ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGHIÊN CỨU TRA CỨU THÔNG TIN TIẾNG VIỆT VỚI PHẢN HỒI LIÊN QUAN

NGUYỄN ĐỨC TOÀN

Thái Nguyên, 2015

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả luận văn

Nguyễn Đức Toàn

LỜI CẢM ƠN

Để hoàn tất một luận văn thạc sĩ yêu cầu sự tập trung, sự cố gắng và độc lập nghiên cứu. Bản thân tôi sau những năm tháng học tập vất vả và nghiên cứu cũng đã cố gắng để hoàn thành được luận văn này. Tôi luôn ghi nhận những sự đóng góp giúp đỡ,sự ủng hộ, sự hỗ trợ nhiệt tình của những người bên cạnh mình, nhân đây tôi muốn gửi lời cảm ơn sâu sắc nhất tới họ.

Lời cảm ơn trân trọng đầu tiên tôi muốn dành tới TS Nguyễn Hữu Quỳnh, người đã dìu dắt và hướng dẫn tôi trong suốt quá trình làm luận văn, sự chỉ bảo và định hướng của thầy giúp tôi tự tin nghiên cứu những vấn đề mới và giải quyết bài toán một cách khoa học.

Tôi xin trân trọng cảm ơn Ban giám hiệu, Bộ phận sau Đại học, Phòng đào tạo, phòng khảo thíĐại học công nghệ thông tin và truyền thông – Đại học Thái Nguyên, đã tạo các điều kiện cho chúng tôi được học tập và làm khóa luận một cách thuận lợi.

Lời cảm ơn sâu sắc muốn được gửi tới các thầy giáo, cô giáo đã dạy dỗ và mở ra cho chúng tôi thấy chân trời tri thức mới, hướng dẫn chúng tôi cách khám phá và làm chủ công nghệ mới.

Tôi muốn gửi lời cảm ơn chân thành đến tập thể lớp CH12D đã cùng tôi đi qua những tháng ngày miệt mài học tập, cùng chia sẻ những niềm vui nỗi buồn, động viên tôi đi qua những khó khăn, để tôi vững bước vượt qua những vất vả, quyết tâm hoàn thành luận văn này.

Tôi xin trân trọng cảm ơn bố mẹ, vợ, con tôi đã mang tới tất cả niềm tin, định hướng và theo dõi tôi suốt chặng đường đời. Nâng đỡ và đến bên tôi những giây phút khó khăn nhất của cuộc sống.

Tôi xin chân thành cảm ơn Ban giám hiệu, Phòng đào tạo và CTHS và đồng nghiệpTrường trung cấp Y tế Nam Định, những người đã tạo điều kiện và giúp đỡ tôi trong công việc và học tập để tôi có thể theo học và hoàn thành khóa luận tốt nghiệp.

Thái Nguyên, ngày tháng 06 năm 2015

MỤC LỤC

DANH MỤC CÁC HÌNH

PHẦN MỞ ĐẦU

1. Đặt vấn đề

Trong thời đại ngày nay, thông tin là nhu cầu thiết yếu đối với mọi người trên mọi lĩnh vực. Hằng ngày có hàng triệu văn bản, trang web được đưa lên Internet, làm giàu cho hệ thống tài nguyên khổng lồ này. Tuy nhiên, chúng ta không thể sử dụng thông tin trong hệ thống thông tin khổng lồ này nếu chúng ta không tổ chức và khai thác nguồn tài nguyên này một cách hợp lí.

Trên thực tế, đã có khá nhiều hệ thống thực hiện công việc này theo những phương pháp khác nhau, tuy chưa đạt được hiệu quả tối ưu nhưng cũng phần nào đáp ứng được các yêu cầu thông tin cho người sử dụng. Mỗi phương pháp khác nhau đều thể hiện được những điểm mạnh riêng của nó và việc lựa chọn phương pháp nào phụ thuộc vào những mục đích, yêu cầu và tiêu chí riêng đặt ra. Tuy nhiên, việc khai thác nguồn dữ liệu này vẫn còn là một bài toán khó.

Kỹ thuật tra cứu thông tin đã và đang được nghiên cứu, phát triển trong nhiều lĩnh vực khác nhau như y tế, giáo dục, kinh tế...Những kiến thức liên quan đến tra cứu thông tin là rất rộng và tổng hợp, bao gồm thuật toán, cấu trúc dữ liệu, cơ sở dữ liệu, các hệ thống phân tán, tính toán song song, tổ chức file, data mining.

Để nâng cao chất lượng của các kết quả tra cứu, phản hồi liên quan được kết hợp vào hệ thống tra cứu thông tin. Ý tưởng của phản hồi liên quan (RF- Relevance Feedback) là bao gồm người dùng tham gia vào quá trình tra cứu để cải tiến tập kết quả cuối cùng. Cụ thể, người dùng đưa phản hồi về sự liên quan của các tài liệu trong một tập các kết quả ban đầu. Phản hồi liên quan có thể đi qua một hay nhiều vòng lặp của sự sắp xếp này. Quá trình sử dụng ý tưởng có thể khó để tính một truy vấn tốt khi chúng ta không biết

toàn bộ tập tài liệu, nhưng dễ đánh giá các tài liệu cụ thể. Trong ngữ cảnh như thế, phản hồi liên quan cũng có thể hiệu quả trong theo dõi nhu cầu thông tin của người dùng: xem một số tài liệu có thể dẫn người dùng cải tiến hiểu thông tin mà họ đang tìm.

Vì những lý do trên tôi đã chọn đề tài **"Nghiên cứu tra cứu thông tin tiếng Việt với phản hồi liên quan"**.

2. Mục tiêu của luận văn

Nghiên cứu phương pháp sử dụng phản hồi liên quan để nâng cao độ chính xác của tra cứu văn bản (lấy thông tin của người dùng để nâng cao độ chính xác).

3. Các đóng góp của luận văn

- Nghiên cứu một số phương pháp tra cứu đối với văn bản tiếng Việt.

- Sử dụng kỹ thuật phản hồi liên quan nhằm nâng cao hiệu năng của hệ thống tra cứu văn bản tiếng Việt.

- Trên cơ sở phương pháp đã được nghiên cứu, luận văn tiến hành xây dựng hệ thống tra cứu thông tin và ứng dụng trong tra cứu thông tin tiếng Việt.

4. Bố cục của luận văn

Chương 1: Tổng quan về tra cứu thông tin

Chương 2: Tra cứu thông tin tiếng việt sửa dụng phản hồi liên quan

Chương 3: Ứng dụng tra cứu thông tin văn bản tiếng Việt

Kết luận và hướng phát triển

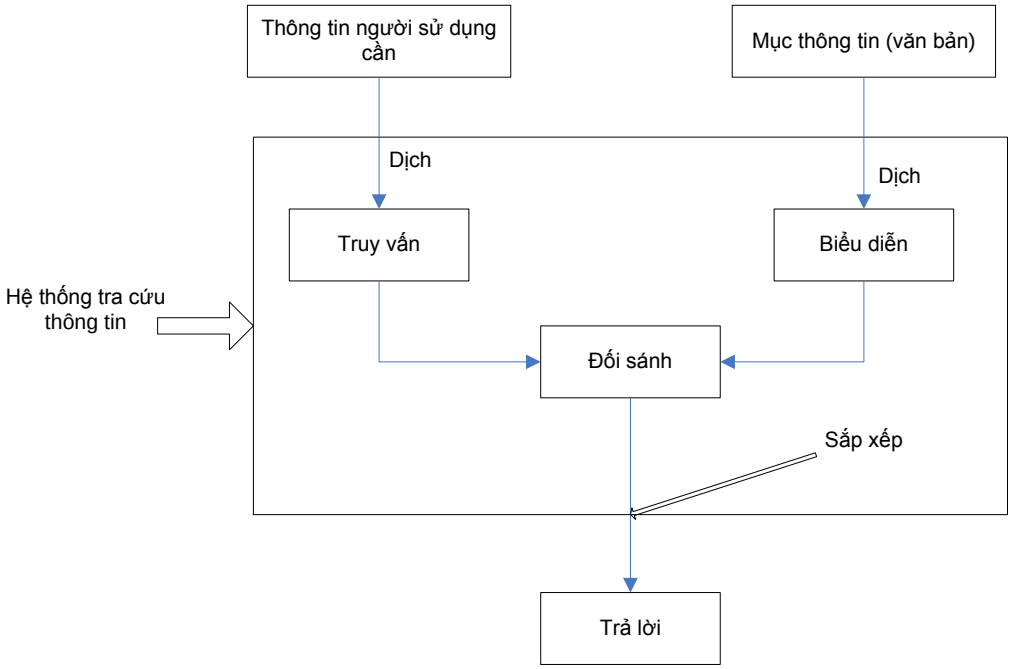
Tài liệu tham khảo

Chương 1 : TỔNG QUAN VỀ TRA CỨU THÔNG TIN

1.1.Tra cứu thông tin

Tra cứu thông tin là một nhánh của khoa học máy tính nhằm mục tiêu lưu trữ và cho phép truy cập nhanh một lượng thông tin lớn. Thông tin này có thể là văn bản, đa phương tiện hoặc âm thanh [2,3].

Một hệ thống tra cứu thông tin là hệ thống có thể lưu trữ, tra cứu các mục thông tin. Hiện nay, nhiều hệ thống tra cứu các mục phi văn bản dựa trên các tìm kiếm mô tả văn bản. Các mục văn bản thường được xem như là các tài liệu, sách, bài báo, ...Các hệ thống tra cứu thông tin thực tế nhất lưu trữ và cho phép tra cứu các tài liệu hoặc thông tin văn bản. Tuy nhiên, đây không phải là một nhiệm vụ dễ dàng, vì các tập tài liệu trong hệ thống tra cứu thông tin thường phải xử lý vài chục ngàn hoặc vài chục triệu tài liệu.



Hinh 1.1 Tổng quan hệ thống tra cứu thông tin

Người sử dụng truy cập hệ thống tra cứu thông tin bằng việc tạo một truy vấn (gửi yêu cầu vào hệ thống). Sau đó hệ thống tra cứu thông tin tra cứu tất cả các tài liệu liên quan đến yêu cầu truy vấn [2,3]. Đối với mục tiêu này, trong pha ban đầu, các tài liệu được phân tích để cung cấp một biểu diễn của

nội dung: quá trình này được gọi là "đánh chỉ số". Lúc đầu tài liệu được phân tích, một đại diện mô tả tài liệu được lưu trữ, trong khi bản thân tài liệu cũng được lưu trữ. Để biểu diễn các nhu cầu thông tin, người sử dụng tạo truy vấn trong ngôn ngữ truy vấn của hệ thống. Yêu cầu truy vấn được đối sánh với các mục để xác định các tài liệu liên quan đến người sử dụng.

Phản hồi đối với một truy vấn, hệ thống tra cứu thông tin có thể cung cấp hoặc một trả lời chính xác hoặc danh sách phân hạng các tài liệu chứa thông tin liên quan đến truy vấn. Kết quả phụ thuộc vào mô hình được chọn của hệ thống, mô hình boolean cho ra một trả lời chính xác, trong khi các mô hình khác (áp dụng lược đồ đối sánh từng phần) cho ra một danh sách các tài liệu được phân hạng sao cho tài liệu nào tương tự nhất được xếp hạng ở trên. Lược đồ một hệ thống tra cứu thông tin được thể hiện như Hình 1.1.

1.2.Các thành phần của hệ thống tra cứu thông tin

Trọng tâm của hệ thống tra cứu thông tin là so sánh truy vấn với mỗi tài liệu trong tập hợp. Điều này thu được bằng chức năng tính điểm, có đầu vào là biểu diễn của các tài liệu và truy vấn. Chi tiết hàm tính điểm và biểu diễn của các tài liệu phụ thuộc vào mô hình tra cứu được sử dụng. Chuyển đổi truy vấn, từ đầu vào thành biểu diễn, được thực hiện ngay khi được nhập vào bởi người dùng. Với các tài liệu trong cơ sở dữ liệu, chuyển đổi là một quá trình ngoại tuyến được thực hiện một lần.

Xét một tập hợp C chứa N tài liệu và T thuật ngữ duy nhất. Mỗi tài liệu trong tập hợp này được biểu thị bởi d_i và tính toán biểu diễn cho mỗi d_i là ánh xạ một chiều $d_i \Rightarrow \mathbf{d_i}$. Dạng gốc của $d_i$là một chuỗi các từ. Biểu diễn $\mathbf{d_i}$có thể được xem như một chuỗi T trọng số tương ứng với mức độ đối với mỗi thuật ngữ mô tả tài liệu.

Nếu mỗi tài liệu trong tập hợp đượcxem như một chuỗi các trọng số, bản thân tài liệu có thể được biểu diễn bằng một ma trận tài liệu-thuật ngữ, ở