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

HÀ THANH THỦY

TÌM KIẾM VĂN BẢN THEO NỘI DUNG VÀ ỨNG DỤNG

Luận văn thạc sĩ khoa học máy tính

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

Thái nguyên – 2012

LỜI CẢM ƠN

Để hoàn thành chương trình cao học, tôi đã nhận được sự hướng dẫn, giúp đỡ và góp ý nhiệt tình của quý thầy cô trường Đại học Công nghệ thông tin - Truyền thông, Đại học Thái Nguyên.

Trước hết, tôi xin chân thành cảm ơn quí thầy cô trường Đại học Công nghệ thông tin - Truyền thông, đặc biệt là những thầy cô đã tận tình dạy bảo cho tôi suốt thời gian học tập tại trường.

Tôi xin bày tỏ lòng biết ơn sâu sắc tới **PGS.TS. Đặng Văn Đức** người thầy đã dành rất nhiều thời gian, tâm huyết và sự tận tình giúp đỡ, hướng dẫn tôi trong suốt quá trình nghiên cứu để hoàn thành luận văn này.

Đồng thời, tôi xin chân thành cảm ơn Sở Giáo dục và đào tạo tỉnh Thái Nguyên, Ban Giám hiệu trường THPT Lương Ngọc Quyến đã tạo điều kiện giúp đỡ tôi về mọi mặt để tôi học tập và hoàn thành tốt khóa học.

Cuối cùng tôi xin chân thành cảm ơn gia đình và bạn bè, những người đã động viên, khuyến khích tôi trong suốt quá trình học tập và nghiên cứu.

Mặc dù đã có nhiều cố gắng hoàn thiện luận văn bằng tất cả sự nhiệt tình và năng lực của mình, tuy nhiên vẫn không thể tránh khỏi những thiếu sót, tôi rất mong nhận được những đóng góp quí báu của quí thầy cô và các bạn.

Thái Nguyên, ngày 20 tháng 6 năm 2012

Học viên

Hà Thanh Thủy

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.

Thái Nguyên, ngày 20 tháng 6 năm 2012

Học viên

Hà Thanh Thủy

DANH MỤC CÁC KÍ HIỆU, CHỮ CÁI VIẾT TẮT

Từ gốc	Giải nghĩa
CSDL	Cơ sở dữ liệu
DBMS (DataBase Management System)	Hệ quản trị Cơ sở dữ liệu
IR (Information Retrieval)	Truy tìm thông tin
IDF(Inverse Document Frequency)	Tần số xuất hiện tài liệu phù hợp
LSI(Latent Semantic Indexing)	Chỉ số hóa ngữ nghĩa ẩn
MMDBMS (Multimedia Database Management System)	Hệ quản trị cơ sở dữ liệu đa phương tiện
SVD(Singular Value Decomposition)	Kỹ thuật tách giá trị đơn
TF (Term Frequency)	Tần số xuất hiện thuật ngữ

DANH MỤC HÌNH VẼ

MỤC LỤC

MỞ ĐẦU

Công nghệ thông tin trên thế giới đang phát triển rất nhanh trong giai đoạn hiện nay. Những tiến bộ của khoa học công nghệ thông tin được áp dụng phục vụ công tác nghiên cứu khoa học, công tác quản lý, sản xuất và phục vụ đời sống con người hàng ngày. Càng ngày, người ta càng nhận thấy tính hiệu quả, tiện ích của khoa học trí tuệ đã từng bước thay thế lao động thủ công, giảm bớt thời gian lao động, tiết kiệm chi phí và tạo nên rất nhiều tiện ích khác. Có thể khẳng định rằng: công nghệ thông tin đã mở ra một kỉ nguyên mới, kỉ nguyên của tự động hoá và đã tạo ra một bước nhảy vọt của nền khoa học thế giới cũng như nền văn minh nhân loại

Trong xu thế phát triển chung của xã hội cũng như định hướng của Đảng và Nhà nước ta trong việc áp dụng công nghệ thông tin trong tất cả các lĩnh vực đời sống, với sự phát triển mạnh mẽ của công nghệ thông tin, tất cả các lĩnh vực đời sống trong xã hội đã tạo ra một khối lượng dữ liệu khổng lồ. Trong rất nhiều tình huống, chúng ta phải tìm ra những thông tin cần thiết từ kho dữ liệu khổng lồ đã có ấy. Tuy nhiên, vì khối lượng dữ liệu lớn, vì thời gian hạn hẹp cho nên nhiều khi việc tìm kiếm dữ liệu gặp rất nhiều khó khăn.

Do đó, cần có các hệ thống tìm kiếm thông tin để hỗ trợ người dùng tìm kiếm nhanh và hiệu quả những thông tin mà họ quan tâm. Việc tìm tòi nghiên cứu ứng dụng những thuật toán giúp cho việc tìm kiếm dữ liệu được nhanh chóng, tiết kiệm thời gian, có hệ thống và khoa học là một việc làm hết sức cần thiết trong giai đoạn hiện nay.

Văn bản là một trong số các dạng của dữ liệu đa phương tiện, nó được quan tâm từ hàng nghìn năm trước trong việc tổ chức sắp xếp và lưu trữ. Tài liệu văn bản chiếm đa số trong mọi cơ quan tổ chức, đặc biệt là trong thư viện

và còn được sử dụng để mô tả các dạng khác của dữ liệu đa phương tiện như video, audio, hình ảnh. Số lượng tài liệu văn bản ngày càng lớn và có vai trò vô cùng quan trọng, vì thế việc lưu trữ, xử lý và truy tìm thủ công trước đây không thể hoặc khó có thể thực hiện được.

Vì vậy mục tiêu của luận văn này nhằm tìm hiểu một số kỹ thuật tìm kiếm văn bản theo nội dung trong cơ sở dữ liệu đa phương tiện nhằm đáp ứng được những nhu cầu cấp thiết của thời đại bùng nổ thông tin điện tử.

Trên thực tế, đã có nhiều công trình nghiên cứu về vấn đề này được công bố ở cả trong và ngoài nước. Mục tiêu luận văn với đề tài *"Tìm kiếm văn bản theo nội dung và ứng dụng"* mà tôi hướng tới là nghiên cứu một số kỹ thuật/phương pháp mới, thử đánh giá so sánh và ứng dụng vào môi trường cụ thể.

Đối tượng và phạm vi nghiên cứu

Hệ thống đa phương tiện là một vấn đề phức tạp và rộng lớn, do vậy phạm vi nghiên cứu của luận văn chỉ giới hạn trong việc sử dụng một số kỹ thuật tìm kiếm văn bản theo nội dung, sau đó phát triển chương trình demo ứng dụng tìm kiếm văn bản theo nội dung.

Hướng nghiên cứu của đề tài

- Nắm vững qui trình thiết kế CSDL đa phương tiện, trong đó CSDL văn bản là thành phần quan trọng.
- Nghiên cứu một số kỹ thuật tìm kiếm văn bản theo nội dung như: mô hình tìm kiếm Bool, mô hình tìm kiếm không gian vector, mô hình tìm kiếm theo xác suất, kỹ thuật chỉ mục ngữ nghĩa tiềm ẩn (Latent Sematic Indexing-LSI).
- Nghiên cứu các độ đo phù hợp để đánh giá hiệu năng hệ thống

- Xây dựng thử nghiệm chương trình demo tìm kiếm văn bản theo nội dung trên cơ sở bộ thư viện dtSearch.

Phương pháp nghiên cứu

- Tổng hợp tài liệu từ nhiều nguồn khác nhau.
- Phân tích, liệt kê, so sánh, đối chiếu, trực quan, thực nghiệm,…

Cấu trúc luận văn

Ngoài phần mở đầu giới thiệu ý nghĩa của chủ đề nghiên cứu và phần kết luận nêu lên các kết quả chính đã đạt được, luận văn gồm các chương sau đây:

Chương I: Giới thiệu tổng quan về hệ thống tìm kiếm thông tin theo nội dung.

Chương II: Một số kỹ thuật tìm kiếm thông tin văn bản theo nội dung.

Chương III: Nghiên cứu thử nghiệm thư viện tìm kiếm văn bản dtSearch.