Tạp chí Tin học và Điều khiển học, T.15, S.4 (1999), 50–54

PHƯƠNG PHÁP XÂY DỰNG CÔNG CỤ TRA CỨU VÀ TÌM KIẾM NHANH CÁC TRANG WEB TRÊN INTERNET/INTRANET

NGÔ TRẦN ANH

Abstract. Various aspects of design for a fast seeking and searching utility for Web pages, especially which are written in Vietnamese, are discussed. A brief description of moludes in an implementation of such utility is also given.

1. ĐẶT VẤN ĐỀ

Nhu cầu tìm kiếm nhanh trong quá trình tra cứu tại liệu điện tử nói chung và trên Internet/Intranet nói riêng trở nên bức thiết khi mà lượng thông tin ngày càng trở nên đồ sộ, đồng thời, nhiều chủ đề mới, thuật ngữ mới, khái niệm mới cũng phát sinh không ngừng, dẫn đến việc khai thác thông tin nhiều khi không thể theo những tiêu chí hay chủ đề định sẵn từ trước được. Máy chủ cung cấp dịch vụ phải làm việc hết công suất mỗi khi có những yêu cầu tìm kiếm như vậy. Các nhà cung cấp dịch vụ đều có công cụ trợ giúp khách hàng tìm kiếm thông tin, tối thiểu là trên Website của chính mình. Tuy nhiên, do đặc thù của ngôn ngữ tiếng Việt, việc tìm kiếm thông tin bằng tiếng Việt dùng công cụ có sẵn vốn không được thiết kế dùng cho tiếng Việt nên không mang lại kết quả như mong muốn.

2. PHÂN TÍCH YÊU CẦU

2.1. Yêu cầu về tổng thể hệ thống

Công cụ tra cứu và tìm kiếm nhanh một mặt phải tuân thủ những nguyên tắc hoạt động của Web Server để không làm ảnh hưởng tới hoạt động chung của toàn bộ Website. Do đó công cụ đó phải làm việc hài hòa trong tổng thể Kiến trúc Client/Server trên Internet (Inetnet Based Client Server Architecture) cũng như Kiến trúc đa phương tiện phân tán (Distributed Multimedia Architecture) vì nội dung trên Internet/Intranet ngày nay không thể tách rời văn bản với âm thanh và hình ảnh và những yêu cầu khắc nghiệt khác về thời gian xử lý trong hoàn cảnh đường truyền đang ngày càng trở nên quá tải. Mặt khác, để hoàn thành được mục tiêu "nhanh, đầy đủ và chính xác", công cụ tìm kiếm phải tận dụng hết tài nguyên của máy, bằng mọi cách để tìm ra được những nguồn thông tin cần thiết theo yêu cầu của khách hàng hay người sử dụng trong một thời gian càng nhanh càng tốt, một mặt để giải quyết tốt yêu cầu của người dùng, mặt khác cũng là để giải phóng tài nguyên máy cho người dùng khác với những yêu cầu mới. Đó cũng chính là mục tiêu mà kiến trúc đa phương tiện phân tán đặt ra trong thiết kế của mình (từ những vấn đề về cách đặt tên, chất lượng dịch vụ, đồng bộ hóa, phân chia thời gian, chuẩn hóa, cơ động,... đến những vấn đề lưu trữ, truy xuất thông tin và quản lý mạng).

2.2. Yêu cầu về tính mở

Những nghiên cứu trên lĩnh vực xử lý song song và hệ điều hành thời gian thực đang giành được sự quan tâm rất lớn [1]. Kết quả mới nhất về Máy song song ảo PVM (Parallel Virtual Machine) mở ra một khả năng cho phép tiến hành những giao tác song song trên mạng và tạo ra sức mạnh tổng hợp của toàn bộ máy tính tham gia vào mạng (Cuộc trình diễn của IBM với 17 máy

Pentium dùng Linux đọ sức với siêu máy tính Cray-1 là một ví dụ minh họa [7]). Một số mở rộng để một hệ điều hành đa năng (General Purpose OS) có một số khả năng của một hệ điều hành thời gian thực (Real-Time OS) được tung ra [6, 8, 9], thậm chí phiên bản thời gian thực tương thích với một hệ điều hành đa năng do cùng một nhà sản xuất đưa ra [10], chứng tỏ yêu cầu xử lý nhanh ngày càng trở nên cấp bách hơn, mặc dù tốc độ xử lý của CPU đã tăng vọt và kiến trúc đa xử lý cũng đã có những tiến bộ vượt bậc cùng với bước đột phá về kiến trúc cơ sở dữ liệu (Máy đánh cờ Big Blue là một minh chứng cụ thể [7]). Công cụ cũng phải đáp ứng yêu cầu về tính mở [1].

2.3. Yêu cầu về thực tế sử dụng

Sử dụng bất cứ trình duyệt Web nào:

Sử dụng trình duyệt Web có ưu điểm là việc khai thác hệ thống ứng dụng sẽ không bị phụ thuộc vào hệ điều hành. Trình duyệt Web là ứng dụng đang trở thành chuẩn trên thực tế. Từ một máy trạm dùng một hệ điều hành bất kỳ, bằng trình duyệt Web trên hệ điều hành đó, có thể truy xuất thông tin trên máy chủ chạy bất kỳ một hệ điều hành nào khác. Như vậy, chỉ cần tập trung phát triển hệ ứng dụng trên nền một hệ điều hành duy nhất (chẳng hạn như Windows NT 4.0) mà không cần phải phát triển công cụ khai thác trên máy trạm.

Hỗ trợ thêm những loại văn bản khác (ngoài HTML)

Đó là những văn bản được soạn bằng phần mềm chế bản điện tử thông dụng như WinWord, WordPerfect, Ventura,... Thậm chí chỉ là văn bản có dạng đơn giản (text thuần túy) được soạn bằng bất cứ công cụ nào (Edlin, C, Pascal, Basic, FoxPro, Bked,...). Mặc dù khả năng chuyển đổi về dạng chuẩn HTML là có, nhưng vì một lý do nào đó mà rất nhiều văn bản không nên chuyển đổi (như do font chữ, hình ảnh và âm thanh cùng nằm chung trong một tổng thể, văn bản được khóa bằng mật khẩu, v.v...).

Hỗ trợ nhiều bảng mã khác nhau

Mặc dù có công cụ chuyển đổi mã (do đặc thù ở Việt Nam tồn tại quá nhiều bảng mã) nhưng có lý do không nên chuyển đổi (việc chuyển đổi có thể làm mất thông tin hoặc thời gian/không gian lưu trữ, nhất là nếu văn bản đó đang được khai thác bởi nhiều ứng dụng khác nhau).

Lưu ý đến đặc điểm của ngôn ngữ tiếng Việt

* Không có quy ước chính tả thống nhất (hòa - hòa).
* Trình tự sắp xếp tiếng Việt (đặc biệt với tên riêng).

Xuất phát từ những yêu cầu trên, trong khuôn khổ xây dựng Hệ thống quản lý lưu trình công việc và bản tin trên mạng WAN [5], tác giả đã nghiên cứu về công cụ tra cứu tìm kiếm văn bản trên Internet/Intranet và đã phát triển thành công công cụ đó trên nền hệ điều hành Windows NT.

3. THIẾT KẾ

3.1. Những nghiên cứu và khảo sát chuẩn bị cho thiết kế các module

Những kỹ thuật làm tăng tốc độ truy xuất văn bản trên Internet

Chia nhỏ nội dung: Nội dung văn bản (kể cả hình ảnh) được chia thành những đơn vị nhỏ, sử dụng FRAMESET và FRAME, vừa dễ truyền đi trên mạng, vừa tạo điều kiện cho trình duyệt hiển thị từng phần của trang Web mà không cần phải chờ nhận hết cả trang. Việc chia nhỏ nội dung đến bao nhiêu và khi nào nên chia, khi nào không nên... dẫn đến một tình huống mà muốn giải quyết được cần phải giải một bài toán tối ưu hoặc fuzzy mà tác giả không đi sâu thêm trong khuôn khổ của bài báo này.

Caching: Sử dụng các lệnh "meta" liên quan đến thời hạn hiệu lực như "REFESH", "EXPIRES" để hỗ trợ trình duyệt Web lưu trữ một cách tối ưu và cập nhật có hiệu quả các văn bản liên quan với nhau [3]. Vấn đề này có thể đưa đến một bài toán fuzzy.

Những kỹ thuật làm tăng tốc độ xử lý

Tiền xử lý ngay từ máy trạm: Dùng JavaScript/JavaApplet để tiền xử lý những yêu cầu của người sử dụng nhằm loại bỏ tối đa khả năng lỗi về cú pháp hoặc những yêu cầu hàm chứa mâu thuẫn.

Khả năng song song hóa quá trình xử lý: Do đặc điểm rất quan trọng của quá trình tìm kiếm là tính độc lập giữa các văn bản, giữa các thư mục và giữa các máy tham gia vào mạng nên khả năng tiến hành song song các tiến trình là rất cao. Điều này có lợi cho việc mở rộng hệ thống khi cần thiết mà không ảnh hưởng đến hoạt động của toàn hệ.

Caching: Ngoài những tác dụng như đã trình bày trên đây, caching còn làm tăng tốc độ xử lý. Bản thân hệ điều hành thì có cơ chế cache đĩa (disk cache), phần cứng có hỗ trợ thêm cache bộ nhớ trong và bản thân CPU cũng có internal cache. Tuy nhiên, nếu có thêm phần cache do chính phần mềm ứng dụng tự quản lý dưới sự hỗ trợ của hệ thống thì tốc độ tìm kiếm sẽ nhanh lên rất nhiều (ví dụ: câu lệnh lặp lại ở một người sử dụng hay trùng lặp với người sử dụng khác thì kết quả cuối cùng được trả lời lại ngay nếu chưa có gì thay đổi trong hệ thống hoặc chỉ cần kiểm tra những gì thay đổi rồi kết hợp với kết quả trước thành kết quả mới [4]. Việc lựa chọn thuật toán cache tối ưu ở đây cũng dẫn đến một bài toán fuzzy mà tác giả không đi sâu trong khuôn khổ của bài báo này.

Các đặc thù của các tệp văn bản

Ngoài tệp dạng plain-text ra, các dạng văn bản thông dụng thường có ba loại như sau:

Dạng Word document: được soạn bằng WinWord hoặc WordPad. Nội dung được đặt nằm giữa phần header và định dạng, thường bị lặp lại và chồng chéo nếu văn bản được sửa chữa nhiều lần và được lưu với tùy chọn là Allow Fast Save.

Dạng Rich Text Format: cũng soạn bằng WinWord/WordPad nhưng thích hợp cho việc truyền qua E-mail với máy chủ chỉ hỗ trợ SMTP hoặc độ dài ký tự là 7 bit. Nội dung nằm xen lẫn các lệnh định dạng và điều khiển khác. Các ký tự tiếng Việt sẽ được lưu dưới dạng ký pháp \'xx với xx là mã hexadecimal của ký tự đó. Khi tìm kiếm những văn bản này cần phải chuyển đổi về dạng chuẩn.

Dạng siêu văn bản HTML: thường được soạn bằng Frontpage hoặc những công cụ phần mềm khác. Trên Internet/Intranet người ta dùng loại văn bản này là chủ yếu. Nội dung nằm xen lẫn những tag định dạng (giữa dấu < và > như <B> và </B> để chỉ chữ đậm,...). Ký tự tiếng Việt có thể được lưu dưới ba dạng khác nhau (xem giải thích thêm trong phần thiết kế). Đặc biệt trong văn bản thường có chứa những kết nối sang văn bản khác nằm cùng Website hoặc ở một Website khác. Phải đảm bảo cho những kết nối địa chỉ tương đối không bị lạc vị trí khi hiển thị văn bản trong kết quả tìm kiếm vì tiến trình tìm kiếm thường không nằm cùng thư mục với những văn bản được tìm (dùng khai báo meta "BASE HREF").

Các đặc thù của phiên bản và loại trình duyệt

Giữa Netscape Navigator và Internet Explorer có sự khác biệt về cách thể hiện cũng như khả năng kết nối (ví dụ: với Windows NT Challenge thì chỉ có Interner Explorer là vào được). Nhưng giữa các phiên bản khác nhau của Netscape Navigator hay Internet Explorer cũng khác nhau về chức năng về tiện ích. Điều này cần được lưu ý trong khi thiết kế để có thể chia ra từng trường hợp ứng xử cụ thể phù hợp cho từng trình duyệt khác nhau.

Các yêu cầu thường gặp từ phía người sử dụng

Tìm kiếm dưới dạng biểu thức lô-gích. Ví dụ: tìm văn bản có chứa "Hà Nội" *hoặc* "Thăng Long"...

Tìm kiếm *phân biệt / không phân biệt* chữ hoa / chữ thường.

Hiển thị toàn bộ kết quả hay mỗi lần chỉ hiện một số nhất định văn bản thỏa mãn để người sử dụng đỡ sốt ruột trong khi chờ đợi hoặc tự quyết định có tìm tiếp nữa hay không.

Tìm kiếm theo giới hạn về thời gian phát hành của văn bản.

Với những văn bản trong Hệ thống quản lý trình công việc và bản tin thì cho phép người dùng tra cứu ngược trở lại văn bản đó do đâu phát hành, ai ký,... như trong Hệ thống quản lý lưu trình công việc và bản tin, có lưu ý cơ chế bảo mật các tài liệu.

3.2. Thiết kế các module

Module duyệt các văn bản theo yêu cầu tìm

Duyệt tìm cây thư mục tính từ gốc ảo (vitual root directory) xuống các thư mục con từng file và áp dụng biểu thức tìm. Cách này cho phép tối thiểu thời gian đọc và chuyển đổi file, đồng thời cho phép tiến hành tìm kiếm song song theo cơ chế của PVM vì các file là những đơn vị độc lập. Quyền truy nhập và cơ chế bảo mật được tôn trọng.

Module phân biệt dạng văn bản

Việc phân biệt dạng văn bản không chỉ căn cứ vào kiểu file mà còn theo nội dung bên trong vì có thể có file kiểu là *.DOC nhưng chỉ chứa plain-text (*.TXT) hoặc Rich Text Format (*.RTF). ngoài ra, để xác định văn bản dùng bảng mã nào có thể dùng thuật toán fuzzy ([2]) mà ở đây, tác giả không đi sâu.

Module quy đổi dạng HTML về dạng chuẩn

Đây là vấn đề gây nhiều phiền phức nhất trong quá trình tìm kiếm vì trong trang văn bản dạng HTML có thể song song tồn tại cả ba dạng biểu diễn các ký tự mang dấu của tiếng Việt: dạng nguyên mã (ví dụ: ấ), dạng thực thể ký tự - character entity (ví dụ: Ê) và dạng mã số (ví dụ: Ê). Như vậy, các từ **Tuấn** hay **TuÊn** và **TuÊn** thực chất là một. Mặt khác, còn những tag định dạng trang (nằm giữa < và >, hoặc là comment) phải bỏ qua trong khi so sánh. Đặc biệt, white space trong trang văn bản dạng HTML cũng được biểu diễn bằng nhiều dạng khác nhau (nhiều dấu cách liên tiếp được coi là một, dấu Carriage Return và Line Feed đứng một mình hay cả hai đều coi là một dấu cách). Do đó, không thể áp dụng cách tìm bằng đối sánh từng byte với biểu thức tìm được mà trước hết phải quy đổi văn bản và biểu thức tìm về nội dung dạng chuẩn.

Module phân tích cú pháp của câu lệnh tìm kiếm và phân rã thành những biểu thức tìm primitive cùng với các toán tử logic (AND, OR, NOT)

Bằng đoạn trình viết bằng Javascript trong trang Web đầu tiên, câu lệnh tìm kiếm đã được kiểm tra sơ bộ về cú pháp nên khi được chuyển sang máy chủ, câu lệnh được phân rã thành những biểu thức tìm primitive (không thể phân rã hơn được nữa). Từng biểu thức đó được tìm kiếm trong từng tệp (do ba modules ở trên duyệt và chọn ra) để rồi kết hợp với nhau bằng những toán tử lô-gich. Tuy nhiên, ở đây có áp dụng qui tắc tính tắt (short cut) để tiết kiệm thời gian (chỉ cần tìm thấy một trong các biểu thực primitive nối với nhau bằng OR thỏa mãn thì quá trình tìm kiếm dừng và cho kết quả là tìm thấy). Mặt khác, việc tìm kiếm được thực hiện theo phương thức "qui đổi về dạng chuẩn đến đâu, thực hiện tìm kiếm ngay đến đó" nên tiết kiệm được thời gian do không phải đổi lại với từng biểu thức primitive.

Module đưa ra kết quả

Trả lại kết quả tìm kiếm dưới dạng HTML cho trình duyệt WEB, riêng với loại văn bản có dạng HTML thì cho phép chỉ rõ (highlight) những vị trí thỏa mãn điều kiện tìm và thực hiện những chuyển đổi cần thiết cho phù hợp với ký pháp chuẩn của HTML (điền thông tin meta để các hyperlink tương đối không bị lạc địa chỉ; đổi "\" thành "/", ký tự đặc biệt thành dạng &#nnn;...).

Module phục vụ theo dõi và quản lý truy xuất thông tin (nếu cần thiết)

Module này không còn cần thiết nữa trên nền Widown NT - Sercive Pack 3 - Option Pack 4.0 nên tác giả không đi sâu thêm. Mục đích của module là nhằm theo dõi và quản lý tất cả những dữ liệu liên quan đến truy xuất thông tin, phục vụ cho công tác thống kê, để từ đó có thể điều chỉnh những bất hợp lý trong phân phối tài nguyên, đồng thời khai thác được thông tin về những gì đang

được quan tâm... để kịp thời bổ sung những thông tin mới hay dọn dẹp những thông tin không còn giá trị, v.v...

4. KẾT LUẬN

Sau khi thử nghiệm công cụ được phát triển trên cơ sở những phân tích và thiết kế đã trình bày ở trên tại một đơn vị ứng dụng Hệ thống quản lý lưu trình công việc và bản tin [5], cho thấy công cụ phục vụ tốt cho công tác tra cứu và tìm kiếm văn bản, kể cả văn bản do những ứng dụng khác tạo ra (như Lotus Notes). Còn nhiều vấn đề mới nảy sinh, trong đó có việc tìm kiếm theo âm tiết, tìm kiếm theo từ đồng nghĩa... Ngoài ra, còn nảy sinh yêu cầu cần phải cấm sao chép, in ấn khi xem những tài liệu mật dẫn đến việc phải xây dựng một trình duyệt riêng. Chúng tôi đang tiếp tục nghiên cứu và tìm cách giải quyết.

TÀI LIỆU THAM KHẢO

[1] Cooling J. E., *Software Design for Real-time Systems*, Chapman and Hall, 1991.

[2] Nguyễn Cát Hồ, Trần Đình Khang, Trần Hoàng Yến, et al. - Hệ suy luận ngôn ngữ LINGRES - Báo cáo đề tài 1993, Viện Công nghệ thông tin.

[3] Eric Ladd, Jim O'Donnel, Using HTML 3.2., Java 1.1. and CGB - Que Corp., 1996.

[4] Vũ Duy Lợi, Kỹ thuật caching trên WEB server, Tài liệu seminar, 1999.

[5] Nguyễn Văn Tam et al., Hệ thống quản lý lưu trình công việc và bản tin trên mạng WAN.

[6] INtime (real-time Windows NT), iRMX - Radisys Corp., Tài liệu trên internet.

[7] PCWorld Vietnam, tháng 4/1999, trang 27.

[8] Real-time Extension (RTX) for Windows NT - VenturCom, Inc, Tài liệu trên internet.

[9] RTXDOS - Technosoftware AG, Tài liệu trên internet.

[10] Windows CE ,Microsoft Corp., Tài liệu trên internet.

Nhận bài ngày 12 - 11 - 1998
Nhận lại sau khi sửa ngày 2 - 4 - 1999

Viện Công nghệ thông tin