương đồng của n tài liệu đem thử

– Chạy thuật toán:

B1. Bộ phát sinh ngẫu nhiên cho ra số thứ tự i của tập tin P_i được chọn, đọc nội dung, tính tần suất từ → hình thành vector VP_i văn bản cho P_i.

B2. Tính $\cos(VP_i,X_j)$ trong đó X_j là một trang trong tập tích cực so với ngưỡng. Tính hệ số trang $SigmaP_i$

B3. So sánh $SigmaP_i$ với τ để đưa ra quyết định. Trong quy trình kiểm thử ta thực hiện hai công việc sau:

Nếu $SigmaP_i >= \tau$ thì:

tăng biến đếm TrenNguong lên 1 đơn vị

ngược lại

tăng biến đếm DuoiNguong lên 1 đơn vị

Đưa giá trị $SigmaP_i$ vào danh sách lưu trữ (dùng cho việc vẽ lưu đồ)

4.2.2. Thử nghiệm trên mạng:

4.2.2.1. Dịch vụ mạng Proxy:

4.2.2.1.1. PROXY là gì?

PROXY trong tiếng Anh nó có nghĩa là "người được ủy nhiệm, ủy quyền". Đây là một chương trình quản lý server, nhằm tăng cường tính bảo mật và hiệu quả của server.

Proxy có chức năng của một tường lửa (firewall), nhưng có thêm tiện ích sử dụng bộ đệm (cache) để lưu trữ data. Proxy hoạt động như một cổng (gateway) với khả năng bảo mật cấp tường lửa giữa mạng LAN và Internet. Nó sẽ ngăn chặn việc

người dùng net truy cập tới các địa chỉ "nhạy cảm". Bởi vậy, khi kết nối với một proxy server (như Cinet), bạn phải chấp nhận việc truy cập vào các website dưới quyền kiểm soát của proxy

4.2.2.1.2. Một số ưu thế của Proxy Server:

- ❒ Bảo mật bức tường lửa (Firewall Security): Bảo vệ mạng nội bộ trong khi cho phép kết nối với Internet với chế độ truy nhập thời gian thực.
- ❒ Lưu giữ dữ liệu (High Performance Caching): Nhờ dùng cache để lưu giữ data nên sẽ giúp tăng tốc độ truy nhập Internet.
- ❒ Quản lý và kiểm soát: Ngăn chặn việc truy nhập vào các website mà ban quản trị ISP không muốn các khách hàng của mình truy cập tới. Hỗ trợ các công cụ quản lý được tập trung hóa và không đắt tiền.

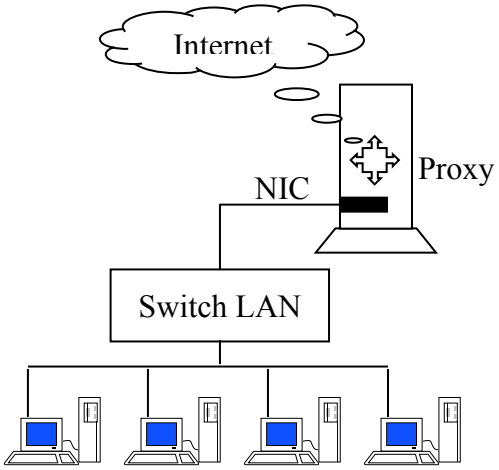
Tùy theo cấu hình proxy của từng server mà nó cho phép người sử dụng dùng những tiện ích thứ ba nào khác browser:

- Web Proxy thì cho phép xài các protocol: HTTP, FTP, Gopher, & SSL

- WinSock Proxy: HTTP, FTP, Telnet, RealAudio, VDOLive,... (bất kỳ ứng dụng nào tương thích WinSock 1.1)

- SOCKS Proxy: HTTP, FTP, Telnet (chỉ dùng chuẩn TCP), các ứng dụng tương thích SOCKS.

4.2.2.2. Hình vẽ hệ thống mạng thử nghiệm:



Hình 4.4. Sơ đồ mạng thử nghiệm cài đặt proxy

4.2.2.3. Hoạt động:

– Khi một máy bên trong gởi yêu cầu truy xuất qua một địa chỉ URL thì thành phần tường lửa kiểm soát ngõ ra chặn lấy yêu cầu (URL) kiểm tra trong Blacklist hay IP xem sự tồn tại của URL hay IP trên. Nếu có thì trả ngay kết quả là trang bị khóa không cho truy xuất. Nếu không thì lấy trang theo IP về và xử lý bằng thuật toán lọc web theo nội dung, xử lý để đưa ra quyết định có cho trang đó đi qua không?

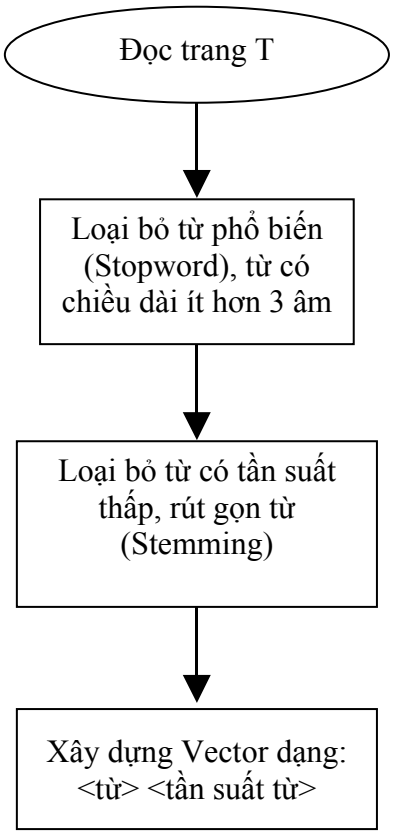
– Khi có một truy xuất từ ngoài vào mạng bên trong thành phần lọc web sẽ chặn lấy các gói "ráp" thành tập tin rồi tiến hành dùng thuật toán để xử lý cho kết quả là truy xuất đó có được phép đi vào mạng bên trong không?

4.3. Giải thuật cải tiến và lưu đồ:

4.3.1. Lưu đồ từng bước:

4.3.1.1. Lưu đồ và mô tả cải tiến bước 1

Đây là lưu đồ tạo vector tập huấn luyện đồng thời cũng là quy trình tạo vector cho trang P trong quá trình xét duyệt.



Hình 4.5. Lưu đồ bước 1

Cải tiến trong bước này:

– Xây dựng thư viện các từ khóa cho phân lớp (từ chuyên ngành trong lĩnh vực)

– Trong quá trình tính tần suất từ dựa trên những từ trong thư viện này. Thời gian sẽ rút ngắn lại do phạm vi của các từ được giới hạn.

– Việc lưu trữ các phần tử trong tập huấn luyện T_s lưu theo dạng vector (<từ>, <tần suất từ>) vì thế lần sau bỏ qua giai đoạn tính vector tần suất từ.

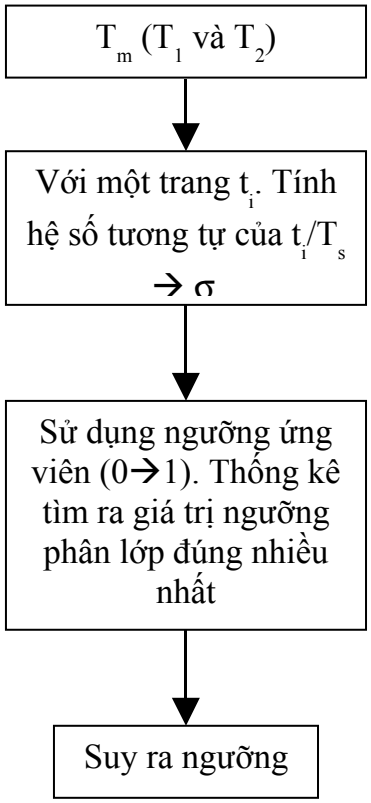
Sử dụng CSDL Access để lưu tập huấn luyện:

– Tạo một CSDL Access gồm các các bảng với mối quan hệ đủ để diễn giải cho một tài liệu và biểu diễn được vector của tài liệu.

(Xem mô hình quan hệ và tổ chức CSDL trình bày ở mục 4.1.2.1)

– Với mô hình này thì việc khởi tạo và bổ sung sẽ nhanh hơn và không phải duy trì một tập huấn luyện dạng file nữa.!!!

4.3.1.2. Lưu đồ và mô tả cải tiến bước 2



Hình 4.6. Lưu đồ bước 2

Tập mẫu thử T_m chứa các trang đã được phân loại đúng, các trang này nằm bên trong và bên ngoài lớp cấm (tức là những trang cấm không cho phép truy cập và những trang cho phép truy cập).

* Tổ chức dữ liệu cho T_m:

Dữ liệu của tập thử được lưu trữ dưới dạng vector tần suất và được xác định rõ lớp văn bản mà nó được phân loại chính xác cũng như xác định nó có thuộc lớp cấm không?

* Một số xử lý trong bước 2:

i) Tính hệ số tương tự từng thành viên trong T_m đối với T_s. Thao tác tính toán này dùng lại hàm tính hệ số tương tự ở bước 3. Cấu trúc dữ liệu dùng để lưu trữ các giá trị tương đồng là một mảng C.

ii) Xây dựng một dãy ứng viên với sai số định trước. Đem từng ứng viên so với hệ số phân loại trong mảng C. Để xác định ứng viên nào phân loại đúng nhiều nhất chọn làm ngưỡng cho hệ thống.

* Lưu ý:

– Bước 2 này được cài đặt sau khi đã cài đặt bước 3, vì trong bước 2 có dùng đến hàm tính hệ số trang trong bước 3.

– Tập mẫu thử bố trí trong cơ sở dữ liệu của tập huấn luyện, nên chỉ cần khởi tạo kết nối cơ sở dữ liệu một lần.

– Trong quá trình tính hệ số trang ta chọn ra hệ số trang SigmaMin và SigmaMax để dùng cho việc duyệt mảng ứng viên giới hạn từ SigmaMin đến SigmaMax thay vì từ 0 đến 1. Như thế, số lượng ứng viên sẽ giảm đi rất nhiều. Trường hợp xấu nhất SigmaMin = 0 và SigmaMax = 1. Tuy nhiên, vẫn có nhiều trường hợp SigmaMin > 0 và SigmaMax < 1 hơn là trường hợp đặc biệt kể trên.

4.3.1.3. Lưu đồ và mô tả cải tiến bước 3:

– Bước này thực hiện sau khi hoàn tất bước 1, tức là đã hình thành được vector tài liệu (Từ, tần suất) và đã tính xong ngưỡng τ cho hệ thống.

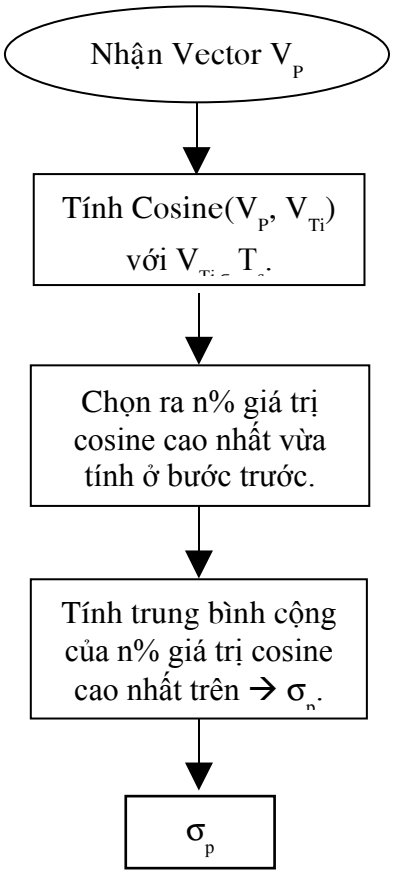
– Công việc tính hệ số trang này còn dùng cho bước 2, vì trong bước 2 dùng bước 3 để tính hệ số trang từng thành viên trong mẫu thử T_m so với tập huấn luyện T_s để tìm ra ngưỡng cho hệ thống.

– Các bước thực hiện như sau:

Đầu vào: Tập huấn luyện T_s, Vector của trang web là V_p:

Đầu ra: hệ số trang P là σ_p.

– Sơ đồ như sau:



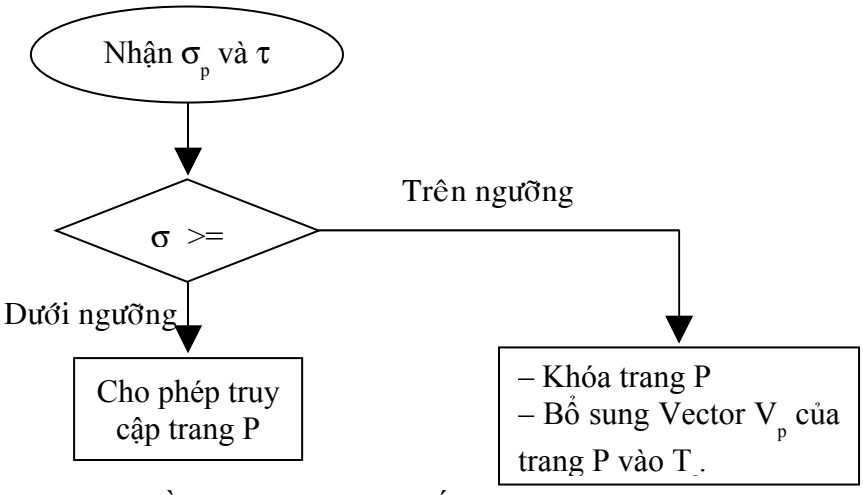
Hình 4.7. Lưu đồ bước 3

* Một số xử lý trong bước 3:

– Tính cosine theo công thức $\cos(X,Y) = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$, trong đó X và Y là hai vector của hai văn bản P_X và P_Y. Hai vector này có thứ tự các từ được sắp xếp theo đúng thứ tự tương ứng với nhau. Trong khi tính cos, các giá trị này được lưu vào trong một mảng C. Đến cuối cùng, mảng C này được sắp xếp theo thứ tự giảm dần.

– Từ mảng C tính trung bình cộng của n% phần tử đầu tiên → σ_p. Trong đó n% có được xác định bằng cách dựa vào số phân lớp con trong tập huấn luyện. Ví dụ: một phân lớp có 2 lớp con thì n% = 50%.

* Xử lý đưa ra quyết định cho toàn hệ thống: giai đoạn này chỉ là bước so sánh giá trị σ_p với ngưỡng τ đã tính ở bước 2. Lưu đồ như sau:



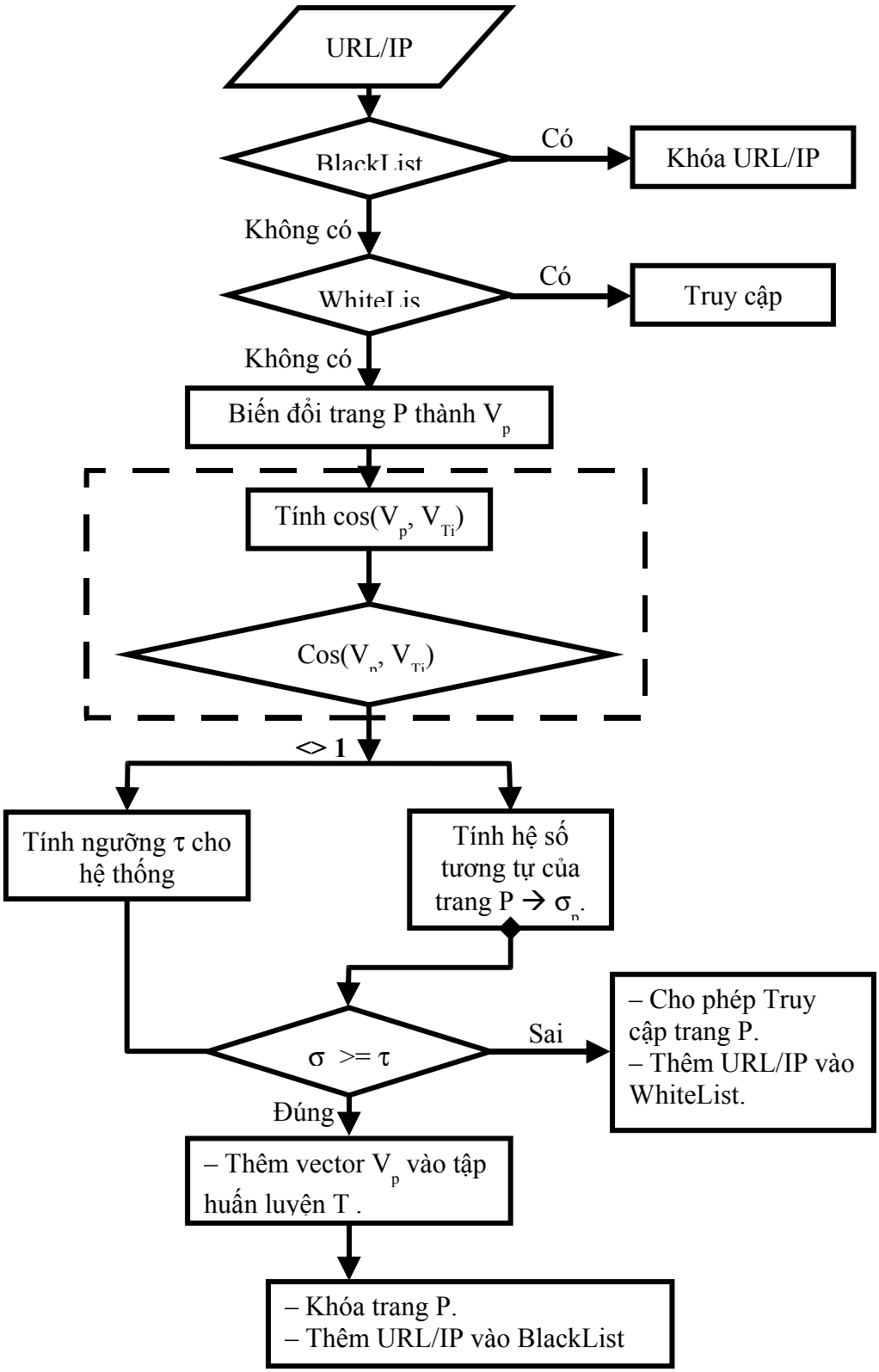
Hình 4.8. Lưu đồ bước so sánh hệ số trang và ngưỡng

Kết quả:

– Cho phép truy cập trang P: quá trình xem xét của hệ thống đưa ra kết luận là địa chỉ URL/IP mà client yêu cầu không chứa những tiềm ẩn nguy hiểm.

– Không cho phép truy cập: là trang web có tiềm ẩn nguy hiểm, trang web này sẽ được bổ sung vào tập huấn luyện dưới dạng Vector, đồng thời địa chỉ IP hay URL mà client yêu cầu sẽ được bổ sung vào blacklist. Như vậy, cơ sở dữ liệu cho hệ thống sẽ được bổ sung thêm tự động trong quá trình vận hành hệ thống.

4.3.2. Các lưu đồ cho thuật toán:



Hình 4.9. Lưu đồ cài đặt ứng dụng

4.4. Cài đặt:

4.4.1. Cài đặt Proxy:

Trong bài này người viết chọn một Proxy nguồn mở, tải về từ địa chỉ http://jerome.jouvie.free.fr/ hay (http://topresult.tomato.co.uk/~jerome/), tác giả Jérôme JOUVIE, bản quyền 2004 – 2006. Phiên bản gốc được tác giả của chương trình viết Proxy này (Ông Jérôme JOUVIE) xây dựng thành dịch vụ lọc IP.

Sau khi nghiên cứu bản gốc, người viết luận văn này chỉnh sửa thành bộ lọc nội dung bằng cách:

– Biến đổi bộ phận tiếp nhận yêu cầu Client theo dạng IP thành tiếp nhận các URL hay IP gởi lên từ Client.

– Thêm thuật toán so sánh văn bản: sau khi bắt được URL hay IP từ dưới Client gởi lên Proxy sẽ tải trang web (nếu có) về so sánh với tập huấn luyện và đưa ra quyết định có cho phép người dùng truy cập đến trang web đó không.

* Yêu cầu:

– Đảm bảo chức năng của một dịch vụ mạng (proxy)

– Ngăn chặn được các trang web cấm

– Bổ sung được tư liệu cho tập huấn luyện

– Không làm trì trệ mạng

4.4.2. Mô tả chi tiết các bước thuật toán:

Bước 1: Biến đổi tài liệu thành vector:

* Mô tả các công đoạn:

1. Dữ liệu được chuyển từ proxy qua là một tập tin văn bản html, nội dung của file này là các tag định dạng HTML và nội dung thể hiện trên trang web. Xây dựng hàm xử lý loại bỏ các tag HTML. Kết quả sau cùng là một file văn bản chỉ còn lại nội dung chính của trang Web.

2. Sau khi nhận tập tin phần xử lý tiếp theo là lọc bỏ các từ thuộc stopword, vì các từ này không có ảnh hưởng đến nội dung chính của văn bản. (các từ như the, to, for, sine, …)

3. Văn bản sau khi loại bỏ các từ dạng stopword còn lại các từ chính, trong đó có những từ được biến dạng theo các thì tiếng Anh như: quá khứ, tiếp diễn, danh từ, … do đó ta dùng phương pháp suy từ gốc (Stemming) để thực hiện rút ngắn các từ để giảm số chiều của vector văn bản xuống thấp hơn nhằm đẩy nhanh tốc độ tính toán.

4. Vector hóa văn bản: xây dựng cấu trúc dữ liệu cho vector và chọn cấu trúc lưu trữ. Mỗi vector được lưu trữ dưới dạng mảng cấu trúc gồm 2 thành phần là: từ cần lưu, số lần từ đó xuất hiện trong văn bản gọi là tần suất từ. Như đã trình bày ở phần Vector thì mỗi tài liệu là một vector nhiều chiều, mỗi từ của nó đặc trưng cho một chiều của hệ trục đa chiều.

* Thuật toán:

i1. Bỏ các tag HTML:

– Đọc file html

– Duyệt qua các tag: đặc điểm của các tag là bắt đầu bằng dấu mở ngoặc nhọn "<" và kết thúc bằng dấu đóng ngoặc nhọn ">" giữa hai dấu này là từ khóa của loại tag đó (có 2 loại mở tag và đóng tag có dấu "/"). Chạy từ dấu "<" đến dấu đóng ">" xác định vị trí đầu và vị trí cuối, thay thế nó bằng khoảng trắng. Nội dung của trang web còn giữ lại, lưu thành tập tin văn bản F1.txt, chuyển qua công đoạn tiếp theo.

i2. Bỏ các từ Stopword:

– Xây dựng cấu trúc dữ liệu chứa các từ thuộc stopword, và phương thức tìm kiếm trong cấu trúc này (phục vụ cho việc xác định một từ có thuộc về nó không?). Đọc tập tin stopword.txt (lấy từ mạng internet tại địa chỉ www.stoplist.com: hội thống kê ngôn ngữ) nạp vào cấu trúc đã xây dựng trên.

– Đọc tập tin văn bản, duyệt qua từng từ, kiểm tra xem nó có nằm trong stopword không, nếu có thì loại bỏ thay bằng khoảng trắng.

– Còn lại các từ không phải stopword, lưu tập tin F1.txt lại để chuyển qua bước xử lý tìm từ gốc.

i3. Tìm từ gốc (stemming):

– Dùng thuật toán steming lấy tại địa chỉ http://www.comp.lancs.ac.uk/computing/research/stemming/Links/implementations.htm của tác giả Chris O'Neill cài đặt.

– Sau khi chạy thuật toán này với tập tin F1.txt ta sẽ được tập tin F1.txt với các từ được rút gọn.

i4. Thống kê, tạo vector của tập tin văn bản F1.txt:

– Xây dựng cấu trúc dữ liệu cho vector: dạng mảng động và mỗi phần tử giữ một cấu trúc lưu trữ gồm có hai trường <word, tần suất>.

– Đọc tập tin F1.txt, lấy từng từ tìm trong mảng vector: nếu có tăng số lượng (tần suất) lên 1. Ngược lại, tạo một phần tử mảng mới với dữ liệu là từ đó và tần suất bắt đầu bằng 1.

– Kết quả là một vector tần suất, là vector dùng để tính toán ở các bước sau.

Bước 2: Tính ngưỡng cho hệ thống

Gồm các giai đoạn sau:

i1. Xây dựng một tập dữ liệu T_x (tập mẫu thử) mới gồm có các vector của tài liệu cho phép và các tài liệu bị cấm: Trong CSDL sử dụng Table Vector_Sample, trong bảng này chứa các tài liệu được lưu ở dạng vector, với cấu trúc <DocNo, Term, TF, Side>. (Side là thuộc tính cho biết tài liệu đó có trạng thái là trong hay ngoài lớp cấm.

i2. Lấy từng tài liệu trong T_x (đặt tên tài liệu này là P_x) để tính hệ số tương tự $SigmaP_x$ của văn bản P_x so với T_s. Vì trong T_x có nhiều tài liệu nên sẽ tạo thành một dãy các hệ số tương tự. (D_{Sigma}).

i3. Xây dựng mảng các ứng viên ngưỡng $X_{threshold}$ với các giá trị từ 0 đến 1, chọn sai số cho mỗi $SigmaP_x$ là 1/10000. Lấy từng phần tử trong D_{Sigma} đem so với từng phần tử trong mảng ứng viên $X_{threshold}$. Chọn ra giá trị trong D_{Sigma} xuất hiện nhiều nhất, chọn làm ngưỡng của hệ thống.

* Thuật toán:

B1: Chọn ra danh mục văn bản trong tập T_x → nhận được danh sách L.

B2: Với mỗi L_i chọn ra Vector tài liệu V_i. Đem V_i tính hệ số tương tự (SigmaV) so với tập T_s cho ra mảng Sigma[i].

B3: Thống kê tính ra ngưỡng cho hệ thống: là giá trị Sigma[x] có sự phân loại đúng cho nhiều tài liệu nhất.

* Xử lý:

Xây dựng danh sách L (tên tài liệu)

Với mỗi tài liệu trong L

{

Chọn ra vector V_i trong T_x;

Tính Sigma(V_i, T_s) → s_p;

Gán vào mảng S[i] ← s_p;

}

Chọn một giá trị S[m] là giá trị phân loại chính xác nhiều tài liệu nhất.

Quá trình chọn một giá trị ngưỡng: dùng một dãy các giá trị ứng viên ngưỡng từ 0 đến 1. Tùy vào việc chọn sai số cho hệ thống mà thời gian tính toán để chọn lựa ứng viên ngưỡng thích hợp cho hệ thống.

Bước 3: Quy trình xét trang P

Trang P được biến đổi thành vector V_p ở bước 1 chuyển qua, dùng V_p đó để thực hiện tính toán tìm ra hệ số tương tự.

* Mô tả các công đoạn:

i1. Tính cosine(V_p, V_{Ts}) kết quả trả về một mảng các giá trị cos và số hiệu văn bản.

– Cấu trúc lưu trữ kết quả sự tương tự của V_p đối với từng thành viên trong T_s: là một mảng cấu trúc với 2 thuộc tính:

- ❒ cos: lưu giá trị cosine(V_p, V_{Ts})
- ❒ Mã số văn bản: là số hiệu văn bản (dùng cho công đoạn tiếp theo).

i2. Chọn n% giá trị tương tự cao nhất từ mảng kết quả đã tính ở công đoạn 1, trong đó n% được xác định dựa trên số phân lớp phụ của văn bản chứa trong tập huấn luyện.

i3. Tính hệ số lớp của trang P (σ_p): dựa vào danh sách n% giá trị tương tự cao nhất đã tạo ở công đoạn 2. Tính σ_p theo công thức (*) gồm các công đoạn sau:

- ❒ Tính tổng giá trị cosine của n% giá trị cao nhất → Tổng S
- ❒ Tính số tài liệu theo n% (số lượng) → Tổng k
- ❒ Tính trung bình cộng σ_p tức là hệ số lớp của P: lấy S/k.

i4. Đem σ_p mới tính công đoạn trên so sánh với ngưỡng của hệ thống:

– Nếu σ_p < ngưỡng τ thì cho trang web P qua để đến người dùng.

– Ngược lại, báo kết quả cấm truy cập và bổ sung trang V_p vào tập huấn luyện.

* Thuật toán:

i1. Tính cosine:

– Dữ liệu đưa vào là vector V_p đã được tạo ra từ bước 1.

– Sử dụng cấu trúc vector <Word, Tần suất> để lưu trữ từng vector văn bản trong T_s. Một mảng C các giá trị cosine để lưu trữ cosine của từng cặp V_P và V_Y

– Duyệt qua T_s để chọn ra toàn bộ các tài liệu đang có trong đó, hình thành danh sách các tài liệu L.

– Với từng tài liệu trong L chọn ra vector V_Y.

– Sử dụng một cấu trúc mảng lưu thông tin gồm có 3 yếu tố: Từ, Tần suất P, tần suất Y. Dùng phép toán hội để tạo ra thông tin cho mảng này từ V_p và V_Y. Sau đó tính cosine theo công thức:

$$\cos ine(V_p, V_Y) = \frac{\sum P_i * Y_i}{\sqrt{\sum P_i^2 * \sum Y_i^2}}$$

lưu vào mảng C.

i2. Chọn n% giá trị cosine cao nhất từ mảng C:

– Sắp xếp mảng C theo giá trị giảm dần

– Tạo một mảng C’ tương tự cấu trúc C

– Duyệt qua n% (so với tổng số phần tử C) chọn ra các giá trị cao đến thấp và sao chép qua C’. Như vậy trong C’ chứa các giá trị v là giá trị cosine đã tính ở bước trước như được xếp từ cao đến thấp.

i3. Tính hệ số lớp của P (σ_p): là tính trung bình cộng của mảng C’.

– Dựa trên mảng C’ tính tổng của n% giá trị cosine cao nhất:

– Lắp ghép vào công thức → $\sigma_p = \frac{\sum_{v \in C'} v}{|T_s| \ x \ n\%}$

i4. So sánh (σ_p) với ngưỡng τ:

– Đây là phép so sánh 2 số double nên dùng

if (σ_p) < τ then

Trả kết quả về cho client (cho phép truy cập đến URL)

Else

{

Thông báo trang web cấm truy cập đến với client

Ghi vector V_P vào T_s.

Duyệt qua mảng V_P, lấy từng phần tử ghi vào CSDL

}

4.4.3. Mã chương trình:

Chương trình được lập trình bằng ngôn ngữ Java với tập tin cơ sở dữ liệu TrainingSetX.mdb chứa số liệu các vector của tập huấn luyện và tập mẫu thử cũng như các dữ liệu khác cho chương trình như blacklist, whitelist.

4.4.4. Một số cải tiến trong chương trình:

– Cải tiến về cách lưu trữ dữ liệu học: gồm của tập huấn luyện, tập mẫu thử và thông tin hiện tại của hệ thống. Cải tiến này nhằm làm giảm thời gian đọc và biến đổi nội dung một trang web về dạng vector (càng nhiều trang web huấn luyện thì thời gian giải quyết càng lâu).

– Theo thuật toán nguyên thủy thì có phân tích ngẫu nhiên các link có trong trang web đang xét. Việc này dẫn đến vấn đề phải giải quyết là độ sâu của một phép đệ quy, vì các trang web ngày nay có rất nhiều liên kết nên sẽ tốn nhiều thời gian để đi đến một kết luận. Do đó, trong quá trình xây dựng chương trình người viết đã bỏ qua giai đoạn này. Khi một trang web hiển thị trên màn hình, người dùng click vào một link thì địa chỉ URL của nó sẽ gởi đến proxy, công việc xét tiếp theo là của proxy thực hiện kiểm duyệt như một trang web thông thường.

– Cải tiến trong xét chọn ứng viên làm ngưỡng: có hai vấn đề cải tiến

- Trong thuật toán nguyên thủy đề nghị chọn dãy ngưỡng từ 0 đến 1, nếu như độ phân giải 10^{-4} thì sẽ có 10,000 ứng viên đem xét, có những ứng

viên không có cơ hội để chọn làm ngưỡng vì nó nằm ngoài giới hạn trên và giới hạn dưới của dãy các hệ số trang σ_p. Để rút ngắn dãy số lượng ứng viên làm như sau: Trong quá trình tính hệ số trang của các phần tử trong mẫu thử (T'_s) so với tập huấn luyện (T_s) ta chọn giá trị nhỏ nhất (MinSig) và lớn nhất (MaxSig) của dãy hệ số trang. Sau đó, trong quá trình tìm ngưỡng ta chọn các ứng viên từ MinSig đến MaxSig. Trường hợp xấu nhất khi MinSig = 0 và MaxSig = 1 trở về trường hợp được đề xuất trong thuật toán nguyên thủy.

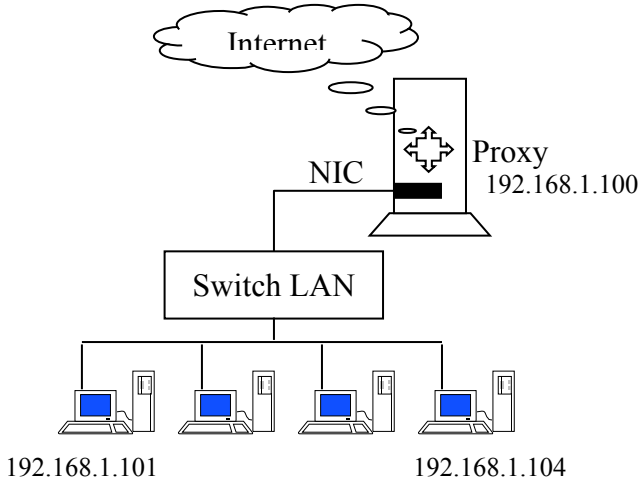
- Quá trình xét một giá trị ứng viên chọn làm ngưỡng bằng cách xác định ứng viên đó phân loại đúng bao nhiêu tài liệu (bên trong và bên ngoài lớp cấm), chọn ra giá trị ứng viên nào phân loại đúng nhiều nhất làm ngưỡng hệ thống. Một cải tiến nhằm làm tăng tốc độ làm việc là trong lúc tính hệ số tương đồng sắp xếp các giá trị σ_p giảm dần. Xây dựng hàm tìm kiếm tương đối để tìm vị trí của một giá trị ngưỡng trong dãy hệ số tương đồng từ vị trí mới tìm đó xác định được số lượng tài liệu có hệ số tương đồng nằm trên và nằm dưới giá trị ứng viên. So sánh số tài liệu vừa xác định đó với số lượng tài liệu thật hiện có và tìm ra giá trị nào gần đúng nhất để suy ra ứng viên đó là ngưỡng.

4.5. Thử nghiệm và đánh giá:

4.5.1. Môi trường thử nghiệm – cấu hình các dịch vụ:

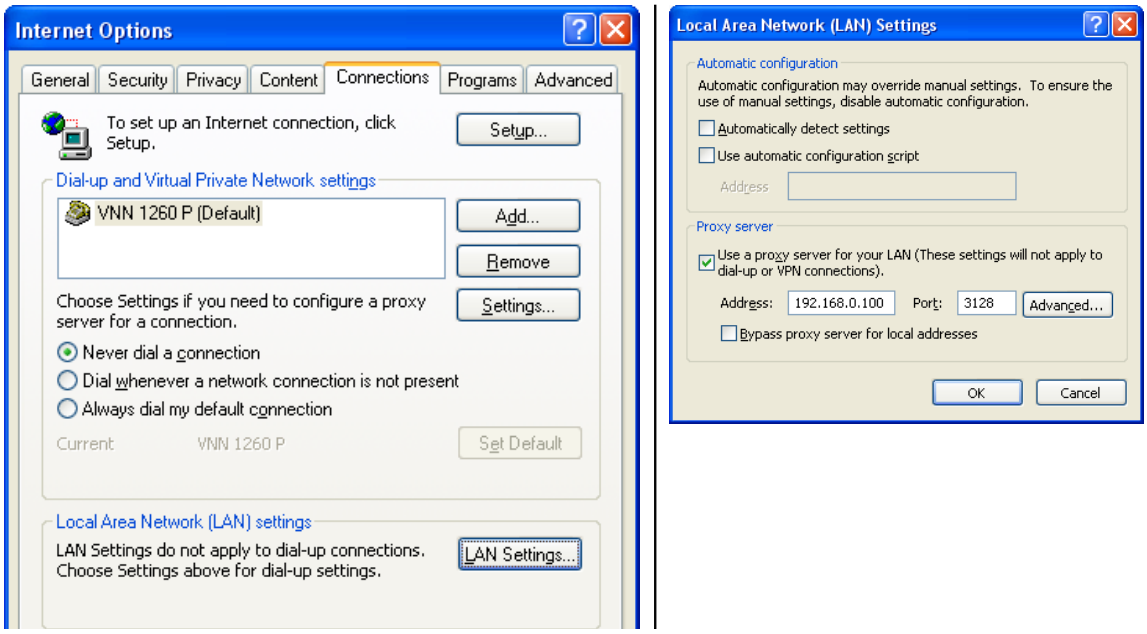
– Môi trường thử nghiệm: proxy được viết bằng ngôn ngữ Java và được biên dịch thành mã máy chạy trong môi trường Windows. Khi khởi động nó sẽ là một dịch vụ chạy ẩn bên dưới.

– Cấu hình dịch vụ: Vì đây là một proxy trong môi trường thử nghiệm nên các máy kết nối trong mạng phải điều chỉnh proxy để hướng luồng truy cập mạng thông qua proxy (khi cài đặt bộ lọc web trên firewall kiểm soát cổng ra vào mạng thì không cần làm thao tác này). Giả sử sơ đồ mạng LAN như sau:



Hình 4.10. Mô hình mạng LAN thử nghiệm bộ lọc proxy phân loại văn bản

Điều chỉnh như sau: Vào Internet Option, chọn Connections, LAN Settings…



Hình 4.11. Điều chỉnh địa chỉ proxy trên máy client

Thực hiện điều chỉnh proxy server cho toàn bộ các máy trong mạng. Như vậy, toàn bộ các máy trong mạng sẽ hướng đến proxy server trước khi nhận được "lệnh" truy cập vào một trang web.

4.5.2. Phương pháp thử nghiệm - Một số thử nghiệm:

4.5.2.1. Dữ liệu dùng cho thử nghiệm:

Tổ chức dữ liệu thử nghiệm gồm hai loại, chủ yếu tải từ website theo chủ đề kiểm thử (trang web có nội dung sex và tình cảm). Có 3 thành phần:

– Nguồn làm tập huấn luyện: dùng trong việc huấn luyện T_s, có 378 trang (lấy từ website **http://www.girl-directory.com/erotic-stories.php**)

– Nguồn làm tập mẫu thử dùng để huấn luyện tập sample gồm có 2 tập con:

- Tập thứ nhất (T'_1) là các trang web bên trong lớp cấm, những trang web này được phân loại chính xác thông qua con người, những trang này có cùng chủ đề.
- Tập thứ hai (T'_2) là các trang được phân loại chính xác là ngoài lớp cấm, những trang web này không cùng chủ đề với lớp cấm.

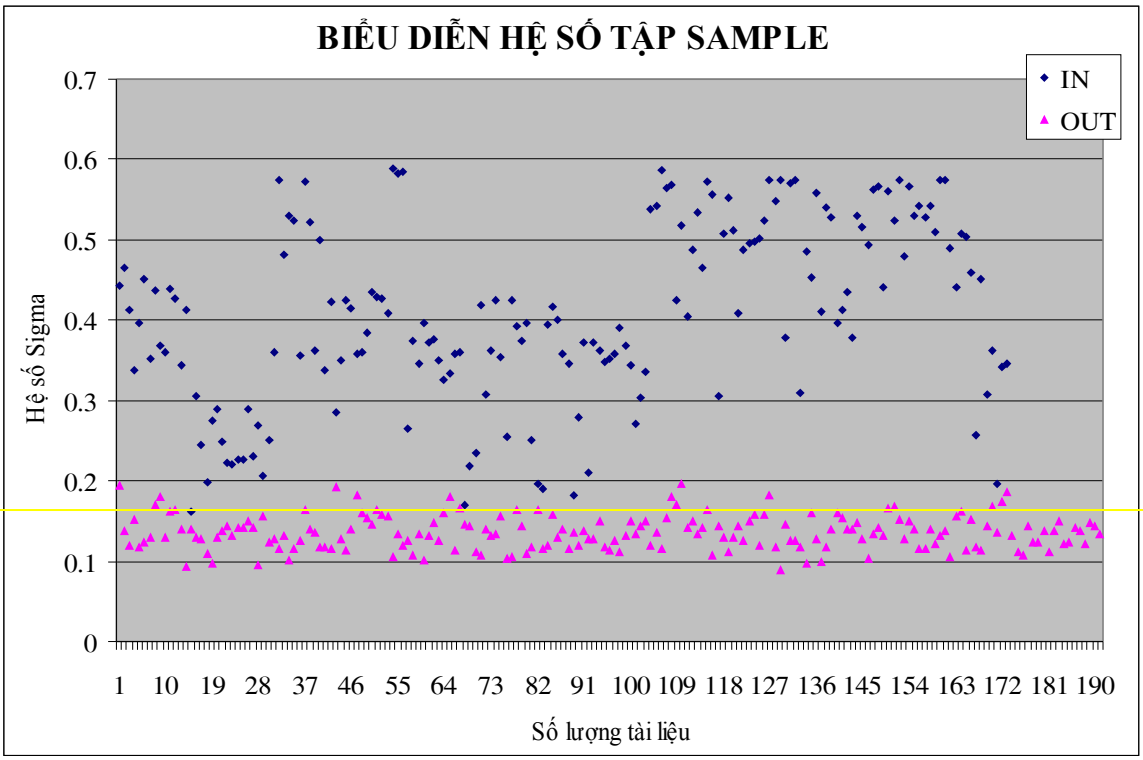
Cả hai tập con trên được lưu trữ trên máy dạng các trang web, các trang web này được tải về từ Internet T'_1 từ www.bondage.com, www.pinkflamengo.com, www.freesexstory.com. Thành phần T'2 là những trang web truyện tình cảm được tải về địa chỉ http://acacia.pair.com và nhiều trang khác được tìm kiếm bằng từ khóa "Love Story" hay "romance" hay "love-affair".

Tổng số lượng $T_s' = T'_1 + T'_2 = 173 + 191 = 364$ trang

4.5.2.2. Kiểm thử hệ số trang của T_s':

Sử dụng tập T_s' để tính ra ngưỡng giới hạn cho hệ thống theo thuật toán trong bước 2. Trong đó có giai đoạn tính hệ số tương tự của các phần tử trong T_s' để tạo ra tập hợp các giá trị dùng cho việc xác định ngưỡng.

Với một bộ dữ liệu thử T_s' ta có biểu đồ biểu diễn các giá trị như sau:



Hình 4.12. Biểu đồ biểu diễn các giá trị hệ số tương tự của T_s' so với T_s.

Biểu đồ trên (hình 4.12) cũng cho thấy việc lựa chọn ngưỡng là quan trọng như thế nào. Vì một giá trị ngưỡng nếu ta chọn quá cao sẽ bỏ sót rất nhiều tài liệu. Chẳng hạn theo hình 4.12 ta chọn ngưỡng $\tau = 0.3$ thì số lượng trang bị bỏ sót rất nhiều (hệ số tương tự của những trang thuộc lớp cấm biểu diễn bằng hình thoi). Nếu chọn ngưỡng $\tau = 1.5$ thì số trang cho phép truy cập bị chặn lại rất nhiều (hệ số tương tự của những trang không thuộc lớp cấm biểu diễn bằng hình tam giác).

Tập thử gồm có: 173 trang thuộc lớp cấm (IN) và 191 trang ngoài lớp cấm (OUT). Kết quả chạy chương trình:

Phân lớp	Số lượng	Phân loại	Tỉ lệ % sai số	Ngưỡng σ_P
IN	173	182	4.95%	0.16849
OUT	191	182	4.71%	

4.5.2.3. Thử nghiệm trên giao diện với σ_p:

Nguồn thử gồm 500 trang web có cả bên trong và bên ngoài lớp cấm, được tải từ website www.bondage.com và www.lovestory.com với số lượng xác định cho từng loại.

Phát sinh ngẫu nhiên 200 địa chỉ trang P_i lấy trong nguồn thử. Chạy chương trình và ghi nhận kết quả để tính toán theo 2 công thức sau:

* Tính tỉ lệ trang có nội dung cấm bị bỏ sót:

Gọi:

- M_{IN} tổng số trang lớp cấm đem thử
- N_{IN} tổng số trang lớp cấm bị khóa (chặn đúng)

$$\%block = \frac{N_{IN}}{M_{IN}}\%$$

* Tính tỉ lệ trang có nội dung trang không cấm bị bỏ sót:

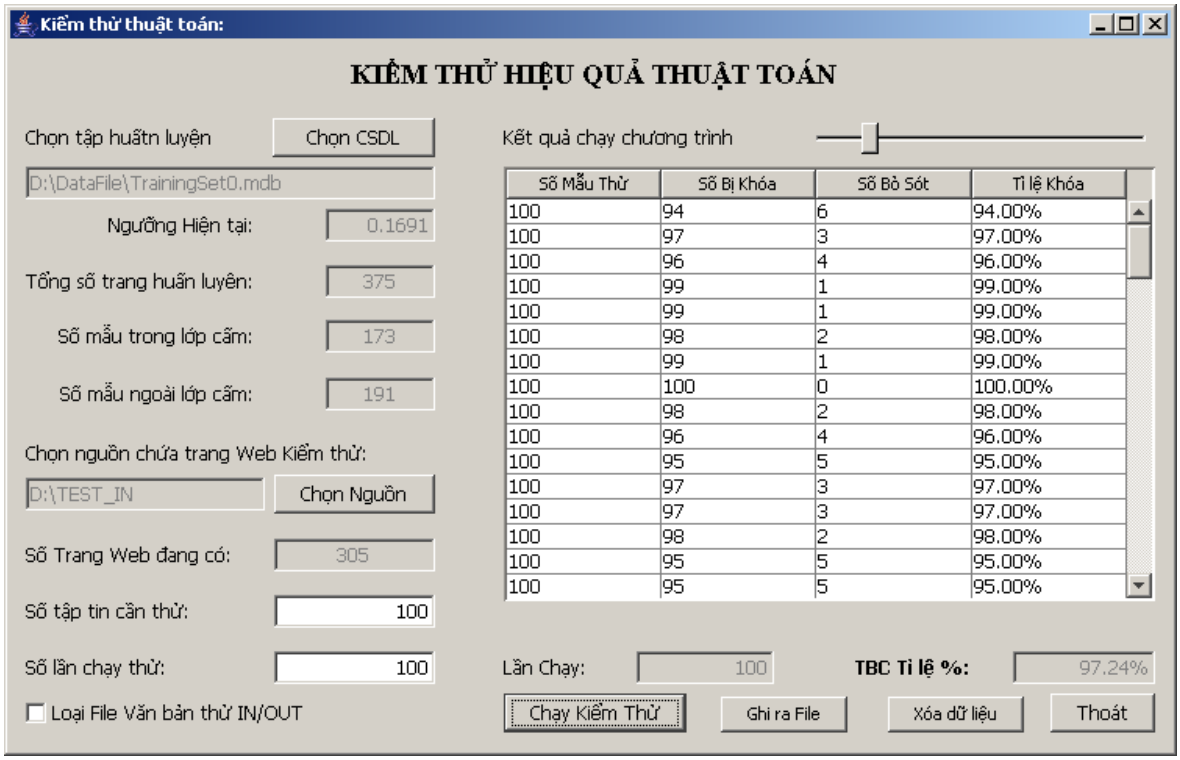
Gọi:

- P_{OUT}: Tổng trang ngoài lớp cấm đem thử
- Q_{OUT}: Tổng trang ngoài bị khóa (chặn sai)

$$\%overblock = \frac{Q_{OUT}}{P_{OUT}}\%$$

Thử nghiệm trên giao diện: sử dụng một tập các địa chỉ các trang web hay IP dùng thử nghiệm. Chọn ngẫu nhiên nhiều địa chỉ gởi đến proxy, nhận kết quả và thống kê.

Giao diện:



Hình 4. 13. Màn hình thử nghiệm thuật toán

Hình 4.13. Thực hiện kiểm thử 100 lần với bộ TEST_IN gồm những trang thuộc loại truyện Sex gồm 305 trang (truyện), tỉ lệ bình quân là 97.24% (sai số 2.76%), kết quả này có nguyên nhân ở cách dùng từ và lối hành văn của tác giả viết truyện. Những từ rất bình thường nhưng nó diễn tả cho lối hành văn thô tục.

Hoạt động:

– Khởi động: chọn tập huấn luyện, bấm nút Chọn… mở thư mục chứa các tập tin Access dùng lưu trữ thông tin đã huấn luyện. Đọc thông tin về tập huấn luyện, tập mẫu thử (Sample: số trang nằm trong lớp cấp, số trang ngoài lớp cấm), tính giá trị ngưỡng τ.

– Chọn nguồn chứa các trang Web kiểm thử: bấm nút Chọn Nguồn để mở cửa sổ chọn thư mục, có 2 thư mục: TEST_IN chứa trang web sex (lớp cấm) và thư mục TEST_OUT chứa các trang web chuyện tình cảm (Love Story: không thuộc lớp cấm).

– Cho biết số trang web cần thử vào khung văn bản.

– Bấm nút Chạy Kiểm Thử để thực hiện chương trình, quá trình thực hiện sẽ phát sinh ngẫu nhiên không trùng các trang Web có trong thư mục nguồn được chọn.

– Quy trình thực hiện với một trang P: tính bước 1 cho ra kết quả là một vector văn bản (vector tần suất). Chuyển sang bước 3 để tính hệ số trang P so với tập huấn luyện T_s cho kết quả σ_p. Đem so sánh với ngưỡng τ và kết luận.

Sau đây là bảng ghi nhận kết quả chạy thử: cho ngưỡng hệ thống là 0.1691

1. Đối với những trang thuộc lớp cấm: (10 mẫu đầu tiên)

Ngưỡng	0.1691				
Lần Chạy	**Số Mẫu Thử**	**Số Mẫu Bị Khóa**	**Số Bỏ Sót**	**Tỉ lệ khóa**	**Tỉ lệ bỏ sót**
1	100	98	2	98.00%	2.00%
2	100	99	1	99.00%	1.00%
3	100	98	2	98.00%	2.00%
4	100	99	1	99.00%	1.00%
5	100	98	2	98.00%	2.00%
6	100	97	3	97.00%	3.00%
7	100	97	3	97.00%	3.00%
8	100	97	3	97.00%	3.00%
9	100	98	2	98.00%	2.00%
10	100	99	1	99.00%	1.00%
...	...	...			
			Max	100.00%	5.00%
			Min	95.00%	0.00%
			AVG	97.71%	2.29%

Bảng 4.1: Kết quả 100 lần thử 100 trang chọn ngẫu nhiên thuộc lớp cấm

2. Đối với những trang không thuộc lớp cấm:

Ngưỡng	0.1691				
Lần Chạy	**Số Mẫu Thử**	**Số Mẫu Bị Khóa**	**Số Bỏ vượt qua**	**Tỉ lệ khóa**	**Tỉ lệ bỏ vượt qua**
1	100	9	91	9.00%	91.00%
2	100	23	77	23.00%	77.00%
3	100	17	83	17.00%	83.00%
4	100	18	82	18.00%	82.00%
5	100	18	82	18.00%	82.00%
6	100	23	77	23.00%	77.00%
7	100	19	81	19.00%	81.00%
8	100	17	83	17.00%	83.00%
9	100	21	79	21.00%	79.00%
10	100	16	84	16.00%	84.00%
...	...	...			
			Max	25.00%	91.00%
			Min	9.00%	75.00%
			AVG	17.62%	82.38%

Bảng 4.2: Kết quả 100 lần thử 100 trang chọn ngẫu nhiên không thuộc lớp cấm

4.5.2.3. Thử nghiệm trên mô hình mạng với σ_p:

Dùng kiểm tra một số địa chỉ web, xem kết quả trả về và thống kê. Trên mô hình này gồm có bộ phát sinh ngẫu nhiên để chọn các trang web, bộ xử lý, tính toán và cho kết quả của trang web đó, đem so với ngưỡng hệ thống.

Tổ chức kiểm thử: kỹ thuật và thực nghiệm:

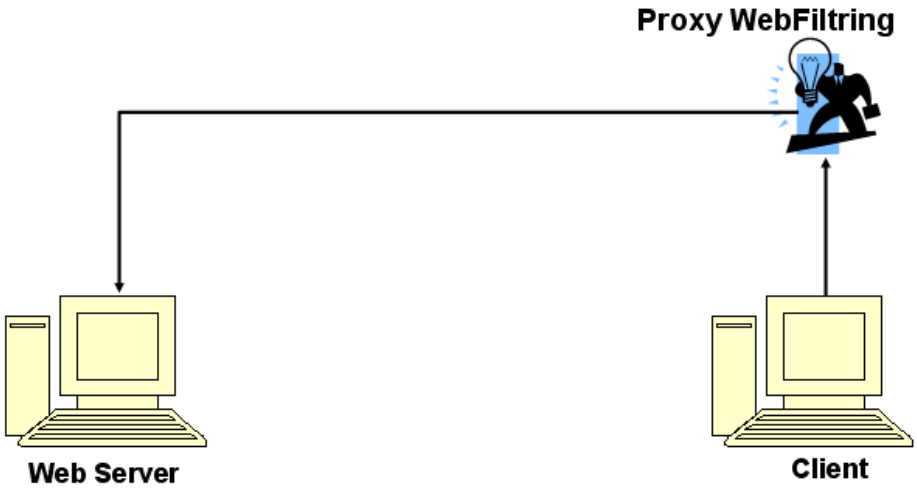
Điều kiện thử nghiệm: Chọn 2 máy tính

– Xây dựng một web server IIS trên máy B, với một thư mục chứa các trang web thuộc lớp cấm và một thư mục chứa các trang không thuộc lớp cấm để máy A truy cập vào. Như vậy trên máy B sẽ có hai thư mục ảo tương ứng với hai địa chỉ

trỏ đến thư mục ảo cho ra n liên kết đến các trang lớp cấm và m liên kết đến các trang không thuộc lớp cấm.

– Tại máy A chạy dịch vụ proxy có cài thuật toán lọc web và đồng thời cũng là máy dùng để truy cập web server trên máy B qua trình duyệt.

Mô hình:



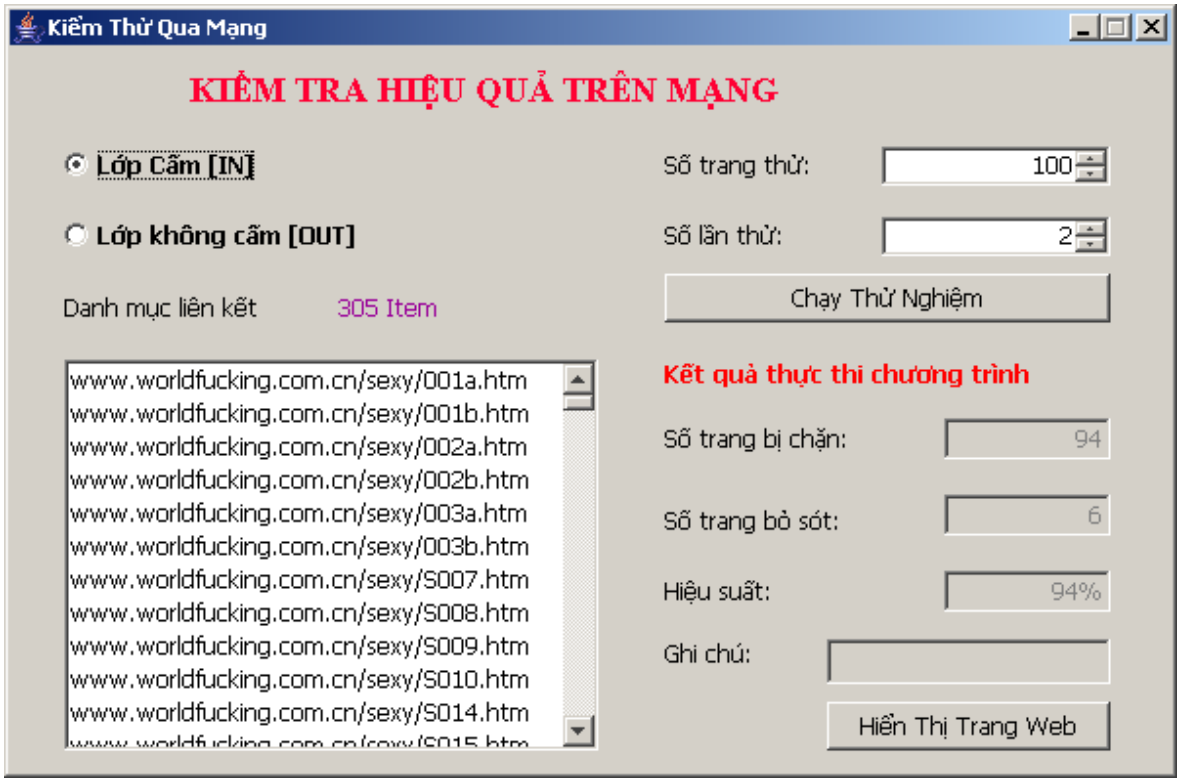
Hình 4. 14. Mô hình kiểm thử Proxy Web Filter trên mạng LAN

Thử nghiệm:

– Tắt proxy: dùng trình duyệt đánh địa chỉ của máy B cùng với một trang web trong số những trang web trong thư mục ảo trên máy B. Trình duyệt sẽ hiển thị trang web theo địa chỉ yêu cầu (bất kể lớp cấm hay không cấm).

– Mở proxy: có 2 thử nghiệm:

- ❒ Chạy thủ công qua trình duyệt: đánh vào một số địa chỉ trang cấm và không cấm để biết được hiệu quả của bộ lọc trên proxy.
- ❒ Chạy hàng loạt: cách này chỉ dùng trong thống kê, chọn số lượng địa chỉ cần kiểm tra và chọn loại trang (cấm hay không cấm) để kiểm tra.



Hình 4. 15. Màn hình kiểm tra hiệu quả trên mạng

Giả sử: danh mục của hai loại trang web được lưu trong cơ sở dữ liệu, khi chọn loại lớp sẽ tự động nạp vào danh sách liên kết.

Tiến hành kiểm thử:. Thực thi chương trình:

– Khi khởi động chương trình, bên dưới sẽ tự động kết nối đến proxy

– Người dùng muốn kiểm tra loại nào thì click chuột vào loại đó trong 2 mục chọn radio. Khi chọn một loại thì các địa chỉ liên kết sẽ tự động phát sinh và đưa vào danh sách (list).

– Chọn số trang (số địa chỉ) muốn thử, mặc nhiên chương trình chọn 100.

– Bấm nút chạy thử nghiệm: chương trình sẽ phát sinh ngẫu nhiên không trùng số liên kết có trong danh sách và gởi xuống proxy xử lý. Sau khi xử lý xong kết quả được thể hiện trong những ô bên dưới.

* Ghi nhận kết quả một số lần thử nghiệm:

TT	Số Mẫu Thử	Trên Ngưỡng	Dưới Ngưỡng	Tỉ Lệ Trên	Tỉ Lệ dưới
1	100	92	8	92%	8%
2	100	84	16	84%	16%
3	100	89	11	89%	11%
4	100	91	9	91%	9%
5	100	90	10	90%	10%
6	100	91	9	91%	9%
7	100	94	6	94%	6%
8	100	91	9	91%	9%
9	100	89	11	89%	11%
10	100	88	12	88%	12%
	Max	95	17	95.00%	17.00%
	Min	83	5	83.00%	5.00%
	AVG	90.68	9.32	90.68%	9.32%

Bảng 4.3. Ghi nhận kết quả thử nghiệm qua mạng

4.5.3. Đánh giá mức độ hiệu quả:

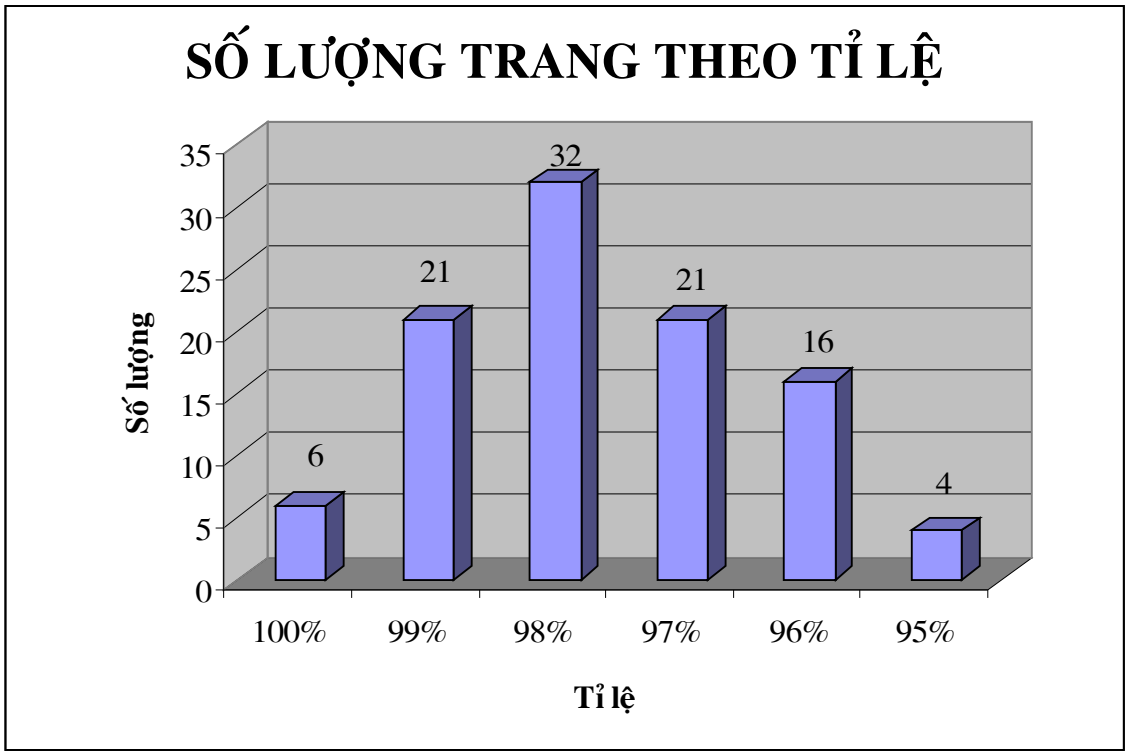
4.5.3.1. Xét về hiệu quả:

Theo bảng 4.1 và bảng 4.2 ta có thể nhận xét như sau:

– Số liệu trên bảng 4.1: Sử dụng bộ kiểm thử gồm 305 trang web truyện sex, mức độ chặn thành công trung bình 97.71% hệ thống tương đối ổn định và số lượng chặn đúng nhiều nhất ở mức 98%.

Mức độ	Số Lượng	Tỉ lệ
100%	6	6%
99%	21	21%
98%	32	32%
97%	21	21%
96%	16	16%
95%	4	4%

Bảng 4.4: Thống kê số lượng trang khóa đúng theo tỉ lệ

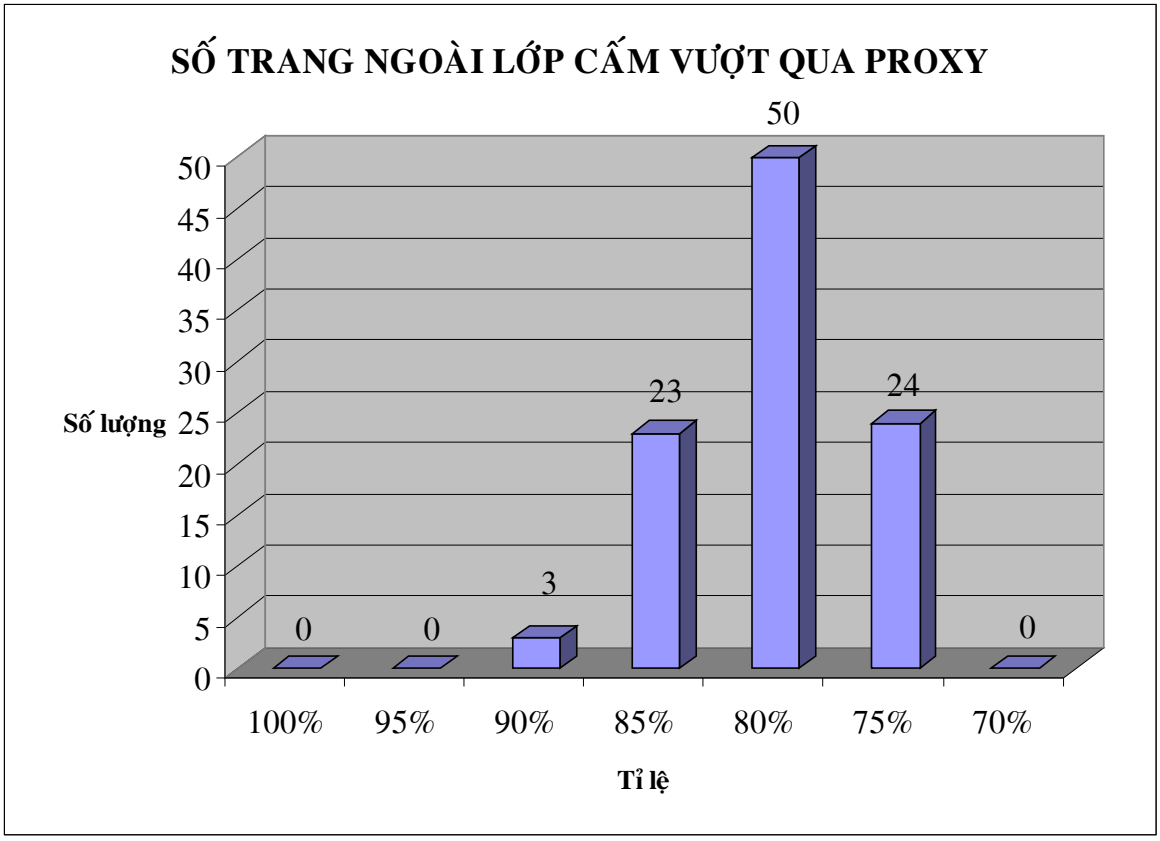


Hình 4. 16. Biểu đồ so sánh số lượng theo tỉ lệ trang bị khóa

– Số liệu trên bảng 4.2: Sử dụng bộ kiểm thử gồm 428 trang web truyện tình cảm (love story), mức độ khóa sai trung bình 17.62% hệ thống tương đối ổn định và số lượng vượt qua nhiều nhất ở mức lớn hơn hay bằng 80% và bé hơn 85%.

Mức độ	Số lượng	Tỉ lệ
>95% → 100%	0	0.0%
>90% → 95%	0	0.0%
>85% → 90%	3	3.0%
>80% → 85%	23	23.0%
>75% → 80%	50	50.0%
>70% → 75%	24	24.0%
>65% → 70%	0	0.0%

Bảng 4.5: Thống kê số lượng trang khóa đúng theo tỉ lệ



Hình 4.17: Biểu đồ biểu diễn số lượng trang qua được ngưỡng τ

Đối với thể loại truyện tình cảm, do một số từ tác giả sử dụng có tần suất lớn và trùng lặp với một số từ trong thể loại truyện sex, do đó nó có ảnh hưởng đến hệ số tương tự, làm cho hệ số tương tự của nó lớn và hệ thống so sánh với ngưỡng cho ra kết quả là trang bị khóa!.

4.5.3.2. Xét về thời gian thực thi::

– Thời gian trong giai đoạn chuẩn bị tập huấn luyện T_s và tập mẫu T_s'. Đây là giai đoạn tốn nhiều thời gian do các tiến trình sau: (i) Tạo vector tài liệu: đọc văn bản HTM, xử lý bỏ các thẻ định dạng, loại stopword, thống kê → hình thành vector. Ghi vào cơ sở dữ liệu (bảng lưu vector tài liệu hay Vector_Sample).

– Thời gian thực thi: Tính từ khi nhận vào một trang P đến khi ra hệ số tương tự của trang P ta có các bước sau: (ii): Tính cosine của vector tài liệu với từng vector trong tập huấn luyện, tính trung bình cộng của n% tài liệu có độ tương đồng cao nhất ➔ Hệ số tương đồng (Sigma).

– Thời gian tính ngưỡng: (iii) Đối với giai đoạn tính ngưỡng τ thời gian để hệ thống tính ra ngưỡng lớn nhất vì trong tập mẫu thử có n trang thì nó sẽ thực hiện n lần giai đoạn (ii) để cho ra danh sách sigma của từng trang của tập mẫu thử. Ngoài ra còn có thêm công đoạn chọn ứng viên làm ngưỡng, nếu chọn sai số càng nhỏ (độ chính xác lớn) thì số lần tính càng tốn nhiều thời gian. Ví dụ trong bài này, người viết chọn sai số 10^{-4} tức là sẽ có 10,000 ứng viên.

– Thời gian thực thi trung bình cho xét duyệt một trang:

- Trong một thực nghiệm kiểm tra chạy kiểm thử 100 lần x 100 trang là 92 giây; tương đương với xử lý một trang là 9.2 mili giây.
- Kiểm thử qua hệ thống mạng một lần chạy là truy cập đến 100 lần x100 trang là 105 giây; tương đương với xử lý một trang là 10.5 mili giây.

– Đánh giá về thời gian thực thi chỉ là ước tính vì vấn đề đánh giá phụ thuộc vào rất nhiều yếu tố chẳng hạn như băng thông mạng, hệ thống máy tính, độ dài của trang web đem kiểm thử.

4.5.3.3. Độ phức tạp của thuật toán:

Có thể xét độ phức tạp của các bước như sau:

B1: Biến đổi trang web thành vector văn bản: trong bước này chủ yếu dùng các phép toán tuyến tính, thống kê, một vòng lặp nên độ phức tạp là O_n.

B3: Tính hệ số trang: bước này tốn nhiều thời gian ở công đoạn tính cosine của vector trang web với từng trang trong tập huấn luyện. Hoạt động của bước này như sau: đọc ra một trang T_x tính cosine với V_P. Mã giả cho bước này như sau:

For (each T_i in T_S)

{

Read T_i.

$Cos(T_i,V_p)$ ➔ $vcos_i$

}

Tính trung bình cộng của n% cos_i cao nhất ➔ hệ số trang. Trong phép tính trung bình cộng này có thể tốn chi phí cho việc sắp xếp.

B2: Tính ngưỡng hệ thống: chia làm hai giai đoạn:

– Giai đoạn tính hệ số trang trong tập mẫu thử (sample): dùng lại bước 3 để tính do đó: nếu có m trang trong tập mẫu thử thì chi phí thời gian cho nó là m x t_{b3}. Còn độ phức tạp: thêm một vòng lặp duyệt qua tập mẫu T_S, trong quá trình tính hệ số trang của từng phần tử cho ra tập S_S.

– Giai đoạn chọn ứng viên: đây là giai đoạn tốn kém nhiều thời gian nhất, tùy thuộc vào việc chọn độ phân giải của ứng viên (sai số) mà số lượng ứng viên sẽ nhiều hay ít, đồng thời nó cũng cho biết số lần thực hiện tìm ứng viên. Xét một ứng viên ta cần phải thống kê trên tập các phần tử trong tập S_S và chọn ra ứng viên phân loại đúng nhiều nhất chọn làm ngưỡng.

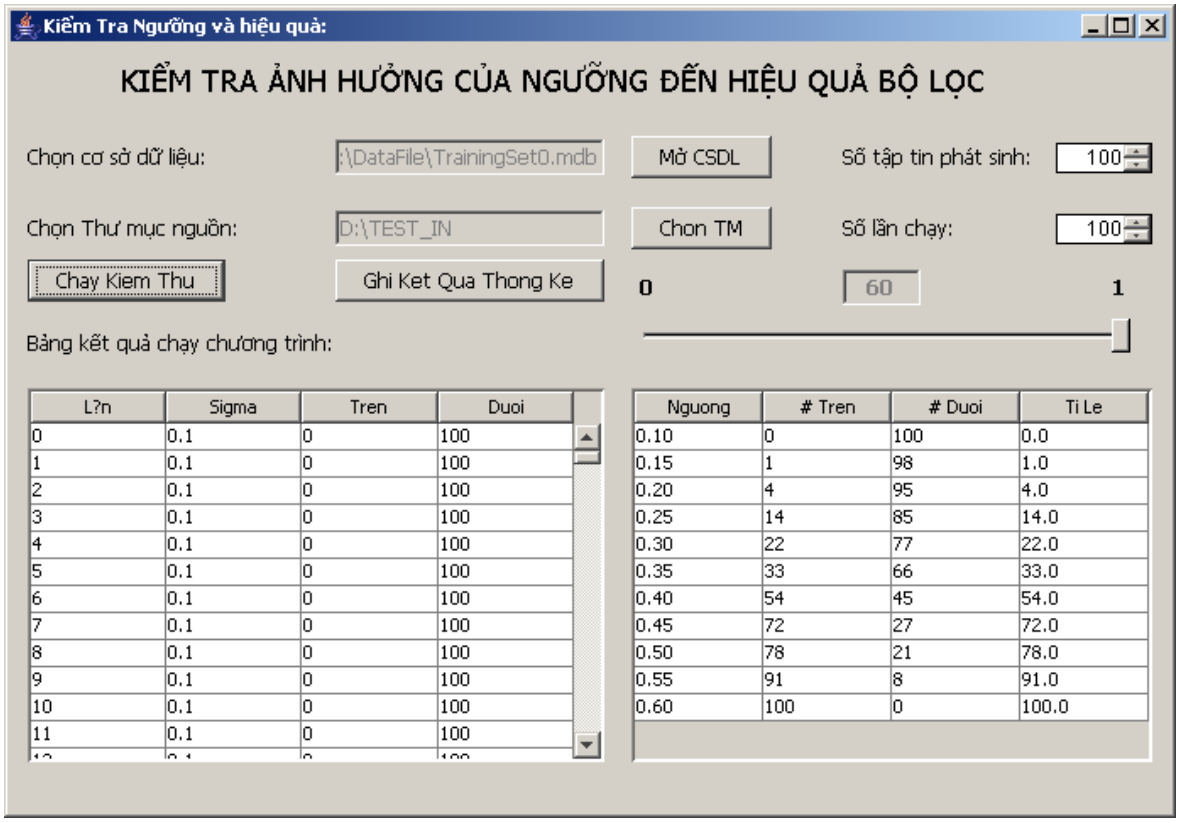
Có thể nói, độ phức tạp của bước 2 này gồm $\mathbf{m \times O_n \times O_n^2}$, tương ứng với nó là thời gian tính toán cũng tăng lên.

4.5.3.4. Ý nghĩa của việc xác định ngưỡng:

* Ảnh hưởng của giá trị ngưỡng đến hiệu quả của thuật toán:

Xét trên bộ thử gồm các trang web thuộc loại cấm (chủ đề sex, TEST_IN), với giá trị ngưỡng τ thay đổi từ 0.1 đến 1.0 (giá trị hệ số trang lớn nhất của tập này ≈ 0.6, do đó các giá trị từ 0.6 đến 1.0 không cần xét đến).

Thực hiện thử nghiệm như sau: chọn $\tau \in \varsigma\{0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6\}$, thử nghiệm với 100 trang web cho một lần chạy với ngưỡng τ_i của tập ς.

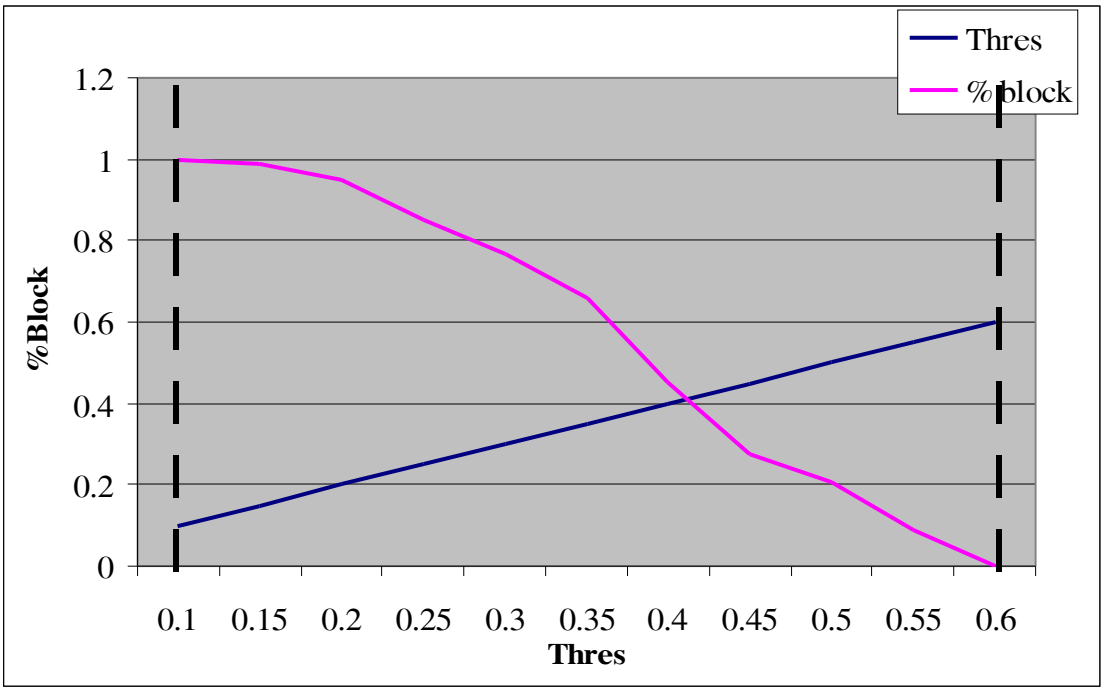


Hình 4.18: Màn hình kiểm thử sự thay đổi ngưỡng τ

Bảng ghi nhận kết quả thử nghiệm và đồ thị biểu diễn:

Thres	% block
0.1	100.00%
0.15	98.90%
0.2	94.88%
0.25	85.11%
0.3	76.74%
0.35	66.11%
0.4	45.44%
0.45	27.58%
0.5	20.86%
0.55	8.75%
0.6	0.00%

Bảng 4.6. Thống kê sự thay đổi ngưỡng τ ảnh hưởng đến hiệu suất lọc web



Hình 4.19: Đồ thị biểu diễn hiệu quả khi thay đổi ngưỡng τ

Tuy nhiên, việc chọn ngưỡng hệ thống không do chủ quan con người chọn mà do chương trình xác định, dựa vào tập mẫu thử (sample set T'_s) và tập huấn luyện (trainingset) đang có để cho ra một giá trị ngưỡng tương ứng với dữ liệu huấn luyện và dữ liệu thử.

Nhận xét:

- ✓ Việc chọn ngưỡng thủ công quá cao hay quá thấp sẽ ảnh hưởng đến hiệu suất làm việc của hệ thống (theo bảng 4.3 và hình 4.14).
- ✓ Tùy thuộc vào tập huấn luyện và tập mẫu thử mà giá trị ngưỡng sẽ thay đổi.
- ✓ Tuy nhiên một giá trị ngưỡng có thể hoạt động tốt với lớp này nhưng có thể không tốt với lớp khác có nguyên nhân xuất phát từ ngôn ngữ và cách dùng ngôn ngữ, hay thể loại (chẳng hạn như trong lớp truyện sex thì những trang web giáo dục giới tính cũng nổi lên vấn đề với cách dùng từ "nhạy cảm" làm chương trình tính ra hệ số trang lớn làm cho tài liệu bị xếp vào lớp cấm!).

Chương 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận:

5.1.1. Kết luận:

Dựa trên sự tìm hiểu kiến thức về khai mỏ dữ liệu bằng phân loại văn bản đưa ra được thuật toán và định hướng xây dựng một bộ lọc web đặt trên một proxy của một hệ thống mạng.

Đối với thuật toán [9, tr3] nguyên bản ban đầu của nhóm tác giả, qua phân tích và đánh giá tôi đã cải tiến hai chi tiết nhằm làm tăng tốc độ làm việc của thuật toán đồng thời làm giảm đi độ phức tạp trong quá trình xác định hệ số của một trang web. Đối với Proxy có source là nguồn mở, nguyên thủy ban đầu là bộ lọc IP, sau khi thêm bộ lọc bằng thuật toán phân loại văn bản thì proxy hoạt động khá ổn định, không gây ra lỗi hay làm chậm hệ thống.

Kết quả đạt được: một proxy có trang bị bộ lọc dùng phương pháp phân loại văn bản. Bộ lọc này đáp ứng được nhu cầu ngăn trang web có nội dung xấu, với tỉ lệ ngăn chặn thành công ở mức hơn 95% và thời gian đáp ứng trả về cho máy yêu cầu trong mạng khoảng 10ms/trang.

Trong quá trình nghiên cứu cũng gặp một số khó khăn như tài liệu nghiên cứu, thời gian thực hiện, điều kiện kiểm nghiệm lý thuyết, nhưng vẫn đảm bảo được nội dung cần nghiên cứu: bộ lọc web sử dụng công nghệ phân loại văn bản.

5.1.2. Khả năng ứng dụng:

– Có thể triển khai trên các mạng LAN, hệ thống kiểm soát luồng truy cập mạng từ các máy client bên trong.

– Có thể biên dịch thành ứng dụng chạy trên máy đơn giúp cho việc sử dụng internet an toàn hơn.

5.1.3. Hạn chế:

Đề tài chỉ nghiên cứu và đề xuất một bộ lọc web bằng phân loại văn bản cho một proxy server nhỏ. Trong vấn đề lọc web, đề tài này chú trọng nhiều đến lọc văn bản, chưa xử lý đến vấn đề lọc hình ảnh cũng như chưa kiểm soát quá trình download film hay những gói nén.

Một khía cạnh khác mà đề tài này chưa đề cập đến là vấn đề văn bản tiếng Việt. Khi nghiên cứu trên tiếng Việt gặp một số vấn đề khó khăn như: ngữ nghĩa và cách dùng từ, bộ stoplist (stopword) của tiếng Việt, mã tiếng Việt được dùng trong văn bản (vì hiện nay có hơn 40 bảng mã khác nhau dùng biểu diễn tiếng Việt trên máy tính),…

5.2. Hướng phát triển:

5.2.1. Kiến nghị hướng phát triển:

Với những hạn chế trên, định hướng cho sự phát triển đề tài như sau:

– Hoàn thiện mô hình lọc web, bổ sung những điểm còn hạn chế để có một bộ lọc hiện đại, đủ sức trang bị cho các tường lửa lớn hơn chẳng hạn như trên các ISP.

– Nghiên cứu thêm thuật toán lọc hình ảnh, kiểm soát download, để hoàn thiện một bộ lọc.

– Tăng cường thêm các Hueristic để tăng tốc độ làm việc của thuật toán nhằm làm giảm thời gian xử lý, tránh ùn tắc mạng do "proxy" gây ra. Trong luận văn này người viết đã thêm hai danh sách URL đã xác định: một danh sách URL cho phép truy cập và danh sách còn lại là không cho truy cập, nhờ vậy hạn chế lưu lượng đi qua bộ xử lý phân loại web.

– Ngoài ra, một nguyên nhân làm cho thuật toán chậm đi là do thuật toán đã dùng toàn bộ tập huấn luyện để tính toán. Do số lượng phần tử trong tập huấn luyện nhiều và thường được bổ sung nên thời gian tính toán chậm dần. Hướng giải quyết nhằm làm cân đối thời gian tính toán để đi đến quyết định cũng như tính ngưỡng giới hạn cho hệ thống là trang bị thêm bộ xử lý tập huấn luyện nhằm chọn ra một

tập ngưỡng (nhỏ hơn tập huấn luyện) để việc so sánh và tính toán diễn ra nhanh hơn. Hướng đề xuất này là dùng một thuật toán máy học (chẳng hạn như Bayes) trang bị cho hệ thống có vai trò "học" trên tập dữ liệu huấn luyện để đưa ra tập ngưỡng.

– Tăng thêm tốc độ làm việc của hệ thống bằng cách cải tạo thành bộ lọc web phân tán trên mạng. (Chẳng hạn như áp dụng kỹ thuật mobile agent tận dụng xử lý phân tán). Đa dạng hóa lĩnh vực lọc web (hiện tại chỉ nghiên cứu một lĩnh vực).

– Cải tiến thuật toán, xây dựng bộ lọc tiếng Việt: bộ lọc tiếng Việt cần một số yêu cầu sau: bộ stopword tiếng Việt, hàm nhận biết mã tiếng Việt, bộ chuyển đổi về mã Unicode (hay một chuẩn nào đó), tập huấn luyện tiếng Việt, một thành phần rất quan trọng đó là ngữ nghĩa và cách dùng từ trong văn bản tiếng Việt, …

– Trang bị thêm bộ phận cảm ngữ cảnh cho việc nhận dạng loại văn bản (đâu là trang web sex cần chặn và đâu là trang web giáo dục giới tính có thể cho phép xem).

5.2.2. Thảo luận:

Vấn đề lọc web ngày nay được nhiều người quan tâm, vì vấn đề an toàn đối với người dùng mạng. Việc trang bị một bộ lọc web an toàn hiệu quả và có chức năng tự động cập nhật dữ liệu theo quá trình làm việc là cần thiết. Bộ lọc web sử dụng phân loại văn bản (trong khai mỏ văn bản) đáp ứng được yêu cầu trên.

Trong luận văn này, người viết đã nghiên cứu và cài đặt minh họa bộ lọc web sử dụng phân loại văn bản cho một proxy. Vẫn còn một số vấn đề chưa giải quyết được và có đề xuất hướng phát triển trong tương lai.

TÀI LIỆU THAM KHẢO

1. GS.TSKH Hoàng Kiếm (2004), Tập bài giảng chuyên đề Công Nghệ Tri thức và ứng dụng, ĐHQG TPHCM.
2. TS Đỗ Phúc (2004), Tập bài giảng chuyên đề Khai phá dữ liệu và Nhà kho dữ liệu – ĐHQG TPHCM.
3. Dr. Edel Garcia (2005), Term Vector Theory and Keyword Weights. (www.miislita.com/term-vector/term-vector-1.html)
4. Dr. Edel Garcia (2005-Bản cập nhật trên mạng 11-9-2006), Term Vector Fast Track.
5. Dr. Edel Garcia (5-9-2006-Bản cập nhật trên mạng 11-9-2006), A Linear Algebra Approach to Term Vectors.
8. Miller David W. (2001), Automatic Text Classification through Machine Learning.
9. Rongbo Du, Reihaneh Safavi-Naini and Willy Susilo (2003), **Web Filtering Using Text Classification**, Centre for Communication Security School of Information Technology and Computer Science University of Wollongong, Australia.
10. Rosen-Zvi Michal (2001), Text Classification - University of California.
11. Sebastiani Fabrizio (Jan.2004), Text Classification for Web Filtering.
12. Stern Benjamin A. (5/12/2003), Web Filtering Technology Assessment.
13. www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html (webpage) - Tính cosine.
14. WordHoard team - Comparing texts (wordhoard.northwestern.edu/ userman/analysis-comparingtexts.html).

(*): Bài báo chính dùng nghiên cứu luận văn này.