XÂY DỰNG VÀ PHÁT TRIỂN MỘT HỆ THỐNG SEARCH ENGINE HỖ TRỢ TÌM KIẾM THEO TỪ KHÓA TIẾNG VIỆT

Trương Mỹ Dung, Đỗ Hoàng Cường
Khoa Công nghệ thông tin, Trường Đại học Khoa học tự nhiên - ĐHQG-HCM
(Bài nhận ngày 20 tháng 10 năm 2003, hoàn chỉnh sửa chữa ngày 31 tháng 12 năm 2003)

TÓM TẮT: *Bài báo nhằm trình bày vấn đề xây dựng và phát triển một Hệ thống Search Engine (S.E) tìm kiếm theo từ khóa Tiếng Việt.*

1. Mở đầu

Hiện nay, InterNET đã trở thành một Siêu Xa lộ Thông tin, cung cấp thông tin cho mọi người, ở mọi nơi, trong mọi ngành, mọi lãnh vực. Hiện nay trên thế giới có rất nhiều SEARCH ENGINE chẳng hạn như GOOGLE (xem [2], [3], [5]), YAHOO, ALLTHEWEB, ALTA VISTA (xem [4]),... có khả năng tìm kiếm trên nhiều ngôn ngữ khác nhau, nhưng với Tiếng VIÊT vẫn có hạn chế. Và trong nước cũng có vài SEARCH ENGINE chẳng hạn như NETNAM (xem [6]), VINASEEK (xem [7]).

Vấn đề đặt ra là chọn các phương pháp và công cụ nào thích hợp cho việc tìm kiếm và chọn lọc thông tin tiếng Việt cần thiết theo từ khóa sao cho đáp ứng được nhu cầu, trong thời gian cho phép. Mục tiêu chính của bài báo nhằm trình bày Hệ thống SEARCH ENGINE (S.E) tìm kiếm theo từ khóa Tiếng Việt, dựa trên ứng dụng công nghệ ORACLE TEXT.

2. Giới thiệu và so sánh một số S.E thông dụng hiện nay

2.1. Bảng tóm tắt đặc trưng một số S.E trong và nước ngoài

Trích từ tài liệu thống kê của Infopeople Project, cập nhật ngày 29/01/2003 (xem [11]), ta có đặc trưng của một số S.E nước ngoài theo Bảng sau:

BẢNG 1. TÓM TẮT ĐẶC TRƯNG CỦA MỘT SỐ S.E NƯỚC NGOÀI

S.E	Cơ sở d ữ li ệu	Toán tử Logic	Khả năng tìm kiếm	Hỗn hợp
Google **www.google.com** Hỗ trợ tìm kiếm nâng cao. Hệ thống danh mục theo đề mục, mở ()	Toàn bộ văn bản của các trang Web, PDF, MS.Office, PostScript, Lotus WordPerfect, ...	AND *(mặc định)* OR.	Dùng * để rút gọn (thay thế từ trong cụm từ). Dùng " " tìm cụm t ừ Tìm theo vùng Tìm các trang liên quan	Hỗ trợ kiểm tra chính tả. Tìm hình ảnh. TIn tóm tắt. Tìm kiếm theo nhiều ngôn ngữ khác nhau.
AllTheWeb **www. allTheWeb. com** Hỗ trợ tìm kiếm nâng cao. Không có Hệ thống danh mục theo đề mục	Toàn bộ văn bản của các trang Web, PDF, MS.Office, PostScript, Lotus WordPerfect, ...	AND *(mặc định)* OR, ANDNOT, RANK	Không rút gọn, Dùng " " tìm cụm từ Tìm theo vùng.	Tìm kiếm mở rộng: hình ảnh và Video.

AltaVista.com Hỗ trợ tìm kiếm nâng cao (rất tốt) Hệ thống danh mục theo đề mục, mở)	Toàn bộ văn bản của các trang Web.	AND *(mặc định)* OR.	Dùng " " tìm cụm từ . Dùng * để rút gọn. Phân biệt chữ Hoa, chữ thường.	Tự động xác định cụm từ và kiểm tra chính tả. Tìm hình ảnh, âm thanh, Video và tin tức. Tìm kiếm theo nhiều ngôn ngữ khác nhau.
Teoma **teoma. com** Hỗ trợ tìm kiếm nâng cao. Phân loạI dựa trên # tiêu đề cụ thể của các trang liên kết	Toàn bộ văn bản của các trang Web.	AND *(mặc định)* Dùng ""để tìm các từ thông dụng.	Không rút gọn Dùng " " tìm cụm từ . Tìm kiếm theo các giớI hạn như ngày, ngôn ngữ, vị trí, độ sâu liên kết của trang WEB.	Có thể gom nhóm các kết quả.

BẢNG 2. TÓM TẮT ĐẶC TRƯNG CỦA MỘT SỐ S.E TRONG NƯỚC.

S.E	Cơ sở d ữ li ệu	Toán tử Logic	Khả năng tìm kiếm	Hỗn hợp
NetNam . Com	Toàn bộ văn bản của các trang Web.	Dùng ""để tìm các từ thông dụng.	Dùng " " tìm cụm từ . Phân biệt chữ Hoa và thường.	Sử dụng từ khóa để lọc các Tìm kiếm.
VinaSeek vinaseek. Com	Toàn bộ văn bản của các trang Web.	Dùng ""để tìm các từ thông dụng.	Dùng " " tìm cụm từ . Phân biệt chữ Hoa và thường.	Sử dụng từ khóa để lọc các Tìm kiếm.

2.2. So sánh và nhận xét

- **Giống nhau**

Các S.E đều được thiết kế theo cùng một qui trình. Nhưng mỗi S.E có giải pháp xử lý khác nhau nên có thể cho kết quả khác nhau. Hiện nay ngày càng nhiều các S.E kết hợp dịch vụ thư mục Web vào trong trang Web của họ.

- **Khác nhau**

Phương pháp lập chỉ mục tự động của các S.E là khác nhau; chẳng hạn Yahoo lập chỉ mục bằng cách dùng phần mềm con nhện bò khắp nơi trên mạng, tạo sự liên kết khá tốt. Chất lượng Lập chỉ mục sẽ ảnh hưởng đến chất lượng việc truy tìm. Một điểm khác biệt lớn giữa các S.E là việc sắp xếp lại các kết quả tìm kiếm được. Các S.E sau khi tìm được những kết quả sẽ thực hiện tác vụ lọc bớt những kết quả trùng hay những kết quả có độ chính xác kém. Sắp xếp các kết quả này theo một trật tự nào đó, như theo độ chính xác của tài liệu.... Mỗi S.E có một cơ sở dữ liệu khác nhau và chiến lược xử lý kết quả khác nhau nên kết quả trả về cho người sử dụng cũng rất khác nhau.

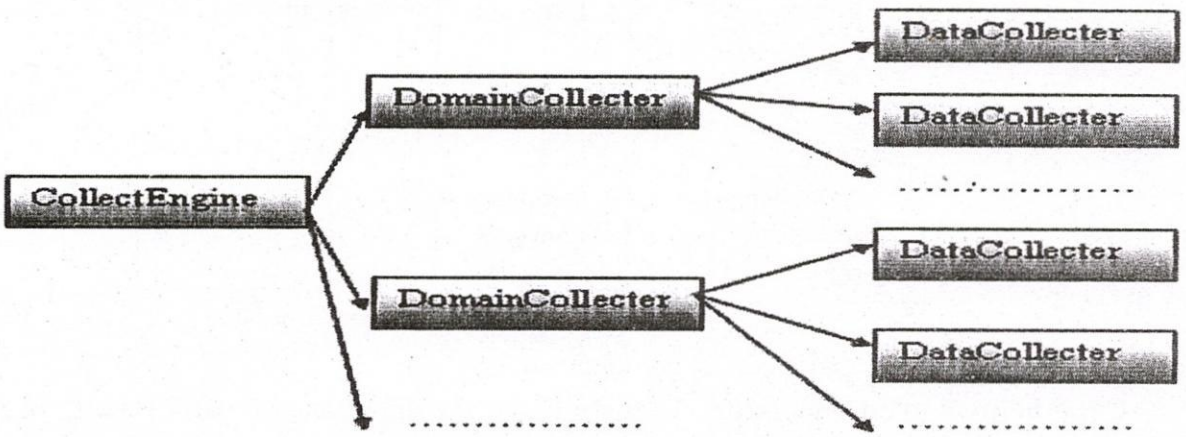
- **Nhận xét**

Hầu hết các S.E lập chỉ mục "bằng tay" đều mang lại kết quả tốt hơn so với lập chỉ mục tự động. Độ đo quan trọng nhất để đánh giá hiệu quả hoạt động của một S.E là chất lượng của kết quả tìm kiếm. Các kết quả hợp lý là các trang chất lượng cao, không có các liên kết bị gãy. Ngoài chất lượng tìm kiếm, một khía cạnh của yêu cầu lưu trữ cần quan tâm là phải sử

dụng hiệu quả bộ nhớ. Ngoài ra cần phải lưu ý đến hiệu quả dò tìm và lập chỉ mục và thời gian tìm kiếm; chẳng hạn, với phiên bản hiện nay của Google trả lời hầu hết các truy vấn từ 1 đến 10 giây.

3. **Thiết kế hệ thống.** Hệ thống gồm ba module:
   1. Module Web Robot. Chức năng chính là thu thập dữ liệu từ các trang Web trên các Website được viết bằng tiếng Việt.
   2. Module ConvertFile. Công việc chính của module này là **thống nhất** lại toàn bộ dữ liệu trước khi chuyển cho Oracle TEXT thực hiện việc lập Chỉ mục (Index). Việc thống nhất này bao gồm:
      a. Chuyển toàn bộ các định dạng (PDF, DOC, EXCEL...) sang dạng HTML.
      b. Chuyển mã của trang (TCVN3, VNI, Unicode...) sang bảng mã Windows-CP1258 (Unicode tổ hợp).
      c. Lưu dữ liệu vào trong CSDL Oracle để index.
   3. Modul truy vấn.

3.1. **Module WebRobot.** Chức năng chính là thu thập dữ liệu từ các trang Web được viết bằng tiếng Việt. Sơ đồ hoạt động của modul này được mô tả như sau:



- **CollectEngine**.Đây là thành phần điều khiển quá trình thu thập thông tin từ Internet. Nó sẽ tạo ra các DomainCollector và quản lý việc truyền tảI (download). Khi CollectEngine hoàn tất công việc cũng là lúc module này kết thúc.
- **DomainCollector.** Có nhiệm vụ thu thập tất cả các trang thông tin bên trong một miền (domain) về.
  - **Đầu vào**: tên miền mà hệ thống cần thu thập về.
  - **Hoạt động:** DomainCollector lấy danh sách tất cả các trang Web đã được đánh dấu thuộc về miền nhưng chưa lấy về. Kế tiếp, DomainCollector sẽ tạo ra các DataCollector và cho các DataCollector này thực hiện việc lấy các trang Web về.
  - **Đầu ra**: miền được tải về.

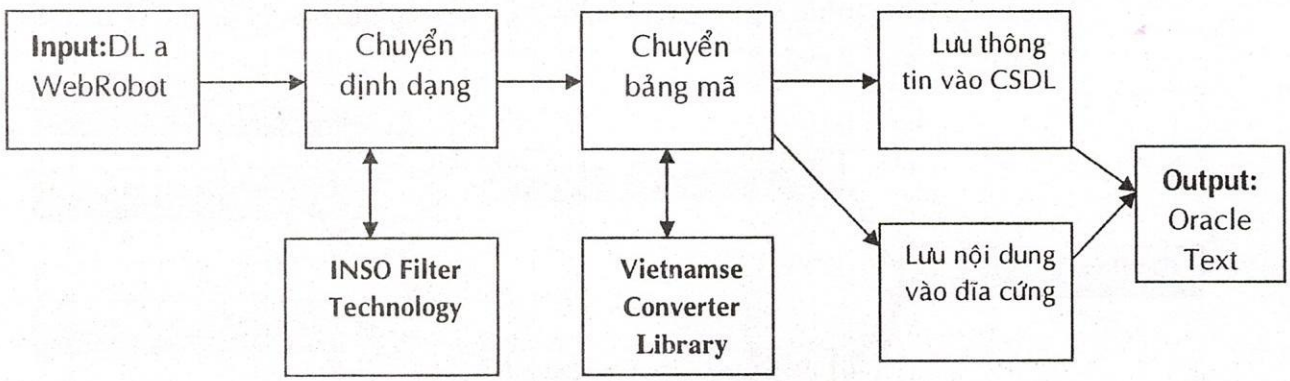
Ngoài ra, DomainCollector còn lưu trữ một Collect Engine (là robot cha của nó) để thực hiện việc giao tiếp.

- **DataCollector.** Bộ thu thập thông tin (các trang WEB). Mục tiêu của DataCollector là thu nhận thông tin từ Internet về, sau đó sẽ lưu trữ chúng xuống cơ sở dữ liệu, đồng thời cũng ghi những dữ liệu này xuống hệ thống tập tin.

- **Đầu vào**: địa chỉ trang WEB cần thu thập.
- **Hoạt động**: Đầu tiên, DataCollector sẽ tạo một kết nối đến domain của trang WEB, sau đó kiểm tra tình trạng của kết nối này. Nếu kết nối ở tình trạng tốt, DataCollector sẽ lấy các thông số về trang WEB như: ContentLength (độ lớn của trang WEB), LastModified (ngày cập nhật cuối cùng của trang WEB) và ContentType (loại của nội dung trang WEB). Nếu ContentType thuộc một trong các loại sau: application/pdf, application/vnd.ms-excel, application/msword, application/msexcel, application/ms-powerpoint, application/postscript, text/rtf, text/html thì mới lấy về. Sau khi lấy về, thì ghi trang WEB này xuống hệ thống tập tin.
- **Đầu ra**: Trang WEB được lấy về.

DomainCollector: robot cha của nó, dùng để thực hiện việc giao tiếp trong khi hoạt động.

3.2. **Module ConvertFile.** Công việc chính của module này là **thống nhất** lại toàn bộ dữ liệu trước khi chuyển cho Oracle TEXT thực hiện việc lập Chỉ mục (Index). Toàn bộ công việc của module này được mô tả qua sơ đồ sau:



Hoạt động của Module này gồm các công đoạn sau:

1. **Công đoạn chuyển định dạng. Chuyển** tất cả các định dạng về dạng HTML, bằng cách sử dụng công nghệ "INSO filter" để lọc tài liệu.
2. **Công đoạn chuyển mã.** Sử dụng UNICODE tổ hợp trong việc lưu trữ và xử lý ngôn ngữ. Xây dựng một thuật giải để nhận dạng bảng mã và sau khi đã xác định bảng mã thực hiện việc chuyển mã.

Thuật giải nhận dạng bảng mã.

- **Các đặc trưng của bảng mã:**

Đặc trưng đầu tiên dùng để nhận biết văn bản là UNICODE. Một văn bản là UNICODE nếu 2 byte đầu của văn bản là FFFEh hoặc FEFFh. Đặc trưng thứ hai của bảng mã là số byte lưu trữ ký tự trong bảng mã. Trong một bảng mã, một ký tự có thể được lưu 1 byte, 2 bytes, 3 bytes, hay có khi là 4 bytes, hoặc biến động, chẳng hạn VNI Windows 2byte, TCVN3 1byte. Đặc trưng thứ ba quan trọng nhất, để nhận biết bảng mã chính là sự ánh xạ khác nhau của ký tự. Ví dụ: chữ "Â" trong bảng mã TCVN3 có giá trị 162, còn trong bảng mã VPS có giá trị là 194. Đa số các ký tự trong các bảng mã khác nhau sẽ có ánh xạ khác nhau, tuy nhiên có một số ký tự khác nhau của các bảng mã khác nhau lại có cùng một giá trị ánh xạ (ví dụ: chữ 'oả' có giá trị trong bảng mã UNICODE là E16F, và mã này sẽ trùng với chữ 'õ' trong bảng mã VNI Windows...). Số lượng những ký tự như vậy, tuy không nhiều nhưng sẽ gây ra sự

nhập nhằng trong việc nhận dạng bảng mã. Để khử sự nhập nhằng này, chúng tôi sử dụng một Heristic sau.

"Khi có sự nhập nhằng về bảng mã của ký tự, ký tự sẽ thuộc về bảng mã của ký tự trước đó".

Với heuristic này, độ chính xác sẽ không thể là 100%, nhưng do hầu hết các văn bản đều sử dụng một bảng mã trong toàn bộ văn bản, nên độ chính xác, theo thống kê của chúng tôi là khoảng >95%.

- **Thuật giải nhận dạng bảng mã:**

B1: Với mỗi bảng mã, làm công việc từ B2, B5.
B2: Đọc một ký tự C ứng với bảng mã.
B3: Nếu đây C là ký tự Western thì qua B6.
B4: Nếu C là ký tự hợp lệ và số byte đọc được >= 1, qua bước 6.
B5: Nếu phát hiện sự nhập nhằng, lấy bảng mã trước rồi ánh xạ C.
B6: Lưu lại bảng mã phát hiện được.
B7: Đưa ký tự C vào kết quả. Quay lại bước 1.

- **Thực hiện việc chuyển mã:**

Sau khi đã xác định được bảng mã, chúng ta đã có thể thực hiện việc chuyển mã. **Ghép nối hai công đoạn chuyển định dạng và chuyển bảng mã,** và lưu dữ liệu vào trong CSDL Oracle để lập chỉ mục.

4. KẾT LUẬN.

Đã xây dựng một Hệ thống S.E theo thiết kế ở Phần 3. Hệ thống S.E này đã được thử nghiệm, kết quả bước đầu như sau:

- Module WebRobot: Việc thu thập và phân tích các trang Web đã được thử nghiệm trên hơn 500 miền khác nhau. Tổng số trang Web tải về trên 80000 trang. Tỷ lệ thành công của WebRobot vào khoảng 96%.
- Module ConvertFile: Số lượng tài liệu chuyển đổi: 22495 tài liệu (HTML, Tài liệu Word (.doc), PDF):
  - Tổng thời gian : 30 phút
  - Số lượng lỗi : 10
- Module Truy vấn:
  - Số lượng tài liệu: 22495 tài liệu.
  - Thời gian truy vấn trung bình cho một tài liệu: 0.1 giây

Đây chỉ mới là kết quả bước đầu trong việc thiết kế và xây dựng một Hệ thống S.E tìm kiếm thông tin theo từ khóa Tiếng Việt. Ngôn ngữ Tiếng Việt rất là phức tạp. Việc xử lý ngữ nghĩa không phải là vấn đề đơn giản. Hơn nữa việc thống nhất trong ngữ pháp Tiếng Việt là một vấn đề còn tranh cải (xem Chi tiết trong [8], [9], [10]). Hướng Phát triển tiếp theo là tiếp tục nghiên cứu sâu hơn để có thể phát triển một Hệ thống S.E cho phép ngườI sử dụng tìm kiếm nhanh thông tin qua từ khóa Tiếng Việt có chú ý đến ngữ nghĩa qua InterNET.

DESIGNING & DEVELOPPING A SEARCH ENGINE BY VIETNAMESE KEYWORDS

Truong My Dung, Do Hoang Cuong
Faculty of IT, University of Natural Sciences, VNU-HCM

ABSTRACT: This word aim to present on the Design of SEARCH ENGINE (S.E) for Searching by Vietnamese Keywords.

TÀI LIỆU THAM KHẢO

[1] *http://www.infopeople.org/search/chart.html*
[2] The Anotomy of a large scale search engine, Google 2001.
[3] *http://www.google.com/webm,aster/4.html*
[4] *http://altavista.com*
[5] *http://www.google.com*
[6] *http://www.panvietnam.com*
[7] *http://www.vinaseek.com*
[8] Nguyễn Thiện Giáp. Từ và nhận diện từ Tiếng Việt NXB Giáo dục Hà Nội, 1996.
[9] Nguyễn Kim Thản. Nghiên cứu ngữ pháp Tiếng Việt. NXB Giáo Dục, 1997
[10] Nguyễn Tài Cẩn. Ngữ Pháp tiếng Việt. NXB ĐạI học Quốc gia Hà Nội, 1998.
[11] Huỳnh Thụy Bảo Trân. Nghiên cứu một số Mô hình Xây dựng Thử nghiệm một SEARCH ENGINE Tiếng Việt, Luận văn Thạc sĩ CNTT, ĐHKHTN, ĐHQG-HCM, 2002.
[12] Report Projet Mitani 1: Vietnamse Search Engine. Truong My Dung, 12/2003