NGHIÊN CỨU HỆ THỐNG THU THẬP VÀ TRÍCH RÚT THÔNG TIN TỪ DỮ LIỆU WEB THEO CHỦ ĐỀ

RESEARCH RETRIEVAL SYSTEM AND INFORMATION EXTRACTION FROM SUBJECT WEB DATA

Mai Mạnh Trừng

Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

Đến Tòa soạn ngày 03/4/2017, chấp nhận đăng ngày 08/5/2017

Tóm tắt: Trong bài báo, tác giả đã nghiên cứu các phương pháp thu thập dữ liệu trên web; thống kê được số lượng dữ liệu và thời lượng để lấy dữ liệu đó về trong việc thu thập dữ liệu web; trích chọn thành công dữ liệu cần khai thác; ứng dụng được cài đặt trên môi trường lập trình web động PHP. Kết quả ứng dụng hoạt động tốt, thu thập cũng như trích chọn thông tin từ dữ liệu trên web theo chủ đề một cách hiệu quả.

Từ khóa: Khai phá dữ liệu, khai phá dữ liệu web, thu thập thông tin, trích chọn thông tin.

Abstract: In this article, the author has studied the methods of collecting data on the web; they have analyzed statistically the quantity of data and the amount of time taken to perform the collection; the required data on the Web has been extracted successfully; the application has been installed in the environment of dynamic web programming with PHP. The application has performed well. It has gotten results in collecting and extracting data from all topics very effectively.

Keywords: Datamining, webmining, information retrieval, information extraction.

1. ĐẶT VẤN ĐỀ

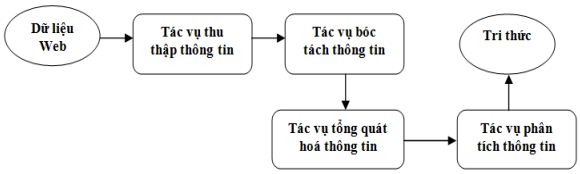
Trong thời đại tràn nghập thông tin, nhu cầu khai thác thông tin hiệu quả là vô cùng cần thiết. Có rất nhiều phương pháp để thu thập thông tin và trích rút thông tin. Phương pháp thu thập thông tin một cách thủ công mà nhiều người đang làm là họ vào các trang web muốn lấy thông tin hoặc vào các trang tìm kiếm sau đó lưu trữ chúng bằng cách download về hoặc người dùng copy nội dung vào một môi trường soạn thảo nào đó để lưu trữ. Đây là cách làm truyền thống nhưng cách làm này thực sự rất mất thời gian. Trong trường hợp chúng ta lấy thông tin theo chủ đề với số lượng dữ liệu lớn thì việc làm này là không khả thi. Do vậy, chúng ta cần thu thập dữ liệu trên web theo chủ đề bằng nhiều phương pháp tối ưu hơn. Việc bóc tách nội dung trên web thường được thực hiện bằng cách sử dụng các crawler hay wrapper. Một wrapper được xem như là một thủ tục được thiết kế để có thể rút trích được những nội dung cần quan tâm của một nguồn thông tin nào đó. Đã có một số công trình nghiên cứu trên thế giới sử dụng nhiều phương pháp tạo wrapper khác nhau để thực hiện rút trích thông tin trên web. Các phương pháp này bao gồm:

- Phân tích mã HTML.
- So sánh khung mẫu.
- Xử lý ngôn ngữ tự nhiên.

2. GIỚI THIỆU THU THẬP VÀ TRÍCH CHỌN THÔNG TIN

2.1. WebMining

Khai phá dữ liệu (DataMining) là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó. Với WebMining là khai phá dữ liệu trên môi trường web. Trên thế giới đã có rất nhiều nghiên cứu về WebMining, phương pháp và công nghệ để xử lý thông tin thu thập từ các nguồn thông tin trên Internet một cách thông minh. WebMining có thể được chia thành bốn tác vụ chính như hình 1.



Hình 1. Các tác vụ của Webmining [6]

2.2. Thu thập thông tin

Tác vụ thu thập thông tin giúp cho người sử dụng có được trang web từ URL hoặc từ yêu cầu mà họ cần. Đối với người sử dụng hiện tại, việc thu thập thông tin thường được thực hiện qua các URL mà người sử dụng đã biết hoặc qua các engine tìm kiếm. Các engine tìm kiếm là các chương trình được viết để có thể truy vấn và thu thập dữ liệu được lưu trong cơ sở dữ liệu (có cấu trúc), trang web (bán cấu trúc) và các văn bản tự do (không có cấu trúc) trên mạng. Hiện tại đã có khá nhiều các engine tìm kiếm mạnh ở thế giới và tại Việt Nam như Google, Altavista, Lycos, Vinaseek, PanVN,... Các engine này ngày càng cố gắng để có thể tương tác với người sử dụng nhiều và thông minh hơn [4].

Như chúng ta đã biết, một hệ thống thu thập thông tin lý tưởng phải là một hệ thống thu thập được những thông tin phù hợp nhất với yêu cầu của người sử dụng (yêu cầu này được diễn giải bằng các câu truy vấn). Đây thật sự là một tác vụ vô cùng phức tạp và khó khăn mà hầu hết các hệ thống thu thập thông tin đều chưa thực hiện được triệt để, phần nhiều có thể kể đến là do tính phi ngữ nghĩa của ngôn ngữ HTML. Hầu hết các hệ thống thu thập thông tin hiện nay đều chú trọng tới tốc độ, số lượng thông tin mà các hệ thống này có thể mang lại cho người dùng với các câu truy vấn tương đối đơn giản.

2.3. Trích chọn thông tin

Một khi thông tin sau khi qua tác vụ thu thập đã được lấy về, việc tiếp theo là phải lấy ra được những thông tin cần thiết và chỉ là những thông tin mà mình cần. Hầu hết các thuật toán bóc tách thông tin hiện nay đều dựa vào các công cụ khác nhau trên nền kỹ thuật "wrapper". Wrapper có thể được hiểu là những hàm để tách thông tin từ các tài nguyên web. Các hàm này được viết dựa trên các luật (quy luật) đã được đúc rút ra sau khi khảo sát các trang web chứa thông tin cần lấy. Các wrapper có thể xây dựng dựa trên rất nhiều quy luật khác nhau và tuỳ thuộc vào mục đích của người sử dụng [1], [2], [5]. Nhưng để trích chọn thông tin một cách hiệu quả các nhà phát triển dùng Crawler. Crawler phát triển rất mạnh chúng hỗ trợ nhiều hàm hiện đại giúp trích chọn thông tin một cách tối ưu hơn.

3. PHƯƠNG PHÁP TRÍCH CHỌN THÔNG TIN

3.1. Phân tích mã HTML

Đây là phương pháp chúng ta truy xuất trực tiếp vào nội dung toàn diện rồi tiến hành bóc tách. Sau đó những đặc tả dữ liệu (meta data) được xây dựng tự động trên nền nội dung đã bóc tách. Sau quy trình khai thác, nội dung sẽ trở thành độc lập với website nguồn, được lưu trữ và tái sử dụng cho những mục đích khác nhau [4], [5].

3.2. So khung mẫu

Phương pháp rút trích thông tin bằng cách so trùng hai trang web được xây dựng trên nền tảng nhận dạng mẫu thực hiện trong việc trích rút nội dung nhằm cung cấp tin tức trên trang web. Phương pháp này không đòi hỏi người sử

dụng phải biết về ngôn ngữ xây dựng wrapper hay phải thay đổi wrapper khi cách trình bày thay đổi do trang web mẫu có thể lấy trực tiếp từ trang chủ và có cùng cách trình bày với trang cần rút trích.

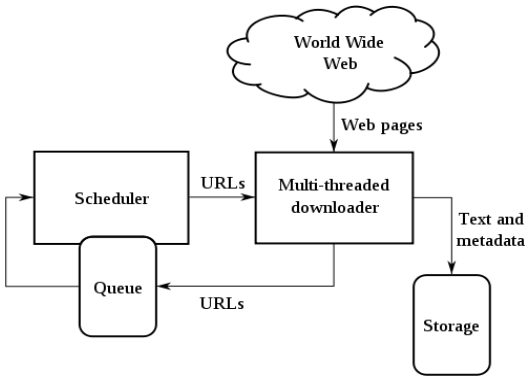
3.3 Xử lý ngôn ngữ tự nhiên

Đây là lĩnh vực các kỹ thuật xử lý ngôn ngữ tự nhiên được áp dụng cho những tài liệu mà thông tin trên đó thường không có một cấu trúc nhất định (như truyện). Các công cụ sử dụng phương pháp này thích hợp cho việc rút trích thông tin trên những trang web có chứa những đoạn văn tuân theo quy luật văn phạm. Một số công cụ sử dụng phương pháp xử lý ngôn ngữ tự nhiên trong việc bóc tách nội dung như: WHISK hay RAPIER [1], [2].

3.4. Web Crawler

Theo định nghĩa trên Wikipedia, thì Web Crawler, Web Spider hay Web Robot là một chương trình hoặc các đoạn mã có khả năng tự động duyệt các trang web khác theo một phương thức tự động, Web Crawler thường được sử dụng để thu thập tài nguyên (tin tức, hình ảnh, video…) trên internet.

Quá trình thực hiện của Web Crawler là Web Crawling hay Web Spidering. Hầu hết các công cụ tìm kiếm online hiện nay đều sử dụng quá trình này để thu thập và cập nhập kho dữ liệu phục vụ nhu cầu tìm kiếm của người dùng.



Hình 2. Kiến trúc Web Crawler

Web Crawler bắt đầu từ danh sách các địa chỉ URL được gọi là hạt giống (seeds), seeds được người dùng nhập vào, đây là những địa chỉ Web mà người dùng muốn thu thập thông tin. Hệ thống sẽ vào địa chỉ này lọc thông tin rồi tìm ra các địa chỉ URL khác (dựa vào những liên kết có bên trong các seeds). Sau đó thêm chúng vào danh sách các địa chỉ đã được duyệt qua gọi là Crawl frontier. Sau đó hệ thống lặp lại quá trình trước đó để duyệt qua những URL mới. Quá trình Crawling sẽ qua rất nhiều địa chỉ Website và thu thập rất nhiều nội dung khác nhau từ địa chỉ thu thập được.

4. CÀI ĐẶT CHƯƠNG CHƯƠNG TRÌNH ỨNG DỤNG

4.1. Yêu cầu thử nghiệm và tập dữ liệu thử nghiệm

- Yêu cầu thử nghiệm:

Mô tả bài toán: Đầu vào là đường link của lĩnh vực cần lấy dữ liệu; Đầu ra là nội dung chính của trang tin tức đã được lọc bỏ các thẻ HTML và các nội dung khác.

- Tập dữ liệu: vnexpress.net được biết đến như một tờ báo online có nhiều độc giả nhất Việt Nam. Ngoài tin tức thời sự, giáo dục, khoa học, Vnexpress còn mở rộng thêm một số các trang web con về công nghệ (thethao.vnexpress.net) và văn hóa giải trí (ngoisao.net); dantri.com.vn là trang web tin tức của Hội Khuyến học Việt Nam. Được thành lập sau vnexpress.net nhưng trang web đã nhanh chóng thu hút được nhiều độc giả vì sự cập nhật thông tin nhanh chóng và chính xác. Ngoài ra còn rất nhiều các trang web tin tức khác cũng có một số lượng độc giả đông đảo như thanhnien.com.vn hay vietnamnet.vn.

4.2. Cài đặt thử nghiệm ứng dụng

- Yêu cầu phần cứng và phần mềm:

Cấu hình phần cứng máy tính sử dụng để cài đặt chương trình.

Bảng 1. Phần cứng chạy ứng dụng

Thành phần	Chỉ số
CPU	Intel® Core™2 Duo 1.8 GHz
RAM	2 GB
OS	Windows 7 Professional
Bộ nhớ ngoài	200 G

Danh mục phần mềm sử dụng trong thực nghiệm:

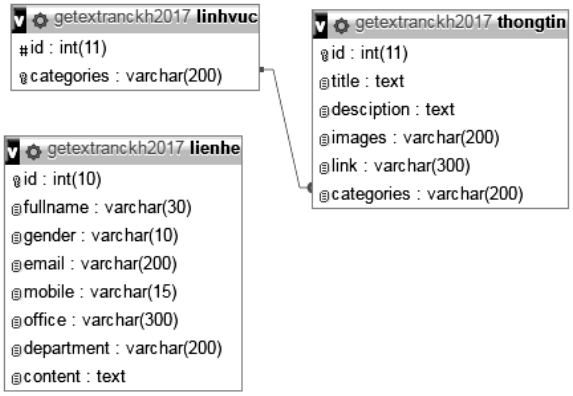
Bảng 2. Phần mềm xây dựng ứng dụng

STT	Tên phần mềm	Hãng
1	Dreawaver 8	Adobe
2	Dom HTML	W3C
3	Thư viện Crawler	Crawler
4	Apache, Mysql	Mã nguồn mở
5	Photoshop cs6	Adobe
6	Trình duyệt firefox	Mozilla

- Giới thiệu cấu trúc chương trình và một số module chính

Thiết kế cơ sở dữ liệu getextranckh2017.

Gồm 3 bảng: Bảng lienhe (liên hệ) để giúp người dùng có nhu cầu liên hệ trao đổi với nhà cung cấp; Bảng linhvuc (lĩnh vực) để lưu những lĩnh vực (chủ đề) muốn thu thập dữ liệu; Bảng thongtin (thông tin) để lưu trữ thông tin bài viết ở các lĩnh vực trên.



Hình 3. Cơ sở sở liệu ứng dụng

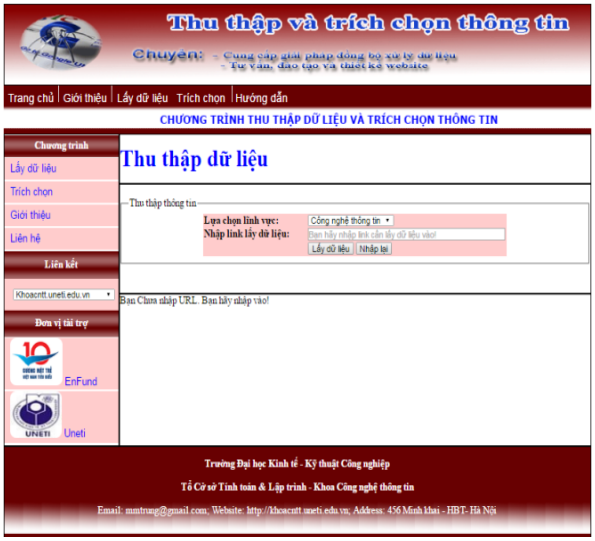
Các chức năng chính của chương trình bao gồm:

- Thu thập các URL cần trích rút nội dung là đầu vào của bài toán trích rút.

- Lưu nội dung này trích rút được vào cơ sở dữ liệu

- Hiển thị kết quả thu được ra màn hình.

- Đánh giá chung về kết quả thu được của chương trình thử nghiệm.



Hình 4. Giao diện trang thu thập dữ liệu

Giao diện này giúp người dùng lựa chọn lĩnh vực, sau đó nhập link (liên kết) cần lấy dữ liệu và kích vào nút Lấy dữ liệu thì dữ liệu sẽ được lấy về và lưu trong cơ sở dữ liệu.

Hình 5. Kết quả thu thập dữ liệu

Khi người dùng kích vào nút lấy dữ liệu thì dữ liệu sẽ được lưu trữ trong cơ sở dữ liệu, ngoài ra, hiển thị tiêu đề của bài viết và link tương ứng như hình 5. Phần phía dưới tổng hợp kết quả sau khi lấy dữ liệu như số link lấy về, kích thước dữ liệu, thời gian thực thi.

Khi dữ liệu lấy về từ web được lưu vào cơ sở dữ liệu. Tiếp theo người dùng muốn khai thác chúng như trích chọn thông tin hay truy vấn dữ liệu thì ta sẽ được kết quả tương ứng như hình 6.

Hình 6. Trang truy vấn và hiển thị dữ liệu

Hình 7. Giao diện trang liên hệ

Để giúp người sử dụng tăng tính tương tác, trao đổi kinh nghiệm cũng như có những thắc mắc thì người sử dụng sẽ nhập thông tin vào form liên hệ như hình 7. Giao diện này được thiết bằng các thẻ form của HTML5, kiểm tra thông tin nhập trước khi liên hệ. Khi thông tin đã nhập đầy đủ vào hợp lệ thì dữ liệu sẽ được lưu trữ trong cơ sở dữ liệu.

5. KẾT LUẬN

Với những khó khăn còn tồn tại trong việc thu thập và tách thông tin, người sử dụng, doanh nghiệp hay tổ chức luôn luôn phải mất rất nhiều tiền bạc, thời gian, công sức cho việc có được đúng những thông tin mình cần từ một kho thông tin khổng lồ là Internet. Sau quá trình nghiên cứu, khảo sát và xây dựng hệ thống, còn là những nghiên cứu tâm huyết nhằm xây dựng một hệ thống thu thập và trích rút thông tin một cách hiệu quả. Chương trình thử nghiệm chạy ổn định đã thống kế được số lượng link, tiêu đề, tính được dung lượng và thời gian thu thập dữ liệu từ web, dữ liệu lưu trữ vào cơ sở dữ liệu, truy vấn tốt dữ liệu. Tôi hy vọng rằng những nghiên cứu và sản phẩm của tôi sẽ được tiếp tục phát triển và thật sự có ích cho giảng dạy, nghiên cứu và những người tiêu dùng tại Việt Nam.

TÀI LIỆU THAM KHẢO

[1] Robert Meusel, Heiko Paulheim, *Linked Data for Information Extraction Challenge 2014*: LD4IE 2014 Linked Data for Information Extraction.

[2] Fabio Benedetti, Sonia Bergamaschi, Laura Po, *Online Index Extraction from Linked Open Data Sources*, LD4IE 2014 Linked Data for Information Extraction.

[3] J. Cowie and Y. Wilks, *Information extraction*, 2013.

[4] Tin Huynh, Kiem Hoang, *Automatic Metadata Extraction from sciencetific papers*. Proceeding of IT@EDU, Phan Thiet, VietNam, 2010.

[5] G. Pant, K. Tsioutsiouliklis, J. Johnson, C.L. Giles: *Panorama: Extending Digital Libraries with Topical Crawlers*. Proc. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004).

Thông tin liên hệ: **Mai Mạnh Trừng**

Điện thoại: 0912355022 - Email: mmtrung@uneti.edu.vn

Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp