LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành nhất đến ***PGS.TS Đặng Văn Đức,*** người đã tận tình hướng dẫn, giúp đỡ em trong suốt thời gian thực hiện luận văn này.

Con cảm ơn Cha, Mẹ và gia đình, những người đã dạy dỗ, khuyến khích, động viên con trong những lúc khó khăn, tạo mọi điều kiện cho chúng con nghiên cứu học tập.

Em cảm ơn các thầy, cô trong Viện Công Nghệ Thông Tin Hà Nội cùng các thầy cô trong Khoa Công nghệ thông tin – ĐH Thái Nguyên đã dìu dắt, giảng dạy em, giúp em có những kiến thức quý báu trong những năm học qua.

Cảm ơn các bạn đã tận tình động viên đóng góp ý kiến cho luận văn của tôi.

Mặc dù đã cố gắng hết sức cùng với sự tận tâm của thầy giáo hướng dẫn song do trình độ còn hạn chế, nội dung đề tài còn mới mẻ nên Luận văn khó tránh khỏi những thiếu sót. Em rất mong nhận được sự thông cảm và góp ý của thầy cô và các bạn.

Thái Nguyên, tháng 11/2008

Học viên

Phạm Thị Ngọc

MỤC LỤC

DANH MỤC CÁC TỪ TIẾNG ANH VÀ VIẾT TẮT

Từ gốc	Giải nghĩa
Cluster-based	Cơ sở cụm
CSDL	Cơ sở dữ liệu
DBMS (Database Management System)	Hệ quản trị cơ sở dữ liệu
MDBMS (Multimedia Database Management System)	Hệ quản trị cơ sở dữ liệu đa phương tiện
Doc	Tài liệu
Docs	Nhiều tài liệu
DSS (Decision Support Systems)	Hệ hỗ trợ ra quyết định
Exact match	Đối sánh chính xác
IMS (Information Management System)	Hệ quản lý thông tin
Index	Chỉ mục
IR (Information Retrieval)	Truy tìm thông tin
IRS (Information Retrieval System)	Hệ truy tìm thông tin
LSI (Latent Semantic Indexing)	Chỉ mục ngữ nghĩa tiềm ẩn
MultiMedia	Truyền thông da phương tiện
Precision	Độ chính xác
QAS (Question Anser System)	Hệ trả lời câu hỏi
Query	Truy vấn
Term	Thuật ngữ (từ)
Ranking	Sắp xếp
Record	Bản ghi
Recall	Khả năng tìm thấy
SC (Similarity Coeficient)	Độ tương quan
SVD (Singular Value Decomposition)	Kỹ thuật tách giá trị đơn
Text-partern	Mẫu văn bản
The Term Discrimination Value	Giá trị phân biệt từ
The Signal – Noise Ratio	Độ nhiễu tín hiệu

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH, ĐỒ THỊ

MỞ ĐẦU

Cùng với sự phát triển nhanh chóng của công nghệ tin học thì khối lượng dữ liệu đa phương tiện (Multimedia) được thu thập và lưu trữ dưới dạng số ngày càng nhiều dẫn tới việc tìm kiếm dữ liệu đa phương tiện trở nên khó khăn vì vậy cần có các hệ thống tìm kiếm thông tin (Information Retrieval) hỗ trợ người dùng tìm kiếm một cách chính xác và nhanh chóng các thông tin mà họ cần trên kho tư liệu khổng lồ này.

Hiện nay có một số hệ thống tìm kiếm như *GoogleDesktop, DTSearch, Lucene,* tuy nhiên các hệ thống này sử dung các kỹ thuật tìm kiếm đơn giản nên hiệu quả còn chưa cao. Vì vậy mục tiêu của luận văn này nhằm tìm hiểu một số kỹ thuật nâng cao tìm kiếm thông tin, cụ thể ở đây là tìm kiếm văn bản theo nội dung trong cơ sở dữ liệu đa phương tiện nhằm đáp ứng nhu cầu cấp thiết của thời đại bùng nổ thông tin điện tử hiện nay.
Bố cục của luận văn gồm các phần sau:

+ CHƯƠNG 1: TỔNG QUAN VỀ HỆ QUẢN TRỊ CSDL ĐA PHƯƠNG TIỆN: Phần này sẽ giới thiệu tổng quan về hệ quản trị CSDL đa phương tiện.

+ CHƯƠNG 2: MỘT SỐ KỸ THUẬT CHỈ MỤC VÀ TÌM KIẾM VĂN BẢN

- Trình bày các v ấn đềvề hệ tìm kiếm thông tin.

- Trình bày kỹ thuật cơ sở chỉ mục văn bản trên cơ sở mô hình Bool và mô hình vector.

+ CHƯƠNG 3: MỘT SỐ KỸ THUẬT NÂNG CAO HIỆU NĂNG TÌM KIẾM VĂN

- Trình bày cơ sở lý thuyết về một số kỹ thuật chỉ mục nâng cao.

- Giới thiệu kỹ thuật chỉ mục nâng cao LSI.

+ CHƯƠNG 4: PHÁT TRIỂN CHƯƠNG TRÌNH THỬ NGHIỆM: Chương này phát triển chương trình thử nghiệm áp dụng kỹ thuật chỉ mục và kỹ thuật tìm kiếm văn bản theo nội dung trong cơ sở dữ liệu đa phương tiện.

+ KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN: Trình bày các kết quả đạt được trong luận văn và nêu phương hướng phát triển của đề tài trong tương lai.
+ TÀI LIỆU THAM KHẢO và PHỤ LỤC: Trình bày các thông tin liên quan đến luận văn.

CHƯƠNG 1: TỔNG QUAN HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU ĐA PHƯƠNG TIỆN (MDBMS)

Trung tâm của một hệ thống thông tin đa phương tiện (MULTIMEDIA) chính là hệ quản trị CSDL MULTIMEDIA (MDBMS - Multimedia Database Management System). Theo truyền thống, một CSDL bao gồm một bộ các dữ liệu có liên quan về một thực thể cho trước hoặc một hệ quản trị CSDL (DBMS) là một bộ các dữ liệu có liên quan đến nhau với một tập hợp các chương trình được dùng để khai báo, tạo lập, lưu trữ, truy cập và truy vấn CSDL. Tương tự như vậy, chúng ta có thể xem một CSDL MULTIMEDIA là một tập các loại dữ liệu Multimedia như văn bản, hình ảnh, video, âm thanh, các đối tượng đồ hoạ…. Một hệ quản trị CSDL MULTIMEDIA cung cấp hỗ trợ cho các loại dữ liệu MULTIMEDIA trong việc tạo lập, lưu trữ, truy cập, truy vấn và kiểm soát.

Sự khác nhau của các kiểu dữ liệu trong CSDL MULTIMEDIA có thể đòi hỏi các phương thức đặc biệt để tối ưu hoá việc lưu trữ, truy cập, chỉ số hoá và khai thác. MDBMS cần phải cung cấp các yêu cầu đặc biệt này bằng cách cung cấp các cơ chế tóm tắt bậc cao để quản lý các kiểu dữ liệu khác nhau cũng như các giao diện thích hợp để thể hiện chúng.

1.1 Mục đích của MDBMS

Một MDBMS cung cấp một môi trường thích hợp để sử dụng và quản lý các thông tin CSDL MULTIMEDIA. Vì vậy, nó phải hỗ trợ các kiểu dữ liệu MULTIMEDIA khác nhau bên cạnh việc phải cung cấp đầy đủ các chức năng của một DBMS truyền thống như khai báo và tạo lập CSDL, khai thác dữ liệu, truy cập và tổ chức dữ liệu, độc lập dữ liệu, tính riêng, toàn vẹn dữ liệu, kiểm soát phiên bản. Các chức năng của MDBMS cơ bản tương tự như các chức năng của DBMS, tuy nhiên, bản chất của thông tin MULTIMEDIA tạo ra các đòi hỏi mới. Bằng cách sử dụng các chức năng tổng quát của DBMS chúng ta có thể trình bày mục đích của MDBMS như sau:

- **Sự thống nhất**: bảo đảm rằng một dữ liệu không phải tạo lại khi các chương trình khác nhau đòi hỏi dữ liệu đó.

- **Độc lập dữ liệu:** Đảm bảo sự tách rời giữa CSDL và các chức năng quản trị từ các chương trình ứng dụng.

- **Điều khiển nhất quán:** đảm bảo sự toàn vẹn của CSDL MULTIMEDIA thông qua các quy tắc được áp dụng trên các giao dịch đồng thời.

- **Sự tồn tại:** bảo đảm các đối tượng dữ liệu tồn tại qua các giao dịch khác nhau cũng như các yêu cầu của chương trình.

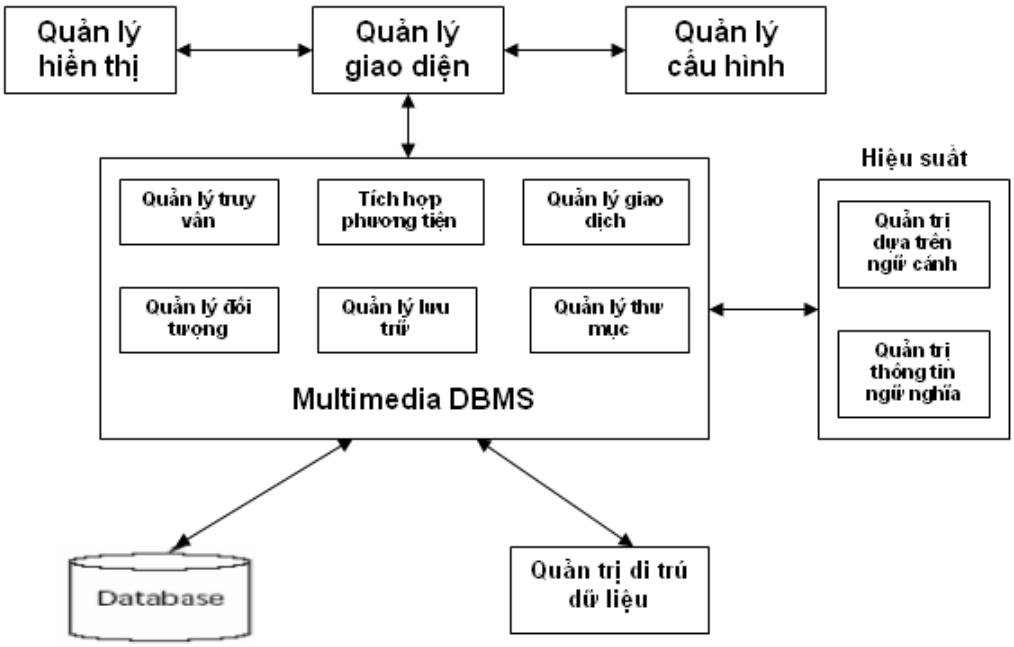
- **Tính riêng:** ngăn chặn các truy cập và sửa chữa các dữ liệu được lưu trữ một cách trái phép.

- **Kiểm soát sự toàn vẹn:** bảo đảm sự toàn vẹn của CSDL từ một giao dịch này sang một giao dịch khác thông qua việc áp đặt các ràng buộc.

- **Khả năng phục hồi:** phải có các phương thức cần thiết để đảm bảo rằng kết quả của các giao dịch thất bại không làm ảnh hưởng đến dữ liệu lưu trữ.

- **Hỗ trợ truy vấn**: bảo đảm các cơ chế truy vấn phù hợp với dữ liệu MULTIMEDIA.

- **Kiểm soát phiên bản:** tổ chức và quản lý các phiên bản khác nhau của các đối tượng lưu trữ có thể được yêu cầu bởi các ứng dụng.



Hình1.1. Kiến trúc bậc cao cho một MDBMS đáp ứng các yêu cầu cho dữ liệu MULTIMEDI

Đối với việc điều khiển nhất quán, một giao dịch là một chuỗi các hướng dẫn được thực thi một cách hoàn toàn hoặc không hoàn toàn, đối với trường hợp không hoàn toàn CSDL sẽ được khôi phục lại trạng thái trước đó, việc đưa ra được một cơ chế tương ứng đảm bảo cho việc nhất quán là một vấn đề khó khăn đối với CSDL MULTIMEDIA. Các CSDL quan hệ truyền thống sử dụng một bản ghi hoặc một bảng duy nhất như là một đơn vị nhất quán. CSDL MULTIMEDIA thường sử dụng một đối tượng đơn lẻ (hoặc đối tượng ghép) như là một đơn vị logic của truy cập. Như vậy một đối tượng MULTIMEDIA đơn lẻ có thể tạo thành đơn vị nhất quán.

Đối với vấn đề lưu trữ, một phương thức đơn giản là lưu trữ các tệp MULTIMEDIA trong các tệp tương ứng của hệ điều hành. Tuy nhiên với đặc thù là dung lượng lớn, các dữ liệu MULTIMEDIA là cho chi phí triển khai theo cách thức này trở nên tốn kém. Hơn nữa, hệ thống cũng cần phải lưu trữ các metadata MULTIMEDIA và có thể cả các đối tượng MULTIMEDIA tổng hợp. Vì vậy, hầu hết các MDBMS phân loại thành 2 phần là cố định và tạm thời và chỉ lưu trữ các dữ liệu cố định sau khi các giao dịch được cập nhật. Các dữ liệu tạm thời