Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG CÔNG CỤ NGĂN CHẶN VIỆC TRUY CẬP WEB ĐEN (HÌNH ẢNH, NỘI DUNG)

Huỳnh Bé Thơ và Trương Quốc Định[1]
[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Building tool to prevent access to black web (image, content)

Từ khóa:
Lọc web, Máy học vector hỗ trợ (SVM), phân lớp văn bản, phân lớp hình ảnh

Keywords:
Web filtering, Support vector machine learning, text classification, image classification

ABSTRACT

Web filtering is used to prevent access to black web pages (web pages have desirable content or images). In this paper, we apply classification method with Support vector machine learning (SVM) to build a web filtering tool that is integrated with 2 filters: text filter – use text classification method and image filter – use image classification method. With two filters, this tool can prevent user access to desirable content web pages or remove desirable images when web page is displayed on the browser.

TÓM TẮT

Công cụ lọc web (web filtering) được sử dụng để ngăn chặn việc truy cập đến các trang web đen (trang web mang nội dung hoặc hình ảnh không mong muốn). Trong bài báo này, chúng tôi ứng dụng phương pháp phân lớp với Máy học vector hỗ trợ (SVM) để xây dựng một công cụ lọc web được tích hợp 2 bộ lọc: bộ lọc văn bản – phân lớp văn bản (text classification) và bộ lọc image – phân lớp hình ảnh (image classification). Với hai bộ lọc này, công cụ có thể cấm người dùng truy cập đến trang web có nội dung văn bản không mong muốn hoặc loại bỏ các hình ảnh không mong muốn khi hiển thị web lên trình duyệt.

1 GIỚI THIỆU

Cùng với sự phát triển của công nghệ ADSL và sự cạnh tranh giữa các nhà cung cấp dịch vụ dẫn đến Internet ngày càng phổ biến và dễ tiếp cận. Không chỉ có thể tiếp xúc với môi trường Internet tại cơ quan, trường học, quán net mà ngay tại gia đình, nhà trọ cũng thật dễ dàng sở hữu một đường truyền ADSL hay cáp quang tốc độ cao. Không thể phủ nhận Internet là kho tri thức khổng lồ, một công cụ đắc lực hỗ trợ cho việc học tập, nghiên cứu nhưng ngược lại nó cũng chứa nhiều mối nguy hiểm tiềm ẩn bên trong, cụ thể là sự lan tràn của các trang web chứa nội dung, hình ảnh không lành mạnh, không phù hợp với thuần phong mỹ tục của một số quốc gia...

Các chương trình lọc web trước đây thường sử dụng các công nghệ mang tính thủ công: ngăn chặn dựa trên từ khóa (keyword), chặn theo địa chỉ liên kết (URL hay IP),… Các công nghệ này đã lỗi thời bởi vì không phải bất cứ trang web nào có chứa các từ ngữ "nhạy cảm" đều là trang web khiêu dâm và cũng thật khó để ngăn chặn theo URL hoặc IP khi số lượng các trang web sex quá lớn và tăng thêm liên tục. Trong ngữ cảnh đó, xu hướng chung của các phần mềm lọc web đen ngày nay là dựa trên phân tích nội dung trang web sử dụng kỹ thuật khai mỏ dữ liệu (data mining).

Đi theo xu hướng chung đó, chúng tôi đã xây dựng một hệ thống lọc web dựa trên kỹ thuật phân lớp, cụ thể là phân lớp văn bản và hình ảnh với SVM.

Phần tiếp theo của bài báo được tổ chức như sau. Phần 2 trình bày các nghiên cứu liên quan và hướng tiếp cận của bài báo. Phần 3 giới thiệu các bước nghiên cứu bao gồm: bài toán phân lớp (phần 3.1), máy học Vector hỗ trợ SVM (phần 3.2), mô hình hóa văn bản (phần 3.3), cuối cùng là biểu diễn ảnh bằng đặc trưng SIFT và mô hình Bag of Words. Kết quả nghiên cứu trình bày ở phần 4. Kết luận và đề xuất được giới thiệu trong phần 5.

2 NGHIÊN CỨU LIÊN QUAN

2.1 Nghiên cứu liên quan

Các hệ thống lọc web dựa trên kỹ thuật phân lớp được phát triển gần đây thường dựa trên 3 hướng tiếp cận: phân lớp văn bản, phân lớp hình ảnh và phân lớp dựa trên sự kết hợp nhiều yếu tố (văn bản, hình ảnh/video). Theo xu hướng thứ nhất có thể kể đến nghiên cứu của Du *et al*.2003 [5], Kim *et al*.2006 [8] hoặc Santos *et al*.2012 [11]. Du và các đồng sự đã đề xuất một hệ thống lọc web sử dụng thuật toán tính độ tương đồng của vector văn bản với tập học để phân loại trang web khiêu dâm và không khiêu dâm. Một nghiên cứu khác của Kim phát triển hệ thống phân loại trang web đen thành nhiều cấp bậc bằng phương pháp phân lớp văn bản với máy học SVM. Nghiên cứu của Santos áp dụng giải thuật DMC (dynamic Markov compression) phân lớp văn bản để lọc các website khiêu dâm.

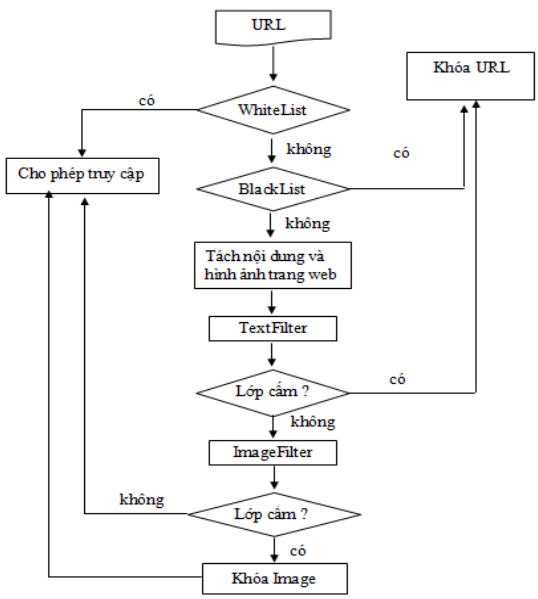
Phân loại trang web dựa theo kỹ thuật phân lớp hình ảnh cũng có nhiều nghiên cứu. Jiao *et al.*2011 [12] nghiên cứu ứng dụng kỹ thuật phân lớp hình ảnh sử dụng đặc trưng màu sắc và SVM cho hệ thống ngăn chặn các website khiêu dâm. Nghiên cứu của Zhao. 2010 [13] đề xuất phương pháp kết hợp đặc trưng màu sắc, kết cấu và đặc trưng SIFT cho thuật toán phân lớp hình ảnh với SVM ứng dụng cho hệ thống phát hiện ảnh khỏa thân trên web.

Cuối cùng là phương pháp kết hợp thông tin văn bản và hình ảnh/videos để nhận diện web khiêu dâm. W.Hu *et al*. 2011 [6] đề xuất phân loại trang web dựa trên đặc trưng được kết hợp giữa văn bản, hình ảnh và video. Trang web được đại diện bởi vector $W = (w_1, w_2,..., w_t, w_{t+1}, w_{t+2}, ..., w_{t+7})$ với w_i $(1 \le i \le t)$ là trọng số các đặc trưng văn bản và w_{t+j} $(1 \le j \le 7)$ đại diện cho các đặc trưng thứ j của image/videos. Nghiên cứu của M. Hammami *et al*.2003 [7] thực hiện kết hợp đặc trưng văn bản và hình ảnh để xây dựng đặc trưng cho trang web, sau đó quyết định xem trang web này có thuộc web cấm hay không bằng giải thuật cây quyết định (decision tree). Saikat Sen. 2010 [10] giới thiệu một hệ thống lọc web sử dụng giải thuật phân lớp Naïve Bayes để phân loại trang web dựa vào các đặc trưng url, tiêu đề, từ khóa và nội dung (văn bản, hình ảnh) của trang web.

2.2 Hướng tiếp cận của bài báo

Tất cả các nghiên cứu trên đều thực hiện lọc web tiếng Anh và chỉ có chức năng lọc web (cho truy cập hay không), không có chức năng lọc ảnh khiêu dâm (vẫn truy cập được nhưng các ảnh khiêu dâm sẽ không hiển thị).

Trong bài báo này, chúng tôi giới thiệu một hệ thống lọc web dựa trên kỹ thuật phân lớp văn bản và hình ảnh với SVM. Tuy nhiên không thực hiện phân loại trang web bằng cách kết hợp 2 yếu tố nội dung văn bản và hình ảnh trang web mà là xây dựng 2 bộ lọc. Bộ lọc text (textfilter) thực hiện phân lớp văn bản tiếng Việt để phân loại trang web và bộ lọc image (imagefilter) dựa trên kỹ thuật phân lớp hình ảnh, thực hiện chức năng "lọc" hình ảnh trang web. Hình 1 thể hiện sơ đồ hoạt động của hệ thống.



Hình 1: Sơ đồ hoạt động của hệ thống

3 CÁC BƯỚC NGHIÊN CỨU

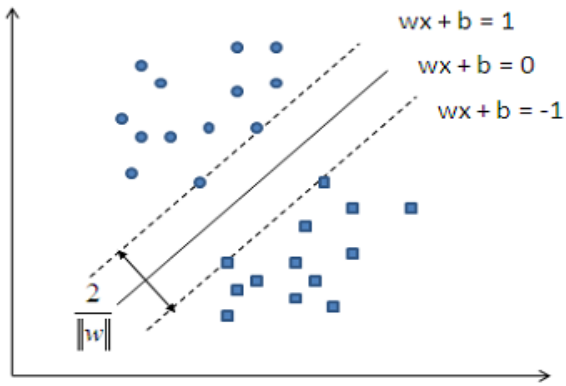
3.1 Bài toán phân lớp

Bài toán phân lớp (phân loại) là một bài toán kinh điển trong lĩnh vực khai mỏ dữ liệu. Mục tiêu của bài toán là xây dựng một mô hình phân lớp dựa trên tập dữ liệu học có nhãn (lớp) [14]. Ví dụ cho sẵn một tập dữ liệu các trang web được gán nhãn là web khiêu dâm hay bình thường, vấn đề là cần một

phương pháp huấn luyện để xây dựng một mô hình phân lớp từ tập dữ liệu mẫu này sau đó dùng mô hình này dự đoán lớp của những trang web mới (chưa biết nhãn).

3.2 Máy học Vector hỗ trợ - SVM

Phương pháp SVM [15] ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. SVM được đánh giá là 1 trong 10 giải thuật quan trọng của khai mỏ dữ liệu [14]. Các ứng dụng thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...).



Hình 2: Phân lớp tuyến tính với SVM

Bài toán cơ bản của SVM là bài toán phân loại hai lớp: Cho trước n điểm trong không gian d chiều (mỗi điểm thuộc vào một lớp kí hiệu là +1 hoặc –1, mục đích của giải thuật SVM là tìm một siêu phẳng (hyperplane) phân hoạch tối ưu cho phép chia các điểm này thành hai phần sao cho các điểm cùng một lớp nằm về một phía với siêu phẳng này. Hình 2 minh họa phân lớp với SVM trong mặt phẳng.

Xét tập dữ liệu mẫu có thể tách rời tuyến tính $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ với $x_i \in R^d$ và $y_i \in \{\pm 1\}$. Siêu phẳng tối ưu phân tập dữ liệu này thành hai lớp là siêu phẳng có thể tách rời dữ liệu thành hai lớp riêng biệt với lề (margin) lớn nhất. Tức là, cần tìm siêu phẳng H: y = w.x + b = 0 và hai siêu phẳng H1, H2 hỗ trợ song song với H và có cùng khoảng cách đến H. Với điều kiện không có phần tử nào của tập mẫu nằm giữa H_1 và H_2, khi đó:

w.x + b >= +1 với y = +1

w.x + b >= -1 với y = -1

Kết hợp hai điều kiện trên ta có:

y(w.x + b) >= 1.

Khoảng cách của siêu phẳng H_1 và H_2 đến H là ||w||. Ta cần tìm siêu phẳng H với lề lớn nhất, tức là giải bài toán tối ưu tìm $\min_{w,b}\|w\|$ với ràng buộc y(w.x + b) >= 1. Người ta có thể chuyển bài toán sang bài toán tương đương nhưng dễ giải hơn là $\min_{w,b}\frac{1}{2}\|w\|^2$ với ràng buộc y(w.x + b) >= 1. Lời giải cho bài toán tối ưu này là cực tiểu hóa hàm Lagrange:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i(y_i(<w,x_i> + b) - 1) \quad (1)$$

Trong đó α là các hệ số Lagrange, α≥0. Sau đó người ta chuyển thành bài toán đối ngẫu là cực đại hóa hàm W(α):

$$\max_\alpha W(\alpha) = \max_\alpha (\min_{w,b} L(w,b,a)) \quad (2)$$

Từ đó giải để tìm được các giá trị tối ưu cho w,b và α. Về sau, việc phân loại một mẫu mới chỉ là việc kiểm tra hàm dấu sign(wx +b).

Giải thuật SVM cơ bản giải quyết được bài toán phân lớp tuyến tính, tuy nhiên nếu ta kết hợp SVM với phương pháp hàm nhân (kernel – based method), sẽ cho phép giải quyết một số bài toán phi tuyến bằng cách ánh xạ dữ liệu vào một không gian có số chiều lớn hơn. Không có bất kỳ thay đổi cần thiết nào về mặt giải thuật, việc duy nhất cần làm là thay thế các tích vô hướng của hai vector u.v bởi hàm nhân K(u,v). Chúng ta có một số hàm nhân cơ bản được dùng phổ biến được cho trong Bảng 1.

Bảng 1: Một số hàm nhân thường được dùng

Kiểu hàm	Công thức
Tuyến tính	K(u,v) = u.v
Đa thức bậc d	$K(u,v) = (u.v + c)^d$
Radial Basis Function	$K(u,v) = \exp(-\gamma\|u-v\|^2)$

Trong giới hạn nghiên cứu này, chúng tôi không đi sâu vào giải thuật SVM. Các mô hình phân loại trong nghiên cứu được thực hiện nhờ vào sự hỗ trợ của công cụ LibSVM [23].

3.3 Mô hình hóa văn bản

Để có thể thực hiện phân lớp văn bản với Máy học vector hỗ trợ, mỗi văn bản cần được biểu diễn dưới dạng vector với các thành phần (chiều) của vector này là các trọng số của các từ chỉ mục. Khái niệm "từ chỉ mục" ở đây theo nghĩa là một chuỗi kí tự liên tiếp nhau trong văn bản, không nhất thiết phải là một từ có nghĩa trong ngôn ngữ [15]. Như vậy, giai đoạn đầu tiên trong việc vector hóa văn bản là thực hiện việc tách rời các từ. Tập hợp tất cả

các từ để biểu diễn văn bản được rút ra từ tập các văn bản đang xét gọi là tập đặc trưng.

Việc tách các từ này một cách chính xác có ảnh hưởng rất lớn đến kết quả phân loại. Trong những năm gần đây, có nhiều công trình nghiên cứu về tách từ tiếng Việt. Tiêu biểu là công cụ vnTokenizer được phát triển trong đề tài VLSP của tác giả Lê Hồng Phương [22]. Công cụ này tách từ cho độ chính xác là 97%. Hình 3 minh họa một đoạn văn bản được tách từ bởi VnTokenizer.

Ngoài_ra /, / có / một / vấn_đề / nảy_sinh / trong / khi / tách / từ / là / việc / xuất_hiện / các /từ / mới /,/ đây / là / vấn_đề / không_thể / bỏ_qua / khi / ngôn_ngữ / là / luôn_luôn / thay_đổi / và / sinh_ra / các / từ / mới /./ Trong khi / từ_điển / (/ dành / cho /xử_lý / ngôn_ngữ_tự_nhiên /) / không_thể / cập_nhật /hết / được /./

Hình 3: Ví dụ tách từ với VnTokenizer

Rõ ràng rằng, các từ trong văn bản có mức độ quan trọng khác nhau đối với văn bản và cả đối với các văn bản khác trong tập văn bản cần phân loại. Một số từ như từ nối, dấu chấm câu, ký hiệu đặc biệt, từ chỉ số lượng ("và", "các", "những, "mỗi",…). không mang tính phân biệt trong khi phân loại. Để giảm bớt số lượng đặc trưng, nâng cao tốc độ tính toán, các từ gọi là stopword này cần được loại bỏ. Tuy nhiên các stopword này có số lượng không đáng kể, cần thiết phải áp dụng một giải thuật giúp chọn lựa các đặc trưng thật sự hữu ích cho việc phân lớp. Một vài thuật toán giúp lựa chọn các đặc trưng [9] như: ngưỡng tần suất văn bản (Document Frequency thresholding - DF), độ lợi thông tin (Information Gain - IG), thông tin tương hỗ (Mutual Information - MI), phương pháp thống kê χ^2 (CHI), độ mạnh của từ và một số phương pháp khác. Trong bài báo này, chúng tôi đã áp dụng phương pháp ngưỡng tần suất văn bản DF cho nghiên cứu này bởi tính đơn giản và hiệu quả của nó. Phương pháp này tính tần suất văn bản (DF) cho mỗi đặc trưng và loại bỏ những đặc trưng có tần suất văn bản nhỏ hơn ngưỡng cho trước. Các đặc trưng có tần suất văn bản thấp sẽ mang ít thông tin phân loại và thường là dữ liệu nhiễu.

Sau khi loại bỏ các stopword, các từ có tần số DF thấp, tập đặc trưng còn lại đó là tập hợp các từ "quan trọng" còn lại để biểu diễn văn bản. Việc phân loại văn bản sẽ dựa trên tập đặc trưng này.

Kế đến, mỗi văn bản trong tập đang xét sẽ được biểu diễn bởi các trọng số của từ. Cách tính trọng số của từ được áp dụng theo công thức trọng số TF*IDF (tf-idf weighting) [1]. Trọng số tf-idf được tính như sau:

$$TF\text{-}IDF_{t,d} = TF_{t,d} * \log(N/ DF_t) \quad (3)$$

– $TF_{t,d}$: là số lần xuất hiện của từ t trong văn bản d.

– DF_t : số lượng văn bản có chứa từ t.

– N: tổng số văn bản trong tập dữ liệu đang xét.

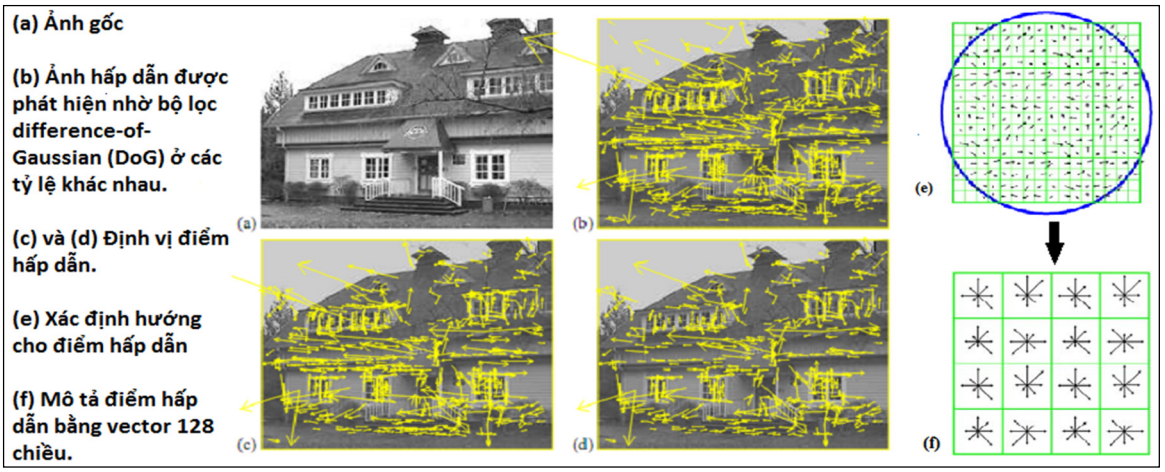
3.4 Biểu diễn ảnh bằng đặc trưng SIFT và mô hình BoW (Bag of Words)

Giống như một từ trong một văn bản text, một tấm ảnh cũng có thể xem là tập hợp các điểm hấp dẫn cục bộ hoặc các điểm nổi bật, là những vùng nhỏ (small regions) chứa nhiều thông tin cục bộ của ảnh – còn gọi là các đặc trưng. Có nhiều đặc trưng có thể được sử dụng để biểu diễn cho ảnh [16], trong đó có đặc trưng cục bộ. Người ta thường chia đặc trưng cục bộ thành 2 loại là những điểm trích xuất được từ điểm "nhô ra" (salient points) của ảnh và đặc trưng SIFT [2] được trích chọn từ các điểm hấp dẫn Haris (interest points). Điểm hấp dẫn này được mô tả bởi vector 128 chiều (gọi là các sift descriptor). Các vector này "bất biến" với việc thay đổi tỉ lệ ảnh, quay ảnh, đôi khi là thay đổi điểm nhìn và thêm nhiễu ảnh hay thay đổi cường độ chiếu sáng của ảnh. Hình 4 mô tả các bước rút trích đặc trưng SIFT.

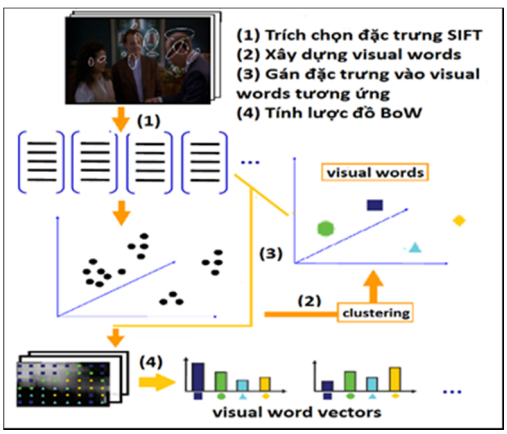
Bước kế tiếp, chúng tôi dùng mô hình bag of words (BoW) [3,4,20,21,25] để thực hiện kết tập các đặc trưng cục bộ SIFT. Mô hình này dùng một giải thuật (ví dụ như k-means) gom nhóm các đặc trưng cục bộ SIFT để xây dựng các visual words. Tập các visual words này gọi là codebook. Sau đó gán các đặc trưng cục bộ trên mỗi tấm ảnh vào visual words gần nhất. Khoảng cách Euclid thường được sử dụng để tính khoảng cách từ đặc trưng đến visual words gần nhất. Khoảng cách Euclid được tính bằng công thức sau:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Ảnh sẽ biểu diễn ảnh bằng một vector với các thành phần là các visual words. Các vector này còn được gọi là các lược đồ BoW (BoW histograms). Trọng số của các visual words được tính theo **tần suất xuất hiện của visual word** trong mỗi ảnh. Hình 5 mô tả các bước tạo mô hình BoW để biểu diễn ảnh.



Hình 4: Các bước rút trích đặc trưng SIFT



Hình 5: Mô hình bag of words (BoW)

4 KẾT QUẢ VÀ THẢO LUẬN

Chúng tôi cài đặt một http proxy có tích hợp 2 bộ lọc TextFilter (lọc văn bản) và ImageFilter để thực hiện chức năng ngăn chặn việc truy cập vào các trang web đen. Khi người dùng truy cập vào trang web có nội dung cấm, công cụ sẽ chuyển trình duyệt về trang thông báo cấm truy cập hoặc sẽ hiển thị trang web nhưng "lọc" lại các ảnh khỏa thân (nếu có). Tùy vào số lượng ảnh có trên trang web, thời gian (được tính từ lúc bắt đầu nhập URL vào address bar đến lúc load xong toàn bộ trang web) để trình duyệt có tích hợp proxy với 2 bộ lọc load một trang web có thể chậm hơn vài giây so với trình duyệt thông thường.

4.1 Đánh giá bộ lọc văn bản - TextFilter

Tập dữ liệu cho bộ lọc TextFilter là 2518 mẫu tin được tải về từ internet. Các mẫu tin này thuộc 2 chủ đề (khiêu dâm và không khiêu dâm - thường). Chúng tôi dùng 1518 mẫu tin làm tập huấn luyện (dữ liệu học) và 1000 mẫu tin để kiểm tra (xem Bảng 2).

Bảng 2: Tập dữ liệu văn bản

STT	Tên tập	Khiêu dâm	Bình thường
1	Tập huấn luyện	900	618
2	Tập kiểm tra	500	500

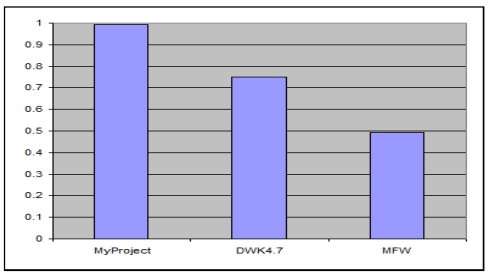
Tập huấn luyện được đặt trong 2 thư mục riêng biệt (0NonAdult: chứa mẫu tin bình thường và 1Adult: chứa mẫu tin khiêu dâm). Sử dụng thư viện VietTokenizer [22] để tách câu thành các "từ" riêng biệt. Tổng số đặc trưng thu được sau khi tách từ là hơn 50.000 đặc trưng. Thực hiện loại bỏ stopword và các đặc trưng có tần số DF<3, cho kết quả còn lại là 19.587 đặc trưng. Biểu diễn các mẫu tin bằng vector trọng số TF*IDF. Nhãn của các mẫu tin sẽ được gán tự động theo thứ tự của thư mục chứa mẫu tin. Trong trường hợp này, các trang web thuộc nhóm bình thường có nhãn là 0 và các trang khiêu dâm có nhãn là 1.

Để phân lớp văn bản bằng SVM, chúng tôi sử dụng bộ thư viện LibSVM [23] với công cụ *grid.py* giúp lựa chọn các tham số tối ưu cho giải thuật SVM. Bằng nghi thức kiểm tra chéo (5-fold) trên tập học, *grid.py* đã tìm ra 2 tham số (C = 32 và gamma = 0.0001220703125) sau khi tính được kết quả phân loại cao nhất là 99,0119% trên tập huấn luyện.

Thực hiện kiểm nghiệm trên tập kiểm tra cho kết quả phân lớp đạt độ chính xác là 93,65%.

Để đánh giá hiệu quả thực tế của bộ lọc, chúng tôi thực hiện so sánh khả năng phát hiện các trang web khiêu dâm của công cụ với các phần mềm lọc

web "thuần việt" hiện nay như: DWK 4.7 (Tác giả Vũ Lương Bằng – Công ty Điện Thoại Đông), MiniFireWall 4.0 - MFW (Tác giả Huỳnh Ngọc Ẩn – Phòng Tin học, bưu điện Đồng Tháp). Các chương trình này ứng dụng kỹ thuật lọc theo URL hoặc keyword để chặn các trang web. Dữ liệu để so sánh là 122 mẫu tin lấy ngẫu nhiên từ 122 website khác nhau từ công cụ tìm kiếm Google với các từ khóa tìm kiếm là "truyện người lớn", "truyện khiêu dâm", "truyện sex", "truyện 18+". Kết quả so sánh thể hiện ở biểu đồ 1.



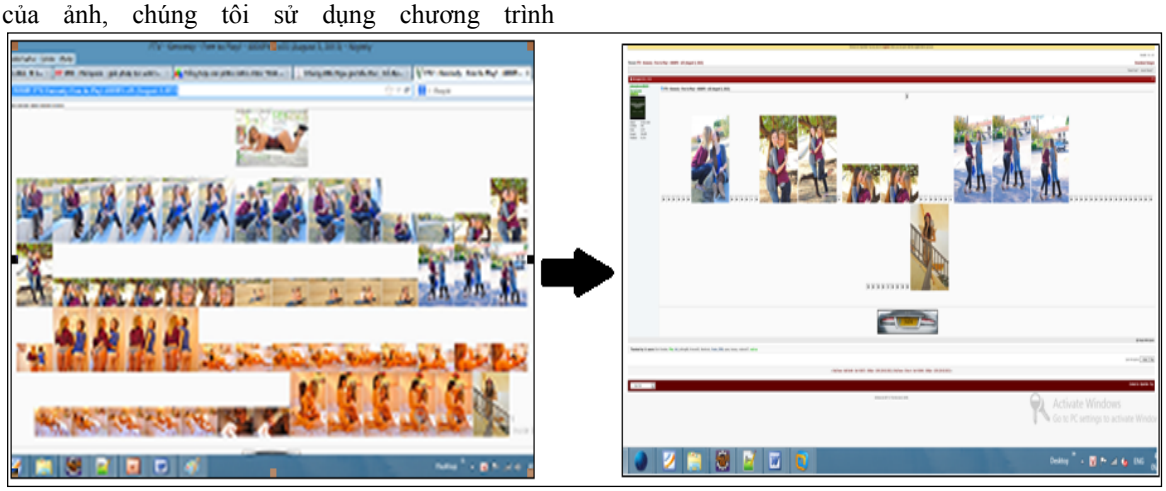
Biểu đồ 1: So sánh khả năng phát hiện website truyện khiêu dâm của công cụ (MyProject) với DWK và MFW

Kết quả này cho thấy việc sử dụng kỹ thuật phân lớp văn bản để phân loại, phát hiện và ngăn chặn các trang web đen thật sự hiệu quả hơn các kỹ thuật khác.

4.2 Đánh giá bộ lọc hình ảnh ImageFilter

Mô hình phân lớp cho bộ lọc ImageFilter được huấn luyện từ tập dữ liệu gồm 1905 ảnh (1066 ảnh thuộc nhóm ảnh khiêu dâm và 839 ảnh không khiêu dâm – bình thường). Tương tự như dữ liệu văn bản cho bộ lọc TextFilter, tập dữ liệu ảnh này cũng được đặt vào 2 thư mục khác nhau để dễ dàng trong việc gán nhãn. Để tìm vector đặc trưng SIFT của ảnh, chúng tôi sử dụng chương trình SiftDemoV4 [24] của David G. Lowe. Các vector này sẽ được gom nhóm bằng giải thuật k-means để xây dựng các visual words và thực hiện vector hóa tập dữ liệu ảnh với các visual words tìm được. Chúng tôi thí nghiệm với số lượng các visual words là 200, 300, 500, 1000, 2000, 3000 và N (Với N là căn bậc hai của tổng số đặc trưng). Ta thu được mô hình phân lớp với SVM cho độ chính xác ổn định nhất là (76.3255% và 80.1084%) ứng với số visual words là (200 và 3000). Từ thực nghiệm và nghiên cứu của [17] cho thấy quy mô của tập từ vựng có ảnh hưởng rất lớn đến tốc độ tính toán của hệ thống. Vì vậy với một ứng dụng chạy thời gian thực như trong nghiên cứu này, chúng tôi ưu tiên chọn mô hình với số visual words là 200, dù độ chính xác của mô hình này thấp hơn mô hình với số visual words 3000. Hình 6 cho ví dụ về kết quả phân lớp ảnh của bộ lọc ImageFilter. Ảnh bên trái là một trang web có chứa ảnh khiêu dâm và bên phải là trang web được hiển thị lên trình duyệt sau khi bộ lọc loại bỏ các ảnh khiêu dâm.

Trên thực tế các nghiên cứu về phát hiện ảnh khiêu dâm dựa vào đặc trưng SIFT có thể cho kết quả phân lớp rất khác nhau, tùy vào nhiều yếu tố như dữ liệu học, giải thuật phát hiện điểm đặc trưng cục bộ,... Nghiên cứu của Lopes *et al.*2009 [18] cho kết quả nhận diện ảnh khiêu dâm sử dụng đặc trưng SIFT với SVM là 65±3%. Nghiên cứu khác của Steel *et al.* [19], cho kết quả đánh giá dựa vào chỉ số TPR (true positive rates) là (0.58 và 0.66) tương ứng với FPR (false positive rates) tại (0.1 và 0.2), cao hơn so với phương pháp phân lớp dựa vào màu da có giá trị TPR là (0.49 và 0.61). Kết quả này cho thấy, với độ chính xác 76.3255% như trong bài báo này vẫn có thể chấp nhận được.



Hình 6: ví dụ về chức khóa các ảnh khiêu dâm của công cụ

5 KẾT LUẬN VÀ ĐỀ XUẤT

5.1 Kết luận

Trong bày viết này chúng tôi trình bày một hướng tiếp cận trong việc xây dựng công cụ lọc web chạy thời gian thực giúp ngăn chặn sự truy cập đến các trang web chứa thông tin và hình ảnh khiêu dâm, đồi trụy. Từ thực nghiệm cho thấy việc áp dụng kỹ thuật phân lớp văn bản tiếng Việt với SVM cho kết quả phân loại với độ chính xác cao (hơn 90%). Tuy nhiên kỹ thuật này thực sự bị hạn chế về tốc độ tính toán do số chiều vector quá lớn.

Hiện tại công cụ chỉ giới hạn trong việc ngăn chặn các trang web khiêu dâm, tuy nhiên việc mở rộng phạm vi lọc của công cụ sang các chủ đề khác như lọc các trang web có nội dung phản động, bạo lực được thực hiện tương đối dễ dàng (kể cả đối với người dùng không có nhiều kiến thức về khai mỏ dữ liệu) nhờ vào quy trình huấn luyện mô hình tự động.

5.2 Đề xuất

– Cần hoàn thiện proxy, bổ sung các giao thức còn thiếu.

– Cần nghiên cứu áp dụng một giải thuật lựa chọn đặc trưng văn bản thật hiệu quả để thu gọn tập đặc trưng, tăng tốc độ tính toán.

– Nghiên cứu thuật toán giúp kiểm soát download phim và những gói nén.

– Xây dựng thêm bộ lọc văn bản tiếng Anh.

– Nghiên cứu kết hợp trích chọn đặc trưng sift với đặc trưng màu da, huấn luyện mô hình với các tham số của SVM như nghiên cứu của Lopes et al.2009 [18] hoặc Do. 2011 [25] để tăng hiệu quả cho bộ lọc hình ảnh.

TÀI LIỆU THAM KHẢO

1. Christopher D.Manning, Prabhakar Raghavan, Hinrich Schütze , 2009. An Introduction to Information Retrieval. Cambridge University Press Cambridge, England: 117 – 119.
2. David G.Lowe, 2004. "*Distinctive Image Features from Scale-Invariant Keypoints*". International Journal of Computer Vision: 91-110.
3. Jun-Yang, Yu-GangJiang, 1997. "Evaluating Bag-of-Visual-Words Representations in Scene Classification". WOODSTOCK '97 El Paso, Texas, USA: 2 - 3.
4. Thomas Deselaers, Lexi Pimenidis, Hermann Ney, . "Bag-of-Visual-Words Models for Adult Image Classification and Filtering", 2008. 19th International Conference on Pattern Recognition ICPR : 1 – 4.
5. Rongbo Du, Reihaneh Safavi-Naini and Willy Susilo, 2003. "Web Filtering Using Text Classification". The 11th IEEE International Conference on Networks: 325 – 330.
6. W.Hu, H.Zuo, Ou Wu, Yunfei Chen, 2011. "Recognition of Adult Images, Videos, and Web Page Bags". ACM Transactionson Multimedia Computing, Communications and Applications,Vol.7S,No.1: 1 – 28.
7. Mohamed Hammami, Youssef Chahir, Liming Chen, 2003. Combining Text and Image Analysis in the Web filtering System "WEBGUARD". International Conference WWW/Internet: 611 – 618.
8. Youngsoo Kim, Taekyong Nam, Dongho Won, 2006. Text Classification for Harmful Web Document. Computational Science and Its Applications - ICCSA 2006: 545 – 551.
9. Yiming Yang, Jan O. Pedersen, 1997. A comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th international conference on Machine Learning: 412 – 420.
10. Saikat Sen, 2010. Adult Website Classifier. CS229 Machine Learning Course Project, Stanford University, USA. http://cs229.stanford.edu/proj2010/Stanford project CS229.Saikat.Sen.pdf
11. I. Santos, P. Galán-García, A. Santamaría-Ibirika, B. Alonso-Isla, I. Alabau-Sarasola, and Pablo G. Bringas, 2012. Adult Content Filtering through Compression-based Text Classication. CISIS/ICEUTE/SOCO Special Sessions, volume 189 of Advances in Intelligent Systems and Computing: 281-288.
12. Feng Jiao, Wen Gao, Lijuan Duan, Guoqin Cui, 2011. Detecting Adult Image using Multiple Features. Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on vol 3: 378 - 383.
13. Zhicheng Zhao, 2010. Combining multiple SVM classifiers for adult image recognition. Network Infrastructure and Digital Content, 2010 2nd IEEE International Conference on: 149 – 153.
14. Đỗ Thanh Nghị, 2012. Khai mỏ dữ liệu. NXB Đại học Cần Thơ.

15. Trần Cao Đệ, Phạm Nguyên Khang, 2012. Phân loại văn bản với Máy học vector hỗ trợ và Cây quyết định. Tạp chí Khoa học 2012 (21a): 52-63.
16. Nguyễn Thị Hoàn, 2010. Phương pháp trích chọn đặc trưng ảnh trong thuật toán Học máy tìm kiếm ảnh áp dụng trong bài toán tìm kiếm sản phẩm. Khóa luận Tốt nghiệp Đại học, Đại học Quốc gia Hà Nội: 13 – 20.
17. Ana P. B.Lopes, Sandra E.F.de Avila, Anderson N. A. Peixoto, Rodrigo S. Oliveira, Marcelo de M. Coelho and Arnaldo de A. Araújo, 2009. Nude Detection in Video using Bag-of-Visual-Features. Computer Graphics and Image Processing (SIBGRAPI): 224 – 231.
18. Ana P. B.Lopes, Sandra E.F.de Avila, Anderson N. A. Peixoto, Rodrigo S. Oliveira, Marcelo de M. Coelho and Arnaldo de A. Araújo, 2009. A Bag-of-Features Approach Based on HUE-SIFT Descriptor for Nude Detection. In Proceedings of the 17th European Signal Processing Conference, Glasgow, Scotland, 2009.
19. Steel, C.M.S, 2012. The Mask-SIFT Cascading Classifier for Pornography Detection. Internet Security (WorldCIS): 139 – 142.
20. Bag of visual words model: recognizing object categories. http://www.robots.ox.ac.uk/~az/icvss08_az_bow.pdf
21. Rob Fergus, 2012. Recognition - Bag of words models http://cs.nyu.edu/~fergus/teaching/vision_2012/9_BoW.pdf
22. L. H. Phương, N. T. M Huyền và V. L. Xuân, 2010. VnTokenizer 4.1.1 – Tách từ tiếng Việt http://vlsp.vietlp.org:8080/demo/?page=resources
23. Chang, C.C., Lin, C.J, 2001. LIBSVM – a library for support vector machines http://www.csie.ntu.edu.tw/~cjlin/libsvm
24. David Lowe, 2005. SiftDemoV4 – SIFT Keypoint Detector http://www.cs.ubc.ca/~lowe/keypoints/
25. T-N.Do, 2011. Detection of Pornographic Images Using Bag-of-Visual-Words and Arcx4 of Random Multinomial Naïve Bayes. 4th International Conference on Theories and Applications of Computer Science, vol.49 of Journal of Science and Technology, Special Issue on Theories and Application of Computer Science: 13 – 24.