ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
CHƯƠNG TRÌNH THẠC SĨ CNTTQM

Dương Đình Dũng

TÓM LƯỢC LUẬN VĂN

ĐỀ TÀI:

ỨNG DỤNG PHÂN LOẠI VĂN BẢN XÂY DỰNG BỘ LỌC WEB

Chuyên ngành : **KHOA HỌC MÁY TÍNH**
Mã số : 60 48 01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HƯỚNG DẪN KHOA HỌC:
Tiến sĩ Nguyễn Văn Hiệp

Thành phố Hồ Chí Minh – 2007

TÓM TẮT LUẬN VĂN CAO HỌC

Đề tài:

ỨNG DỤNG PHÂN LOẠI VĂN BẢN XÂY DỰNG BỘ LỌC WEB

Học viên thực hiện: **Dương Đình Dũng** Lớp: **Cao khọc khóa 1**

Giảng viên hướng dẫn: **TS. Nguyễn Văn Hiệp (ĐH BK TP. HCM)**

1. Tổng quan

Ngày nay, số người dùng Internet và các dịch vụ chạy trên Internet ngày càng nhiều và Ineternet được xem như là một phương tiện để tiếp nhận và truyền tải thông tin. Đặc biệt là Web và Mail, số người truy cập và sử dụng dịch vụ này nhiều nhất.

Tuy nhiên, cũng có những người sử dụng phương tiện Internet để truyền bá những thông tin không lành mạnh và cũng có những đối tượng tham gia vào việc truy cập những thông tin này.

Chính vì lý do này mà buộc các nhà quản trị mạng phải tìm cách ngăn chặn việc truy cập đến những trang web có nội dung không lành mạnh đó. Sự hình thành và phát triển các giải pháp lọc web ngày càng nhiều, trong đó xuất hiện nhiều hãng sản xuất phần mềm lọc web và có cả lý thuyết về công nghệ lọc web.

Bộ lọc Web có 2 ứng dụng lớn như sau:

– Bảo vệ chống truy cập nhưng nội dung bất hợp lệ: bộ lọc web được dùng để chống truy cập đến những trang có các hình ảnh, nội dung không lành mạnh được đặt ra bởi các quy định dùng Internet.

– Ngăn chặn việc lạm dụng mạng: đây là chức năng chống lại việc lạm dụng tài nguyên đường truyền của cơ quan đều làm những việc riêng như tải nhạc, phim, tài liệu không tốt… làm giảm năng suất hoạt động trên mạng của cơ quan.

2. Vấn đề nghiên cứu

– Trong đề tài này tác giả 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 của trang web, đánh giá thông qua một tập huấn luyện để đưa ra quyết định có cho truy cập trang web đó không?

– Nội dung chính là so sánh hai văn bản bằng công thức cosine và quan hệ giữa một văn bản với tập huấn luyện đang có gọi là hệ số trang.

3. Cách giải quyết vấn đề

3.1. Lý do chọn khai mỏ văn bản:

Chọn cách thức thực hiện bằng khai mỏ văn bản (phân loại văn bản) có một số ưu điểm sau:

– Phân tích nội dung trang web.

– Triển khai dễ dàng và hiệu quả.

– Cơ động trong vấn đề cập nhật thông tin cho tập dữ liệu mẫu.

3.2. Đặc điểm:

– Sử dụng tập huấn luyện để làm cơ sở so sánh.

– Sử dụng tập mẫu thử để xác định ngưỡng cho hệ thống

– 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.

– Dùng giá trị ngưỡng và hệ số trang để đưa ra quyết định.

3.3. Các nghiên cứu có liên quan

3.3.1. Blacklist và Whitelist:

Có hai danh sách riêng biệt các website 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.

3.3.2. 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 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.

3.3.3. Hệ thống phân loại (Rating systems):

Một hệ thống phân loại điển hình là PICS (Platform for Internet Content Selection) có thể thực hiện phân loại các Web site. Có 2 cách tiếp cận theo dạng phân loại các site, self-rating: cách này những trang Web được phát hành tự phát sinh thông tin phân loại của riêng chúng. Third-party rating, một sự phụ thuộc vào thành phần thứ ba dùng để đánh giá trang web và phát hành kết quả. Các thông tin này có thể dùng cho 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 phân loại nên seft-rating cũng không tin cậy và chính xác.

3.4. So sánh với các hướng tiếp cận khác

– 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 đánh lừa bằng cách cố ý đánh vần sai những keyword là kỹ thuật để vượt qua vấn đề này dẫn đến kết quả năng suất tính toán cao và gia tăng số lượng tích cực sai. Cuối cùng là các hệ thống phân loại (rating systems) không cung cấp thông tin đáng tin cậy.

4. Giải thuật

4.1. Mô tả giải thuật và cấu trúc dữ liệu

– 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" là thành viên của lớp đó sẽ bị chặn và những trang còn lại "không giống" sẽ cho qua.

Phần lớn những hệ thống phân loại văn bản truyền thống đòi hỏi tập huấn luyện gồm có hai lớp:

– Tập tích cực (positive) là các tài liệu có cùng đặc điểm với một lớp (lớp cấm)

– Tập tiêu cực (negative) là những văn bản không có cùng đặc điểm với một lớp (không phải lớp cấm).

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" nhiều lĩnh vực.

4.2. Trình bày giải thuật

* Vector hóa văn bản:

– Mỗi văn bản được biển diễn như một vector tần suất từ, độ dài của vector là N, vì thế chỉ có tần suất N những từ phổ biến sẽ được giữ lại. Sự giống nhau giữa 2 văn bản được đo bằng thuật ngữ COSINE của góc giữa hai vector, văn bản càng giống nhau thì góc càng nhỏ do đó COSINE sẽ lớn và ngược lại, văn bản càng xa nhau thì góc giữa hai vector càng lớn, COSINE của nó càng nhỏ.

* Công thức tính COSINE

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

Với X, Y là hai vector của hai văn bản

* Các bước thuật toán:

B1: Chuyển đổi thành vector: loại bỏ tag HTML, bỏ từ stopword, rút gọn từ (stemming), thống kê từ → vector tần suất.

B2: Tính ngưỡng cho hệ thống: dùng tập T' gồm các trang nằm bên trong và bên ngoài lớp cấm. Tính hệ số trang (xem bước 3) từng thành viên T' so với T. Sử dụng ngưỡng ứng viên tìm giá trị ngưỡng τ, là giá trị phân lớp T' đúng nhất theo hệ số trang đã tính.

B3: Tính hệ số trang P so với T và đưa ra quyết định:

– Tính $\cos(V_p, V_{Ti})$ với $\forall T_i \in$ T. Lưu vào dãy C

– Từ dãy C chọn ra n% giá trị cos cao nhất (n% phụ thuộc vào số phân lớp con trong T) → S

– Hệ số trang σ_p có được bằng cách tính trung bình cộng các giá trị trong S, theo công thức sau:

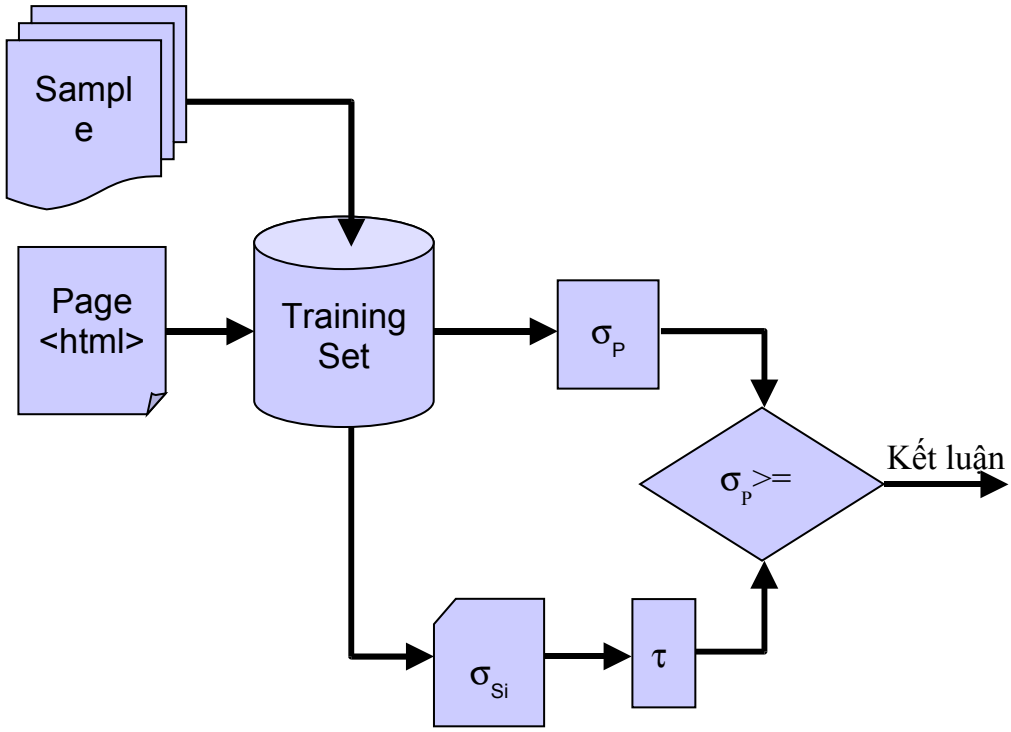
$$\sigma_p = \frac{\sum_{v \in S} v}{|T| \times n\%}$$

* So sánh và quyết định:

Nếu $\sigma_p >= \tau$ thì trang P sẽ bị cấm và bổ sung P vào trong T.

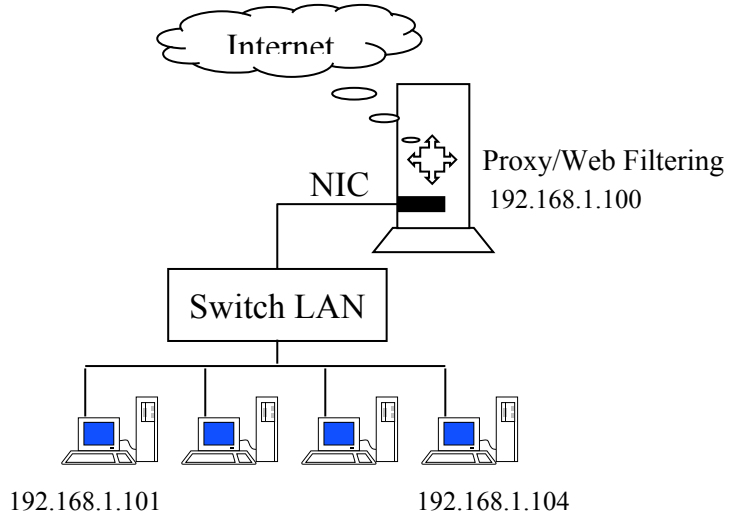
Ngược lại cho truy cập.

4.3. Sơ đồ thuật toán:



Sơ đồ thuật toán phân loại văn bản ứng dụng cho bộ lọc web.

4.4. Mô hình proxy trang bị bộ lọc web bằng phân loại văn bản



Mô hình mạng trang bị bộ lọc web với proxy web Filtering

4.5. Mô tả hoạt động của hệ thống

– Các máy trong mạng được điều chỉnh hướng về máy chủ đại diện (proxy server)

– Khi có một yêu cầu gởi lên từ một máy trong mạng, Proxy sẽ làm đại diện đi ra Internet để lấy trang web về và xử lý tại Proxy theo thuật toán lọc web bằng phân loại văn bản. Nếu thỏa điều kiện cho truy cập nó sẽ gởi quyền truy cập đến trang web đó về máy gởi yêu cầu. Ngược lại, một trang báo lỗi sẽ được gởi đến máy yêu cầu thông báo không truy cập được.

5. Cài đặt

5.1. Lưu đồ cài đặt có cải tiến

– Bổ sung thêm 2 tập liên kết (link hay URL) loại blacklist và whitelist

– Khi nhận một yêu cầu URL từ dưới gởi lên, hệ thống sẽ lấy URL đó tìm trong blacklist, nếu có sẽ gởi thông báo cấm truy cập đến client yêu cầu đó. Ngược lại hệ thống sẽ tìm trong whitelist, nếu URL đó có trong whitelist nó sẽ cho truy cập, ngược lại hệ thống sẽ tiến hành xét trang theo lưu đồ trong mục 4.3.

5.2. Ngôn ngữ cài đặt

– Chương trình được cài đặt bằng ngôn ngữ Java và biên dịch thành mã máy thi hành trong môi trường Windows.

– Cơ sở dữ liệu được dùng lưu trữ và làm cấu trúc dữ liệu cho chương trình là Access.

6. Thử nghiệm

6.1. Tư liệu thử nghiệm:

– 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**)

– Tập mẫu thử (sample): gồm 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. 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. Tổng số lượng $T_s' = T'_1 + T'_2 = 173 + 191 = 364$ trang

6.2. Phương pháp thử nghiệm:

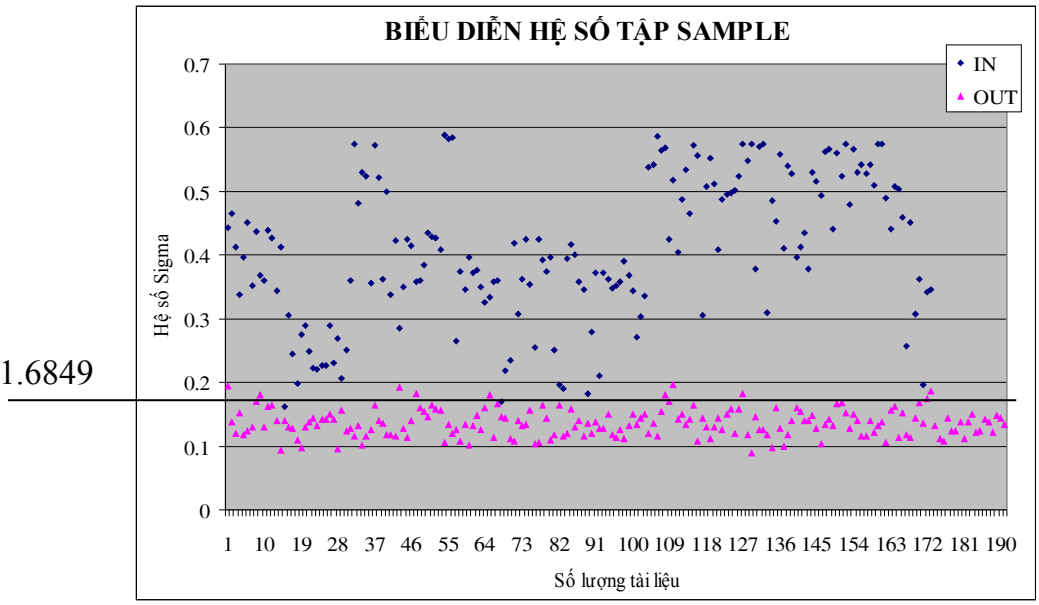
– Bộ phát sinh URL tự động gởi đến proxy: với một bộ dữ liệu T_s, sẽ cho ra một giá trị ngưỡng τ, chọn ngẫu nhiên n liên kết trong tập M liên kết có sẵn (n đủ lớn) cho qua proxy kiểm tra đánh giá về thời gian thực thi cũng như về hiệu quả làm việc của bộ lọc.

– Tổ chức bộ dữ liệu truy cập trên web server IIS: 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.

6.3. Biểu đồ phân lớp hệ số tương tự:

– Biểu đồ điểm (point) biểu diễn hệ số tương tự của tập T’ so với T (dựa trên tập dữ liệu giới thiệu trong mục 6.1).



Phân lớp hệ số trang các phần tử bên trong (IN) và bên ngoài (OUT) lớp cấm của tập mẫu thử so với ngưỡng hệ thống.

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%	

6.4. Công thức đo tỉ lệ:

Công thức tính:

* 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)

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

7. Đóng góp của luận văn

7.1. Giá trị thực tiễn

Làm cơ sở cho những nghiên cứu tiếp theo để hoàn thiện một bộ lọc web đa năng: có thể kiểm soát hình ảnh, download, xây dựng bộ lọc đa lĩnh vực.

Đưa ra một mô hình ứng dụng dựa trên cơ sở "Công nghệ Tri thức" có thể áp dụng vào mạng máy tính.

7.2. 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.

8. Kết luận

8.1. Kết quả đạt được:

– Xây dựng một bộ lọc nội dung web bằng phương pháp phân loại văn bản.

– Đưa ra mô hình lọc web trang bị cho một mạng LAN thông qua proxy kiểm soát thông tin bằng bộ lọc nội dung web.

– Xây dựng được ứng dụng minh họa có kiểm thử và đánh giá dựa trên bộ dữ liệu mẫu thử và tập huấn luyện.

– Luận văn trình bày một hướng tiếp cận trong lĩnh vực lọc web, người viết đã chọn một lĩnh vực để kiểm thử đó là trang web sex và không sex, kiểm thử thuật toán cũng như kiểm thử trên mạng thấy rằng hiệu quả ngăn chặn và thời gian kiểm thử để cho ra kết quả là có thể chấp nhận được.

– Còn một số vấn đề cần phải nghiên cứu bổ sung thêm cho thuật toán hoàn thiện hơn, cũng như mở rộng thuật toán cho nhiều lĩnh vực, triển khai trên các tường lửa lớn.

8.2. Khả năng ứng dụng

– Có thể cài đặt thành một proxy cho một mạng máy tính hay có thể biên dịch thành một ứng dụng có thể chạy trên máy đơn.

– Tích hợp vào tường lửa nguồn mở để trang bị cho mạng máy tính lớn hơn.

8.3. Hướng phát triển của đề tài

– Trang bị thêm cơ chế lọc hình ảnh, kiểm soát các tập tin download.

– Mở rộng: phát triển bộ lọc tiếng Việt, bằng cách xây dựng thêm kho tư liệu stoplist tiếng Việt, xây dựng danh mục nhóm từ, cơ chế phân tích ngữ nghĩa.

– Về thuật toán: cải tiến tốc độ làm việc bằng cách tăng cường thêm các Hueristic. Tập huấn luyện cần tối ưu hóa, chẳng hạn xây dựng thêm danh mục từ chuyên cho lĩnh vực (giảm số chiều của vector) hay dùng máy học xây dựng tập ngưỡng dùng để so sánh nhằm giảm thời gian tính toán.

– Mở rộng ứng dụng: nghiên cứu phát triển bộ lọc phân tán.

9. 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. Tính cosine: www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html (webpage).
14. WordHoard team - Comparing texts (wordhoard.northwestern.edu/ userman/analysis-comparingtexts.html).

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