BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

LUẬN VĂN THẠC SỸ KHOA HỌC

NGHIÊN CỨU VÀ ĐÁNH GIÁ CÁC HỆ TRUY XUẤT THÔNG TIN

NGÀNH: CÔNG NGHỆ THÔNG TIN
MÃ SỐ:

CAO THỊ THU HƯƠNG

Người hướng dẫn khoa học: PGS.TS. NGUYỄN THANH THUỶ

HÀ NỘI - 2006

LỜI CẢM ƠN

Em xin chân thành gửi lời cảm ơn sâu sắc tới Thầy giáo hướng dẫn, **PGS.TS.Nguyễn Thanh Thuỷ** người đã có những hướng dẫn tận tình, quý báu giúp em hoàn thành luận văn này.

Em cũng xin cảm ơn các Thầy Cô khoa Công nghệ Thông tin trường Đại học Bách Khoa Hà Nội đã truyền đạt kiến thức quý báu trong khoá học này.

Cuối cùng xin cảm ơn gia đình và cơ quan nơi đang công tác đã tạo điều kiện thuận lợi để tôi hoàn thành khoá học này.

Hà nội, tháng 10 năm 2006

Cao Thị Thu Hương

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG

Chương 1: TỔNG QUAN VỀ HỆ TRUY XUẤT THÔNG TIN

1.1. Lịch sử truy xuất thông tin và hệ thống truy xuất thông tin

Truy xuất thông tin có một lịch sử lâu đời gắn liền với các thư viện và trung tâm tìm kiếm thông tin. Trước đây, khi máy tính và internet chưa ra đời, những người có nhu cầu thông tin ngoài việc nhờ sự trợ giúp thông tin từ bạn bè, người thân còn có thể tìm đến thư viện hoặc các trung tâm thông tin để tìm kiếm thông tin cần thiết. Cách biểu diễn, lưu trữ, tổ chức và phổ biến thông tin của thư viện được xem là cách làm truyền thống của một hệ thống truy xuất thông tin. Khi tiếp nhận các yếu tố thông tin hay tài liệu mới, thư viện sẽ tiến hành phân tích yếu tố thông tin đó. Sau đó, những mô tả thích hợp sẽ được chọn ra để mô tả, phản ánh nội dung của yếu tố thông tin đó. Dựa trên những mô tả này, mỗi yếu tố thông tin sẽ được phân loại theo những thủ tục đã được thiết lập rồi xát nhập vào tập hợp các yếu tố thông tin đã tồn tại. Các thủ tục này được tạo ra để hệ thống hóa các yêu cầu (các yêu cầu được thiết kế để thay thế cho một nhu cầu thông tin) và để so sánh những yêu cầu, truy vấn đó với mô tả của các yếu tố thông tin đã lưu trữ.

Việc so sánh này chính là cơ sở để quyết định các yếu tố thông tin thích hợp với câu truy vấn tương ứng. Cuối cùng, một cơ chế tìm kiếm và phổ biến thông tin sẽ được dùng để trả các yếu tố thông tin cần thiết đến người sử dụng hệ thống. Tuy nhiên, phải xem xét vấn đề nảy sinh về vị trí thật sự của một yếu tố thông tin mới được thêm vào trong tập hợp tài liệu. Có nhiều cơ chế tiếp cận khác nhau để giải quyết vấn đề này nhưng chúng đều liên quan đến cách tổ chức vật lý hoặc luận lý các yếu tố thông tin. Trong thư viện, cách tổ

chức vật lý chính là việc lập chỉ mục cho tài liệu, tức là sự sắp xếp các con số của các quyển sách, cách đánh số thường được quy định bởi các thư viện lớn. Những quyển sách sẽ được đặt vào những vị trí xác định dựa vào những con số này. Ngoài ra, cách tổ chức luận lý dữ liệu phải được thêm vào với cách tổ chức vật lý để giúp người sử dụng truy xuất thông tin dễ dàng hơn. Chẳng hạn, những quyển sách ấn bản về truy xuất thông tin có thể được xác định bằng cách nhìn vào danh mục các chủ đề của thư viện với thuật ngữ cần tìm là "truy xuất thông tin". Một khi ta tìm thấy thuật ngữ thích hợp, các thẻ số kế tiếp nhau sẽ xác định những quyển sách liên quan đến chủ đề đang tìm kiếm. Những quyển sách này phụ thuộc vào các con số và chúng sẽ được tìm thấy tại những vị trí xác định. Bên cạnh đó, mỗi khi muốn thay đổi thuật ngữ chủ đề của sách, chúng ta không cần thay đổi vị trí của sách trên kệ sách; tức là, các yếu tố thông tin có thể được tổ chức luận lý lại bằng cách thay đổi danh mục thư viện mà không cần thay đổi sắp xếp vật lý.

Xã hội ngày càng phát triển, do đó thông tin rất đa dạng phong phú. Bài toán đặt ra là chúng ta phải làm sao để quản lý được số lượng thông tin khổng lồ một cách có hiệu quả. Từ đó dẫn đến nhu cầu làm giảm một lượng các yếu tố thông tin đến một kích thước có thể quản lý, các yếu tố thông tin còn lại được xem là có liên quan nhiều nhất đến lĩnh vực tìm kiếm. Mặt khác, chúng ta rất khó dự đoán mẫu, trạng thái phát triển tương lai của thông tin, hoặc nếu có thể dự đoán thì tỉ lệ rủi ro rất cao. Khó khăn tiếp theo trong việc tổ chức thông tin hiệu quả là ước muốn giữ những yếu tố liên quan gần nhau. Ví dụ, những chủ đề liên quan đến nhiều lĩnh vực như phân tích hệ thống (nó liên quan đến khoa học máy tính, vận trù học, kỹ thuật học, khoa học quản lý, giáo dục và các hệ thống thông tin) không thể để gần nhau được mà phải để riêng ra theo từng lĩnh vực. Đây thực sự là một khó khăn. Còn rất nhiều khó khăn nữa, chẳng hạn các khó khăn trong phân loại, so sánh tài liệu, yếu tố thông

tin, lập chỉ mục, đánh số cho tài liệu. Những khó khăn này sẽ không được giải quyết nếu không có sự ra đời của máy tính. Quả thật, nhờ có máy tính mà việc lưu trữ, truy xuất thông tin trở nên dễ dàng hơn. Máy tính có thể thao tác trên tất cả các loại thông tin và có thể lưu trữ một cách nhanh chóng một số lượng thông tin khổng lồ. Ngoài ra, cơ chế truy xuất thông tin trên máy tính có thể rất nhanh chóng và hiệu quả tùy thuộc mô hình cài đặt, thuật toán của cơ chế đó. Cơ chế tìm kiếm này cũng khá giống với cơ chế truy xuất thông tin của thư viện. Trước hết, dựa trên ngôn ngữ chỉ mục và các yếu tố thông tin đại diện cho nội dung của tài liệu, tập tài liệu sẽ được biểu diễn dưới dạng tập hợp các chỉ mục đại diện cho tập tài liệu đó. Trong khi đó, nhu cầu truy xuất thông tin được biểu diễn dưới dạng câu truy vấn có cấu trúc hoặc không cấu trúc mà máy có thể hiểu được. Sau đó, máy sẽ so sánh hai dạng biểu diễn trên, biểu diễn tài liệu và biểu diễn câu truy vấn, để biết được tài liệu nào phù hợp với truy vấn nào. Sau khi so sánh, máy sẽ định vị được vị trí vật lý của yếu tố thông tin cần tìm kiếm và phổ biến nó đến người sử dụng. Đây là cơ chế tìm kiếm chung cho mọi hệ thống truy xuất thông tin. Tuy nhiên, cách đây không quá 20 năm, sau khi máy tính ra đời, các hệ thống truy xuất thông tin chủ yếu được sử dụng trong phòng thí nghiệm để tìm kiếm một kho ngữ liệu sách và tài liệu. Mặc dù chúng không bao hàm các phương pháp toán phức tạp, nhưng khi Internet phát triển, kỹ thuật tìm kiếm chủ yếu trên World Wide Web chính là các kỹ thuật truy xuất thông tin. Quả thật, các hệ thống truy xuất thông tin ngày càng phát triển về thuật toán, kỹ thuật truy xuất thông tin nhờ có sự ra đời của Internet. Vì nhu cầu truy xuất thông tin của con người trên Internet là một nhu cầu phổ biến, thiết thực, không thể thiếu nên các nhà phát triển hệ thống truy xuất thông tin cũng phải nỗ lực để mang lại hiệu năng, hiệu quả cho người sử dụng.

Chúng ta thấy rõ ràng là nghiên cứu truy xuất thông tin có truyền thống tập trung vào truy xuất thông tin dạng văn bản (Text Retrieval) hay tài liệu văn bản (Document Retrieval). Trong một thời gian dài, truy xuất thông tin gần như đồng nghĩa với tìm kiếm tài liệu hay tìm kiếm văn bản. Trong thời gian gần đây, các viễn cảnh ứng dụng mới như ứng dụng trả lời câu hỏi (Question Answering), ứng dụng nhận dạng chủ đề (Topic Detection), hay ứng dụng lưu vết (tracking) trở thành các lĩnh vực hoạt động mạnh mẽ trong nghiên cứu truy xuất thông tin. Càng ngày, ranh giới giữa cộng đồng truy xuất thông tin hay cộng đồng truy xuất thông tin và các cộng đồng nghiên cứu xử lý ngôn ngữ tự nhiên, cộng đồng nghiên cứu cơ sở dữ liệu trở nên mờ nhạt khi các cộng đồng này cùng nhau phát triển các lĩnh vực quan tâm chung, ví dụ như trả lời câu hỏi, tóm tắt và truy xuất thông tin từ các tài liệu có cấu trúc.

Một lĩnh vực phát triển khác mà các kỹ thuật truy xuất thông tin đang kế tục và phát huy, đó là truy xuất thông tin không văn bản hay còn gọi là truy xuất thông tin đa phương tiện. Loại hình tìm kiếm này sẽ dựa trên rút trích tự động các phần văn bản hay lời nói của các tài liệu đa phương tiện, sau đó được xử lý bởi các kỹ thuật truy xuất thông tin dựa văn bản (text-based IR techniques). Tuy nhiên, người ta ngày càng quan tâm đến sự phát triển các kỹ thuật phơi bày cụ thể thông tin phương tiện truyền thông rồi tích hợp chúng với các phương pháp tìm kiếm đã được thiết lập tốt hơn là cách rút trích chúng.

Trong phạm vi đề tài, sẽ quan tâm nhiều đến truy xuất thông tin trên văn bản.

1.2. Hệ truy xuất thông tin

1.2.1. Khái niệm về hệ truy xuất thông tin

Theo lý thuyết, hệ thống truy xuất thông tin là một hệ thống thông tin. Nó được sử dụng để lưu trữ, xử lý, tra cứu, tìm kiếm, và phổ biến các yếu tố thông tin đến người sử dụng. Hệ thống truy xuất thông tin thường thao tác với các dữ liệu dạng văn bản và không có sự giới hạn về các yếu tố thông tin trong văn bản. Hệ thống thông tin bao gồm một tập hợp các yếu tố thông tin, một tập các yêu cầu và các cơ chế tìm kiếm để quyết định yếu tố thông tin nào liên quan đến các yêu cầu. Theo nguyên tắc, mối quan hệ giữa các câu truy vấn và tài liệu có được từ sự so sánh trực tiếp. Nhưng trên thực tế, sự liên quan giữa các câu truy vấn và tài liệu xác định không phải được quyết định trực tiếp mà gián tiếp bằng cách: các tài liệu, yếu tố thông tin phải chuyển sang ngôn ngữ chỉ mục trước khi xác định mức độ liên quan.

Sau đây là định nghĩa về hệ truy xuất thông tin của một số tác giả:

Salton (1989):

"Hệ truy xuất thông tin xử lý các tập tin lưu trữ và những yêu cầu về thông tin, xác định và tìm từ các tập tin những thông tin phù hợp với những yêu cầu về thông tin. Việc truy xuất những thông tin đặc thù phụ thuộc vào sự tương tự giữa các thông tin được lưu trữ và các yêu cầu, được đánh giá bằng cách so sánh các giá trị của các thuộc tính đối với thông tin được lưu trữ và các yêu cầu về thông tin".

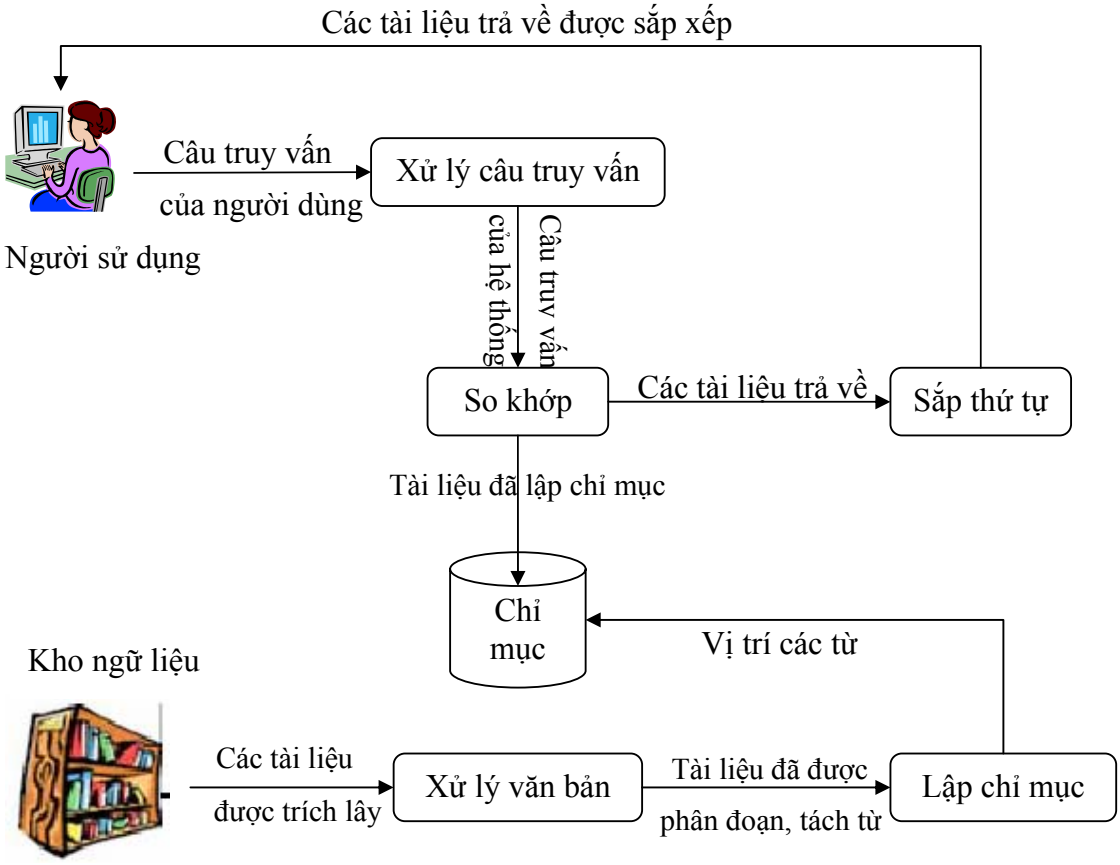
Kowalski (1997):

"Hệ truy xuất thông tin là một hệ thống có khả năng lưu trữ, truy xuất và duy trì thông tin. Thông tin trong những trường hợp này có thể bao gồm văn bản, hình ảnh, âm thanh, video và những đối tượng đa phương tiện khác".

Một cách một cách đơn giản hệ thống truy xuất thông tin là một hệ thống hỗ trợ cho người sử dụng tìm kiếm thông tin một cách nhanh chóng và dễ dàng. Người sử dụng có thể đưa vào những câu hỏi, những yêu cầu (dạng ngôn ngữ tự nhiên) và hệ thống sẽ tìm kiếm trong tập các tài liệu (dạng ngôn ngữ tự nhiên) đã được lưu trữ để tìm ra những tài liệu có liên quan, sau đó sẽ sắp xếp các tài liệu theo mức độ liên quan giảm dần và trả về cho người sử dụng.

1.2.2. Cách thức hoạt động của hệ thống truy xuất thông tin

Hình 1.1 minh họa cấu trúc, cách hoạt động cơ bản của một hệ thống truy xuất thông tin cổ điển.



Hình 1.1: Hệ thống truy xuất thông tin theo cơ chế cổ điển

1. Ở giai đoạn đầu tiên, giai đoạn tiền xử lý, tài liệu thô của ngữ liệu được xử lý thành các **tài liệu được tách từ, phân đoạn** (**tokenized documents**) và sau đó được lập chỉ mục thành một danh sách các **vị trí của từ** (**postings per terms**).
2. Ở giai đoạn thứ hai, người sử dụng đưa ra một câu truy vấn (phi cấu trúc bằng ngôn ngữ tự nhiên) mô tả nhu cầu thông tin của họ. Hệ thống truy xuất thông tin sẽ biểu diễn câu truy vấn này thành những câu truy vấn có hoặc không có cấu trúc mà máy có thể hiểu được. Hệ thống truy xuất thông tin bắt đầu thực hiện chất vấn, đối chiếu để tìm ra tài liệu, các yếu tố thông tin có thể trả lời và liên quan đến câu truy vấn. Các thủ tục được dùng để quyết định các yếu tố thông tin có liên quan đến câu truy vấn đều dựa trên biểu diễn của các câu truy vấn và các yếu tố thông tin có chứa các thành phần ngôn ngữ chỉ mục.
3. Cuối cùng, các tài liệu, yếu tố thông tin được tìm thấy được hiển thị thành một danh sách tài liệu và được sắp xếp theo **thứ tự liên quan** (**ranked retrieved documents**). Thông thường, những tài liệu, yếu tố thông tin có liên quan nhiều nhất được xếp trên những tài liệu ít liên quan hơn. Tùy vào các hệ thống truy xuất thông tin khác nhau mà chúng hiển thị thông tin liên quan theo những cách khác nhau. Chẳng hạn, có hệ thống chỉ hiển thị tên tiêu đề và đường dẫn đến tài liệu đó, hoặc có hệ thống vừa hiển thị tên, đường dẫn, vừa hiển thị một ít nội dung liên quan đến câu truy vấn, hoặc có những hệ thống phục vụ truy xuất thông tin trên mạng thì thêm vào các liên kết đến các trang web khác nhau.

Nhiều hệ thống thông tin còn có cả cơ chế cho phép người sử dụng cung cấp phản hồi đến chất lượng của kết quả trả về. Sử dụng phản hồi, hệ thống cố gắng thích ứng và nỗ lực tìm ra những kết quả tốt nhất cho câu truy vấn.

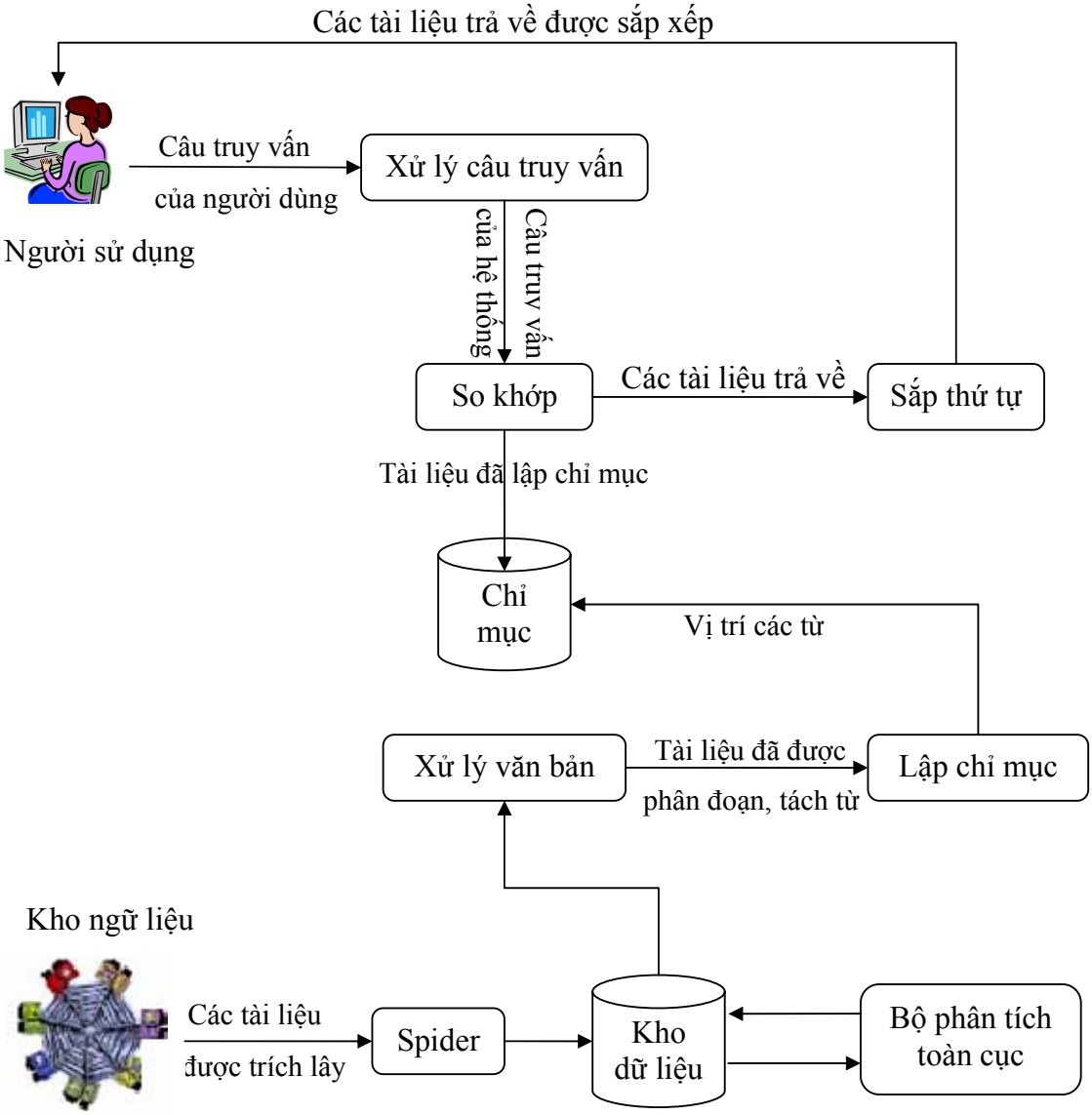
Việc lập chỉ mục trong giai đoạn tiền xử lý về nguyên tắc thì giống nhau đối với từng hệ thống nhưng về thuật toán, cách thức thì khác nhau. Nguyên tắc lập chỉ mục: Tài liệu hay yếu tố thông tin phi cấu trúc khi thêm mới sẽ được hệ thống truy xuất thông tin chuyển sang một thể đặc biệt, đó là ngôn ngữ chỉ mục. Việc chuyển đổi thành phần thông tin thành ngôn ngữ chỉ mục được thực hiện thủ công, hay tự động hoặc cả hai và nó được gọi là tiến trình lập chỉ mục. Tiến trình lập chỉ mục này được thực hiện dựa trên các yếu tố thông tin đại diện cho nội dung của tài liệu. Do đó, kết quả của tiến trình này là một tập chỉ mục đại diện cho tài liệu đó.

1.2.3. Các phương tiện truy xuất thông tin

Hình 1.2 minh họa cấu trúc cơ bản của các phương tiện tìm kiếm. Một phương tiện tìm kiếm là một hệ thống truy xuất thông tin, tuy nhiên, nó không giống hoàn toàn với hệ thống truy xuất thông tin cổ điển đã mô tả ở trên. Sự khác biệt giữa các hệ thống truy xuất thông tin cổ điển và các phương tiện tìm kiếm bắt nguồn từ sự khác biệt nguồn gốc dữ liệu, có nghĩa là một kho lưu trữ khép kín được định nghĩa tốt trái ngược với World Wide Web. Vì không có cách tiếp cận trực tiếp đến các tài liệu trên Web (như là có trong kho ngữ liệu thư viện), phương tiện tìm kiếm phải cần đến thành phần **crawler**. Thành phần phần mềm này chịu trách nhiệm lấy các trang web về và lưu trữ chúng trong một kho nội bộ. Cơ chế **crawling** đưa ra các thách thức công nghệ liên quan đến hiệu năng của quá trình và đến sự liên quan của tài liệu – vì các trang web là động, nên **crawler** phải giữ cho kho nội bộ luôn được cập nhật hằng ngày.

Việc **crawling** các tài liệu ngoài Web thì không đủ bởi vì dữ liệu web gồm có nhiều thông tin dư thừa. Phân tích toàn cục có trách nhiệm loại bỏ dữ liệu không quan trọng như các trang Web giống nhau và các trang bao gồm

sách báo không lành mạnh. Ngoài ra, phân tích toàn cục cũng chịu trách nhiệm tính toán toàn cục được dùng trong các hệ thống truy xuất thông tin như sắp xếp thứ tự trang (thứ tự trang hầu hết được xác định bởi những trang có liên kết với nó và những trang nó liên kết tới).



Hình 1.2: Cơ chế tìm kiếm của Search Engine

1.3. So sánh truy xuất thông tin cổ điển và truy xuất thông tin trên Web

Bảng dưới đây biểu diễn sự khác biệt giữa các hệ thống truy xuất thông tin cổ điển (IR cổ điển) và các hệ thống truy xuất thông tin trên Web (Web IR).

Bảng 1.1: So sánh IR cổ điển với Web IR

	IR cổ điển	***Web IR***
Kích thước	Lớn	Khổng lồ
Chất lượng dữ liệu	Sạch, không trùng lặp	Lộn xộn, trùng lặp
Tỉ lệ thay đổi dữ liệu	Hiếm	Liên tục
Khả năng truy cập dữ liệu	Có thể	Truy cập một phần
Đa dạng định dạng	Đồng nhất, cùng nguồn gốc	Rất đa dạng
Tài liệu	Văn bản	HTML
# liên quan	Nhỏ	Lớn
Kỹ thuật IR	Dựa nội dung	Dựa liên kết

Khối lượng dữ liệu trong một hệ thống IR cổ điển khá lớn, trong khi đó khối lượng dữ liệu này trong hệ thống Web IR là khổng lồ. Khác biệt lớn nhất trong khối lượng dữ liệu, chính là các thứ tự của lượng, ảnh hưởng đến phần cứng được đòi hỏi (một máy tính thì không bao giờ đủ, bộ nhớ không thể chứa toàn bộ dữ liệu) và các thuật toán (các định nghĩa hiệu năng của thời gian và không gian bị thay đổi). Một khác biệt nữa là khác biệt của dữ liệu. Trong hệ thống IR cổ điển dữ liệu được làm sạch, trong khi đó dữ liệu trên Web IR thì phức tạp, cả hai đều do sự trùng lắp vô ý và do các spam có dụng ý tăng thứ hạng của trang đó hoặc chỉ tạo sự lộn xộn.

Như đã đề cập ở trên, sự thay đổi dữ liệu trong IR cổ điển là không thường xuyên, do đó nó thường được lập chỉ mục 1 lần. Ngược lại, dữ liệu trên Web thì thay đổi thường xuyên nên chỉ mục cũng cần được cập nhật. Hơn nữa, tính khả truy cập của dữ liệu là không quan trọng trong Web IR.

Tài liệu trong IR cổ điển thường đồng nhất về định dạng còn tài liệu trong Web IR gồm nhiều loại khác nhau: bất cứ ai cũng có thể tạo một trang web trong bất kì định dạng nào và bất kì ngôn ngữ nào.

Một điểm khác biệt quan trọng nữa là tài liệu web không thường xuyên được viết ở dạng văn bản thô như trong tài liệu IR cổ điển. Trang Web thường được viết bằng HTML (Hypertext Markup Language), vừa có những lợi ích và bất lợi đối với hệ thống truy xuất thông tin : một mặt, nó bao gồm dữ liệu có cấu trúc giúp việc phân tích dễ dàng hơn ; mặt khác, nó thường không chứa nhiều văn bản (hệ thống IR dựa trên thứ này), do đó khó phân loại hơn.

Kết quả trả về trong Web IR cũng nhiều hơn so với IR cổ điển, do đó khó để sắp thứ tự danh sách kết quả hơn.

Và cuối cùng, IR cổ điển sử dụng kĩ thuật sắp thứ tự chỉ **dựa trên nội dung** (content-based). Tuy nhiên, kĩ thuật này không thể áp dụng với Web IR. Đây là một kĩ thuật thông dụng trước khi Google giới thiệu kĩ thuật sắp thứ tự mới **dựa trên liên kết** (link-based). Kĩ thuật sắp thứ tự dựa trên liên kết sử dụng **siêu liên kết** (hyperlink) giữa các tài liệu web để sắp thứ tự các trang web một cách hiệu quả và chắc chắn hơn.

1.4. So sánh truy xuất thông tin với truy xuất dữ liệu

Một hệ thống truy xuất thông tin không phải là một hệ thống truy xuất dữ liệu. Bảng dưới đây trình bày một số thuộc tính khác nhau giữa hệ thống truy xuất thông tin và hệ thống truy xuất dữ liệu.

Bảng 1.2: Sự khác nhau giữa hệ truy xuất thông tin và hệ truy xuất dữ liệu.

	Truy xuất thông tin	*Truy xuất dữ liệu*
Dữ liệu	Văn bản tự do, không cấu trúc	Các bảng dữ liệu, có cấu trúc
Truy vấn	Từ khóa, ngôn ngữ tự nhiên	SQL, đại số quan hệ
Kết quả	Liên quan tương đối, xấp xỉ. Sắp xếp theo mức độ liên quan	Liên quan chính xác. Không sắp xếp
Truy cập	Những người không phải chuyên gia	Người sử dụng có kiến thức hoặc các tiến trình tự động

Hệ thống truy xuất thông tin thu thập tài liệu dựa trên yêu cầu thông tin của người dùng. Câu truy vấn trên dữ liệu không có cấu trúc (thường là dạng văn bản tự do), sử dụng từ khóa hoặc ngôn ngữ tự nhiên và do vậy có thể được viết bởi người dùng không thông thạo. Vì cú pháp của câu truy vấn không được định nghĩa chính xác nên kết quả có thể bao gồm các kết hợp không chính xác và thứ tự liên quan hay tương quan (relevance) của chúng chỉ là gần đúng.

Hệ thống truy xuất dữ liệu thu thập một tập hợp các tài liệu phù hợp về mặt cú pháp với câu truy vấn của người sử dụng. Câu truy vấn trên dữ liệu có cấu trúc (thường là bảng trong cơ sở dữ liệu) và thường sử dụng một ngôn ngữ truy vấn được định nghĩa hoàn chỉnh như là SQL hay đại số quan hệ. Người sử dụng phải quen thuộc với cú pháp và hiểu được ngữ nghĩa của ngôn ngữ truy vấn. Vì vậy, câu truy vấn thường được viết bởi người am hiểu hoặc một quá trình tự động. Kết quả trả về bao gồm tất cả các tài liệu chính xác phù hợp với ngữ nghĩa của câu truy vấn, thứ tự bất kì.

1.5. So sánh IRS với các hệ thống thông tin khác

Hệ truy xuất thông tin cũng tương tự như nhiều hệ thống xử lý thông tin khác. Hiện nay các hệ thống thông tin quan trọng nhất là: hệ quản trị cơ sở

dữ liệu (DBMS), hệ quản lý thông tin (MIS), hệ hỗ trợ ra quyết định (DSS), hệ trả lời câu hỏi (QAS) và hệ truy xuất thông tin (IR).

Hệ quản trị cơ sở dữ liệu (DBMS)

Bất cứ hệ thống thông tin nào cũng dựa trên một tập các mục được lưu trữ (gọi là cơ sở dữ liệu) cần thiết cho việc truy cập. Do đó hệ quản trị cơ sở dữ liệu đơn giản là một hệ thống được thiết kế nhằm thao tác và duy trì điều khiển cơ sở dữ liệu.

DBMS tổ chức lưu trữ các dữ liệu của mình dưới dạng các bảng. Mỗi cơ sở dữ liệu được lưu trữ thành các bảng khác nhau. Mỗi cột trong bảng là một thuộc tính duy nhất đại diện cho bảng, nó không được trùng lặp và ta gọi đó là khóa chính. Các bảng có mối liên hệ với nhau thông qua các khóa ngoài. DBMS có một tập các lệnh để hỗ trợ cho người dùng sử dụng truy vấn đến dữ liệu của mình. Vì vậy muốn truy vấn đến CSDL trong DBMS ta phải học hết các tập lệnh này. Nhưng ngược lại nó sẽ cung cấp cho ta các dữ liệu đầy đủ và hoàn toàn chính xác. Hiện nay DBMS được sử dụng rộng rãi trên thế giới. Một số DBMS thông dụng: Access, SQL Server, Oracle.

Hệ quản lý thông tin (IMS)

Hệ quản lý thông tin là hệ quản trị cơ sở dữ liệu nhưng có thêm nhiều chức năng về việc quản lý. Những chức năng quản lý này phụ thuộc vào giá trị của nhiều kiểu dữ liệu khác nhau. Nói chung bất kỳ hệ thống nào có mục đích đặc biệt phục vụ cho việc quản lý thì ta gọi đó là hệ quản lý thông tin.

Hệ hỗ trợ ra quyết định (DSS)

Hệ hỗ trợ ra quyết đinh sẽ dựa vào các tập luật được học, từ những luật đã học rút ra những luật mới, sau khi gặp một vấn đề nó sẽ căn cứ vào tập các luật để đưa ra những quyết định thay cho con người.

Hệ thống này đang được áp dụng nhiều cho công việc nhận dạng và chẩn đoán bệnh.

Hệ trả lời câu hỏi (QAS)

Hệ trả lời câu hỏi cung cấp việc truy cập đến các thông tin bằng ngôn ngữ tự nhiên. Việc lưu trữ cơ sở dữ liệu thường bao gồm một số lượng lớn các vấn đề liên quan đến các lĩnh vực riêng biệt và các kiến thức tổng quát. Câu hỏi của người dùng có thể ở dạng ngôn ngữ tự nhiên. Công việc của hệ trả lời câu hỏi là phân tích câu truy vấn của người dùng, so sánh với các tri thức được lưu trữ và tập hợp các vấn đề có liên quan lại để đưa ra câu trả lời thích hợp.

Tuy nhiên, hệ trả lời câu hỏi vẫn đang ở giai đoạn thử nghiệm. Việc xác định ý nghĩa của ngôn ngữ tự nhiên dường như vẫn là chướng ngại lớn để có thể sử dụng rỗng rãi hệ thống này.

Bảng 1.3: So sánh hệ truy xuất thông tin với các hệ thống khác

	IRS	**DBMS**	**QAS**	**MIS**
Tìm kiếm	Nội dung trong các tài liệu	Các phần tử có kiểu dữ liệu đã được định nghĩa	Các sự kiện rõ ràng	Giống DBMS nhưng hỗ trợ thêm những thủ tục (tính tổng, tính trung bình, phép chiếu,…)
Lưu trữ	Các văn bản ngôn ngữ tự nhiên	Các phần tử dữ liệu ở dạng bảng	Các sự kiện rõ ràng và các kiến thức tổng quát	
Xử lý	Các câu truy vấn không chính xác	Các câu truy vấn có cấu trúc	Các câu truy vấn không giới hạn	

Chương 2: XÂY DỰNG MỘT HỆ TRUY XUẤT THÔNG TIN

2.1. Một số mô hình xây dựng một hệ truy xuất thông tin

Mục tiêu của các hệ truy xuất thông tin là trả về các tài liệu càng liên quan đến câu hỏi càng tốt. Vì thế người ta đã đưa ra rất nhiều mô hình tìm kiếm nhằm tính toán một cách chính xác độ tương quan này.

Sau đây là một số mô hình tìm kiếm cơ bản:

2.1.1.Mô hình không gian vector

Mô hình không gian vector tính toán độ tương quan giữa câu hỏi và tài liệu bằng cách định nghĩa một vector biểu diễn cho mỗi tài liệu, và một vector biểu diễn cho câu hỏi. Mô hình dựa trên ý tưởng chính là ý nghĩa của một tài liệu thì phụ thuộc vào các từ được sử dụng bên trong nó. Vector tài liệu và vector câu hỏi sau đó sẽ được tính toán để xác định độ tương quan giữa chúng. Độ tương quan càng lớn chứng tỏ tài liệu đó càng liên quan tới câu hỏi.

Giả sử một tập tài liệu chỉ gồm có hai từ là t_1 và t_2. Vector xây dựng được sẽ gồm có 2 thành phần: thành phần thứ nhất biểu diễn sự xuất hiện của t_1, thành phần thứ hai biểu diễn sự xuất hiện của t_2. Cách đơn giản nhất để xây dựng vector là đánh 1 vào thành phần đó nếu nó xuất hiện, và đánh 0 nếu từ đó không xuất hiện. Giả sử tài liệu chỉ gồm có 2 từ t_1. Ta biểu diễn cho tài liệu này bởi một vector nhị phân như sau: <1,0>. Tuy nhiên, biểu diễn như vậy không cho thấy được tần số xuất hiện của mỗi từ trong tài liệu. Trong trường hợp này, vector được biểu diễn như sau: <2,0>

Đối với một câu hỏi đã cho, thay vì chỉ căn cứ so sánh các từ trong tài liệu với tập các từ trong câu hỏi, ta nên xem xét đến tầm quan trọng của mỗi từ. Ý tưởng chính là một từ xuất hiện tập trung trong một số tài liệu thì có trọng số cao hơn so với một từ phân bố trong nhiều tài liệu. Trọng số được tính dựa trên tần số tài liệu nghịch đảo (*Inverse Document Frequency*) liên quan tới các từ được cho:

n: số từ phân biệt trong tập tài liệu

tf_{ij}: số lần xuất hiện của từ t_j trong tài liệu D_i (tần số)

df_j: số tài liệu có chứa từ t_j

$idf_j = \log_{10} \frac{d}{df_j}$ trong đó d là tổng số tài liệu

Vector được xây dựng cho mỗi tài liệu gồm có n thành phần, mỗi thành phần là giá trị trọng số đã được tính toán cho mỗi từ trong tập tài liệu. Các từ trong tài liệu được gán trọng số tự động dựa vào tần số xuất hiện của chúng trong tập tài liệu và sự xuất hiện của mỗi từ trong một tài liệu riêng biệt. Trọng số của một từ tăng nếu từ đó xuất hiện thường xuyên trong một tài liệu và giảm nếu từ đó xuất hiện thường xuyên trong tất cả các tài liệu. Để tính trọng số của từ thứ t_j trong tài liệu D_i, dựa vào công thức:

$$d_{ij} = tf_{ij} * idf_j$$

d_{ij} : là trọng số của từ t_j trong tài liệu D_i

Đối với hệ thống tìm kiếm thông tin theo mô hình vector, mỗi tài liệu là một vector có dạng: $D_i(d_{i1}, d_{i2},\ldots, d_{in})$. Tương tự, câu truy vấn Q cũng là một vector có dạng: $Q(w_{q1}, w_{q2},\ldots, w_{qn})$

W_{qj}: là trọng số của từ t_j trong câu truy vấn Q.

Độ tương quan (SC: Similarity Coeficient) giữa câu truy vấn Q và tài liệu D_i được tính như sau:

$$SC(Q, D_i) = \sum_{j=1}^{n} w_{qj} * d_{ij}$$

2.1.2.Tìm kiếm Boolean

Mô hình tìm kiếm Boolean khá đơn giản. Câu hỏi đưa vào được cho dưới dạng biểu thức Boolean. Nghĩa là phải thỏa:

- Ngữ nghĩa rõ ràng
- Hình thức ngắn gọn

Mô hình liên quan (relevance) cơ bản nhất trong hệ thống truy xuất thông tin cổ điển là mô hình Đại số Bool. Một tài liệu được định nghĩa là một vector boolean d trong $\{0,1\}^k$ (trọng lượng boolean) trong đó d_i =1 khi d_i có mặt trong d. Một câu truy vấn được định nghĩa là một công thức boolean q trên các tokens: q: $\{0,1\}^k \rightarrow \{0,1\}$. Nghĩa là, q là một hàm sao cho với một vector trong $\{0,1\}^k$ cho trước biểu diễn một tài liệu, thì hàm sẽ trả về một giá trị boolean phụ thuộc vào độ liên quan giữa tài liệu và câu truy vấn. Hàm tính độ liên quan được định nghĩa đơn giản bằng cách áp dụng hàm này trên một tài liệu, f(d, q) = q(d).

Ví dụ, một câu truy vấn trong mô hình boolean có thể là “Micheal Jordan” AND (Not basketball). Lợi ích chính của mô hình boolean là tính đơn giản cho người sử dụng. Tuy nhiên, hàm tính độ liên quan của nó quá tồi khi nó chỉ trả về một giá trị boolean.

Do các từ hoặc xuất hiện hoặc là không xuất hiện, nên trọng số $w_{ij} \in \{0, 1\}$. Giả sử đưa vào một câu hỏi dạng biểu thức Boolean như sau: t_1 AND t_2. Sau khi tìm kiếm ta xác định được các tài liệu liên quan đến t_1 là $\{d_1, d_3, d_5\}$ và các tài liệu liên quan đến t_2 là $\{d_3, d_5, d_7\}$. Như vậy với phép AND, các tài liệu thoả yêu cầu của người dùng là $\{d_3, d_5\}$.

Phương pháp này có một số khuyết diểm như sau:

- Các tài liệu trả về không được sắp xếp (ranking).
- Câu hỏi tìm kiếm đòi hỏi phải đúng định dạng của biểu thức Boolean gây khó khăn cho người dùng.
- Kết quả trả về có thể là quá ít hoặc quá nhiều tài liệu.

2.1.3.Tìm kiếm Boolean mở rộng

Mô hình tìm kiếm Boolean không hỗ trợ việc sắp xếp kết quả trả về bởi vì các tài liệu hoặc thoả hoặc không thoả yêu cầu Boolean. Tất cả các tài liệu thoả mãn đều được trả về. Ở đây chưa có ước lượng nào được tính toán mức độ liên quan của chúng đối với câu hỏi.

Mô hình tìm kiếm Boolean mở rộng ra đời nhằm hỗ trợ việc sắp xếp (ranking) kết quả trả về dựa trên ý tưởng cơ bản là đánh trọng số cho mỗi từ trong câu hỏi và trong tài liệu. Giả sử một câu hỏi yêu cầu (t_1 OR t_2) và một tài liệu D có chứa t_1 với trọng số w_1 và t_2 với trọng số w_2. Nếu w_1 và w_2 đều bằng 1 thì tài liệu nào có chứa cả hai từ này sẽ có thứ tự sắp xếp cao nhất. Tài liệu nào không chứa một trong hai từ này sẽ có thứ tự sắp xếp thấp nhất. Ý tưởng đơn giản là tính khoảng cách Euclide từ điểm (w1, w2) tới gốc:

$$SC(Q, D_i) = \sqrt{(w_1)^2 + (w_2)^2}$$

Với trọng số 0.5 và 0.5, $SC(Q, D_i) = \sqrt{(0.5)^2 + (0.5)^2} = 0.707$

SC cao nhất nếu w_1 và w_2 đều bằng 1. Khi đó: $SC(Q, D_i) = \sqrt{2} = 1.414$

Để đưa SC vào khoảng [0, 1], SC được tính như sau:

$$SC(Q_{t1 \vee t2}, d_i) = \frac{\sqrt{(w_1)^2 + (w_2)^2}}{\sqrt{2}}$$

Công thức này giả sử là câu hỏi chỉ có toán tử OR. Đối với toán tử AND, thay vì tính khoảng cách tới gốc, ta sẽ tính khoảng cách đến điểm (1, 1). Câu hỏi nào càng gần đến điểm (1, 1) thì nó càng thoả yêu cầu của toán tử AND:

$$SC\left(Q_{t1 \wedge t2}, d_i\right) = 1 - \frac{\sqrt{\left(1-w_1\right)^2 + \left(1-w_2\right)^2}}{\sqrt{2}}$$

2.1.4.Mô hình xác suất

Mô hình tìm kiếm xác suất tính toán độ tương quan giữa câu hỏi và tài liệu dựa vào xác suất mà tài liệu đó liên quan đến câu hỏi. Các xác suất được áp dụng để tính toán độ liên quan giữa câu hỏi và tài liệu. Các từ trong câu hỏi được xem là mối để xác định tài liệu liên quan. Ý tưởng chính là tính xác suất của mỗi từ trong câu hỏi và sau đó sử dụng chúng để tính xác suất mà tài liệu liên quan đến câu hỏi.

2.1.5. Đánh giá chung về các mô hình

Mô hình Boolean được xem là mô hình yếu nhất trong các mô hình bởi vì như đã trình bày nó có rất nhiều nhược điểm.

Theo kinh nghiệm của Salton và Buckley, nhìn chung mô hình vector làm tốt hơn mô hình xác suất.

2.2. Các bước xây dựng một hệ truy xuất thông tin

2.2.1. Tách từ tự động cho tập các tài liệu.

Đối với tiếng Anh, việc tách từ đơn giản chỉ dựa vào khoảng trắng. Tuy nhiên đối với tiếng Việt, giai đoạn này tương đối khó khăn. Cấu trúc tiếng Việt rất phức tạp, không chỉ đơn thuần dựa vào khoảng trắng để tách từ. Hiện

nay có rất nhiều công cụ dùng để tách từ tiếng Việt, mỗi phương pháp có ưu, khuyết điểm riêng. Ở đây ta xét một số phương pháp hay sử dụng trong tiếng Việt.

Các phương pháp tách từ tiếng Việt:

Phương pháp fnTBL (Fast Transformation – Based Learning)

Ý tưởng chính của phương pháp học dựa trên sự biến đổi (TBL) là để giải quyết một vấn đề nào đó ta sẽ áp dụng các phép biến đổi, tại mỗi bước, phép biến đổi nào cho kết quả tốt nhất sẽ được chọn và được áp dụng lại với vấn đề đã đưa ra. Thuật toán kết thúc khi không còn phép biến đổi nào được chọn. Hệ thống fnTBL gồm hai tập tin chính:

- Tập tin dữ liệu học (Training): Tập tin dữ liệu học được làm thủ công, đòi hỏi độ chính xác. Mỗi mẫu (template) được đặt trên một dòng riêng biệt.
- Tập tin chứa các mẫu luật (rule-template): mỗi luật được đặt trên một dòng, hệ thống fnTBL sẽ dựa vào các mẫu luật để áp dụng vào tập tin dữ liệu học.

Phương pháp Longest Matching

Phương pháp Longest Matching dựa vào từ điển có sẵn.

Theo phương pháp này, để tách từ tiếng Việt ta đi từ trái sang phải và chọn từ có nhiều âm tiết nhất mà có mặt trong từ điển rồi tiếp tục cho các từ kế tiếp cho đến hết câu. Với cách này, ta dễ dàng tách được chính xác các ngữ/câu như: “hợp tác| mua bán”, “thành lập| nước| Việt Nam| dân chủ| cộng hoà”,…Tuy nhiên, phương pháp này sẽ tách từ sai trong trường hợp như: “học sinh| học sinh| học”, “một| ông| quan tài giỏi”, “trước| bàn là| một| ly nước”,…

Kết hợp giữa fnTBL và Longest Matching

Chúng ta có thể kết hợp giữa hai phương pháp fnTBL và Longest Matching để có được kết quả tách từ tốt nhất. Đầu tiên ta sẽ tách từ bằng Longest Matching, đầu ra của phương pháp này sẽ là đầu vào cho phương pháp fnTBL học luật.

2.2.2. Lập chỉ mục cho tài liệu

Sau khi có được tập các từ đã được trích, ta sẽ chọn các từ để làm từ chỉ mục. Tuy nhiên, không phải từ nào cũng được chọn làm từ chỉ mục. Các từ có khả năng đại diện cho tài liệu sẽ được chọn, các từ này được gọi là key word, do đó trước khi lập chỉ mục sẽ là giai đoạn tiền xử lý đối với các từ trích được để chọn ra các key word thích hợp. Ta sẽ loại bỏ danh sách các từ ít có khả năng đại diện cho nội dung văn bản dựa vào danh sách gọi là stop list. Đối với tiếng Anh hay tiếng Việt đều có danh sách Stop list.

Lập chỉ mục bao gồm các công việc: Xác định các từ có khả năng đại diện cho nội dung của tài liệu và đánh trọng số cho các từ này, trọng số phản ánh tầm quan trọng của từ trong một tài liệu.

Lập chỉ mục cho tài liệu sẽ được xem xét cụ thể ở chương sau.

2.2.3. Tìm kiếm

Mục đích của tìm kiếm là cho phép ánh xạ giữa một yêu cầu riêng biệt của người dùng và các item trong cơ sở dữ liệu thông tin trả lời yêu cầu đó. Người dùng sử dụng các câu truy vấn tìm kiếm để giao tiếp mô tả các thông tin được yêu cầu với hệ thống.

Người dùng nhập câu hỏi và yêu cầu tìm kiếm, câu hỏi mà người dùng nhập vào cũng sẽ được xử lý, nghĩa là ta sẽ tách từ cho câu hỏi. Phương pháp tách từ cho câu hỏi cũng nên là phương pháp tách từ cho các tài liệu thu thập

được để đảm bảo sự tương thích. Sau đó, hệ thống sẽ tìm kiếm trong tập tin chỉ mục để xác định các tài liệu liên quan đến câu hỏi của người dùng.

2.2.4. Sắp xếp các tài liệu trả về (Ranking)

Các tài liệu sau khi đã xác định là liên quan đến câu hỏi của người dùng sẽ được sắp xếp lại, bởi vì trong các tài liệu đó có những tài liệu liên quan đến câu hỏi nhiều hơn. Hệ thống sẽ dựa vào một số phương pháp để xác định tài liệu nào liên quan nhiều nhất, sắp xếp lại (ranking) và trả về cho người dùng theo thứ tự ưu tiên.

Chương 3: LẬP CHỈ MỤC

3.1. Khái quát về hệ thống lập chỉ mục

Các trang tài liệu sau khi thu thập về sẽ được phân tích, trích chọn những thông tin cần thiết (thường là các từ đơn, từ ghép, cụm từ quan trọng) để lưu trữ trong cơ sở dữ liệu nhằm phục vụ cho nhu cầu tìm kiếm sau này.

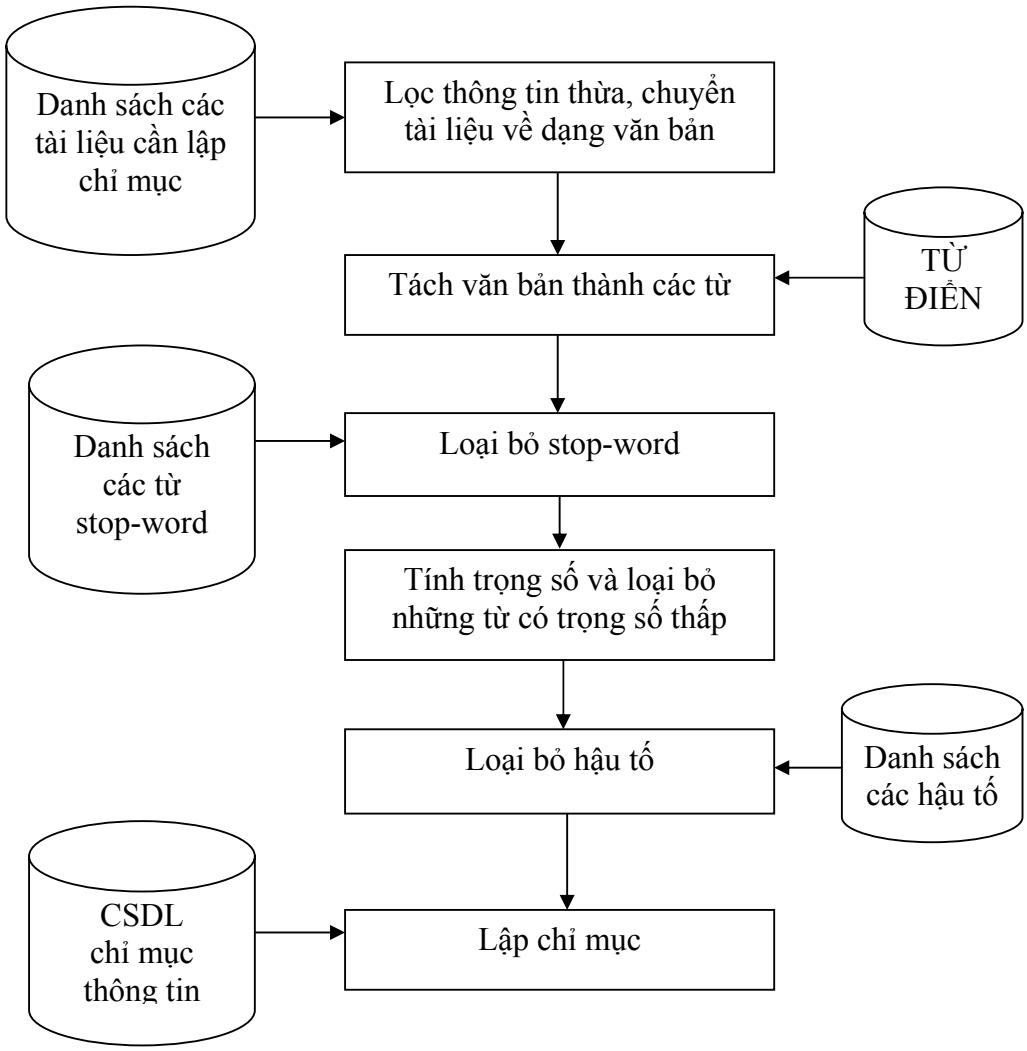
Một cách để tăng tốc độ tìm kiếm thông tin lên là tạo chỉ mục cho các tài liệu. Tuy nhiên, việc lập chỉ mục có một nhược điểm lớn, đó là khi thêm một tài liệu mới, phải cập nhật lại tập tin chỉ mục. Nhưng đối với hệ thống tìm kiếm thông tin, chỉ cần cập nhật lại tập tin chỉ mục vào một khoảng thời gian định kỳ. Do đó, chỉ mục là một công cụ rất có giá trị.

Lập chỉ mục bao gồm các công việc sau:

- Xác định các từ có khả năng đại diện cho nội dung của tài liệu
- Đánh trọng số cho các từ này, trọng số phản ánh tầm quan trọng của từ trong một tài liệu.

Lập chỉ mục là quá trình phân tích và xác định các từ, cụm từ thích hợp cốt lõi có khả năng đại diện cho nội dung của tài liệu. Như vậy, vấn đề đặt ra là phải rút trích ra những thông tin chính, có khả năng đại diện cho nội dung của tài liệu. Thông tin này phải "vừa đủ", nghĩa là không thiếu để trả ra kết quả đầy đủ so với nhu cầu tìm kiếm, nhưng cũng phải không dư để giảm chi phí lưu trữ và chi phí tìm kiếm và để loại bỏ kết quả dư thừa không phù hợp. Việc rút trích này chính là việc lập chỉ mục trên tài liệu. Trước đây, quá trình này thường được các chuyên viên đã qua đào tạo thực hiện một cách "thủ công" nên có độ chính xác cao. Nhưng trong môi trường hiện đại ngày nay, với lượng thông tin khổng lồ thì việc lập chỉ mục bằng tay không còn phù hợp, phương pháp lập chỉ mục tự động mang lại hiệu quả cao hơn.

Mô hình xử lý tổng quát của một hệ thống được trình bày như sau:



Hình 3.1: Lưu đồ xử lý cho hệ thống lập chỉ mục

3.2. Xác định mục từ quan trọng cần lập chỉ mục

Mục từ hay còn gọi là mục từ chỉ mục, là đơn vị cơ sở cho quá trình lập chỉ mục. Mục từ có thể là từ đơn, từ phức hay một tổ hợp từ có nghĩa trong một ngữ cảnh cụ thể. Ta xác định mục từ của 1 văn bản dựa vào chính nội

dung của văn bản đó, hoặc dựa vào tiêu đề hoặc tóm tắt nội dung của văn bản đó.

Hầu hết việc lập chỉ mục tự động bắt đầu với việc khảo sát tần số xuất hiện của từng loại từ riêng rẽ trong văn bản. Nếu tất cả các từ xuất hiện trong tập tài liệu với những tần số bằng nhau, thì không thể phân biệt các mục từ theo tiêu chuẩn định lượng. Tuy nhiên, trong văn bản ngôn ngữ tự nhiên, tần số xuất hiện của từ có tính thất thường, Do đó những mục từ có thể được phân biệt bởi tần số xuất hiện của chúng.

Đặc trưng xuất hiện của từ vựng có thể được định bởi hằng số "thứ hạng - tần số" (Rank_Frequency) theo luật của Zipf :

Tần số xuất hiện * thứ hạng = Hằng số.

Biểu thức luật Zipf có thể dẫn ra những hệ số ý nghĩa của từ dựa vào những đặc trưng của tân số xuất hiện của mục từ riêng lẽ trong những văn bản tài liệu.

Một đề xuất dựa theo sự xem xét chung sau:

1. Cho một tập hợp n tài liệu, trong mỗi tài liệu tính toán tần số xuất hiện của các mục từ trong tài liệu đó.

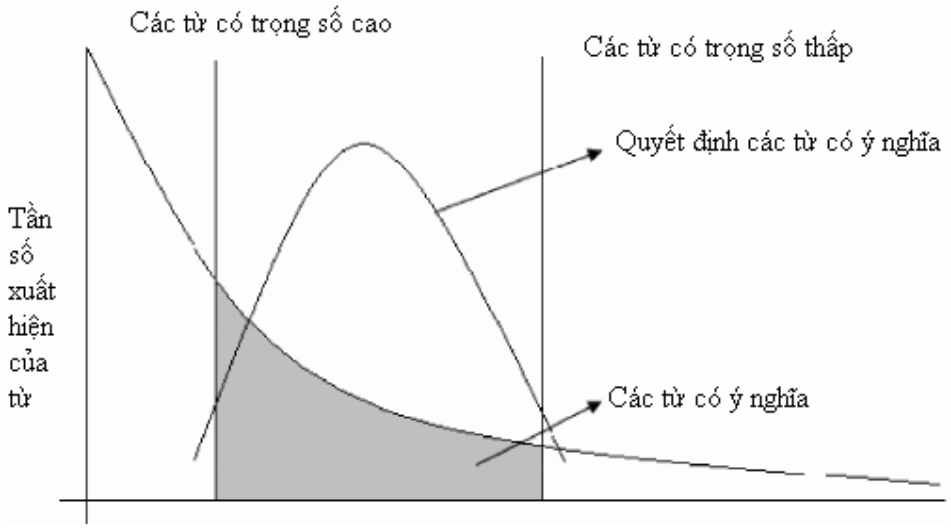
   Kí hiệu F_{ik} (Frequency): tần số xuất hiện của mục từ k trong tài liệu i.
2. Xác định tổng số tập tần số xuất hiện TF_k (Total Frequency) cho mỗi từ bằng cách cộng những tần số của mỗi mục từ duy nhất trên tất cả n tài liệu.

$$TF_k = \sum_{i}^{n} F_{ik}$$

3. Sắp xếp những thứ tự giảm theo tập tần số xuất hiện của chúng. Quyết định giá trị ngưỡng cao và loại bỏ tất cả những từ có tập tần số xuất

hiện cao trên ngưỡng này. Những từ bị loại bỏ là những từ xuất hiện phổ biến ở hầu hết các tài liệu. Đó chính là các Stop-Word.

4. Tương tự, loại trừ những từ được xem là có tần số xuất hiện thấp. Nghĩa là, xác định ngưỡng thấp và loại bỏ tất cả các từ có tần số nhỏ hơn giá trị này. Điều này sẽ loại bỏ các từ ít xuất hiện trong tập tài liệu, nên sự có mặt của các từ này cũng không ảnh hưởng đến việc thực hiện truy vấn.
5. Những từ xuất hiện trung bình còn lại bây giờ được dùng cho việc ấn định tới những tài liệu như những mục từ chỉ mục.



Hình 3.2: Các từ được sắp theo thứ tự

Chú ý: một khái niệm xuất hiện ít nhất hai lần trong cùng một đoạn thì được xem là một khái niệm chính. Một khái niệm xuất hiện trong hai đoạn văn liên tiếp cũng được xem là một khái niệm chính mặc dù nó chỉ xuất hiện duy nhất một lần trong đoạn đang xét. Tất cả những chú giải về những khái niệm chính được liệt kê theo một tiêu chuẩn nhất định nào đó.

Thực tế cho thấy rằng ý tưởng trên khá cứng nhắc, vì nếu loại bỏ tất cả những từ có tần số xuất hiện cao sẽ làm giảm giá trị recall (độ bao phủ), tức giảm hiệu quả trong việc trả về số lượng lớn của những mục tin thích đáng.

Ngược lại, sự loại bỏ những mục từ có tần số xuất hiện thấp có thể làm giảm giá trị của độ chính xác. Một vấn đề khác là sự cần thiết để chọn những ngưỡng thích hợp theo thứ tự để phân biệt những mục từ hữu ích có tần số xuất hiện trung bình trong phần còn lại.

3.3. Một số hàm tính trọng số mục từ

Trọng số của một từ phản ánh tầm quan trọng của từ đó trong tài liệu. Ý tưởng chính là một từ xuất hiện thường xuyên trong tất cả các tài liệu thì ít quan trọng hơn là từ chỉ xuất hiện tập trung trong một số tài liệu.

Trọng số của mục từ: là sự tần xuất xuất hiện của mục từ trong toàn bộ tài liệu. Phương pháp thường được sử dụng để đánh giá trọng số của từ là dựa vào thống kê, với ý tưởng là những từ thường xuyên xuất hiện trong tất cả các tài liệu thì "ít có ý nghĩa hơn" là những từ tập trung trong một số tài liệu.
Ta xét các khái niệm sau:

- Gọi $T=\{t_1, t_2,..., t_n\}$ là không gian chỉ mục, với t_i là các mục từ.
- Một tài liệu D được lập chỉ mục dựa trên tập T sẽ được biểu diễn dưới dạng:

  $T(D)=\{w_1,w_2,...w_n\}$ với w_i là trọng số của t_i trong tập tài liệu D.

Nếu $w_i=0$ nghĩa là t_i không xuất hiện trong D hoặc mục từ t_i ít quan trọng trong tài liệu D ta không quan tâm tới.

T(D) được gọi là vector chỉ mục của D, nó được xem như biểu diễn cho nội dung của tài liệu D và được lưu lại trong cơ sở dữ liệu của hệ thống tìm kiếm thông tin để phục vụ cho nhu cầu tìm kiếm.

Mặc dù T(D) biểu diễn nội dung của tài liệu D nhưng không phải bất cứ từ nào có trong D đều xuất hiện trong T(D) mà chỉ có những từ có trọng lượng (có ý nghĩa quan trọng trong tài liệu D) mới được lập chỉ mục cho D.
Sau đây ta xét một số hàm tính trọng số của mục từ:

3.3.1. Tần số tài liệu nghịch đảo (Inverse Document Frequency)

Đây là phương pháp tính trọng số mà mô hình không gian vector đã sử dụng để tính trọng số của từ trong tài liệu.

N: số từ phân biệt trọng tập tài liệu

$FREQ_{ik}$: số lần xuất hiện của từ k trong tài liệu D_i (tần số từ)

$DOCFERQ_k$: số tài liệu có chứa từ k

Khi đó trọng số của từ k trong tài liệu D_i được tính như sau:

$$WEIGHT_{ik} = FREQ_{ik} * [\log_2(n) - \log_2(DOCFREQ_k)]$$

Trọng số của từ k trong tài liệu D_i tăng nếu tần số xuất hiện của từ k trong tài liệu i tăng và giảm nếu tổng số tài liệu có chứa từ k tăng.

3.3.2. Độ nhiễu tín hiệu (The Signal – Noise Ratio)

Trọng số của từ được đo lường bằng sự tập trung hay phân tán của từ. Ví dụ từ "hardware" xuất hiện 1000 lần nhưng trong 200 tài liệu (tập trung) thì có trọng lượng cao hơn từ "computer" cũng xuất hiện 1000 lần nhưng trong 800 tài liệu.

Một quan điểm tương tự được xem xét đó là dựa vào thông tin để đánh giá tầm quan trọng của từ. Trong thực tế, nội dung thông tin của một đoạn hay một từ có thể xác định dựa vào xác suất xuất hiện của các từ trong văn bản đã cho. Rõ ràng, xác suất xuất hiện của một từ càng cao thì thông tin mà nó chứa càng ít.

Nội dung thông tin của một từ được xác định như sau:

$$INFORMATION = - Log_2 p,$$

trong đó p là xác suất xuất hiện của từ.

Ví dụ: nếu từ "vi tính" xuất hiện 1 lần sau 10000 từ, xác suất xuất hiện của nó là 0.0001, khi đó thông tin của nó sẽ là:

$$\text{INFORMATION} = -\text{Log}_2(0.0001) = 13.278$$

Ngược lại, từ "sẽ" xuất hiện 1 lần sau 10 từ, xác suất xuất hiện của nó là 0.1, khi đó thông tin của nó sẽ là:

$$\text{INFORMATION} = -\text{Log}_2(0.1) = 3.223$$

Nếu một tài liệu có chứa t từ, mỗi từ có xác suất xuất hiện là p_k, thông tin trung bình của tài liệu sẽ là:

$$\text{AVERAGE INFORMATION} = -\sum_{k=1}^{t} p_k \log_2 p_k$$

Ta định nghĩa độ nhiễu NOISE_k của từ k trong tập tài liệu như sau:

$$\text{NOISE}_k = \sum_{i=1}^{n} \frac{\text{FREG}_{ik}}{\text{TOTFREQ}_k} \log_2 \frac{\text{TOTFREQ}_k}{\text{FREG}_{ik}}$$

Độ nhiễu thay đổi nghịch đảo với "sự tập trung" của một từ trong tập tài liệu. Nghĩa là, nếu một từ được phân phối đều trong tất cả các tài liệu thì độ nhiễu của nó càng lớn. Ngược lại, nếu một từ chỉ tập trung trong một số tài liệu nào đó thì độ nhiễu càng nhỏ.

Giả sử, từ k xuất hiện một lần trong mỗi tài liệu ($\text{FREQ}_{ik} = 1$), khi đó độ nhiễu của nó bằng:

$$\text{NOISE}_k = \sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{n}{1} = \log_2 n$$

Ngược lại, giả sử k chỉ xuất hiện trong một tài liệu, khi đó độ nhiễu của nó bằng:

$$\text{NOISE}_k = \sum_{i=1}^{n} \frac{\text{TOTFREQ}_k}{\text{TOTFREQ}_k} \log_2 \frac{\text{TOTFREQ}_k}{\text{TOTFREQ}_k} = 0$$

Hàm số nghịch đảo của độ nhiễu, gọi là độ signal, được tính như sau:

$$\text{SIGNAL}_k = \log_2(\text{TOTFREQ}_k) - \text{NOISE}_k$$

Trọng số của từ k trong tài liệu i được tính bằng cách kết hợp giữa FREQ_{ik} và SIGNAL_k:

$$WEIGHT_{ik} = FREQ_{ik} * SIGNAL_k$$

3.3.3. Giá trị độ phân biệt của mục từ (Term Discrimination Value)

Rõ ràng là kết quả tìm kiếm trở lên không có giá trị khi trả về tập tất cả các tài liệu có trong tập hợp (nghĩa là tập chỉ mục của các tài liệu chứa nhiều từ giống nhau). Độ phân biệt của mục từ là giá trị phân biệt mức độ tương đương giữa các tài liệu. Nếu một mục từ có trong chỉ mục mà làm cho độ tương tự của các tài liệu cao thì nó có độ phân biệt kém (nghĩa là từ này thường xuyên xuất hiện trong các tài liệu) và ngược lại. Như vậy, các mục từ có độ phân biệt cao nên được chọn để lập chỉ mục. Thực chất, việc sử dụng độ phân biệt này cũng cho kết quả tương đương với việc sử dụng tần số nghịch đảo và tỉ lệ tín hiệu nhiễu.

Một chức năng khác để xác định tầm quan trọng của một từ là tính giá trị phân biệt của từ đó. Gọi $SIMILAR(D_i, D_j)$ là độ tương quan giữa cặp tài liệu D_i, D_j. Khi đó, độ tương quan trung bình của tập tài liệu là:

$$AVGSIM = CONSTANT \sum_{i=1}^{n} \sum_{i \neq j}^{n} SIMILAR(D_i, D_j)$$

Gọi $AVGSIM_k$ là độ tương quan trung bình của tập tài liệu khi bỏ từ k. Rõ ràng, nếu từ k xuất hiện thường xuyên trong tập tài liệu thì khi bỏ từ k, độ tương quan trung bình sẽ giảm. Ngược lại, nếu từ k chỉ tập trung trong một số tài liệu, khi bỏ từ k, độ tương quan trung bình sẽ tăng lên.

Giá trị phân biệt $DISVALUE_k$ của từ k được tính như sau:

$$DISCVALUE_k = (AVGSIM)_k - AVGSIM$$

Trọng số của từ k trong tài liệu thông tin được tính bằng cách kết hợp giữa $FREQ_{ik}$ và $DISCVALUE_k$:

$$WEIGHT_{ik} = FREQ_{ik} * DISCVALUE_k$$

Phép tính $DISCVALUE_k$ cho tất cả những mục từ k, những mục từ có thể được xếp theo thứ tự giảm của giá trị phân biệt $DISCVALUE_k$. Những mục từ chỉ mục có thể thuộc một trong ba nhóm dựa theo giá trị độ phân biệt của chúng như sau:

- Độ phân biệt tốt đối vơi $DISCVALUE_k$ dương, những mục từ có độ phân biệt cao.
- Đối với $DISCVALUE_k$ gần bằng 0, độ phân biệt giữa các tài liệu không khác nhau khi thêm vào hay bớt đi những mục từ đó.
- Độ phân biệt yếu khi $DISCVALUE_k$ âm, những mục từ có độ phân biệt thấp (độ tương tự cao).

3.4. Lập chỉ mục cho tài liệu tiếng Anh

Một quá trình đơn giản để lập chỉ mục cho tài liệu có thể được mô tả như sau:

- Trước hết, xác định tất cả các từ tạo thành tài liệu. Trong tiếng Anh, chỉ đơn giản là tách từ dựa vào khoảng trắng.
- Loại bỏ các từ có tần số xuất hiện cao. Những từ này chiếm khoảng 40-50% các từ, như đã đề cập trước đây, chúng có độ phân biệt kém do đó không thể sử dụng để đại diện cho nội dung của tài liệu. Trong tiếng Anh, các từ này có khoảng 250 từ, do đó, để đơn giản có thể lưu chúng vào từ điển gọi là Stop List.
- Sau khi loại bỏ các từ có trong Stop List, xác định các từ chỉ mục "tốt". Trước hết cần loại bỏ các hậu tố đưa về từ gốc, ví dụ các từ như: analysis, analyzing, analyzer, analyzed, analysing có thể chuyển về từ gốc là "analy". Từ gốc sẽ có tần số xuất hiện cao hơn so với các dạng thông thường của nó. Nếu sử dụng từ gốc làm chỉ mục, ta có thể thu được nhiều tài liệu liên quan hơn là sử dụng từ ban đầu của nó.

Đối với tiếng Anh, việc loại bỏ hậu tố có thể được thực hiện dễ dàng bằng cách sử dụng danh sách các hậu tố có sẵn (Suffix List).

Sau khi có được danh sách các từ gốc, sử dụng phương pháp dựa vào tần số (frequency – based) để xác định tầm quan trọng của các từ gốc này. Chúng ta có thể sử dụng một trong các phương pháp đã được đề cập ở trên như: tần số tài liệu nghịch đảo (Inverse Document Frequency), độ nhiễu tín hiệu ($SIGNAL_k$), độ phân biệt từ ($DISCVALUE_k$).

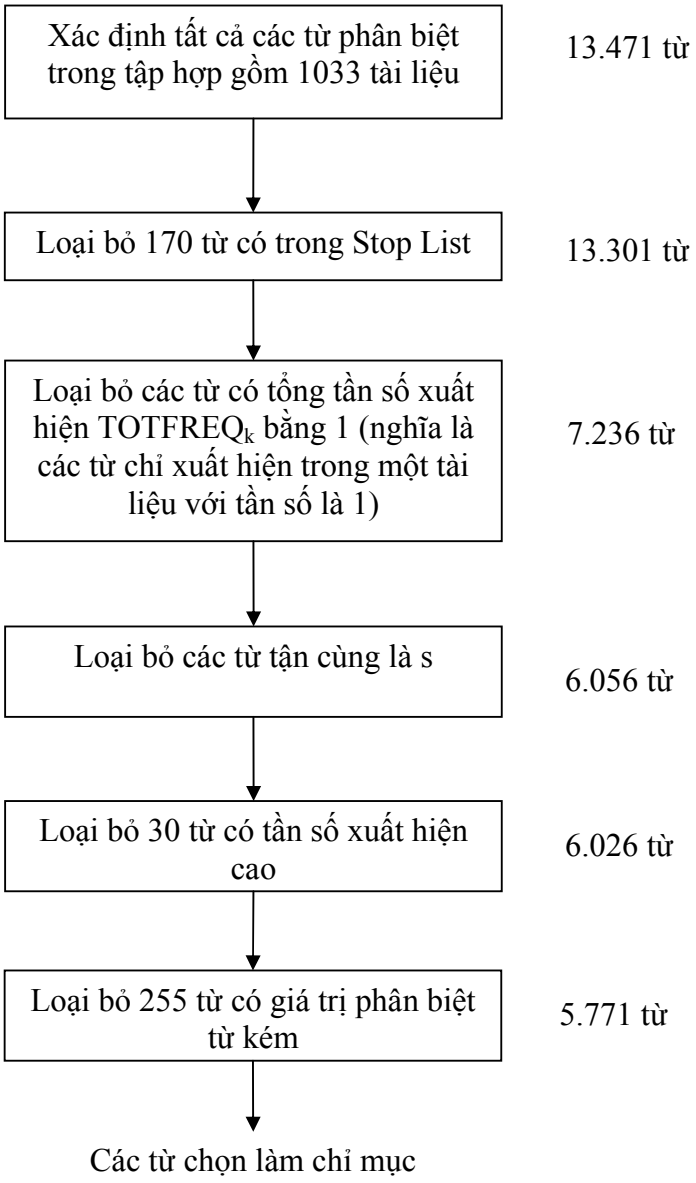
Trong hệ thống chỉ mục có trọng số, trọng số của một từ được sử dụng để xác định tầm quan trọng của từ đó. Mỗi tài liệu được biểu diễn là một vector: $D_i = (d_{i1}, d_{i2},\ldots, d_{in})$ trong đó d_{ij} là trọng số của từ j trong tài liệu D_i.

Giả sử có 1033 tài liệu nói về y học. Quá trình lập chỉ mục đơn giản được thực hiện theo hình 3.3 (trong đó chỉ loại bỏ hậu tố tận cũng là s).

Quá trình stemming: Trong quá trình lập chỉ mục Tiếng Anh, Stemming là quá trình lược bỏ các suffix (phần hậu tố/tiếp vĩ ngữ) của các từ. Việc nằm làm tăng giá trị recall của chương trình, làm cấu trúc cây từ điển chính xác và gọn nhẹ hơn, đương nhiên hiệu quả truy vấn cũng cao hơn.

Ví dụ: studies, studying, studied là các biến thể khác nhau của từ gốc study, nếu không có giai đoạn stemming này thì tất cả các từ này đều được lập chỉ mục và bổ sung vào cây từ điển nếu nó chưa có. Rõ ràng điều này là khuyết điểm lớn của chương trình.

Có nhiều thuật toán phổ biến cho việc loại bỏ phần đuôi của một từ tiếng Anh, thông thường đều dựa vào danh sách các hậu tố để đối chiếu.



Hình 3.3: Quá trình chọn từ làm chỉ mục

3.5. Lập chỉ mục cho tài liệu tiếng Việt

Lập chỉ mục cho tài liệu tiếng Việt cũng tương tự như cho tiếng Anh. Tuy nhiên có vài điểm khác biệt sau:

- Giai đoạn tách từ trong tiếng Anh chỉ đơn giản dựa vào khoảng trắng, còn tiếng Việt là ngôn ngữ đơn lập, một từ có thể có nhiều tiếng. Giả sử sau giai đoạn tách từ, ta sẽ thu được một danh sách các từ riêng biệt.
- Đối với tiếng Việt, không phải qua giai đoạn loại bỏ hậu tố.

Lập chỉ mục cho tài liệu tiếng Việt gồm các bước sau:

- Xác định các từ riêng biệt trong tài liệu
- Loại bỏ các từ có tần số cao (Trong tiếng Việt, cũng như tiếng Anh, ta có một danh sách Stop List chứa những từ không thể là nội dung của văn bản như: và, với, những, gì, sao, nào,…).
- Loại bỏ các từ có trọng số thấp
- Các từ thu được sẽ được chọn làm các từ chỉ mục.

3.5.1. Khó khăn cho việc lập chỉ mục tiếng Việt

Các điểm khó khăn khi thực hiện quá trình lập chỉ mục cho tài liệu tiếng Việt so với tài liệu tiếng Anh cần phải giải quyết :

- Xác định **ranh giới** giữa các từ trong câu. Đối với tiếng Anh điều này quá dễ dàng vì khoảng trắng chính là ranh giới phân biệt các từ ngược lại tiếng Việt thì khoảng trắng không phải là ranh giới để xác định các từ mà chỉ là ranh giới để xác định các tiếng.

- Chính tả tiếng Việt còn một số điểm chưa thống nhất như sử dụng "**y" hay "i**" (ví dụ "quý" hay "quí"), **cách bỏ dấu** ("lượng" hay "lượng"), **cách viết hoa** tên riêng("Khoa học Tự nhiên" hay "Khoa Học Tự Nhiên")... đòi hỏi quá trình hiệu chỉnh chính tả cho văn bản cần lập chỉ mục và cho từ điển chỉ mục.

- Tồn tại nhiều **bảng mã tiếng Việt** đòi hỏi khả năng xử lý tài liệu ở các bảng mã khác nhau. Cách giải quyết là đưa tất cả về bảng mã chuẩn của hệ thống.

- Sự phong phú về nghĩa của một từ (**từ đa nghĩa**). Một từ có thể có nhiều nghĩa khác nhau trong những ngữ cảnh khác nhau nên việc tìm kiếm khó có được kết quả với độ chính xác cao.

- **Từ đồng nghĩa hoặc từ gần nghĩa:** có nhiều từ khác nhau nhưng lại có cùng ý nghĩa. Do đó, việc tìm kiếm theo từ khoá thường không tìm thấy các websites chứa từ đồng nghĩa hoặc gần nghĩa với từ cần tìm. Vì vậy, việc tìm kiếm cho ra kết quả không đầy đủ.

- Có quá nhiều từ mà mật độ xuất hiện cao nhưng không mang ý nghĩa cụ thể nào mà chỉ là những từ nối, từ đệm hoặc chỉ mang sắc thái biểu cảm như những từ láy. Những từ này cần phải được xác định và loại bỏ ra khỏi tập các mục từ. Nó giống như **stop-word** trong tiếng Anh.

- Các văn bản có nội dung chính là một vấn đề cụ thể, một đề tài nghiên cứu khoa học nhưng đôi khi trọng số của các từ chuyên môn này thấp so với toàn tập tài liệu. Vì vậy, một số thuật toán tính trọng số bỏ sót những trường hợp như vậy. Kết quả là các từ chuyên môn đó không được lập chỉ mục.

Trong các vấn đề trên, việc **xác định ranh giới từ** trong câu là quan trọng nhất vì nó ảnh hưởng lớn đến hiệu quả của quá trình lập chỉ mục (**nếu quá trình tách từ sai có nghĩa là nội dung của câu bị phân tích sai**) và **cũng là vấn đề khó khăn nhất**. Các vấn đề còn lại chỉ là thuần tuý về mặt kỹ thuật mà hầu như chúng ta có thể giải quyết một cách triệt để.

3.5.2. Đặc điểm về từ trong tiếng Việt

Tiếng Việt là ngôn ngữ đơn lập. Đặc điểm này bao quát tiếng Việt cả về mặt ngữ âm, ngữ nghĩa, ngữ pháp. Khác với các ngôn ngữ Ấn-Âu, mỗi từ là một nhóm các ký tự có nghĩa được cách nhau bởi một khoảng trắng. Còn tiếng Việt và các ngôn ngữ đơn lập khác, thì khoảng trắng không phải là căn cứ để nhận diện từ.

Tiếng:

- Trong tiếng Việt trước hết cần chú ý đến đơn vị xưa nay vẫn quan gọi là tiếng. Về mặt ngữ nghĩa, ngữ âm, ngữ pháp đều có giá trị quan trọng.

- Sử dụng tiếng để tạo từ có hai trường hợp:

  - ✓ Trường hợp một tiếng: đây là trường hợp một tiếng được dùng làm một từ, gọi là từ đơn. Tuy nhiên không phải tiếng nào cũng tạo thành một từ.

  - ✓ Trường hợp hai tiếng trở lên: đây là trường hợp hai hay nhiều tiếng kết hợp với nhau, cả khối kết hợp với nhau gắn bó tương đối chặt chẽ, mới có tư cách ngữ pháp là một từ. Đây là trường hợp từ ghép hay từ phức.

Từ:

Có rất nhiều quan niệm về từ trong tiếng Việt, từ nhiều quan niệm về từ tiếng Việt khác nhau đó chúng ta có thể thấy đặc trưng cơ bản của "từ" là sự hoàn chỉnh về mặt nội dung, từ là đơn vị nhỏ nhất để đặt câu.

Người ta dùng "từ" kết hợp thành câu chứ không phải dùng "tiếng" do đó quá trình lập chỉ mục bằng cách tách câu thành các "từ" cho kết qua tốt hơn là tách câu bằng "tiếng".

3.5.3. Việc tách từ

Việc xác định từ trong tiếng Việt là rất khó và tốn nhiều chi phí. Do đó, cách đơn giản nhất là **sử dụng từ điển được lập sẵn**. Tách tài liệu thành các từ, loại bỏ các từ láy, từ nối, từ đệm, các từ không quan trọng trong tài liệu. Một câu gồm nhiều từ ghép lại. Tuy nhiên, trong một câu có thể có nhiều cách phân tích từ khác nhau.

Ví dụ : xét câu "Tốc độ truyền thông tin sẽ tăng cao" có thể phân tích từ theo các cách sau:

Tốc độ / truyền/ thông tin / sẽ / tăng cao.
Tốc độ / truyền thông / tin / sẽ / tăng cao.

Hiện đã có nhiều giải pháp cho vấn đề này với kết quả thu được rất cao. Tuy nhiên thời gian, chi phí tính toán, xử lý lớn không thích hợp cho việc lập chỉ mục cho hệ thống tìm kiếm thông tin vì số lượng tài liệu phải xử lý là rất lớn.

Cách giải quyết: **lập chỉ mục cho các từ có thể có trong một tài liệu**. Ví dụ câu trên ta nên lập xem xét các từ : tốc độ, truyền, truyền thông, thông tin, tin, sẽ, tăng cao.

Sau đó sẽ **dùng ngưỡng** chặn để loại bỏ các từ, giả sử từ "truyền thông" không phải là một từ xuất hiện thật sự trong tài liệu (chỉ có được do sự kết hợp ngẫu nhiên từ "truyền" và "thông tin") thì xác suất xuất hiện của từ này trong tài liệu sẽ không cao nên khi tính toán trọng lượng thì từ này sẽ bị

loại bỏ. Một từ trong tiếng Việt là sự kết hợp của hai hay nhiều tiếng. Phương pháp xác định một từ được ghép lại thông qua nhiều tiếng dựa trên việc xem xét độ gắn kết (cohesion) giữa chúng:

$$Cohension(n_{ij}) = size\_factor*pair\_freq_{ij}/(n_i*n_j)$$

Trong đó:

size_factor: kích thước tập chỉ mục

$pair\_freq_{ij}$: tần số xuất hiện từ

n_i, n_j : tần số xuất hiện tiếng i, j

Hai tiếng có khả năng tạo thành một từ cao khi chúng thường xuất hiện chung với nhau, nghĩa là cohension của chúng cao.

Phương pháp này không tách từ chính xác hoàn toàn nhưng có thể chấp nhận trong hệ thống tìm kiếm thông tin vì trong quá trình lập chỉ mục **chỉ cần xác định đúng các từ có trọng lượng cao**, trong trường hợp việc tách từ là sai thì từ sai chỉ được lập chỉ mục khi nó có trọng lượng cao, **việc lập chỉ mục một từ sai sẽ làm tăng chi phí lưu trữ nhưng không ảnh hưởng lớn tính chính xác kết quả tìm kiếm** vì dù sao từ này cũng có trọng lượng lớn.

Còn trong trường hợp một từ ghép được tách thành nhiều từ đơn ví dụ từ “thông tin” khi được lập chỉ mục sẽ luôn có 3 từ “thông”, “tin”, “thông tin”. Điều này gây ảnh hưởng đến tính chính xác của việc lập chỉ mục vì thực sự các từ “thông”, “tin” không cần thiết lập chỉ mục. Ta giải quyết vấn đề này bằng cách nếu từ “thông tin” được lập chỉ mục thì khi đó số lần xuất hiện của các từ “thông” và “tin” sẽ được tính toán lại bằng cách **trừ đi các trường hợp đã xuất hiện trong từ “thông tin”** để tính toán trọng lượng cho các từ đơn. Nếu từ đơn “tin” chỉ luôn xuất hiện trong từ “thông tin” thì số lần xuất hiện

của từ "tin" và "thông tin" là bằng nhau nên khi lập chỉ mục cho từ "thông tin" thì số lần xuất hiện riêng của từ đơn "tin" sẽ bằng 0 nên không được lập chỉ mục.

3.6. Lập chỉ mục tự động cho tài liệu

Vấn đề chính của lập chỉ mục tự động là xác định tự động mục từ chỉ mục cho các tài liệu. Trong các ngôn ngữ gốc Ấn – Âu thì tách từ có thể nói là đơn giản vì khoảng trắng là ký tự để phân biệt từ. Vấn đề cần quan tâm là xác định những từ này là từ khoá, có thể đại diện cho toàn bộ nội dung của tài liệu. Loại bỏ các từ stop-word có tần số xuất hiện cao, những từ này thường chiếm đến 40-50% trong số các từ của một văn bản. Những từ này có độ phân biệt kém và không thể sử dụng để xác định nội dung của tài liệu. Trong tiếng Anh, có khoảng 250 từ. Số lượng từ này không nhiều lắm nên giải pháp đơn giản nhất là lưu các từ này vào trong một tự điển, và sau đó chỉ cần thực hiện so sánh từ cần phân tích với từ điển để loại bỏ.

Bước tiếp theo là nhận ra các chỉ mục tốt. Để giảm bớt dung lượng lưu trữ, các mục từ cần được biến đổi về nguyên gốc (step of stemming đối với tiếng Anh), phải loại bỏ đi các tiền tố, hậu tố, các biến thể số nhiều, quá khứ…Giải pháp là sử dụng một danh sách các hâu tố. Trong khi loại bỏ hậu tố thì những **hậu tố dài** được ưu tiên loại bỏ trước, rồi sau đó mới loại bỏ những **hậu tố ngắn** hơn. Sau đây là một số vấn đề khi loại bỏ trong tiếng Anh:

1. Chỉ rõ **chiều dài tối thiểu của một từ gốc** sau khi loại bỏ hậu tố. Ví dụ: việc loại bỏ hậu tố "ability" ra khỏi "computability" hay loại bỏ "ing" ra khỏi "singing" là hợp lý. Tuy nhiên, những hậu tố đó không cần phải loại bỏ trong các từ "ability" và "sing".
2. Nếu nhiều hậu tố được kết hợp vào một gốc thì ta sẽ áp dụng **đệ quy** cho quá trình loại bỏ hậu tố vài lần hoặc lập từ điển hậu tố rồi loại bỏ

những hậu tố dài hơn trước rồi đến các hậu tố ngắn sau. Ví dụ: “effectiveness” → “effective” → “effect”.

3. Trong tiếng Anh, **từ gốc có thể bị biến đổi** sau khi đã loại bỏ hậu tố. Do đó, ta cần phải có những luật nhất định để phục hồi từ gốc. Chẳng hạn loại bỏ một trong hai kí tự trùng nhau của những từ có sự xuất hiện b, d, d, l, m, n, p, r, s, t ở cuối của các từ gốc sau khi đã loại bỏ hậu tố. Ví du như “beginning” → “beginn” → “begin”.
4. Một số **ngoại lệ** phụ thuộc vào ngữ cảnh đặc biệt phải được chú ý, sử dụng các quy tắc cảm ngữ cảnh. Ví dụ: một quy tắc cho hậu tố “allic” chỉ rõ chiều dài cực tiểu của từ gốc là ba và không loại bỏ hậu tố sau “met” hoặc “ryst”, hoặc quy tắc chỉ loại bỏ hậu tố “yl” sau “n” hoặc “r”.

Tóm lại, giải quyết vấn đề hậu tố không quá khó nếu chúng ta có sẵn một danh sách chứa các hậu tố, một danh sách chứa các luật thêm các hậu tố và phục hồi từ gốc sau khi thêm hậu tố.

3.7. Tập tin nghịch đảo tài liệu

3.7.1. Tập tin nghịch đảo

Giả sử câu truy vấn của người sử dụng sau khi lập chỉ mục là một tập các mục từ { t1, t2, .., tn}. Ví dụ: truy vấn “công nghệ phần mềm” sẽ được lập chỉ mục gồm hai từ “công nghệ” và “phần mềm” với giá trị n thường không lớn (2, 3, 4..).

Yêu cầu của người sử dụng là mong muốn tìm kiếm các tài liệu có chứa tất cả các mục từ t1, t2,..., tn. Như thế ta không cần khảo sát tất cả các vector chỉ mục mà chỉ cần tìm các vector nào có chứa t1, t2, ..., tn. Điều này có thể thực hiện dễ dàng bằng cách lưu các nhóm vector (tài liệu) theo từng mục từ.

t1 : 1, 3, 4

t2 : 1, 2, 4, 5

t3 : 2, 4, 5

Nghĩa là:

Mục từ t1 có trong các tài liệu 1, 3, 4.

Mục từ t2 có trong các tài liệu 1,2,4,5

Mục từ t3 có trong các tài liệu 2, 4, 5

Khi đó quá trình tìm kiếm (t1, t3) sẽ được thực hiện theo các bước sau:

Tìm tập các tài liệu có chứa t1, gọi là T1={1,3,4}

Tìm tập tài liệu có chứa t3, gọi là T2={2,4,5}

Tập các tài liệu có chứa cả t1 và t3 là T=T1∩ T2={4}

Tính toán độ tương tự giữa câu truy vấn và các tài liệu có trong tập T

Sử dụng công thức tính độ tương tự :

$Sim(D, Q) = v_i * w_i$, i=1..n

với t_i là mục từ có trong Q

(do w_i=0 với mục từ t_i không có trong Q và w_i =1 nếu t_i có trong Q)

Rõ ràng việc tính độ tương tự chỉ cần tới trọng lượng của các mục từ có trong Q nên để có thể tăng thêm hiệu quả ta sẽ **lưu thêm giá trị trọng lượng** của mục từ trong tập tin nghịch đảo.

t1 : (1, 0.5) (3, 0.7) (4, 0.2)

t2 : (1, 0.4) (2, 0.8) (4, 0.9) (5, 0.1)

t3 : (2, 0.3) (4, 0.2) (5, 0.5)

Nghĩa là mục từ t1 có trong tài liệu 1 với trọng lượng là 0.5, trong tài liệu 3 với trọng lượng là 0.7 v...v...

Khi đó để tìm kiếm cho câu truy vấn (t1, t3) chỉ cần đọc 2 khối dữ liệu của t1 và t3 là đủ (giảm truy xuất đĩa và giảm thời gian xử lý).

Mô hình tập tin nghịch đảo hiện nay được sử dụng rất rộng rãi trong các hệ thống tìm kiếm thông tin vì với cách tổ chức này vì các dữ liệu cần đọc được lưu trữ liên tục nên giảm việc di chuyển đầu đọc của đĩa cứng, cũng như nếu ta lưu lại vị trí bắt đầu của các mục từ thì có thể truy xuất trực tiếp đến vị trí đó để đọc dữ liệu.

Khó khăn: của việc sử dụng tập tin nghịch đảo là khi cần thêm một tài liệu vào mục từ, giả sử cần thêm tài liệu 6 vào mục từ t1.

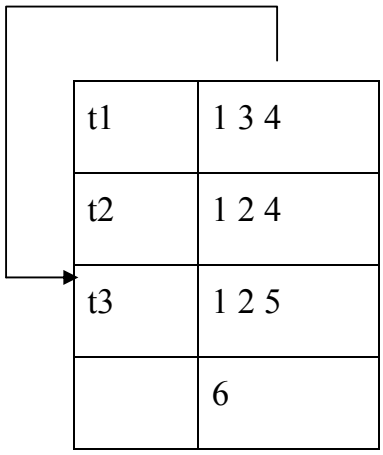
t1 : 1,3,4,6

t2 : 1,2,4,5

t3 : 2,4,5

Với chú ý rằng các khối dữ liệu của t1, t2, t3 được lưu trữ liên tiếp nhau trên đĩa cứng và dung lượng của tập tin nghịch đảo này rất lớn (chứa hàng trăm ngàn mục từ với hàng triệu tài liệu), hơn nữa việc thêm tài liệu này rất thường xuyên (lập chỉ mục cho các Web site mới, cập nhật lại các Web site có thay đổi) cho nên không thể sử dụng phương pháp chèn bằng cách dời dữ liệu ra sau để tạo khoảng trống chèn tài liệu 6 vào.

Cách giải quyết: cấp phát không gian cho các mục từ **theo trang**, khi một mục từ đã chứa hết trang này thì sẽ cấp phát thêm vào cuối tập tin và có một link chỉ đến trang cuối này.



Phương pháp này mặc dù lãng phí không gian cho các trang chưa dùng đến, giả sử có 100.000 mục từ, trang dung lượng là 1K, **dung lượng đĩa lãng phí** lớn nhất là 100.000 K (100 M) **và phải di chuyển đầu đọc** nhiều nhưng **giải quyết được vấn đề thêm tài liệu** cũng như dễ dàng đọc được dữ liệu cần thiết cho một mục từ nào đó (đọc theo các link). Có thể điều chỉnh giữa dung lượng lãng phí và việc phải di chuyển đầu đọc (tính bằng số trang cấp phát cho một mục từ) bằng cách tăng hoặc giảm dung lượng cấp phát cho một trang. Nếu tăng dung lượng cấp phát cho một trang thì sẽ giảm việc di chuyển đầu đọc và ngược lại.

3.7.2. Phân biệt giữa tập tin nghịch đảo và tập tin trực tiếp

Tập tin trực tiếp (Direct File) là tập tin mà chính các mục thông tin đã cung cấp thứ tự chính của tập tin.

Ngược lại, tập tin nghịch đảo (Inverted File) được sắp xếp theo chủ đề, mỗi chủ đề lại bao gồm một tập các mục thông tin.

Giả sử có một tập các tài liệu, mỗi tài liệu chứa danh sách các từ. Nếu một từ xuất hiện trong một tài liệu, ghi số 1. Ngược lại, ghi 0. Khi đó, tập tin trực tiếp và tập tin nghịch đảo sẽ lưu trữ như sau:

Bảng 3.1: Cách tập tin nghịch đảo lưu trữ

	Tài liệu 1	Tài liệu 2	Tài liệu 3
Từ 1	1	0	1
Từ 2	1	1	0
Từ 3	0	1	1
Từ 4	1	1	1

Bảng 3.2: Cách tập tin trực tiếp lưu trữ

	Từ 1	Từ 2	Từ 3	Từ 4
Tài liệu 1	1	1	0	1
Tài liệu 2	0	1	1	1
Tài liệu 3	1	0	1	1

3.7.3. Tại sao sử dụng tập tin nghịch đảo để lập chỉ mục

Trong hệ thống truy xuất thông tin, tập tin nghịch đảo có ý nghĩa rất lớn, giúp việc truy cập đến các mục thông tin được nhanh chóng. Giả sử khi người dùng nhập một câu truy vấn, hệ thống sẽ tách thành 2 từ là "từ 1" và "từ 2". Dựa vào tập tin nghịch đảo, ta dễ dàng xác định được các tài liệu có liên quan đến hai từ này để trả về cho người tìm kiếm. Tuy nhiên, khó khăn chính của tập tin nghịch đảo là khi thêm một tài liệu mới, tất cả các từ có liên quan đến tài liệu này đều phải được cập nhật lại. Ví dụ khi thêm tài liệu 4 có chứa 2 từ "từ 3" và "từ 4" vào tập tin nghịch đảo:

Bảng 3.3 Thêm một tài liệu mới vào tập tin nghịch đảo

	Tài liệu 1	Tài liệu 2	Tài liệu 3	Tài liệu 4
Từ 1	1	0	1	0
Từ 2	1	1	0	0
Từ 3	0	1	1	1
Từ 4	1	1	1	1

Rõ ràng việc này tốn một chi phí lớn nếu tập tin nghịch đảo rất lớn. Trong thực tế, tập tin nghịch đảo tài liệu có thể chứa hàng trăm ngàn từ. Tuy nhiên, trong các hệ thống truy xuất thông tin, người ta chỉ cập nhật lại tập tin tại một khoảng thời gian nhất định kỳ. Vì vậy, tập tin nghịch đảo vẫn được sử dụng để lập chỉ mục.

Chương 4: TRUY XUẤT THÔNG TIN ĐA PHƯƠNG TIỆN

4.1. Truy xuất thông tin đa phương tiện

Truy xuất thông tin truyền thống tập trung vào vào tìm kiếm thông tin dạng văn bản (Text Retrieval) hay tài liệu văn bản (Document Retrieval). Trong một thời gian dài, truy xuất thông tin gần như đồng nghĩa với tìm kiếm tài liệu hay tìm kiếm văn bản. Trong thời gian gần đây, các viễn cảnh ứng dụng mới như ứng dụng trả lời câu hỏi (question answering), ứng dụng nhận dạng chủ đề (Topic detection), hay ứng dụng lưu vết (tracking) trở thành các lĩnh vực hoạt động mạnh mẽ trong nghiên cứu truy xuất thông tin.

Một lĩnh vực phát triển khác mà các kỹ thuật truy xuất thông tin đang kế tục và phát huy, đó là truy xuất thông tin không văn bản hay còn gọi là truy xuất thông tin đa phương tiện. Loại hình tìm kiếm này sẽ dựa trên rút trích tự động các phần văn bản hay lời nói của các tài liệu đa phương tiện, sau đó được xử lý bởi các kỹ thuật truy xuất thông tin dựa văn bản (text-based IR Techniques). Tuy nhiên, người ta ngày càng quan tâm đến sự phát triển các kỹ thuật phơi bày cụ thể thông tin đa phương tiện truyền thông rồi tích hợp chúng với các phương pháp tìm kiếm đã được thiết lập.

Định nghĩa: Truy xuất thông tin đa phương tiện là quá trình làm thỏa mãn các thông tin mà người dùng yêu cầu bởi việc chỉ ra tất cả các văn bản, đồ họa, audio (lời nói liên tục, các hình ảnh hoặc các tài liệu video có liên quan) hoặc vị trí của các tài liệu từ một kho tài liệu.

4.2. Truy xuất audio ngôn ngữ nói

Một người dùng có thể muốn để tìm kiếm trong một kho dữ liệu văn bản lớn, khả năng để tìm kiếm nội dung của các nguồn audio chẳng hạn như lời nói, radio quảng bá và các đoạn hội thoại có thể đánh giá cho một phạm vi các ứng dụng. Một sự phân loại các kỹ thuật được phát triển hỗ trợ cho việc nhận dạng tự động lời nói. Có nhiều ứng dụng trong một phạm vi các lĩnh vực ứng dụng chẳng hạn như xác minh người nói, transcription, điều khiển bằng lời nói,…

4.3. Truy xuất audio

Thêm vào truy cập dựa nội dung tới âm thanh lời nói, truy xuất nhiễu/tiếng động cũng quan trọng trong các lĩnh vực sản xuất âm nhạc và phim/video/. Một hệ thống đã mô tả một sự phân loại tiếng động user-extensible và hệ thống truy xuất, được gọi là Sound Fisher (www.musclefish.com), nó được đưa ra từ một số môn học bao gồm xử lý tín hiệu, Psychoacoustics, nhận dạng tiếng nói, âm nhạc máy tính và các cơ sở dữ liệu đa phương tiện. Các thuật toán đánh chỉ mục hình ảnh sử dụng các vector đặc trưng để tạo chỉ mục và đối sánh các ảnh, tác giả đã sử dụng một vector đo được trực tiếp các đặc trưng âm học (như khoảng thời gian, loudness, pitch, độ sáng-brightness) để lập chỉ mục các âm thanh. Điều này làm cho người sử dụng có thể tìm kiếm các âm thanh trong các phạm vi đặc trưng được chỉ rõ.

4.4. Truy xuất đồ hoạ

Lớp phương tiện quan trọng khác là đồ hoạ, bao gồm các bảng và các đồ thị (ví dụ đồ thị cột, thanh, line, hình tròn, scatter,…). Đồ thị được tạo

thành từ các thành phần dữ liệu chẳng hạn như các điểm, dòng, nhãn. Một ví dụ về một hệ thống truy xuất đồ hoạ là Sagebook được đưa ra bởi trường đại học Carnegie Mellon. Sagebook, có thể bao gồm cả tìm kiếm theo yêu cầu từ các dữ liệu đồ hoạ được lưu trữ. Ta có thể yêu cầu một truy vấn audio trong truy xuất audio. Sagebook hỗ trợ các truy vấn dữ liệu đồ hoạ, việc biểu diễn (ví dụ mô tả nội dung), đánh chỉ số, tìm kiếm và các khả năng thích ứng.

Thêm vào đó, các dữ liệu đồ hoạ được truy xuất có thể được sửa lại cho thích hợp bằng tay. Sagebook chứa một sự biểu diễn bên trong về ngữ nghĩa và cú pháp của các dữ liệu đồ hoạ, bao gồm các quan hệ không gian giữa các đối tượng, mối quan hệ giữa các miền dữ liệu (ví dụ interval, tọa độ 2 chiều), các đồ thị biến thiên và các thuộc tính dữ liệu. Tìm kiếm được thực hiện trong cả các đồ thị và các thuộc tính của dữ liệu, với 3 và 4 chiến lược tìm kiếm luân phiên, theo thứ tự định sẵn để có thể biến đổi mức độ của sự đối sánh. Khi các bộ sưu tập hình ảnh văn bản lớn, có một số các kỹ thuật nhóm các dữ liệu đồ hoạ dựa vào các thuộc tính dữ liệu và đồ thị được thiết kế để có thể phân cụm cho việc trình duyệt các bộ sưu tập.

Sagebook cũng cung cấp các kỹ thuật thích ứng tự động mà có thể sửa đổi các đồ thị được truy xuất (ví dụ việc loại bỏ các thành phần đồ thị) mà không phù hợp với truy vấn đã chỉ ra.

Khả năng truy xuất các đồ thị bởi nội dung có thể đưa ra các khả năng mới trong một phạm vi các miền dựa vào các đồ thị thương mại. Chẳng hạn, các đồ thị hiển thị một quy tắc chiếm ưu thế hơn (predominant) trong các miền chẳng hạn như nghiên cứu bản đồ (địa hình, các đặc trưng), kiến trúc (bản thiết kế nhà), truyền thông và mạng (các router và các liên kết), các hệ thống máy móc (các thành phần và các kết nối) và các kế hoạch vận động cho lực lượng vũ trang (ví dụ: ảnh hưởng và sự phòng thủ che phủ trên các bản đồ). Trong mỗi trường hợp của các trường hợp đó các thành phần của đồ thị,

các thuộc tính của chúng, các quan hệ và cấu trúc có thể được phân tích cho mục đích truy xuất dữ liệu.

4.5. Truy xuất ảnh

Các cuốn sách tăng nhiều hình ảnh - từ hình ảnh trong các trang web tới các bộ sưu tập cá nhân từ các máy ảnh số - được leo thang các yêu cầu truy nhập hình ảnh hiệu quả và hiệu suất cao hơn. Các nhà nghiên cứu đã chỉ rõ các yêu cầu cho việc lập chỉ mục và tìm kiếm không chỉ metadata kết hợp với các hình ảnh (ví dụ: các tên, các chú giải) mà còn truy xuất trực tiếp trong cả nội dung của các hình ảnh. Sự phát triển của các thuật toán đang tập trung vào việc lập chỉ mục tự động cho các đặc trưng visual của hình ảnh (ví dụ: màu, vân, hình dáng) có thể được sử dụng như các nghĩa cho việc truy xuất các hình ảnh trong chủ đề lập chỉ mục thủ công. Tuy nhiên, mục tiêu cuối cùng là dựa vào ngữ nghĩa truy nhập vào hình ảnh.

Lấy thông tin từ dữ liệu ảnh có liên quan đến rất nhiều các lĩnh vực khác, từ những phòng trưng bày tranh nghệ thuật cho tới những nơi lưu trữ tranh nghệ thuật lớn như viện bảo tàng, kho lưu trữ ảnh chụp, kho lưu trữ ảnh tội phạm, cơ sở dữ liệu ảnh về địa lý, y học, ... điều đó làm cho lĩnh vực nghiên cứu này phát triển nhanh nhất trong công nghệ thông tin.

Lấy thông tin từ dữ liệu ảnh đặt ra nhiều thách thức nghiên cứu mới cho các khoa học gia và các kỹ sư. Phân tích ảnh, xử lý ảnh, nhận dạng mẫu, giao tiếp giữa người và máy là những lĩnh vực nghiên cứu quan trọng góp phần vào phạm vi nghiên cứu mới này.

Khía cạnh tiêu biểu của lấy thông tin từ dữ liệu ảnh dựa trên những công bố có sẵn như là những đối tượng nhận thức như màu sắc, vân (texture), hình dáng, cấu trúc, quan hệ không gian, hay thuộc về ngữ nghĩa căn bản như: đối tượng, vai trò hay sự kiện hay liên quan đến thông tin về

ngữ nghĩa quan hệ như cảm giác, cảm xúc, nghĩa của ảnh. Thật ra phân tích ảnh, nhận dạng mẫu, hay xử lý ảnh đóng một vai trò căn bản trong hệ thống lấy thông tin từ ảnh. Chúng cho phép sự trích rút tự động hầu hết những thông tin về nhận thức, thông qua phân tích sự phân bổ điểm ảnh và sự phân tích độ đo.

4.5.1. Truy xuất ảnh dựa vào màu sắc

Màu sắc là vấn đề cần tập chung giải quyết nhiều nhất, vì một ảnh màu thì thông tin quan trọng nhất trong ảnh chính là màu sắc. Hơn nữa thông tin về màu sắc là thông tin người dùng quan tâm nhất; qua đặc trưng màu sắc, có thể lọc được rất nhiều lớp ảnh, thông qua vị trí, không gian, định lượng của màu trong ảnh.

Phương pháp phổ biến để tìm kiếm ảnh trong một tập những ảnh hỗn tạp cho trước là dựa vào lượt đồ màu của chúng. Đây là cách làm khá đơn giản, tốc độ tìm kiếm tương đối nhanh nhưng khuyết điểm là kết quả tìm kiếm lại có độ chính xác không cao. Nhưng đây có thể được xem như là bước lọc đầu tiên cho những tìm kiếm sau. Muốn được kết quả chính xác cao đòi hỏi sự kết hợp đồng thời với vân (texture) và hình dáng (shape). Cho đến nay, để giải quyết vấn đề về màu sắc, cách tiếp cận chính vẫn là dựa vào lượt đồ màu.

4.5.2. Truy xuất ảnh dựa vào vân

Vân (texture), đến này vẫn chưa có một định nghĩa chính xác cụ thể về vân, là một đối tượng dùng để phân hoạch ảnh ra thành những vùng được quan tâm và để phân lớp những vùng đó. Vân cung cấp thông tin về sự sắp xếp về mặt không gian của màu sắc và cường độ của một ảnh. Vân được đặc trưng bởi sự phân bổ không gian của những mức cường độ trong một khu vực

láng giềng với nhau. Vân của ảnh màu và vân đối với ảnh xám là như nhau. Vân gồm nhiều vân gốc hay vân phần tử gộp lại, đôi khi được gọi là texel.

Có những lớp ảnh mà màu sắc không thể giải quyết được, đòi hỏi phải dùng đặc trưng vân. Ví dụ như những ảnh liên quan đến cấu trúc của điểm ảnh như: cỏ, mây, đá, sợi. Vân sẽ giải quyết tốt cho việc tìm kiếm đối với lớp ảnh này.

Trong hầu hết các trường hợp, phân đoạn những ảnh thật ra những texel khó hơn nhiều đối với trường hợp tự nhiên sinh ra những hoa văn thiên nhiên.

Thay vì vậy, việc định lượng về số hay thông tin thống kê bằng số mô tả cho một vân có thể được tính từ chính mức xác, hay mức màu của chúng. Tuy cách tiếp cận này ít trực quan nhưng nó có hiệu suất tính toán cao, hơn nữa cách tiếp cận này cũng phù hợp với đồng thời cho việc phân đoạn vân và phân loại vân.

4.5.3. Truy xuất ảnh dựa vào hình dạng

Màu sắc và vân là những thuộc tính có khái niệm toàn cục của một bức ảnh. Trong khi đó, hình dạng không phải là một thuộc tính của ảnh. Thay vì vậy, hình dạng có khuynh hướng chỉ định tới một khu vực đặc biệt của ảnh. Hay hình dạng chỉ là biên của đối tượng nào đó trong ảnh

Đối với những lớp ảnh cần tìm mà liên quan đến hình dạng của đối tượng thì đặc trưng vân và màu không thể giải quyết được. Ví dụ như tìm một vật có hình dạng ellipse hay hình tròn trong ảnh.

Tìm kiếm theo hình dáng thật sự là một cái đích của hệ thống tìm kiếm dựa vào nội dung muốn đạt tới.

Hình dạng là một cấp cao hơn màu sắc và vân. Nó đòi hỏi sự phân biệt giữa các vùng để tiến hành xử lý về độ đo của hình dạng. Trong nhiều

trường hợp, sự phân biệt này cần thiết phải làm bằng tay. Nhưng sự tự động hóa trong một số trường hợp có thể khả thi. Trong đó, vấn đề chính yếu nhất là quá trình phân đoạn ảnh. Nếu quá trình phân đoạn ảnh được làm một cách chính xác, rõ ràng và nhất là hiệu quả thì sự tìm kiếm thông tin dựa vào hình dạng có thể có hiệu lực rất lớn.

Nhận dạng ảnh hai chiều là một khía cạnh quan trọng của quá trình phân tích ảnh. Tính chất hình dạng toàn cục ám chỉ đến hình dạng ảnh ở mức toàn cục. Hai hình dạng có thể được so sánh với nhau theo tính chất toàn cục bởi những phương pháp nhận dạng theo hoa văn, mẫu vẽ. Sự so khớp hình dạng ảnh cũng có thể dùng những kỹ thuật về cấu trúc, trong đó một ảnh được mô tả bởi những thành phần chính của nó và quan hệ không gian của chúng. Vì sự hiển thị ảnh là một quá trình liên quan đến đồ thị, do đó những phương pháp so khớp về đồ thị có thể được dùng cho việc so sánh hay so khớp. Sự so khớp về đồ thị rất chính xác, vì nó dựa trên những quan hệ không gian hầu như bất biến trong toàn thể các phép biến đổi hai chiều. Tuy nhiên, quá trình so khớp về đồ thị diễn ra rất chậm, thời gian tính toán tăng theo cấp số mũ tương ứng với số lượng các phần tử. Trong việc tìm kiếm dữ liệu ảnh dựa vào nội dung, ta cần những phương pháp có thể quyết định sự giống và khác nhau một cách nhanh chóng. Thông thường, chúng ta luôn đòi hỏi sự bất biến cả đối với kích thước của ảnh cũng như hướng của ảnh trong không gian. Vì vậy, một đối tượng có thể được xác định trong một số hướng. Tuy nhiên, tính chất này không thường được yêu cầu trong tìm kiếm ảnh. Trong rất nhiều cảnh vật, hướng của đối tượng thường là không đổi. Ví dụ như: cây cối, nhà cửa, ...

Độ đo về hình dạng rất nhiều trong phạm vi lý thuyết của bộ môn xử lý ảnh. Chúng trải rộng từ những độ đo toàn cục dạng thô với sự trợ giúp của việc nhận dạng đối tượng, cho tới những độ đo chi tiết tự động tìm kiếm

những hình dạng đặc biệt. Lượt đồ hình dạng là một ví dụ của độ đo đơn giản, nó chỉ có thể loại trừ những đối tượng hình dạng không thể so khớp, nhưng điều đó sẽ mang lại khẳng định sai, vì chỉ như là việc làm của lượt đồ màu. Kỹ thuật dùng đường biên thì đặc hiệu hơn phương pháp trước, chúng làm việc với sự hiện hữu của đường biên của hình dạng đối tượng và đồng thời cũng tìm kiếm những hình dạng đối tượng gần giống với đường biên nhất. Phương pháp vẽ phác họa có thề là phương pháp có nhiều đặc trưng rõ ràng hơn, không chỉ tìm kiếm những đường biên đối tượng đơn, mà còn đối với tập những đối tượng đã được phân đoạn trong một ảnh mà người dùng vẽ hay cung cấp.

Chương 5: ĐÁNH GIÁ CÁC HỆ THỐNG TRUY XUẤT THÔNG TIN

5.1. Lý do để tiến hành đánh giá các hệ thống truy xuất thông tin

Khi nhu cầu truy xuất thông tin phát triển, có rất nhiều mô hình, thuật toán, hệ thống truy xuất thông tin ra đời. Do đó, việc đánh giá các mô hình, thuật toán, hệ thống truy xuất thông tin là điều bắt buộc phải làm.

Chúng ta so sánh một hệ thống (có thể là một hệ thống mới) với các hệ thống khác đã tồn tại về phương diện: tính hiệu quả, chi phí, thời gian, tốc độ xử lý…Hệ thống truy xuất thông tin thường thực hiện hai quá trình: quá trình lập chỉ mục và quá trình tìm kiếm. Mỗi một quá trình sẽ có nhiều phương pháp để thực hiện, đánh giá hệ thống cũng có thể dùng để xác định tính tối ưu của các phương pháp trên.

Lý do khác để tiến hành đánh giá là để so sánh các thành phần của hệ thống. Do hệ thống gồm nhiều thành phần, đánh giá hệ thống để xác định cách mỗi thành phần của hệ thống thực thi để khi có sự thay đổi một thành phần bởi một thành phần khác thì sự thay đổi đó ảnh hưởng đến hệ thống như thế nào, từ đó ta có thể quyết định có nên thay đổi thành phần đó không.

Đánh giá để tìm kiếm thành phần nào là tốt nhất cho hàm xếp thứ tự (dot-product, cosine…); thành phần nào là tốt nhất cho lựa chọn thuật ngữ (loại bỏ stopword, phương pháp lấy gốc từ stemming…); thành phần nào là tốt nhất trong lựa chọn phương pháp đánh giá thuật ngữ (term weighting) như TF, IDF...

So sánh để biết người sử dụng cần danh sách các tài liệu trả về (ranked list) dài cỡ bao nhiêu để họ có thể nhìn dễ dàng nhất. Đánh giá để biết hệ thống nào thật sự tốt, người dùng có thể tin tưởng kết quả trả về được.

5.2. Các tiêu chuẩn được dùng để đánh giá

Hiện nay, trên thế giới có ba tiêu chuẩn được dùng để đánh giá hệ thống truy xuất thông tin:

- Tiêu chuẩn về tính **hiệu quả** tức sự chính xác, tính đầy đủ của kết quả trả về so với mục đích tìm kiếm của người sử dụng, và giá trị vẫn có thể đoán được trong các tình huống khác có nghĩa là khi đưa vào các câu truy vấn khác, tập tài liệu khác thì hệ thống vẫn có thể tìm ra kết quả chính xác.
- Tiêu chuẩn về **hiệu năng**, gồm có tốc độ tìm kiếm của thuật toán, khả năng lưu trữ, thời gian trả về cho người sử dụng, thời gian lập chỉ mục, kích thước chỉ mục…
- Tiêu chuẩn về **khả năng sử dụng hệ thống** tức là có thể nghiên cứu, học hỏi trên hệ thống tìm kiếm, người không biết tin học hay các chuyên gia tin học đều có thể sử dụng hệ thống.

5.3. Các mô hình đánh giá

Có tất cả bốn mô hình đánh giá các hệ thống truy xuất thông tin. Chúng bao gồm: đánh giá hộp kính, đánh giá hộp đen, đánh giá hướng hệ thống, đánh giá hướng người dùng hay còn gọi là đánh giá nghiên cứu người dùng.

- **Đánh giá hộp trắng** (Glass Box Evaluation) : đánh giá hệ thống dựa trên việc đánh giá tất cả mọi thành phần của hệ thống. Có nghĩa là khi biết rõ các thành phần của hệ thống, chúng ta tiến hành đánh giá các thành phần đó.

- **Đánh giá hộp đen** (Black Box Evaluation) : đánh giá hệ thống bằng cách xem hệ thống như là một thực thể hợp nhất, không đánh giá chính xác các thành phần bên trong hệ thống.
- **Đánh giá hướng hệ thống** (System-Oriented Evaluation) là xu hướng đánh giá chính từ khi các hệ thống tìm kiếm và lập chỉ mục tự động được phát triển vào những năm 1960. Một trong những mục đích chính của hướng đánh giá này là kiểm tra các hệ thống tự động cũng như các thủ tục thủ công thực thi như thế nào. Ngoài ra, mô hình này còn đánh giá so sánh các cách thực hiện liên quan đến các ngôn ngữ chỉ mục, xử lý tìm kiếm của hệ thống của các hệ thống khác nhau hay đánh giá so sánh các lược đồ chỉ mục tự động khác nhau. Đánh giá hướng hệ thống có một điểm lợi là điều kiện môi trường kiểm tra được quản lý chặt chẽ, sử dụng phương pháp đánh giá theo lô hay còn gọi là đánh giá dựa trên tập câu truy vấn; có nghĩa là hệ thống truy xuất thông tin lần lượt thực hiện các câu truy vấn, tìm kiếm trên tập dữ liệu đã được xây dựng và ghi lại kết quả những tài liệu nào liên quan đến câu truy vấn nào rồi đem so sánh với Bảng đánh giá liên quan chuẩn (Relevance judgment) đã được xây dựng. Với mỗi câu truy vấn tính toán độ chính xác và độ bao phủ dựa trên kết quả trả về và bảng đánh giá liên quan chuẩn để nhận xét hiệu quả tìm kiếm của hệ thống truy xuất thông tin. Hướng đánh giá này được thực hiện rất phổ biến ở các dự án, hội nghị về nghiên cứu hệ thống truy xuất thông tin như: Cranfield, MEDLARS, SMART, STAIRS và TREC.
- **Đánh giá hướng người dùng** (User Studies Evaluation): Hướng nghiên cứu người dùng ra đời vào những năm 1970 khi mà nhiều hệ thống truy xuất thông tin thương mại ra đời. Mục đích chính của hướng nghiên cứu này là nhằm xác định cách thức tìm kiếm của người sử

dụng. Hướng đánh giá này còn cho phép xem xét hệ thống ở khía cạnh người dùng; tức là đánh giá về mặt tương tác với người sử dụng như giao diện của hệ thống truy xuất thông tin, thời gian hệ thống tìm kiếm đối với một câu truy vấn, mức độ hài lòng của người sử dụng… Hướng nghiên cứu này cho rằng nhu cầu của người dùng được thoả mãn tương đương với hiệu quả của hệ thống. Chỉ khi nhu cầu thông tin người dùng được thỏa mãn, khi ấy truy xuất thông tin mới được gọi là có ích. Hội nghị quốc tế về truy xuất thông tin trong Ngữ cảnh (Information Seeking in Context) được tổ chức như là một diễn đàn cho các nhà nghiên cứu lĩnh vực này khám phá các phương pháp và các kết quả nghiên cứu. Một hội nghị khác mới được thành lập tên là Nhóm Quan tâm Đặc biệt (Special Interest Group - SIG) đến tìm kiếm, nhu cầu và sử dụng thông tin của Xã hội Hoa Kỳ về Khoa học Thông tin (American Society of Information Science). Những hội nghị này cũng tương tự như TREC trong việc cố gắng khuyến khích nghiên cứu hướng người dùng, để phát triển mối liên hệ giữa các nhà nghiên cứu trong kỹ thuật, giáo dục và chính phủ, và để xác định, cải tiến các kỹ thuật tìm kiếm thích hợp. Nhưng các hội nghị này khác nhau ở chỗ các hội nghị mới chưa có phương pháp luận đánh giá chuẩn nào được xúc tiến. Đánh giá hướng người dùng có đóng góp rất lớn đến lĩnh vực truy xuất thông tin. Đóng góp này gồm có việc xác định cách thức truy xuất thông tin của con người, nối liền khoảng cách giữa nhu cầu thông tin giữa các cá nhân và các hệ thống truy xuất thông tin, dẫn đến một thế hệ mới của các hệ thống truy xuất thông tin bao gồm các giao diện đồ hoạ máy tính-người sử dụng.

Hiện nay, trong số bốn mô hình trên thì hai mô hình đánh giá hướng hệ thống và hướng người dùng đang được sử dụng chính và rộng rãi nhất. Vì mô

hình đánh giá hướng người dùng cần có sự hợp tác của rất nhiều người dùng để lấy thông tin phản hồi sau khi sử dụng hệ thống truy xuất thông tin đó hoặc cần phải tham gia trao đổi về hiệu năng tìm kiếm tại các hội nghị. Nhưng các hội nghị dành cho mô hình đánh giá hướng người dùng đa số chưa có một phương pháp luận cụ thể nào dùng để đánh giá.

5.4. Các độ đo dùng để đánh giá

Độ bao phủ (Recall) và độ chính xác (Precision) là 2 đơn vị đo cơ bản nhất để đánh giá chất lượng một hệ thống truy xuất thông tin. Độ bao phủ là tỉ lệ giữa các tài liệu liên quan được trả về trên tổng số các tài liệu liên quan thật sự. Trong khi đó, độ chính xác là tỉ lệ giữa các tài liệu liên quan được trả về trên tổng số tài liệu được trả về.

Có nhiều phương pháp sử dụng một hoặc các độ đo này để tính toán đánh giá, chẳng hạn phương pháp Độ chính xác trung bình (Mean Average Precision–MAP) chỉ sử dụng độ chính xác, không quan tâm đến độ bao phủ. Phương pháp đo dựa trên giá trị đơn Swet's E-Measure hoặc chiều dài tìm kiếm trung bình thì cũng chỉ sử dụng một giá trị để tính toán. Phương pháp tính độ chính xác dựa trên 11 điểm chuẩn của độ bao phủ sử dụng cả hai độ đo độ bao phủ và độ chính xác.

5.4.1. Các khái niệm về độ đo và liên quan

- ***Tính liên quan của tài liệu (relevant):***
  Một tài liệu được gọi là có liên quan khi nội dung của tài liệu đó có đề cập đến vấn đề mà câu truy vấn của người dùng quan tâm.
- ***Độ bao phủ (Recall - R):***
  Cho biết khả năng hệ thống tìm kiếm được những tài liệu có liên quan.
- ***Độ chính xác (Precision - P):***

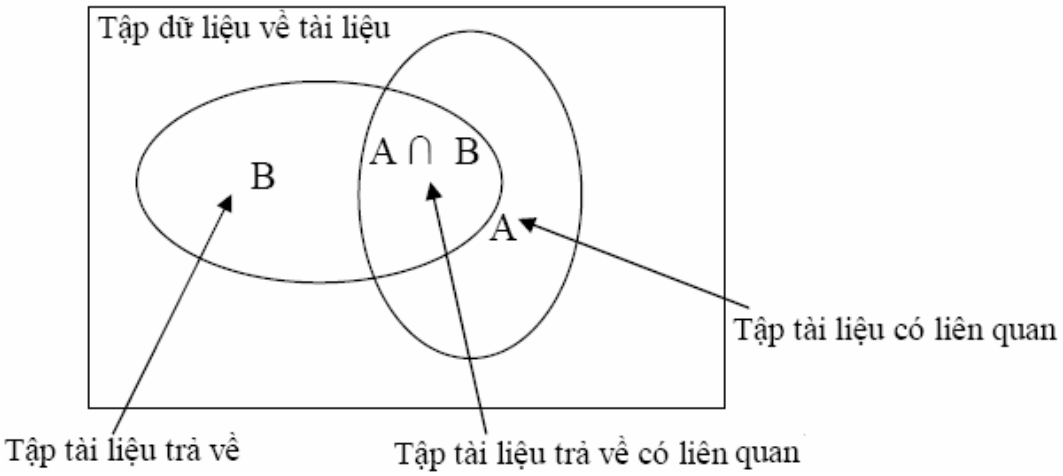
Cho biết khả năng của hệ thống tìm được những tài liệu chính xác

Có liên quan (Relevant) A	Không liên quan (non- relevant) $\overline{A}$	
A ∩ B	$\overline{A}$ ∩ B	B Tìm thấy (retrieved)
A ∩ $\overline{B}$	$\overline{A}$ ∩ $\overline{B}$	$\overline{B}$ Không tìm thấy (not retrieved)

- ***Khả năng loại bỏ: (Fall out - F):***

  Cho biết khả năng của hệ thống loại bỏ những tài liệu không liên quan.

5.4.2. Cách tính độ bao phủ (R) và độ chính xác (P)



Hình 5.1: Tập dữ liệu về tài liệu

Độ bao phủ (R):

$$R = \frac{|A \cap B|}{|B|}$$

Độ chính xác (P):

$$P = \frac{|A \cap B|}{|A|}$$

Khả năng loại bỏ: (Fall out - F):

$$F = \frac{\left|\overline{A} \cap B\right|}{\left|\overline{A}\right|}$$

Mối liên hệ giữa R, P, F:

$$F = \frac{R * G}{R * G + F\left(1 - G\right)}$$

G : là nhân tố tổng quát đo độ dày đặc của tài liệu liên quan trong tập dữ liệu ⇔G cho biết độ liên quan của tài liệu so với câu truy vấn là cao hay thấp:

$$G = \frac{\left|A\right|}{\left|S\right|}$$

Với S là tập tài liệu.

Vấn đề đo độ bao phủ:

Tính độ bao phủ là một vấn đề khó khăn trong việc đánh giá hệ thống tìm kiếm thông tin bởi vì nó liên quan đến việc định giá thủ công tổng số tài liệu liên quan trong tập tài liệu đối với mỗi câu truy vấn (vấn đề tạo bảng liên quan lý thuyết), việc định giá như vậy rất tốn kém nếu tập dữ liệu lớn. Để giải quyết vấn đề này người ta đưa ra phương pháp "pooling". Ý tưởng của phương pháp "pooling" là trong danh sách tài liệu trả về chỉ lấy n tài liệu đầu, n được gọi là chiều dài của "pool".

Việc tạo bảng liên quan lý thuyết áp dụng phương pháp "pooling" được tiến hành như sau: tiến hành tìm kiếm trên nhiều hệ thống áp dụng phương pháp

"pooling", có thể tài liệu liên quan được trả về của một hệ thống là cao, ta tiến hành giao các tập tài liệu liên quan trả về của các hệ thống đó và chỉ lấy n tài liệu đầu.

Bởi vì tập kết quả trả về được sắp xếp theo thứ tự nên độ chính xác và độ bao phủ có thể tính được tại các ngưỡng vị trí thứ tự thứ i tài liệu.

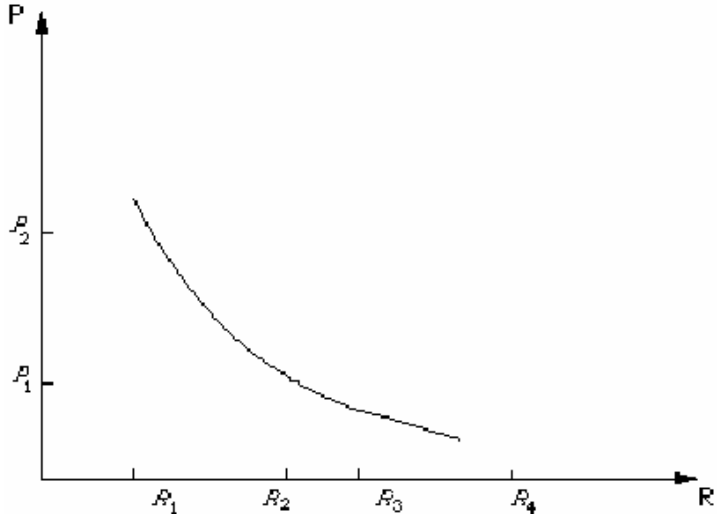
Vấn đề bảng liên quan thực tế:

Đối với cách tính trên ta phải quan niệm về độ liên quan của tài liệu trên 2 mức độ: hoặc là tài liệu có liên quan hoặc là tài liệu không liên quan. Cách quy ước như vậy nhằm làm đơn giản hoá cách đánh giá. Trên thực tế, độ liên quan của tài liệu không chỉ là 2 mức độ mà có thể có nhiều mức độ.

5.5. Phương pháp tính độ chính xác dựa trên 11 điểm chuẩn của độ bao phủ

5.5.1. Đồ thị biểu diễn hiệu suất thực thi hệ thống truy xuất

- Ứng với 1 câu truy vấn được thực hiện bởi hệ thống sẽ có 1 độ bao phủ (R_i), độ chính xác (P_i) cụ thể.
- Với 1 cặp (R_i, P_i) biểu diễn trên hệ trục toạ độ ROP tương ứng với 1 điểm.
- Biểu diễn kết quả của tập câu truy vấn trên ROP ta sẽ có 2 đường cong mô tả hiệu suất thực thi của hệ thống. Đường cong có dạng:



Hình 5.2: Đường cong mô tả hiệu suất thực thi của hệ thống

- Từ đồ thị ta có thể rút ra kết luận: độ bao phủ vả độ chính xác có mối quan hệ gần như tỷ lệ nghịch, khi R tăng thì P có thể sẽ giảm và ngược lại.
- Khi ta cố gắng làm tăng R bằng cách tăng số tài liệu trả về (N), N tăng nên cơ may số tài liệu có liên quan sẽ tăng trên tổng số tài liệu có liên quan so với câu truy vấn trong bảng liên quan chuẩn là không đổi.

  $\Rightarrow$ R sẽ có thể tăng
- Do N tăng có nghĩa là số tài liệu trả về tăng mặc dù số tài liệu có liên quan tăng nhưng không đáng kể so với số tài liệu trả về (lúc này cũng tăng) nên P sẽ giảm.

Nói cách khác, khi cho hệ thống thực thi 1 câu truy vấn mà ta tăng số tài liệu trả về thì kết quả sẽ có được nhiều tài liệu có ích nhiều hơn nhưng số tài liệu không liên quan (tài liệu rác) cũng sẽ tăng.

5.5.2. Đường cong độ bao phủ và độ chính xác RP

Cơ sở tính bảng giá trị cho đường cong RP dựa vào bảng liên quan lý thuyết và danh sách tài liệu liên quan đã được sắp thứ tự do hệ thống truy xuất thông tin trả về (còn gọi là bảng liên quan thực tế).

Xét ví dụ sau:

Thực hiện kiểm tra hệ thống tìm kiếm thông tin với tập câu hỏi. Xét câu hỏi thứ k.

Cách tính như sau:

Tài liệu liên quan được trả về là phần giao của danh sách tài liệu liên quan theo lý thuyết và theo thực tế.

Do đó, tổng số tài liệu liên quan được trả về : 5.

Bảng giá trị R, P tính với n tài liệu được trả về sau:

Bảng 5.1: Bảng giá trị R, P tính với n tài liệu được trả về

n	Doc ID	Liên quan theo lý thuyết ?	Số tài liệu liên quan được trả về	Số tài liệu trả về	Độ bao phủ (R)	Độ chính xác (P)
1	588	true	1	1	1/5=0.2	1/1=1.00
2	589	true	2	2	2/5=0.4	2/2=1.00
3	576	false	2	3	2/5=0.4	2/3=0.67
4	590	true	3	4	3/5=0.6	3/4=0.75
5	986	false	3	5	3/5=0.6	3/5=0.60
6	592	true	4	6	4/5=0.8	4/6=0.67
7	984	false	4	7	4/5=0.8	4/7=0.57
8	988	false	4	8	4/5=0.8	4/8=0.50
9	578	false	4	9	4/5=0.8	4/9=0.44
10	985	false	4	10	4/5=0.8	4/10=0.40
11	103	false	4	11	4/5=0.8	4/11=0.36
12	591	false	4	12	4/5=0.8	4/12=0.33
13	772	true	5	13	5/5=1.0	5/13=0.38
14	990	false	5	14	5/5=1.0	5/14=0.36

Nhìn bảng giá trị trên, ta thấy tại giá trị R=0.6 có 2 giá trị P (P=0.75 và P=0.6) và ngược lại tại giá trị P=1.0 có 2 giá trị R (R=0.2, R=0.4)

Để xây dựng đường cong cho một câu truy vấn ta dùng phương pháp tính nội suy độ chính xác dựa trên 11 điểm chuẩn của độ bao phủ:

Xét các giá trị R tại các điểm chuẩn 0.0, 0.1, 0.2, 0.3, 0.4,0.5,0.6,0.7,0.8,0.9, 1.0.

Tại các vị trí tính giá trị P theo công thức sau:

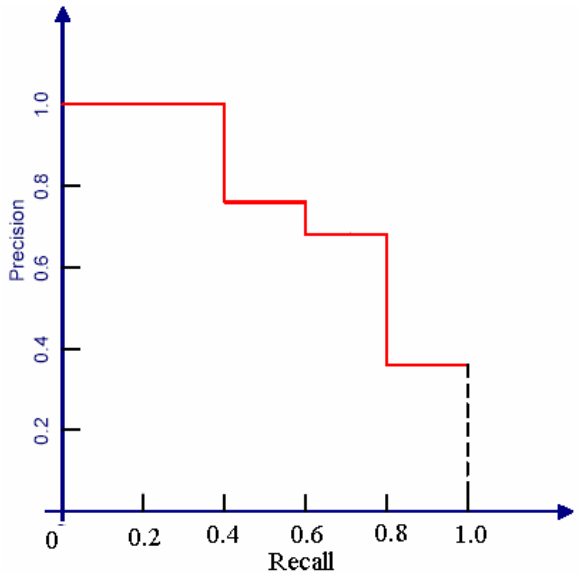
$$P_{R(i)} = \max P_{R(j)} \text{ với } j \geq i$$

Ta có bảng nội suy các giá trị P cho câu hỏi thứ k như sau:

Bảng 5.2: Bảng nội suy các giá trị P cho câu hỏi thứ k

N	ID	Độ bao phủ (R)	Độ chính xác (P)	Độ bao phủ chuẩn hoá	Độ chính xác đã nội suy
1	588	1/5=0.2	1/1=1.00	hoá	1.00
2	589	2/5=0.4	2/2=1.00	0.1	1.00
3	576	2/5=0.4	2/3=0.67	0.2	1.00
4	590	3/5=0.6	3/4=0.75	0.3	1.00
5	986	3/5=0.6	3/5=0.60	0.4	1.00
6	592	4/5=0.8	4/6=0.67	0.5	0.75
7	984	4/5=0.8	4/7=0.57	0.6	0.75
8	988	4/5=0.8	4/8=0.50	0.7	0.67
9	578	4/5=0.8	4/9=0.44	0.8	0.67
10	985	4/5=0.8	4/10=0.40	0.9	0.38
11	103	4/5=0.8	4/11=0.36	1.0	0.38
12	591	4/5=0.8	4/12=0.33		
13	772	5/5=1.0	5/13=0.38		
14	990	5/5=1.0	5/14=0.36		

Đồ thị RP cho câu hỏi thứ k



Hình 5.3: Đồ thị RP cho câu hỏi thứ k

5.5.3. Đường cong RP cho tập truy vấn

Xét tập câu truy vần gồm N câu truy vấn.

Lần lượt tính bảng giá trị RP nội suy như trên (tính P dựa trên 11 điểm chuẩn của R)

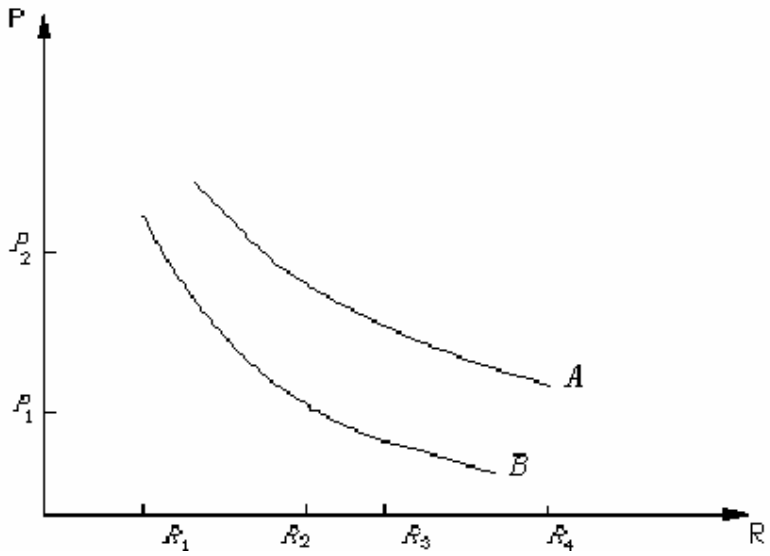
Tính giá trị trung bình P tại các điểm chuẩn của R như sau:

$$\overline{P_{(R)}} = \sum_{=1}^{N} \frac{F_{(R)}}{N}$$

Nhận xét: Phương pháp đánh giá hệ thống dựa vào bảng giá trị RP nội suy không đánh giá một cách chính xác hiệu suất tìm kiếm thông tin của hệ thống truy xuất thông tin bởi vì các giá trị của R, P là các giá trị nội suy.

5.5.4. Đánh giá hệ thống truy xuất thông tin dựa vào đồ thị

Ta tiến hành kiểm tra 2 hệ thống với cùng 1 tập tài liệu mẫu và tập câu truy vấn mẫu. Giả sử đồ thị biểu diễn của 2 hệ thống như sau:



Hình 5.4: Đồ thị biểu diễn 2 hệ thống với cùng 1 tập tài liệu mẫu và tập câu truy vấn mẫu

Nhìn trên đồ thị :

- Đường cong A biểu diễn hiệu suất thực thi của hệ thống A

- Đường cong B biểu diễn hiệu suất thực thi của hệ thống B
- Do đường A nằm trên đường B nên hiệu suất của hệ thống A lớn hơn hệ thống B.

Một cách tổng quát : đường cong nào càng gần về phía góc trên bên phải của hệ trục toạ độ (có nghĩa là độ chính xác và độ bao phủ là lớn nhất) thì đó chính là đường cong biểu diễn hiệu suất thực thi tốt nhất.

Với cách biểu diễn trên đồ thị như vậy ta có thể đánh giá nhiều hệ thống hoặc đánh giá 1 hệ thống trong những điều kiện thực thi khác nhau.

5.6. Sự liên quan giữa câu hỏi và tài liệu

5.6.1. Các độ liên quan

Các độ liên quan gồm có:

- Độ liên quan nhị phân (binary relevance): là độ liên quan chỉ có 2 giá trị: hoặc là có liên quan (relevant: 1), hoặc không liên quan (not relevant: 0).
- Độ liên quan nhiều mức độ (độ liên quan đa cấp độ): (multiple degree relevance, multiple level relevance): độ liên quan được xét ở nhiều mức độ, có nhiều giá trị. Ví dụ độ liên quan 3 mức độ :
    - Mức độ có liên quan (relevant): 2
    - Mức độ liên quan bộ phận (partically relevant): 1
    - Không liên quan (not relevant) : 0

5.6.2. Các vấn đề về độ liên quan

Cơ sở đánh giá hệ thống truy xuất thông tin:

- Một tập tài liệu (document) đại diện
- Một tập chủ đề (topic) đại diện
- Một vài câu truy vấn cho mỗi chủ đề

- Bảng đánh giá độ liên quan của mỗi tài liệu với mỗi chủ đề

Do đó vấn đề cơ bản của việc đánh giá là phải thống nhất quan điểm về mức độ liên quan.

Độ liên quan là một khái niệm đa khía cạnh (multifaceted), đa chiều (multidimensional). Khái niệm về độ liên quan đến nay vẫn là một vấn đề khó khăn trong lĩnh vực khoa học thông tin. Những cuộc nghiên cứu gần đây đã tập trung vào nhân tố ảnh hưởng lên việc đánh giá độ liên quan và chiều (hoặc tiêu chuẩn) của độ liên quan. Có nhiều loại độ liên quan: độ liên quan thuật toán, độ liên quan chủ đề, độ liên quan nhận thức, độ liên quan tình huống, độ liên quan động cơ.

Độ liên quan vốn mang tính chủ quan, đánh giá độ liên quan thường không thống nhất do tính cá nhân và nhân tố thời gian :

- Một tài liệu được đánh giá là có liên quan với tỉ lệ nào đó nhưng đối với người khác tỉ lệ này sẽ khác => độ liên quan phủ thuộc tính cá nhân
- Một tài liệu được đánh giá là có liên quan với tỉ lệ nào đó tại thời điểm t, nhưng tại thời điểm t’ tỉ lệ đó sẽ thay đổi => độ liên quan phụ thuộc nhân tố thời gian. Tuy nhiên sự thay đổi này có thể chấp nhận được do nó tương đối thấp. Trong hầu hết các thử nghiệm đánh giá hệ thống tìm kiếm thông tin (bao gồm cả những thử nghiệm của TREC) người ta thường quan tâm độ liên quan nhị phân (có nghĩa là tài liệu hoặc là được đánh giá là có liên quan (1) hoặc không có liên quan (0)). Ưu điểm của dộ liên quan nhị phân là việc tính toán R, P đơn giản; khuyết điểm là không thể phản ánh được khả năng liên quan của tài liệu ở nhiều mức độ đúng với thực tế.

Trong cách đánh giá tìm kiếm thông tin của TREC, khái niệm "liên quan" là một khái niệm tuyệt đối: một tài liệu hoặc là liên quan hoặc là không liên quan.

Điều giả sử này nhằm làm đơn giản hóa việc tính toán các độ đo. Nhiều cuộc kiểm tra khác đã tiến hành đánh giá với tỷ lệ độ liên quan nhiều mức độ. Độ liên quan 3 cấp độ đã được thực hiện ở Hội nghị NTCIR 1999 (NII-NACSIS Test Collection for IR systems), WEB track của TREC-9.

Độ liên quan 4 cấp được dùng trong NTCIR 2000.

Tỷ lệ độ liên quan của một tài liệu tại vị trí thứ N sẽ được trừ hao, điều này phản ánh một tình trạng là tài liệu trả về càng phía dưới danh sách càng có ít giá trị hơn đối với người sử dụng : mặc dù do mức độ tương quan không giảm nhưng sự trùng lặp thông tin với những tài liệu phía trên cũng làm cho tài liệu phía dưới kém phần giá trị hơn.

Giả sử rằng sự liên quan của một tài liệu là độc lập với các tài liệu khác là không thực tế trong hầu hết các trường hợp. Trong hầu hết các nhiệm vụ tìm kiếm thông tin cơ bản giống như tìm kiếm trên mạng, tìm kiếm câu trả lời cho một câu hỏi đặc biệt nào đó hoặc cho một vài sự tham khảo nào đó, giả sử rằng một người dùng đọc lướt qua các tài liệu được trả về sẽ bắt đầu với tài liệu dễ thấy nhất, nổi bật nhất (ở phía trên danh sách) do đó độ liên quan của tài liệu phía dưới danh sách sẽ phụ thuộc vào những tài liệu đã được đọc. Khả năng một tài liệu chứa những thông tin mới sẽ giảm xuống đến cuối danh sách tài liệu. Sự phụ thuộc này thường được bỏ qua trong những lần nghiên cứu tìm kiếm thông tin.

Ngoài ra việc định giá độ liên quan này mang tính chủ quan. Chúng ta thường có nhiều ý kiến khác nhau về mức độ liên quan. Do đó mức độ liên quan của tài liệu được phân biệt:

- Bảng liên quan đươc định giá do tác giả của tài liệu hay không phải tác giả
- Bảng liên quan được định giá bởi một nhóm đánh giá
- Bảng liên quan được định giá trong cùng điều kiện hay được định giá trong các điều kiện khác nhau.

5.6.3. Đánh giá với độ liên quan nhiều cấp độ

(*Multiple degree relevance or non-binary relevance*)

Trong một vài thử nghiệm về đánh giá độ liên quan nhiều cấp độ chỉ có một vài thí nghiệm thực sự cho thấy lợi ích của việc đánh giá độ liên quan ở nhiều cấp độ khác nhau.

Độ bao phủ (R), độ chính xác (P) là phương pháp cổ điển để đánh giá khả năng thực thi của IR và thường được tính dựa trên việc đánh giá độ liên quan nhị phân. Do đó việc đánh giá độ liên quan nhiều cấp độ chỉ được tiến hành ở bước đầu, sau đó những giá trị mức độ sẽ được qui về 2 giá trị 0, 1 để đánh giá.

Ví dụ : đánh giá độ liên quan được tiến hành 3 mức độ:

- có liên quan (relevant) => ký hiệu A
- liên quan một phần (partically relevant) => ký hiệu B
- không liên quan (not relevant) => ký hiệu C

Mức độ liên quan sẽ được qui về 2 giá trị để tính R, P. Có 2 cách tính:

- A, B mang giá trị 1 (có liên quan) C mang giá trị 0 (không liên quan) hoặc
- A mang giá giá trị 1 (có liên quan) B, C mang giá trị 0 (không liên quan)

Với cách tiến hành như vậy để duy trì mức độ liên quan của tài liệu, định dạng một tập tin đánh giá độ liên quan (relevant judgement) như sau:

topic-ID dumy doc-ID relevant assessment

Trong đó:

topic-ID : chỉ số của chủ đề (topic)

dumy : là trường cho biết tài liệu đó có mức độ liên quan là bao nhiêu (A, hoặc B, hoặc C)

doc-ID : chỉ số tài liệu

relevant assessment: mang giá trị 0 hoặc 1, giá trị đánh giá độ liên quan sau khi được qui về độ liên quan nhị phân.

Một ví dụ khác về đo độ liên quan của tài liệu ở 4 mức độ:

- độ liên quan cao (highly relevant)
- độ liên quan vừa (fairly relevant)
- độ liên quan trung bình (marginally relevant)
- không liên quan (irrelevant)

Tuy nhiên trong các Hội nghị về Đánh giá các hệ thống thông tin gần đây, độ liên quan nhị phân vẫn còn được xem là một cách đánh giá chuẩn, thậm chí nhiều trường hợp đánh giá độ liên quan ở nhiều cấp độ nhưng cũng được qui về đánh giá nhị phân để tính độ bao phủ và độ chính xác. Cách tiến hành này có khuyết điểm là nó không kiểm tra được từng mức độ cụ thể của độ liên quan. Một số người có quan điểm là cách đo độ R và P dựa vào việc đánh giá nhị phân là nên tránh vì cách tính như vậy không quan tâm đến sự thay đổi và độ phức tạp của mức độ liên quan, làm sai lệch tính tự nhiên và thực tế của độ liên quan. Một giải pháp để giải quyết vấn đề này là tổng quát hoá độ R và P.

Dựa vào lý thuyết, thực nghiệm, nghiên cứu, mức độ liên quan của tài liệu thay đổi một cách rõ ràng, một vài tài liệu thì liên quan nhiều hơn, một số khác thì ít hơn. Thật là khó để xác định mức độ liên quan khi tiến hành đánh giá. Điều này còn tuỳ thuộc vào tình huống đánh giá hệ thống của chúng ta.

5.6.4. Phương pháp đo độ bao phủ (R), độ chính xác (P) dựa trên độ liên quan nhiều cấp độ

Phương pháp đo dựa vào độ bao phủ (R) và độ chính xác (P) là một phương pháp truyền thống nhưng độ đo R, P chỉ được tính dựa vào độ liên quan nhị phân.

Đối với trường hợp độ liên quan nhiều cấp độ ta có 2 cách giải quyết sau:

- Qui tất cả mức độ liên quan về 2 giá trị 0, 1 (giống như đưa về độ liên quan nhị phân) => cách này theo Schamber là nên tránh.
- Tổng quát hoá R và P

Độ bao phủ tổng quát và độ chính xác tổng quát:

(generalized, non-binary recall and precision)

Gọi R là tập n tài liệu được phục hồi từ cơ sở dữ liệu tài liệu

D={ d_1, d_2, …, d_N } với một câu truy vấn thuộc về một chủ đề nào đó, R ≤ D

Gọi tài liệu d_i trong cơ sở dữ liệu tài liệu có tỉ lệ độ liên quan là $r(d_i)$

Độ bao phủ tổng quát gR và độ chính xác tổng quát gP được tính theo công thức như sau:

$$gP = \frac{\sum_{d \in D} r(d)}{n}$$

$$gR = \frac{\sum_{d \in R} r(d)}{\sum_{d \in D} r(d)}$$

Cách tính này cũng tương tự tính R, P nhị phân truyền thống, nó cũng cho phép tính R trung bình và P trung bình của tập câu truy vấn, tính P dựa trên R, hoặc tính dựa trên ngưỡng giới hạn số tài liệu trả về và cũng cho phép biểu diễn đường cong PR

Ghi chú: r(d) là một con số thực có giá trị trong khoảng (0.0, 1.0). Ví dụ với mức độ liên quan là 4. Tính r(d)

- Mức độ liên quan cao : 3 => r(d)=3/4
- Mức độ liên quan vừa : 2 => r(d)=2/4
- Mức độ liên quan trung bình : 1 => r(d)=1/4
- Không liên quan :0 => r(d)=0

KẾT LUẬN

Hiện nay có rất nhiều hệ thống truy xuất thông tin (Information Retrieval system) đang tồn tại để trợ giúp con người. Tuy nhiên, khả năng tìm kiếm thông tin của các hệ thống này chắc chắn khác nhau. Do đó, việc đánh giá các hệ thống truy xuất thông tin (Evaluation of Information Retrieval systems) là một nhu cầu không thể thiếu nhằm xác định các hệ thống truy xuất thông tin hiệu quả.

Luận văn đã nghiên cứu các vấn đề về các hệ truy xuất thông tin và đánh giá về các hệ truy xuất thông tin. Việc đánh giá này có ý nghĩa rất lớn đối với sự tồn tại và phát triển của các hệ thống truy xuất thông tin. Nó giúp xác định khả năng tìm kiếm của các hệ thống truy xuất thông tin. Từ đó mà các tổ chức, công ty, trường học tạo ra hệ thống này có thể phát triển, thay đổi hệ thống để đưa ra khả năng tìm kiếm thông tin tốt nhất.

Việc đánh giá hệ truy xuất thông tin (IR) là để biết được điểm mạnh, điểm yếu của từng hệ thống IR mà từ đó ta chọn ra được hệ thống IR tối ưu phục vụ cho nhu cầu tìm kiếm thông tin một cách có hiệu quả.

Tôi hy vọng đề tài này sẽ là một đóng góp nhỏ, có ý nghĩa cho việc nghiên cứu về lĩnh vực truy xuất thông tin.

HƯỚNG PHÁT TRIỂN

Việc nghiên cứu đánh giá các hệ thống tìm kiếm thông tin rất đa dạng với nhiều phương pháp, mô hình đánh giá khác nhau. Những mô hình, phương pháp này đang được tiếp tục nghiên cứu, bàn luận trên thế giới.

Trên cơ sở những phần đã nghiên cứu, đề tài có hướng phát triển về phương pháp đánh giá: Ngoài cách đánh giá dựa vào 11 điểm chuẩn của độ bao phủ, đề tài có thể phát triển thêm các phương pháp đánh giá khác như phương pháp đánh giá dựa trên độ chính xác trung bình nghiêm ngặt (Mean Average Precision – MAP), đo dựa trên giá trị đơn Swet's E-Measure (Single-valued Measure) hoặc chiều dài tìm kiếm trung bình.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

1. Nguyễn Duy Hiệp - Hoàng Minh Ngọc Hải (2004), "**Xây dựng tòa soạn điện tử có hỗ trợ lấy tin từ các website khác**", luận văn cử nhân, trường Đại học Khoa học Tự nhiên.
2. Nguyễn Thị Thanh Hà – Nguyễn Trung Hiếu (2005), "**Xây dựng hệ thống tìm kiếm thông tin tiếng Việt dựa trên các chỉ mục là các từ ghép**", luận văn cử nhân, trường Đại học Khoa học Tự nhiên.

Tiếng Anh:

1. Gerald J.Kowalski, Mark T.Maybury, "**Information Storage and Retrieval System"**, 2004
2. Gerard Salton, Michael J.McGill, "**Introduction to Modern Information Retrieval**", International Student Edition, New York, 1983.
3. William B.Frakes, Ricardo Baeza – Yakes, "**Information Retrieval – Data Structures & Algorithms**", 1992.
4. Ricardo Baeza – Yakes, Berthier Ribeiro-Neto, **"Modern Information Retrieval** ", Addison Press, Anh, 1999.
5. Dong Thi Bich Thuy, Ho Bao Quoc, Marie-France Bruandet, Jean-Pierre Chevallet, "**An approach to Vietnamese Information Retrival**".