Biên soạn: TS. TRƯƠNG QUỐC ĐỊNH (Chủ biên)
TS. NGÔ BÁ HÙNG - TS. TRƯƠNG QUỐC BẢO

GIÁO TRÌNH

CÁC HỆ THỐNG TÌM KIẾM THÔNG TIN VĂN BẢN

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ
2014

BIÊN MỤC TRÊN XUẤT BẢN THỰC HIỆN BỞI
TRUNG TÂM HỌC LIỆU TRƯỜNG ĐẠI HỌC CẦN THƠ

Trương, Quốc Định

Giáo trình các hệ thống tìm kiếm thông tin văn bản / Trương Quốc Định, Ngô Bá Hùng, Trương Quốc Bảo .– Cần Thơ : Nxb. Đại học Cần Thơ, 2014

136 tr. : minh họa ; 24 cm
Sách có danh mục tài liệu tham khảo

1. Information systems
2. Hệ thống tìm kiếm thông tin
I. Nhan đề
II. Ngô, Bá Hùng
III. Trương, Quốc Bảo

004.67 – DDC 22
MFN 189460
Đ312

LỜI GIỚI THIỆU

Nhằm góp phần làm phong phú nguồn tư liệu phục vụ nghiên cứu, học tập cho bạn đọc trong và ngoài ngành Công nghệ thông tin, Nhà Xuất bản Đại học Cần Thơ ấn hành và giới thiệu cùng bạn đọc giáo trình "Các hệ thống tìm kiếm thông tin văn bản" do Tiến sĩ Trương Quốc Định, Tiến sĩ Ngô Bá Hùng và Tiến sĩ Trương Quốc Bảo biên soạn.

Giáo trình gồm 05 chương, giáo trình giúp người đọc hiểu được tìm kiếm thông tin là gì; chỉ mục tài liệu và kết quả tìm kiếm, các cách tìm kiếm thông tin web. Thêm vào đó, cuối mỗi chương còn có nhiều câu hỏi ôn tập hữu ích cho bạn đọc. Giáo trình là tài liệu tham khảo có giá trị cho học viên cao học ngành Hệ thống thông tin và khoa học máy tính cũng như các bạn đọc có quan tâm.

Nhà Xuất bản Đại học Cần Thơ chân thành cám ơn các tác giả và sự đóng góp ý kiến của quý thầy cô trong Hội đồng thẩm định trường Đại học Cần Thơ để giáo trình "Các hệ thống tìm kiếm thông tin văn bản" được ra mắt bạn đọc.

Nhà Xuất bản Đại học Cần Thơ trân trọng giới thiệu đến sinh viên, giảng viên và bạn đọc giáo trình này.

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ

LỜI NÓI ĐẦU

Những năm cuối thập niên 1990 đầu những năm 2000, các bộ máy tìm kiếm thông tin web (web search engine) đã trở nên gần gũi và là địa chỉ mọi người tìm đến khi cần tìm kiếm một thông tin nào đó nhờ vào sự phát triển không ngừng của các kỹ thuật tìm kiếm thông tin. Từ một lĩnh vực nghiên cứu đầy thách thức, tìm kiếm thông tin đã trở nên "*căn bản*" với hầu hết mọi người như là một phương tiện truy cập thông tin hữu ích. Để đáp lại những thách thức khác nhau của nhu cầu truy xuất thông tin, lĩnh vực tìm kiếm thông tin ra đời với mong muốn mang lại những tiếp cận tổng quát cho việc truy xuất, tìm kiếm nhiều dạng thông tin khác nhau. Bắt đầu với các xuất bản khoa học và các bản ghi tài liệu ở thư viện nhưng cũng đã rất nhanh chóng dàn trải ra các định dạng khác của thông tin, đặc biệt là các kho dữ liệu thông tin chuyên dùng của nhà báo, luật sư và bác sĩ. Với sự phát triển không ngừng của Internet và các dịch vụ có liên quan dẫn đến sự bùng nổ về khối lượng thông tin chia sẻ trên Internet, khối thông tin khổng lồ này sẽ trở nên vô nghĩa nếu con người không có cách tiếp cận và tìm thấy những xuất bản phù hợp với nhu cầu thông tin của họ. Chỉ mục một khối lượng khổng lồ thông tin hiện đang được chia sẻ trên Web để phục vụ tìm kiếm lại nó đòi hỏi phải có một kỹ thuật đặc thù, một kiến trúc phần cứng phù hợp. Để đáp ứng nhanh một nhu cầu tìm kiếm của người dùng đòi hỏi thuật toán so khớp phải có độ phức tạp vừa phải. Với một nhu cầu lớn về tìm kiếm thông tin phi cấu trúc trên Web, ngày nay, vì thế phần nhiều các nghiên cứu đã hướng đến tối ưu hóa tốc độ tìm kiếm hơn là tối ưu hóa độ chính xác của giải thuật so khớp.

Giáo trình này nhằm cung cấp cho người đọc các kiến thức tổng quan về tìm kiếm thông tin nói chung, trong đó tập trung chủ yếu vào các hệ thống tìm kiếm thông tin văn bản phi cấu trúc.

Phần đầu của giáo trình mang đến cho người đọc các khái niệm cơ bản nhất liên quan đến lĩnh vực tìm kiếm thông tin, và quan trọng hơn cả là các thành phần chính của một bộ máy tìm kiếm thông tin văn bản. Các kiến thức được đề cập trong phần này bao gồm kiến trúc cơ bản của một

hệ thống tìm kiếm thông tin, các phép đo đánh giá hiệu năng của một bộ máy tìm kiếm thông tin, các tài liệu văn bản được biểu diễn và chỉ mục như thế nào, các kỹ thuật chỉ mục cho phép tối ưu hóa chi phí lưu trữ, các phương pháp đo độ tương thích giữa truy vấn và tài liệu.

Phần tiếp theo của giáo trình trình bày các kỹ thuật cho phép tăng hiệu năng của một hệ thống tìm kiếm như áp dụng kỹ thuật phân tích giá trị đơn để giảm hạng các vec-tơ chỉ mục tài liệu, kỹ thuật phản hồi tương đồng hay mở rộng câu truy vấn để tăng độ bao phủ của kết quả tìm kiếm, các phương pháp cho phép sửa lỗi chính tả xuất hiện trong câu truy vấn. Giáo trình cũng dành một chương để bàn đến các thách thức của việc xây dựng một bộ máy tìm kiếm thông tin web trong đó đặc biệt chú trọng đến kỹ thuật lập chỉ mục phân tán và cách tính độ tương đồng tài liệu – truy vấn.

Nội dung của giáo trình được giảng dạy cho học viên cao học chuyên ngành Hệ thống thông tin với thời lượng 45 tiết lý thuyết. Bên cạnh đó giáo trình cũng có thể là sách tham khảo cho sinh viên, học viên cao học chuyên ngành Khoa học máy tính, các độc giả bước đầu tìm hiểu về lĩnh vực tìm kiếm thông tin, đặc biệt là tìm kiếm thông tin văn bản.

Do thời gian có hạn, giáo trình không thể tránh khỏi những thiếu sót ngoài ý muốn, nhóm tác giả rất mong nhận được những đóng góp chân thành từ quý độc giả để giáo trình được hoàn thiện hơn.

Cần Thơ, tháng 04 năm 2014

NHÓM TÁC GIẢ

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC THUẬT NGỮ

Bag-of-words	Mô hình túi từ
Bibliometric	Đo lường thông tin thư mục
Cross language information retrieval	Tìm kiếm đa ngôn ngữ
Document classification	Phân lớp tài liệu
Document frequency	Tần suất tài liệu
Index term	Từ chỉ mục
Information retireval	Lĩnh vực tìm kiếm thông tin
Information retrieval systems	Các hệ thống tìm kiếm thông tin
Inverse document frequency	Nghịch đảo tần suất tài liệu
Inverted index	Chỉ mục nghịch đảo
LSI – Latent Semantic Indexing	Chỉ mục ngữ nghĩa tiềm ẩn
Ranking algorihtm	Thuật toán xếp hạng
Space of web links	Không gian liên kết web
Stop words	Từ dừng
Term frequency	Tần xuất suất hiện của từ chỉ mục
Term-Document matrix	Ma trận từ chỉ mục – tài liệu
Web search engine	Bộ máy tìm kiếm thông tin web

Chương 1

KHÁI NIỆM VỀ TÌM KIẾM THÔNG TIN

Các nghiên cứu trong giai đoạn thập niên 90 của thế kỷ trước chỉ ra rằng con người phần đông tiếp nhận thông tin từ cộng đồng hơn là tiếp nhận thông tin từ các hệ thống tìm kiếm thông tin (information retrieval system). Điều này cũng dễ hiểu vì ở vào thời điểm đó, công nghệ thông tin (CNTT), Internet chưa phát triển, mọi người vẫn đến các đại lý du lịch để đặt vé máy bay hay đăng ký tour nghỉ dưỡng thay vì tìm kiếm và "book" nó trực tuyến như ngày nay.

Tuy nhiên khoảng 10 năm trở lại đây, khi mà CNTT đã có những bước tiến vượt bậc, Internet phát triển rộng khắp, kỹ thuật tìm kiếm thông tin ngày một hiệu quả thì các bộ máy tìm kiếm thông tin web (web search engine) đã trở thành một "địa chỉ" cung cấp thông tin đáng tin cậy của đa số công dân thời hiện đại.

1.1 TÌM KIẾM THÔNG TIN LÀ GÌ

Ngày nay, khi nói đến tìm kiếm thông tin, mọi người đều nghĩ ngay đến tìm kiếm thông tin "văn bản" (text document) và môi trường thực thi là môi trường web. Tuy nhiên, nhu cầu tìm kiếm thông tin không phải là nhu cầu được phát sinh khi WWW phát triển. Thực tế, khi mà thông tin, tri thức nhân loại được "tích góp" ngày một nhiều thì chúng ta sẽ có nhu cầu tìm kiếm lại một thông tin, một tri thức nào đó mà chúng ta quan tâm, chúng ta mong muốn tìm hiểu tại một thời điểm nào đó.

Thông tin mà chúng ta quan tâm không phải chỉ được lưu trữ dưới dạng văn bản chữ viết, mà theo thời gian thông tin sẽ được lưu trữ dưới nhiều định dạng khác nhau: văn bản thuần, văn bản bán cấu trúc và cấu trúc, hình ảnh, video, … Trong phạm vi của giáo trình này, chúng tôi tập trung giới thiệu các khái niệm, mô hình, kỹ thuật, giải pháp xây dựng bộ máy tìm kiếm thông tin áp dụng cho dữ liệu văn bản với quan niệm văn bản là dữ liệu tuần tự không cấu trúc. Các khái niệm, giải pháp này cơ bản vẫn đúng và là khái niệm nền tảng có thể áp dụng cho các định dạng dữ liệu khác.

Tìm kiếm thông tin bắt nguồn từ nhu cầu tìm lại các xuất bản khoa học, các bản ghi tài liệu trong thư viện (thập niên 1950) và sau đó nhanh chóng phát triển sang các mạng thông tin chuyên ngành về phóng viên, luật sư và bác sĩ. Vào thời gian này, phần lớn các nghiên cứu khoa học trong lĩnh vực tìm kiếm thông tin đều tập trung vào các nguồn dữ liệu trên. Mục tiêu của các nghiên cứu lúc bây giờ là tìm ra cách thức truy cập các nguồn dữ liệu không cấu trúc được lưu trữ bởi nhiều tổ chức, đơn vị khác nhau. Kể từ thập niên 1990 trở đi, khi mà WWW phát triển, các nghiên cứu trong lĩnh vực đã hướng đến việc tìm ra giải pháp giúp người dùng tìm thấy tài liệu, thông tin mà họ quan tâm được xuất bản đâu đó trên thế giới Internet rộng lớn. Phạm vi ứng dụng thay đổi mang đến những thách thức mới cho cộng đồng các nhà khoa học trong lĩnh vực tìm kiếm thông tin. Thách thức lớn nhất đó là việc phải lập chỉ mục một khối lượng khổng lồ các tài liệu (tăng nhanh từng ngày) và cung cấp một kết quả chấp nhận được (người dùng có thể hiểu được vì sau một tài liệu nào đó là một trong số các kết quả mà họ cần tìm) trong một khoảng thời gian rất ngắn (vài giây).

Tìm kiếm thông tin (Information retrieval - IR) vì thế sẽ là tập hợp các giải pháp cho phép "đương đầu" với việc biểu diễn, lưu trữ, cấu trúc và truy xuất lại các "đơn vị" thông tin cần thiết. Việc biểu diễn và cấu trúc thông tin cần được thực hiện theo cách đơn giản mà hiệu quả nhất nhằm giúp người dùng truy cập các phần thông tin mà họ quan tâm (user information need) một cách dễ dàng. Tuy nhiên, để biểu diễn đầy đủ thông tin mà người dùng quan tâm dưới dạng máy tính hiểu được thì không phải là một vấn đề đơn giản.

Để hiểu rõ hơn khó khăn này, chúng ta cùng xét ví dụ sau đây: một người dùng nào đó mong muốn tìm được các tài liệu có thông tin như mô tả bên dưới.

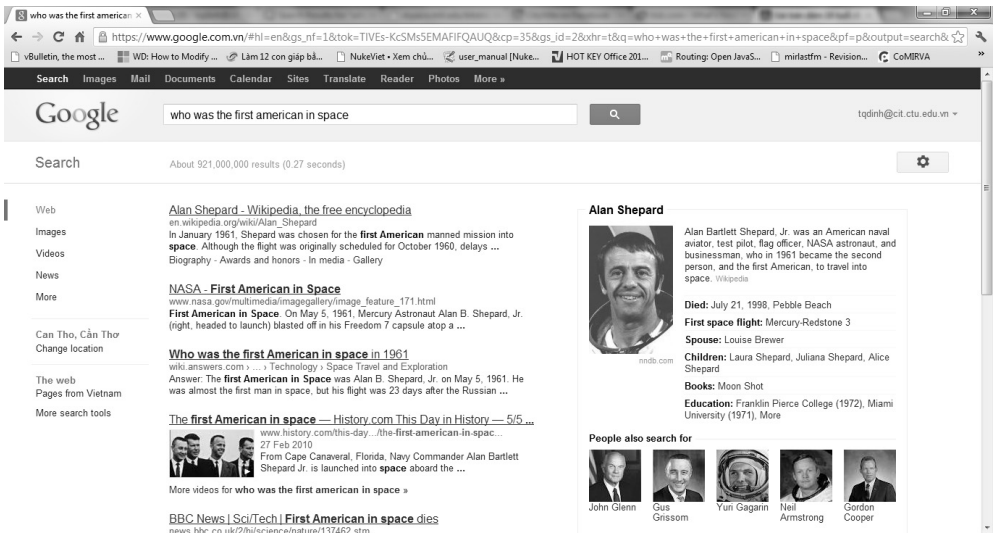
> Tìm các tài liệu cung cấp thông tin về các đội tennis cấp phổ thông trung học với điều kiện : 1- được một trường đại học tại Mỹ đỡ đầu; 2- phải từng tham gia giải đấu NCCA.

Các tài liệu mà người dùng này mong muốn nhận được sẽ phải có, ví dụ như, các thông tin về thứ hạng của đội trong các mùa giải gần nhất, thông tin liên hệ của huấn luyện viên như địa chỉ email, số điện thoại.

Rõ ràng, yêu cầu thông tin của người dùng không thể được sử dụng để truy vấn trực tiếp các bộ máy tìm kiếm thông tin. Để truy vấn các bộ máy tìm kiếm, trước hết, người dùng cần phải "diễn dịch" nhu cầu thông tin của mình dưới dạng câu truy vấn có thể hiểu được bởi các bộ máy tìm

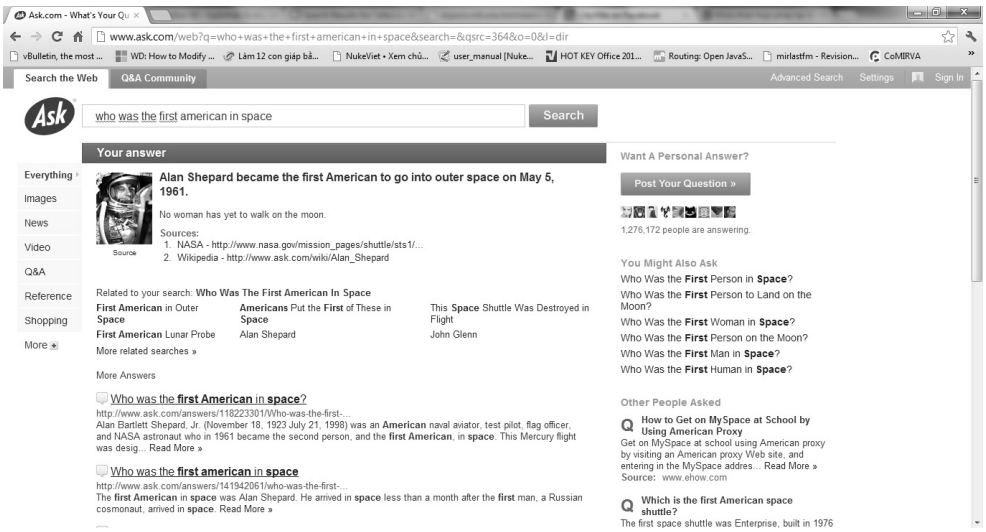
kiếm. Dạng thông dụng nhất của câu truy vấn đó là tập các từ khóa (keywords, index terms) mô tả tóm lược nhu cầu thông tin của người dùng. Khi nhận vào một truy vấn, các bộ máy tìm kiếm sẽ cố gắng truy tìm trong khối tài liệu khổng lồ mà mình "biết", mình quản lý được, các tài liệu được cho là "phù hợp" với câu truy vấn.

Nhu cầu thông tin của người dùng không dừng lại ở việc tìm các tài liệu có chứa đựng thông tin mình cần mà đôi khi nhu cầu thông tin lại là việc tìm tài liệu giải đáp cho một vấn đề nào đó. Chẳng hạn như một người dùng nào đó có nhu cầu thông tin như sau "*Who was the first American in space?*", nếu người dùng sử dụng các hệ thống tìm kiếm thông tin bình thường, Google search chẳng hạn, kết quả như hình 1.1. Nếu người dùng sử dụng các hệ thống đặc thù như Ask[1] chẳng hạn, kết quả trả về như hình 1.2. Phân tích kết quả của 2 hệ thống chúng ta có thể dễ dàng nhận thấy Ask cho kết quả là một câu trả lời và câu trả lời này là duy nhất. Kết quả tìm kiếm của Google cũng cho chúng ta câu trả lời tuy nhiên nó không được thể hiện dưới dạng câu trả lời và thực chất thì đó là các tài liệu mà nội dung của nó có chứa câu hỏi trên. Nhu cầu thông tin dạng này chỉ có thể đáp ứng bởi các hệ thống "Hỏi và Đáp", một phát triển của các hệ thống tìm kiếm.



Hình 1.1 Kết quả truy vấn Who was the first American in space với Google

[1] www.ask.com



Hình 1.2 Kết quả truy vấn Who was the first American in space với Ask

Trong lĩnh vực IR, người ta cũng phân biệt 2 loại tìm kiếm: 1 – Tìm kiếm dữ liệu (data retrieval); 2 – Tìm kiếm thông tin (information retrieval). Trong đó tìm kiếm dữ liệu, thực chất là tìm các tài liệu có chứa càng nhiều các tốt các từ khóa của câu truy vấn trong tập hợp (collection) các tài liệu hiện có. Cách thực hiện này phần lớn không đáp ứng được kỳ vọng của người dùng do người dùng luôn mong tìm được các tài liệu đáp ứng nhu cầu thông tin của họ hơn là các tài liệu phù hợp với câu truy vấn mà họ sử dụng. Điều này phải chăng là nghịch lý? Thực tế thì các hệ thống tìm kiếm dữ liệu cố gắng tìm tất cả các dữ liệu, tài liệu thỏa mãn yêu cầu tìm kiếm được định nghĩa một cách rõ ràng bởi các biểu thức điều kiện hoặc các biểu thức đại số quan hệ chẳng hạn. Đây chính là vấn đề khó khăn đối với người dùng bởi không phải ai cũng có thể diễn đạt chính xác nhu cầu thông tin của mình cũng như sự "mơ hồ" về việc tồn tại thông tin mà họ muốn tìm. Điểm khác biệt chủ yếu giữa hai hướng tiếp cận đó là "data retrieval" nhạy cảm với lỗi. Có nghĩa là nếu trong kết quả trả về của hệ thống "data retrieval" có chứa lỗi thì tác vụ tương ứng xem như thất bại. Ở phía ngược lại, lỗi thường không được phát hiện (người dùng không cảm nhận được lỗi) trong các hệ thống "information retrieval" do người dùng chỉ quan tâm các tài liệu trả về trong danh sách kết quả có phải là cái mà họ muốn tìm hay không.

Trong phạm vi tìm kiếm thông tin văn bản, trước đây, các nghiên cứu tập trung vào mục tiêu lập chỉ mục cho tài liệu và tìm kiếm các tài liệu thích hợp trong tập hợp các tài liệu nguồn. Ngày nay, nghiên cứu về lĩnh vực

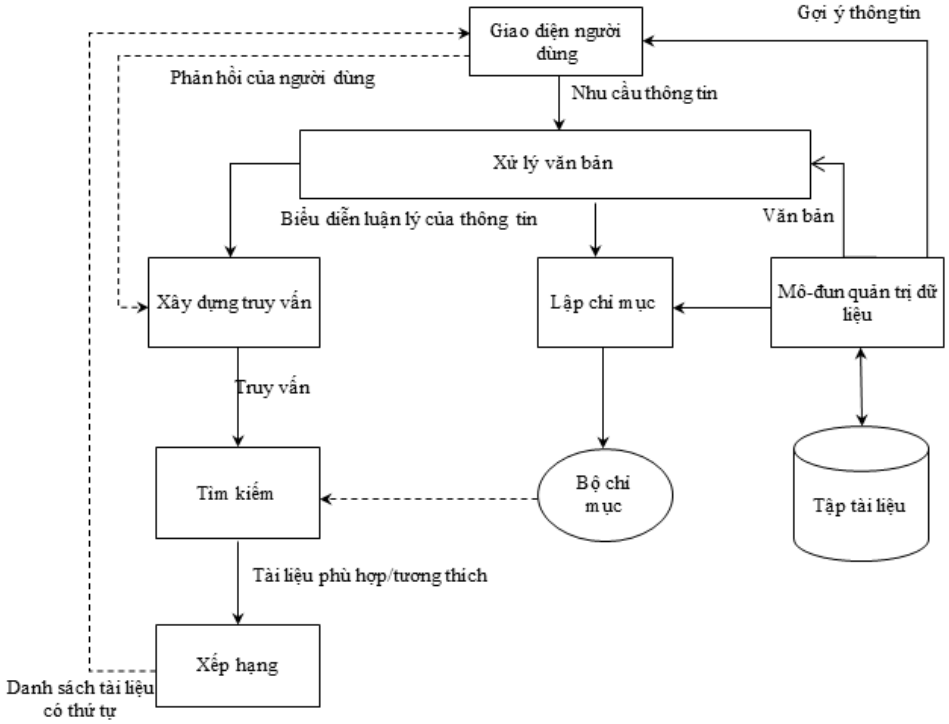
IR đã bao gồm mô hình hóa (cách thức biểu diễn tài liệu), phân lớp tài liệu (document classcification), kiến trúc hệ thống (system architecture), giao diện người dùng (users interface), biểu diễn thông tin (information visualization), lọc (filtering), tìm kiếm đa ngôn ngữ (cross language information retrieval), … Kể từ khi WWW ra đời và phát triển, nó đã là nơi mà mọi người dân từ lớn đến bé, từ tri thức đến bình dân đều có thể tạo và chia sẻ mọi thứ mà họ muốn, đã mang đến cho nhân loại một cách thức chia sẻ và tiếp cận thông tin mới chưa từng có tiền lệ trước đây. WWW đã trở thành "kho chứa" tri thức vô tận của nhân loại và cũng từ đó mang đến những thách thức vô cùng lớn cho cộng đồng các nhà nghiên cứu lĩnh vực IR. Việc tìm được thông tin hữu ích từ WWW, thường xuyên, là một nhiệm vụ khó khăn. Thật vậy, việc tìm thông tin hữu ích từ WWW về góc độ nào đó cũng giống như việc tìm lại một tài liệu lâu ngày không dùng đến mà ta đã quên mất đã lưu trữ nó ở nơi nào trên máy tính cá nhân. Một cách thủ công, chúng ta có thể duyệt qua lần lượt các thư mục lưu trữ trên máy tính để tìm ra nơi "cất giữ" tài liệu mong muốn. Thực tế, thao tác thủ công này đã bao hàm các công việc như duyệt qua mỗi tài liệu trong tập hợp, đánh giá tài liệu nào là tài liệu cần tìm. Cả hai công việc này đều là thách thức lớn cho các hệ thống tìm kiếm thông tin Web. Với môi trường Web, các bộ máy tìm kiếm phải duyệt không gian liên kết web (space of web links) để xây dựng tập tài liệu nguồn, đây là một nhiệm vụ bất khả thi. Để tìm được tài liệu thích hợp, các bộ máy tìm kiếm trước tiên phải "hiểu được" nhu cầu thông tin của người dùng qua tập từ khóa họ dùng cho truy vấn. Tiếp đến phải diễn dịch được thông tin mà mỗi tài liệu mang đến thông qua tập từ chỉ mục nội dung của tài liệu đó. Và sau cùng là thực hiện so khớp để tìm ra kết quả. Thực tế thì với các tài liệu văn bản không cấu trúc, yếu tố để nhận biết thông tin mà tài liệu này "chứa đựng" là khá nghèo nàn. Vì thế, đây cũng là một nhiệm vụ khó khăn bởi tính nhập nhằng của ngôn ngữ tự nhiên. Chính những khó khăn đó đã "thu hút" rất nhiều nghiên cứu từ cộng đồng khoa học trong thời gian khoảng 20 năm trở lại đây.

1.2 KIẾN TRÚC MỘT HỆ THỐNG TÌM KIẾM THÔNG TIN

Con người đã ý thức cần phải lưu trữ tri thức có được để sử dụng về sau từ cách đây khoảng 4 nghìn năm. Ngày qua ngày, số lượng tri thức cất giữ tăng lên, từ đây nảy sinh nhu cầu cần phải định nghĩa được một cấu trúc nào đó để giúp tìm lại nhanh tri thức mình cần. Một khái niệm đã có từ rất lâu đó là sử dụng một nhóm các từ khóa (keywords) hay khái niệm (concepts) biểu diễn cho thông tin mà một tài liệu truyền tải, kỹ thuật này trước đây được biết đến với tên gọi "gán nhãn" (tagging). Đây cũng

là ý tưởng chính của thành phần lập chỉ mục (indexing component) trong các bộ máy tìm kiếm thông tin hiện đại. Trải qua nhiều thập kỷ, việc gán nhãn tài liệu được thực hiện thủ công với kết quả thu được là cây phân cấp các chủ đề, khái niệm. Đến tận ngày nay, một số thủ thư vẫn còn sử dụng hình thức này để phân loại khối tài liệu khổng lồ mà họ quản lý, vì thế đây là một công việc nặng tính chủ quan. Với sự phát triển của ngành kỹ thuật máy tính, việc lập chỉ mục tài liệu đã được thực hiện một cách tự động và có thể nói rằng nó phù hợp với hướng nhìn (point of view) "hệ thống" hơn là hướng nhìn người sử dụng.

Với hướng nhìn hệ thống, xây dựng một hệ thống IR sẽ bao gồm việc xây dựng một thành phần lập chỉ mục hiệu suất cao, thành phần xử lý truy vấn hiệu quả, phát triển các giải thuật xếp hạng (ranking algorithm) tốt nhằm nâng cao chất lượng tập tài liệu kết quả. Với hướng nhìn người dùng, một hệ thống IR cần phải hiểu được cách thức diễn dịch nhu cầu thông tin thành tập từ khóa của người dùng, cần thiết lập được mối quan hệ giữa cách diễn dịch câu truy vấn của người dùng đến tổ chức, thao tác của hệ thống tìm kiếm thông tin. Và vì thế việc quan niệm nhu cầu thông tin chính là tập từ khóa người dùng truy vấn sẽ phá sản hoàn toàn.



Hình 1.3 Kiến trúc tổng quát của hệ thống tìm kiếm thông tin văn bản (Nguồn [Baeza-Yates, 2000])

Đối tượng, phạm vi phục vụ, chức năng, giao diện người dùng sẽ quyết định kiến trúc của hệ thống tìm kiếm thông tin. Ta lấy ví dụ là một hệ thống tìm kiếm sách trong thư viện. Chẳng hạn như trước đây, để tìm một quyển sách nào đó trong thư viện người ta có thể tìm theo tên tác giả và tựa sách. Khi đó thông tin sách sẽ được biểu diễn và lưu trữ dưới dạng CSDL quan hệ, truy vấn của người dùng sẽ được diễn dịch thành câu lệnh truy vấn SQL, hệ thống tìm kiếm sách vì thế sẽ bao gồm giao diện người dùng cho phép nhập vào tên tác giả hoặc tên sách, một thành phần truy vấn CSDL với ngôn ngữ SQL, thành phần giao diện trình bày kết quả. Các thế hệ tiếp nối của hệ thống tìm kiếm thông tin sách ngoài việc phát triển giao diện người dùng, gia tăng các thuộc tính tìm kiếm như bổ sung chủ đề, nhà xuất bản, năm xuất bản, chức năng tìm kiếm đa tiêu chí thì còn bổ sung chức năng tìm kiếm toàn văn (tìm theo nội dung). Chính việc bổ sung chức năng này đã làm thay đổi hoàn toàn kiến trúc của các hệ thống tìm kiếm sách trong thư viện. Hoặc là phải sử dụng các Hệ quản trị CSDL hiện đại có hỗ trợ lập chỉ mục toàn văn và tìm kiếm trên đó hoặc là phải thay đổi cách thức biểu diễn tài liệu, sử dụng định dạng XML chẳng hạn.

Một hệ thống tìm kiếm thông tin văn bản có kiến trúc tổng quát như hình 1.3. Để thực hiện truy vấn, trước tiên hệ thống cần định nghĩa tập dữ liệu văn bản nguồn. Thông thường bao gồm các thành phần: 1- tập dữ liệu sử dụng; 2- các thao tác thực hiện trên văn bản; 3- phương pháp mô hình hóa (cấu trúc và thành phần của văn bản sẽ được rút trích). Trong đó các thao tác trên văn bản sẽ chuyển nội dung văn bản nguồn sang biểu diễn luận lý của nó (cấu trúc biểu diễn dưới dạng máy tính hiểu được).

Thành phần **Quản trị dữ liệu** sẽ lập chỉ mục cho tập tài liệu nguồn, kết quả của quá trình lập chỉ mục sẽ là một tập chỉ mục (index) được tổ chức và lưu trữ theo một cấu trúc nhất định. Cấu trúc tập chỉ mục cần được lựa chọn sao cho có thể thực hiện tìm kiếm nhanh trên một khối lượng dữ liệu khổng lồ. Khá nhiều cấu trúc đã được đề nghị để lưu trữ tập chỉ mục nhưng cấu trúc đơn giản và thông dụng nhất vẫn thường được các bộ máy tìm kiếm thông tin web sử dụng đó là cấu trúc tập chỉ mục nghịch đảo (inverted index). Quá trình lập chỉ mục thường mất rất nhiều thời gian và thông thường được thực hiện từng bước và ngoại tuyến.

Người dùng xác định và cung cấp nhu cầu thông tin của mình cho bộ máy tìm kiếm thông qua giao diện hệ thống. Nhu cầu thông tin này sẽ được xử lý theo cách thức đã tiến hành với nội dung của tài liệu bởi mô-đun xử lý văn bản để thu được dạng biểu diễn luận lý của nó. Tiếp theo đó, tại mô-đun xây dựng truy vấn, các thao tác cần thiết trên biểu diễn

luận lý của nhu cầu thông tin của người dùng có thể được áp dụng để thu được câu truy vấn phù hợp với hệ thống tìm kiếm (sửa lỗi chính tả, mở rộng truy vấn, …). Khi đã có được câu truy vấn hoàn chỉnh, hệ thống sẽ thực hiện tìm kiếm trên câu truy vấn này. Giai đoạn tìm kiếm thực hiện nhanh hay chậm là phụ thuộc vào cấu trúc dữ liệu đã sử dụng để biểu diễn tập chỉ mục cũng như thuật toán so khớp câu truy vấn với từng tài liệu trong tập tài liệu (toàn bộ tập tài liệu hay một phần thích hợp của tạp tài liệu).

Mô-đun xếp hạng sẽ thực hiện sắp xếp các tài liệu kết quả tìm được theo giá trị tương đồng (relevance value) trước khi trả về cho người dùng thông qua giao diện hệ thống. Người dùng sẽ duyệt qua tập xếp hạng các tài liệu kết quả để tìm tài liệu mà mình cho là phù hợp. Một tùy chọn ở giai đoạn này đó là người dùng có thể giúp hệ thống tìm được các tài liệu phù hợp hơn bằng cách xác định các tài liệu phù hợp trong tập kết quả trả về, hệ thống sẽ tạo lại truy vấn một cách tự động và thực hiện lại các bước giống như ban đầu để có được một kết quả tốt hơn.

Kết quả trả về từ các bộ máy tìm kiếm cần được người dùng kiểm chứng để xác định tính phù hợp với nhu cầu thông tin của họ. Các thống kê [Krirch 1998] cho thấy rằng người dùng chỉ duyệt qua trung bình 10 đến 20 kết quả đầu tiên trong số có thể lên đến hàng trăm nghìn kết quả trả về. Vậy điều gì ảnh hưởng đến quyết định của người dung là xem hay không xem nội dung một tài liệu trong danh sách kết quả? Đó chính là cách thức trình bày (display) danh sách tài liệu kết quả. Thật vậy, bên cạnh việc có tập tài liệu nguồn lớn, cấu trúc chỉ mục hiệu quả, giải thuật so khớp tốt thì việc trình bày kết quả cũng là yếu tố quyết định ảnh hưởng đến sự hài lòng của người dùng đối với một hệ thống tìm kiếm. Thông thường với mỗi tài liệu kết quả, các thông tin cần thiết cho việc trình bày là: Tựa đề (title), địa chỉ web của tài liệu (web address), mô tả ngắn gọn về tài liệu (short description). Những thông tin này giúp người dùng hình dung sơ lược về tài liệu kết quả cũng như để lý giải với người dùng vì sao tài liệu xuất hiện trong danh sách kết quả (hình 1.4).