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

ĐỖ TẤT HƯNG

MỘT SỐ KỸ THUẬT TÌM KIẾM VĂN BẢN THEO NỘI DUNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

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

ĐỖ TẤT HƯNG

MỘT SỐ KỸ THUẬT TÌM KIẾM VĂN BẢN THEO NỘI DUNG

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC **CHỦ TỊCH HỘI ĐỒNG**

TS. TRƯƠNG HÀ HẢI **PGS. TS. NGÔ QUỐC TẠO**

Thái Nguyên - 2015

LỜI CAM ĐOAN

Em xin cam đoan : Luận văn thạc sĩ Khoa học máy tính **"Một số kỹ thuật tìm kiếm văn bản theo nội dung"** này là công trình nghiên cứu thực sự của cá nhân em, được thực hiện trên cơ sở nghiên cứu lý thuyết và dưới sự hướng dẫn khoa học của Tiến sĩ Trương Hà Hải, Trường Đại học Công nghệ Thông tin và Truyền thông**.**

Em xin chịu trách nhiệm về lời cam đoan này.

Thái Nguyên, ngày 6 tháng 7 năm 2015

Tác giả

Đỗ Tất Hưng

LỜI CẢM ƠN

Để hoàn thành luận văn, em xin chân thành cảm ơn Trường Đại học Công nghệ Thông tin và Truyền thông, Phòng Đào tạo, các thầy, cô giáo giảng dạy lớp cao học Khoa học máy tính K12E đã quan tâm, tạo điều kiện thuận lợi, tận tình giảng dạy và giúp đỡ em trong thời gian theo học tại trường.

Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến **TS. Trương Hà Hải**, người đã dành nhiều thời gian, tâm huyết hướng dẫn em trong suốt quá trình nghiên cứu và hoàn thành luận văn.

Em cũng xin cảm ơn các cán bộ, giảng viên đồng nghiệp ở Trường Đại học Hùng Vương đã tạo điều kiện về thời gian để em có thể học tập và hoàn thành luận văn.

Mặc dù đã cố gắng hết sức hoàn thiện luận văn, tuy nhiên luận văn vẫn còn nhiều thiếu sót, rất mong sự góp ý quý báu của quí thầy cô và các bạn đồng nghiệp!

Xin trân trọng cảm ơn!

Thái Nguyên, ngày 6 tháng 7 năm 2015

Tác giả

Đỗ Tất Hưng

MỤC LỤC

Trang

LỜI CAM ĐOAN iii
LỜI CẢM ƠN iv
MỤC LỤC v
DANH MỤC BẢNG viii
DANH MỤC HÌNH VẼ ix
MỞ ĐẦU 1
CHƯƠNG I. TỔNG QUAN VỀ CƠ SỞ DỮ LIỆU ĐA PHƯƠNG TIỆN 4
1.1 Cơ sở dữ liệu (CSDL) đa phương tiện 4
1.1.1 Giới thiệu 4
1.1.2 Mục tiêu chính 5
1.1.3 Mô hình dữ liệu đa phương tiện 5
1.1.4 Trích chọn đặc trưng, chỉ mục và đo tính tương tự 7
1.1.5 Hệ thống tìm kiếm thông tin (IR) 13
1.1.6 Xếp hạng tài liệu (Ranking) 19
1.2 Bài toán tìm kiếm văn bản 23
CHƯƠNG II. MỘT SỐ VẤN ĐỀ VỀ TÌM KIẾM VĂN BẢN THEO NỘI DUNG 26
2.1 Mô hình Boolean 26
2.2 Mô hình tìm kiếm không gian vector 27
2.3 Mô hình tìm kiếm theo xác suất 30
2.4 Mô hình tìm kiếm dựa trên cơ sở cụm 30
2.5.1 Ý tưởng cơ bản của LSI 33
2.5.2 Một số khái niệm cơ bản 39
2.5.3 Kỹ thuật phân tích SVD 41
CHƯƠNG III. ỨNG DỤNG THỬ NGHIỆM 55
3.1 Bài toán 55

3.2 Chức năng của chương trình thử nghiệm 57
3.3 Hoạt động cơ bản trong chương trình 65
3.4 So sánh các mô hình tìm kiếm 67
KẾT LUẬN 69
1. Kết luận 69
2. Hướng phát triển 69
TÀI LIỆU THAM KHẢO 71

DANH MỤC TỪ VIẾT TẮT

CSDL	Cơ sở dữ liệu
IDF	Inverse Document Frequency
IR	Information Retrieval
LSI	Latent Semantic Indexing
MDMS	MultiMedia DataBase Manager System
MIRS	Multimedia Information Retrieval System
SVD	Singular value decomposition
TF	Term Frequency

DANH MỤC BẢNG

Bảng 1.1 Ma trận tài liệu - thuật ngữ 21
Bảng 1.2 Ma trận kết quả tài liệu - thuật ngữ TF-IDF 22
Bảng 1.3 Kết quả khoảng cách từ truy vấn Q với các tài liệu 23
Bảng 2.1 Số lần xuất hiện của thuật ngữ trong mỗi tài liệu 43

DANH MỤC HÌNH VẼ

Hình 1.1 Mô hình dữ liệu đa phương tiện 6
Hình 1.2 Mô hình xử lý cho hệ thống lập chỉ mục 11
Hình 1.3 Mô hình tổng quát tìm kiêm thông tin 15
Hình 1.4 Tiến trình truy vấn tài liệu 17
Hình 1.5 Hệ thống IR tiêu biểu 25
Hình 2.1 Sử dụng các khái niệm cho truy vấn 34
Hình 2.2 Các vector văn bản theo mô hình LSI 39
Hình 2.3 Biểu diễn ma trận xấp xỉ Ak có hạng là k 42
Hình 2.4 Biểu đồ 2-D của 16 thuật ngữ và 17 tài liệu từ tập mẫu. 44
Hình 2.5 Sơ đồ SVD của một ma trận hình chữ nhật thuật ngữ- tài liệu 45
Hình 2.6 Sơ đồ của SVD được giảm lược của một ma trận thuật ngữ-tài liệu ..46
Hình 2.7 Đồ thị Recall – Precision của thuật toán LSI 54
Hình 3.1 Kiến trúc mô hình tìm kiếm LSI 65
Hình 3.2 Giao diện cấu hình 66
Hình 3.3 Giao diện tìm kiếm 66
Hình 3.4 Giao diện kết quả tìm kiếm 67

MỞ ĐẦU

Việc tìm kiếm và lưu trữ thông tin từ xa xưa đã được con người chú trọng và quan tâm. Ngày nay, với sự phát triển nhanh chóng của lĩnh vực thông tin và Internet đã tạo ra một khối lượng thông tin vô cùng lớn với sự phong phú, đa dạng và phức tạp của các loại hình như: văn bản, hình ảnh, video, siêu văn bản, đa phương tiện…Vấn đề tìm kiếm thông tin đa phương tiện hiện vẫn được các chuyên gia nghiên cứu trong việc truy tìm thông tin phù hợp với yêu cầu của người sử dụng.

Văn bản là một trong số các dạng của dữ liệu đa phương tiện. Nó đã được quan tâm từ hàng nghìn năm trước trong việc tổ chức, sắp xếp và lưu trữ các loại hình tài liệu. Cho đến nay, tài liệu dưới dạng văn bản vẫn chiếm đa số trong mọi cơ quan, tổ chức, đặc biệt là trong thư viện. Đồng thời, văn bản còn được sử dụng để mô tả các dạng khác của dữ liệu đa phương tiện như video, audio, hình ảnh. Xuất phát từ nhu cầu thực tế sử dụng, số lượng tài liệu văn bản dạng số hóa hiện nay ngày càng lớn và được sử dụng rất phổ biến. Vì vậy việc lưu trữ, xử lý và truy tìm thủ công trước đây đã gặp rất nhiều khó khăn, không thể hoặc khó có thể thực hiện và tìm kiếm được, hoặc có thể tìm kiếm được nhưng hiệu quả không cao. Chính vì vậy, việc tìm kiếm văn bản theo nội dung có vai trò hết sức quan trọng.

Cùng với sự ra đời và phát triển của máy tính, các công cụ xử lý cũng ngày càng hoàn thiện dựa trên những kỹ thuật hiện đại để phục vụ cho nhu cầu đó. Các mô hình truy tìm thường được sử dụng trong phạm vi này, đó là: Đối sánh chính xác, không gian vector, xác suất và trên cơ sở cụm. Song, nhược điểm cơ bản của các mô hình truy tìm thông tin hiện nay là những từ mà người tìm kiếm sử dụng, thường không giống với những từ đã được đánh chỉ mục trong thông tin tìm kiếm. Vấn đề này liên quan nhiều đến hai khía cạnh thực tế: Thứ nhất là tính đồng nghĩa (*synonymy*)- cùng một thông tin