LỜI CẢM ƠN

Tôi xin bày tỏ lòng kính trọng và biết ơn sâu sắc tới PGS.TS ***Đặng Văn Đức***, người đã trực tiếp hướng dẫn, giúp đỡ, động viên tôi 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 con nghiên cứu học tập.

Tôi cũng xin chân thành cảm ơn các thầy cô trong Viện Công nghệ Thông tin, các thầy cô trong khoa Công Nghệ Thông Tin và các bạn bè, đồng nghiệp tại trường Dự bị Đại Học Dân tộc Trung Ương đã giúp đỡ tôi rất nhiều trong quá trình học tập, sưu tầm, tìm tòi tài liệu và trong công tác để tôi có thể hoàn thành bản luận văn này.

Dù đã cố gắng hết sức cùng với sự tận tâm của thầy giáo hướng dẫn song do trình độ còn hạn chế nên khó tránh khỏi những thiếu sót. Rất mong nhận được sự thông cảm và góp ý của thầy cô và các bạn.

Thái Nguyên, tháng 11 năm 2008

Học viên

Lưu Thị Hải Yến

MỤC LỤC

MỤC LỤC CÁC HÌNH VẼ

DANH MỤC CÁC TỪ TIẾNG ANH VÀ VIẾT TẮT

Từ gốc	Nghĩa
IR (Information Retrieval) MIRS (MultiMedia Information Retrieval System)	Truy tìm thông tin Hệ truy tìm thông tin đa phương tiện
MM (MultiMedia)	Truyền thông đa phương tiện
Exact match	Đối sánh chính xác
Cluster-based	Cơ sở cụm
DBMS (DatabaseManagementSystem)	Hệ quản trị cơ sở dữ liệu
Term	Từ
Doc	Tài liệu
Docs	Nhiều tài liệu
Query	Truy vấn
DSS (DecisionSupportSystems)	Hệ hỗ trợ ra quyết định
IMS (InfomationManagementSystem)	Hệ quản lý thông tin
QAS (QuestionAnserSystem)	Hệ trả lời câu hỏi
Text-partern	Mẫu văn bản
Ranking	Xếp loại
SC (Similarity Coeficient)	Độ tương quan
Index	Chỉ mục
Precision	Độ chính xác
Recall	Khả năng tìm thấy

LỜI NÓI ĐẦU

Trong những năm gần đây, sự phát triển mạnh mẽ của CNTT và ngành công nghiệp phần cứng đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Bên cạnh đó việc tin học hoá một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Với một lượng thông tin như vậy thì vấn đề đặt ra là phải làm sao sử dụng chúng vào đúng mục đích và hiệu quả nhất thì cũng là một vấn đề đặt ra hiện nay. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, cần phải có các công cụ hỗ trợ để giúp cho việc tìm kiếm thông tin được nhanh và hiệu quả. Vì vậy mục tiêu của luận văn này nhằm tìm hiểu và xây dựng một hệ thống tìm kiếm thông tin cụ thể là tìm kiếm tài liệu văn bản trên cơ sở phân cụm dữ liệu. Nhằm đáp ứng nhu cầu cấp thiết của thời đại.

Bố cục của luận văn gồm các phần sau:

+ CHƯƠNG 1 - TỔNG QUAN: Giới thiệu chung về hệ thống thông tin đa phương tiện.

+ CHƯƠNG 2 - HỆ TÌM KIẾM THÔNG TIN: Giới thiệu về hệ thống tìm kiếm thông tin (IR), sự khác nhau giữa hệ thống tìm kiếm thông tin và các hệ thống thông tin khác, các mô hình thường gặp trong hệ thống tìm kiếm thông tin.

+ CHƯƠNG 3 - KỸ THUẬT PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG: Khái quát chung về phân cụm, các kiểu dữ liệu trong phân cụm và ứng dụng kỹ thuật phân cụm dữ liệu trong tìm kiếm thông tin.

+ CHƯƠNG 4 - CHƯƠNG TRÌNH DEMO: Cài đặt một chương trình tìm kiếm thông tin trên cơ sở lý thuyết đã trình bày.

+ KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN: Trình bày các kết quả đạt được

và nêu phương hướng phát triển của đề án trong tương lai.

+ TÀI LIỆU THAM KHẢO

CHƯƠNG 1: TỔNG QUAN

1.1. ĐẶT VẤN ĐỀ

Vài năm trước đây, các nghiên cứu và phát triển thuộc lĩnh vực đa phương tiện (MultiMedia) tập trung vào các vấn đề như: truyền thông, authoring và trình diễn đa phương tiện.

Trải qua nhiều năm đã có khối lượng lớn dữ liệu Multimedia (ảnh, video, âm thanh) được thu thập và lưu trữ dưới dạng số, thí dụ:

- Ảnh X quang,
- Các băng hình dạy học…
- Điều tra cảnh sát về các giọng nói trong điện thoại…
- Tài liệu văn bản, …

Nghiên cứu của những năm gần đây tập trung chủ yếu vào: lưu trữ và tìm kiếm hiệu quả dữ liệu đa phương tiện. Tình hình tương tự như hơn 30 năm trước đây khi nhiều dữ liệu text được lưu trữ dưới khuôn dạng máy tính có thể đọc được. Từ đó dẫn tới việc phát triển các hệ thống quản trị cơ sở dữ liệu (DatabaseManagmentSystem) mà ngày nay được sử dụng trong hầu hết các cơ quan, tổ chức. Tuy nhiên hệ quản trị cơ sở dữ liệu không thể quản lý dữ liệu đa phương tiện một cách hiệu quả bởi vì các tính chất dữ liệu văn bản và dữ liệu đa phương tiện là khác nhau. Do vậy, dẫn tới việc nghiên cứu phát triển các kỹ thuật truy tìm và chỉ mục mới trong hệ thống quản trị cơ sơ dữ liệu và việc phát triển hệ thống truy tìm tài liệu văn bản – một phần của dữ liệu đa phương tiện cũng không nằm ngoài xu thế đó.

Luận văn tập trung nghiên cứu cách tìm kiếm văn bản trên cơ sở phân cụm dữ liệu. Mục tiêu chính của phương pháp phân cụm dữ liệu là nhóm các đối tượng tương

tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một lớp là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng.

1.2. HỆ THỐNG THÔNG TIN ĐA PHƯƠNG TIỆN:

Đa phương tiện là gì? Đa phương tiện là tích hợp của văn bản, âm thanh, hình ảnh của tất cả các loại và phần mềm có điều khiển trong một môi trường thông tin số.

Dữ liệu đa phương tiện gồm dữ liệu về :

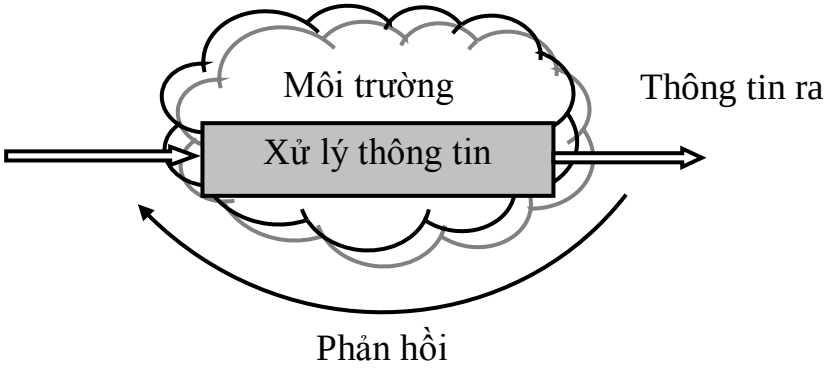
- Văn bản;
- Hình ảnh;
- Âm thanh;
- Hình động.

1.2.1. Khái niệm về đa phương tiện

Con người có nhu cầu diễn tả các trạng thái của mình; và họ có nhiều loại hình thể hiện. Con người có nhu cầu truyền thông, do đó cách thể hiện trên đường truyền rất quan trọng. Trên Internet thông dụng với mọi người, cái đẹp của trang Web phải được thể hiện cả ở nội dung và hình thức.

Đa phương tiện có nhiều loại, những phương tiện công cộng về đa phương tiện: Radio, vô tuyến, quảng cáo, phim, ảnh...

Nhu cầu về tương tác người - máy luôn đặt ra trong hệ thống thông tin. Vấn đề chính về tương tác người - máy không là quan hệ giữa con người với máy tính mà là con người với con người. Con người có vai trò quan trọng trong hệ thống thông tin.



Hình 1.1: Hệ thống thông tin

Định nghĩa

Định nghĩa đa phương tiện (theo nghĩa rộng) là bao gồm các phương tiện: văn bản, hình vẽ tĩnh (vẽ, chụp), hoạt hình (hình ảnh động), âm thanh.

Hay có thể định nghĩa đa phương tiện; *đa phương tiện là kỹ thuật mô phỏng và sử dụng đồng thời nhiều dạng phương tiện chuyển hoá thông tin và các tác phẩm từ các kỹ thuật đó.*

1.2.2. Media

Media (tiếng Latin: *medius,* tiếng Anh: *means, intermediary*) là đề cập đến các loại thông tin hay loại trình diễn thông tin như dữ liệu văn bản, ảnh, âm thanh và video.

Phân loại media: Có nhiều cách phân loại, nhưng cách chung nhất là phân loại trên cơ sở khuôn mẫu (format) vật lý hay các quan hệ media với thời gian. Qui định này dẫn tới hai lớp media: tĩnh (*static*) và động (*dynamic*).

- *Static media*: Không có chều thời gian, nôi dung và ý nghĩa của chúng không phụ thuộc vào thời gian trình diễn. Media tĩnh bao gồm dữ liệu văn bản, đồ họa.

- *Dynamic media*: Có chều thời gian, ý nghĩa và độ chính xác của chúng phụ thuộc vào tốc độ trình diễn. *Dynamic media* bao gồm *annimation, video, audio.* Media động phụ thuộc chặt chẽ vào tốc độ trình diễn. Thí dụ để cảm nhận chuyển động trơn tru, video phải được trình chiếu với tốc độ 25 frame/sec (hay 30 frame/sec phụ thuộc vào loại hệ thống video). Tương tự, khi ta trình diễn (*play*) tiếng nói, âm nhạc, chúng chỉ được cảm nhận tự nhiên khi đạt được tốc độ nhất định, nếu không chúng làm giảm chất lượng và ý nghĩa của âm thanh. Vì các media này phải được trình diễn liên tục và ở tốc độ cố định cho nên chúng còn được gọi là *media liên*

tục. Hay còn gọi chúng là *media đẳng thời (isochronous media)* vì quan hệ giữa các đơn vị media và thời gian là cố định.

1.2.3. Multimedia

Khái niệm *multimedia* (tiếng Latin: *multus*- tiếng Anh: *numerous*) đề cập đến tập hợp các kiểu media được sử dụng chung, trong đó ít nhất có một kiểu media không phải là văn bản (nói cách khác là ít nhất có một media trong đó là ảnh, audio hay video). Khái niệm *multimedia* hiểu theo nghĩa tính từ: thông tin đa phương tiện, dữ liệu đa phương tiện, hệ thống đa phương tiện, truyền thông đa phương tiện, ứng dụng đa phương tiện... Khái niệm dữ liệu đa phương tiện đề cập đến sự biểu diễn các kiểu media khác nhau mà máy tính có thể đọc được. Thông tin đa phương tiện đề cập đến thông tin được truyền đạt bởi các kiểu media. Đôi khi khái niệm dữ liệu đa phương tiện và thông tin đa phương tiện được sử dụng thay thế cho nhau.

1.2.4. CSDL và Hệ quản trị CSDL

Trong nhiều tài liệu thì hai khái niệm CSDL và hệ quản trị CSDL hay được sử dụng thay cho nhau. Ở đây ta sử dụng hai thuật ngữ này như sau:

- Cơ sở dữ liệu - *Database*: Tập hợp bản ghi data hay các mục media.
- Hệ quản trị cơ sở dữ liệu - DBMS: Toàn bộ hệ thống quản trị Database

1.2.5. Truy tìm thông tin tài liệu văn bản

Các hệ thống tự động truy tìm thông tin (*IR - Information Retrieval*) đã được phát triển để quản lý khối lượng lớn tài liệu khoa học từ những năm 40 của thế kỷ XX. Chức năng chính của hệ thống IR là lưu trữ và quản trị khối lượng văn bản lớn theo cách sao cho dễ dàng truy vấn (*query*) tài liệu mà người sử dụng quan tâm. Chú ý rằng đồng nghĩa với IR là *text IR* dù rằng ý nghĩa đầy đủ của khái niệm IR là đề cập đến truy tìm bất kỳ loại thông tin nào.