Multimedia Information Retrieval

Multimedia is everywhere

- # Recent advances in computer technology has precipitated a new era in the way people create and store data.
- # Millions of multimedia documents—including images, videos, audio, graphics, and texts—can now be digitized and stored on just a small collection of CD-ROMs.
- # Internet = a universally accessible multimedia library.
 - △Latest web estimates: 1 billion pages, 20 terabytes of information.

The Need of Digital library

- # The entertainment industry
 - New archives of film and photographs
- # Distance education
- # Telemedicine
 - Collections of medical images
- # Geographic information
- **#** Art gallery and museum
- # etc.

Document types

- # Monomedium
 - text, video, image, music, speech, graph,...
- # multimedia
 - combination of different media
- # hypertext
 - interlinked text document (eg XML, HTML)
- # hypermedia
 - interlinked multimedia documents

The Need of Multimedia Retrieval

Large amount of multimedia data may not be useful if there are no effective tools for easy and fast access to the collected data

Challenges

- △ Amount
- **Access**
- Authority
- Assortment



Multimedia Information Retrieval

Concerns with:

- □ Basic concepts and techniques in retrieving (unstructured) information
- Indexing and similarity-based retrieval of multimedia data
- # What is an information retrieval system?
 - A system used to process, store, search, retrieve and disseminate information items
 - Examples: DBMS, Free-text Systems, Hypermedia Systems etc.

Retrieval or Navigation

- **Retrieval:** Extracting a "document" (or "documents") in response to a query, e.g. keyword search or free text search, search engines on the web
- **Navigation: Moving from one part of the information space to another, typically by following links (hypertext, hypermedia)

Content or Metadata Based Retrieval

- # Metadata based retrieval: widely used for text and non-text media. But assigning metadata (eg key-terms) to non-text media is labour intensive and limiting
- **# Content based retrieval:** uses content of the "documents" for satisfying the query.
 - □ used (fairly!) reliably in text retrieval.
 - content based image and video retrieval is an active research topic. Some commercial products are emerging. It can be reliable in constrained situations.

Information Retrieval (IR)

Information Retrieval

- # Difficult since the data is unstructured
 - ☑ It differs from the DBMS structured record:

Name: <s></s>	Sex: <s></s>	Age: <s></s>	NRIC: <s></s>	
---------------	--------------	--------------	---------------	--

- Information must be analyzed, indexed (either automatically or manually) for retrieval purposes.
- # Examples:







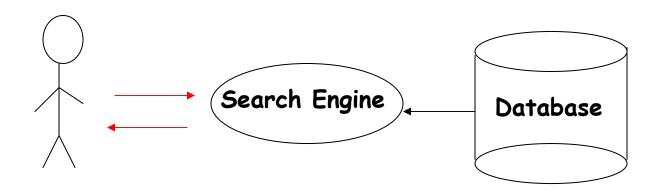
Retrieval Procedure

- # The purpose of an automatic retrieval strategy is to retrieve all the <u>relevant</u> documents whilst at the same time retrieving as few of the non-relevant ones as possible.
- # Simple retrieval procedure:

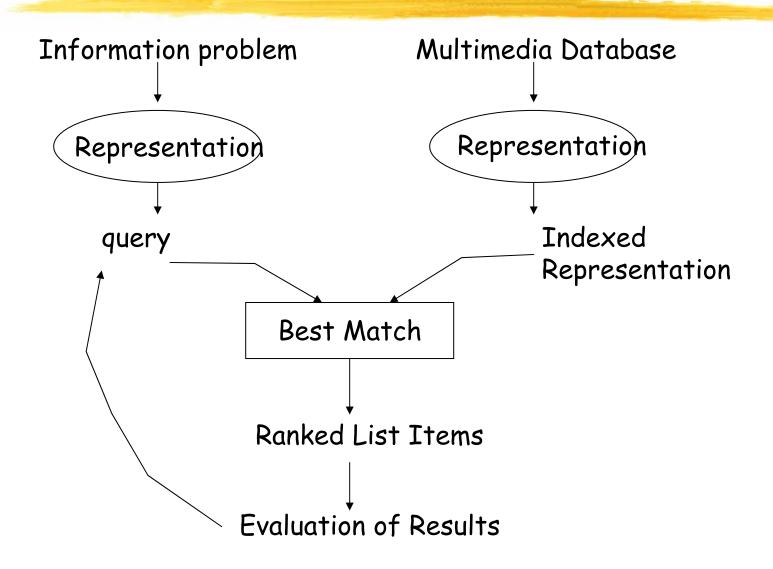
 - Step II: Similarity Evaluation and Ranking
 - Step III: Show the top k retrievals, e.g., k=10 or k=16

Retrieval Procedure Cont...

***Step IV**: User interaction interface, "relevance feedback".



System Overview



Three main ingredients to the IR process

- #1) Text or Documents, 2) Queries, 3) The process of Evaluation
- # For Text, the main problem is to obtain a representation of the text in a form which is amenable to automatic processing.
 - Representation concerns with creating text surrogate which consist of a set of:
 - · index terms
 - or keywords
 - or descriptors

Three main ingredients to the IR process *Cont...*

- # For Queries, the query has arisen as a result of an information need on the part of a user.
 - Query must be expressed in a language understood by the system.
- # Representing information need is very difficult, so the query in IR system is always regarded as approximate and imperfect.

Three main ingredients to the IR process *Cont...*

- #The evaluation process involves a comparison of the texts actually retrieved with those the user expected to retrieve.
- #This leads to some modification, typically of the query through possibly of the information need or even of the surrogates

Example

#Query, "Which films were nominated for the Oscars this year?"

MDR Demo	2
Main manu	
Search page Logout Question: Which films were nominated for oso Find it	
Search Results	
17 documents were found for query 'which films were nominated for oscars this year?'.	
Documents containing query words: film: 221; nomin: 162; oscar: 28; year: 3058;	
Showing results 1 to 5 (Page 1 of 4) ranked by Highest Rank First	
Go to page Previous 1 - Next (2), Sort by Highest rank first - Refresh	
Expand Query	
1. NPR's 'Weekend Edition (Sunday)', >>>>>>> Relevant? Date: 12/03/2000	
Duration: 17 mins 33 secs	
" in the weeks leading up to the awards ceremony and of course make it own choice welcome back in the early and you could talk about this year let's go back to last year's winning dramatic film score the columbia money supply italian composers won the best dramatic score true out of the lab for years but there are no italians nominated this year and any there's also a debate change in the film score category itself this year that's right these last four years we have been listening to the best dramatic score ignoring five other scores that were nominated in the category called best comedy or musical score"	
Keyword Occurences: film: 19 nomin: 5 oscar: 4 year: 13	
Listen to Extract Read Entire Automatic Transcript	

MDR Demo

Hear Original Audio Broadcast

NPR's Weekend Edition (Sunday)

Broadcast 19/03/2000, Story 15

To play a segment of text, select it using the mouse, and let go. You must select more than four words.

Last week we began our annual run up to oscar sunday by listening to some of the nominated film scores

One big difference this year is a change in the rules of the academy of motion picture arts and sciences has combined what had been two categories of film scores into one here to explain and share some of the nominated scores once again is n. p. r.'s andy trudeau high and daily and a little refresher course first please remind us what happened to the best dramatic score and the best comedy amir

The coast guard they're now called the best original score i suppose one way to look at this now is instead of ten instrumental scores five which are competing against each other

They've all been much into a single category of five now will we notice anything different

This year know i'm guessing in future years we will have some interesting juxtapositions of musical styles but this year if you look at the ones we've talked about on the ones we will be talking about the role dramatic scores to make course to some extent i've never understood some of these artificial characterizations when we first did this year the english patient won music like you really are rachel portman one for emma which was uh uh bronte a book which i don't remember being in the comedy section of the literature

Sure you can go to laugh a minute

Hello

But the real well racial portman is one of the five composers with a score nominated for an academy award this year it or score to the movie the cider house rules the other composers and films are thomas newman american beauty john williams angela's ashes john corley on on the red violin and give really are in the talented mr. ripley last week we talked about here in the yard and rachel portman and we heard selections from their *nominated* scores of this week

This week we're going to do john williams has gone almost fourteen nominations lifetime for the oscar and thomas newman was part of a hollywood dynasty we're gonna do the william's first

Williams nominated for angela's ashes this is a movie that was based on a best selling memoir by frank court the *film* starring emily watson and robert carlyle and was directed by alan parker a houston lawyer william



letscape: Espand Query					
le	Edit View	Go	Communicator		
N	<i>IDR</i>	D	emo		
	You may wish to add some of these words to your search. This should improve the accuracy of the search.				
	acade	emi		□acut	angela
	□ash			award	■ beauti
	□ begir	ı		□cider	□comedi
	□direc	t.		director	hollywood
	□ lullat	oi		□monei	movi
	newn	nan		□ plai	□portman
	□rache	el		□riplei	□school
	□sort			soundtrack	□space
	□star			□ trudeau	\square violin
	□want			□win	□wonderfulli
				Search again Clear selection	on

Interface designed by Rich Wareham & Ben Timms Summer 1999

Measures of Effectiveness

The most commonly used measure of retrieval effectiveness are recall and precision

□Recall,

$$R = \frac{No.of\ relevant documents retrived}{No.of\ relevant documents in\ the databas \epsilon}$$

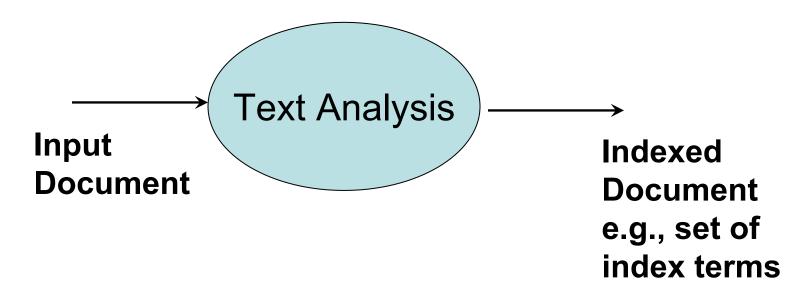
Precision,

$$P = \frac{\text{No.of relevant} documents etrived}{\text{No.of documents etrived}}$$

Measures of Effectiveness Cont...

- # Recall and Precision are based on the assumption that the set of relevant documents for a query is the same, no matter who the user is.
- # Different users might have a different interpretation as to which document is relevant and which is not.
- # Thus, the relevance judgment is usually based on two criterion:
 - □ Ground Truth
 - User subjectivity

Representation of Documents

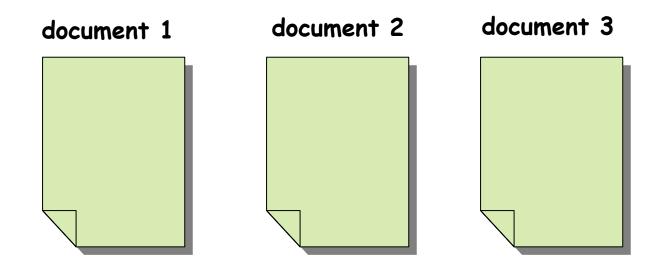


- ****** Text Analysis Methods:
 - Single Document Processing
 - A collection of Documents

Document Modeling by "terms"

#Set of terms:

information retrieval figure example



(1) Single Document Processing

- # Taking a large text document and reducing it to a set of "terms".
- # We need to be able to extract from the document those words or terms that best capture the meaning of the document.
- # In order to determine the importance of a term we will need a measure of term frequency (TF)--- the no. of times a given term occurs in a given document.

A document can be represented by a set of terms and their weights which is called a term vector that can be stored as metadata:

$$D = (T_1, W_1; T_2, W_2; ..., T_n, W_n)$$

$$lpha$$
 Where $w_j = tf_j$

 $w_{\rm j}$ indicates the importance of term j in the document,

tf, gives the no. of occurrences of term j in the document.

Algorithm

- Split the text into manageable chunks.
- II. Remove the stop words. These are very frequently occurring words that have no specific meaning, (e.g., "the", "and", "but", or "large", "small").
- III. Count the number of times the remaining words occur in the chunk.

Example

SAMPLE SEQUENCE 1

More and more application areas such as medicine, maintain large collections of digital images. Efficient mechanisms to efficiently browse and navigate are needed instead of searching and viewing directory trees of image files.

REMOVE STOP WORDS

Application areas medicine collections digital images. mechanisms browse navigate searching viewing directory trees image files.

TERMS

Application (1); area (1); collection (1); image (2); mechanism (1); browse (1); navigate (1)

(2) Processing a Collection of Documents

- # The second technique works on collections of documents.
- # Each document is associated with a term vector as follows:

	Document1	Document 2	Document 3	Document 4
Term 1	1	2	0	0
Term 2	0	2	3	1
Term 3	1	1	2	2
Term		0	1	0
Term				
Term t	1	0	3	1

Term Vector Database

Doc1
$$\vec{d}_1 = [w_{11}, w_{12}, w_{13}, ..., w_{1t}]$$

Doc2 $\vec{d}_2 = [w_{21}, w_{22}, w_{23}, ..., w_{2t}]$
Doc3 $\vec{d}_3 = [w_{31}, w_{32}, w_{33}, ..., w_{3t}]$
Doc4 $\vec{d}_4 = [w_{41}, w_{42}, w_{43}, ..., w_{4t}]$
Doc5 $\vec{d}_5 = [w_{51}, w_{52}, w_{53}, ..., w_{5t}]$
Doc6 $\vec{d}_6 = [w_{61}, w_{62}, w_{63}, ..., w_{6t}]$
...

...

Doc N $\vec{d}_N = [w_{N1}, w_{N2}, w_{N3}, ..., w_{Nt}]$

TFXIDF Model

** A better model for term vector is given by combining term frequency with document frequency:

$$w_{ij} = tf_{ij} \times log(N/df)$$

Where

 W_{ij} indicates the importance of term j in document i tf_{ij} gives the no. of occurrences of terms j in document i

 df_j gives the no. of documents in which term j occurs N gives the no. of document in the collection.

Query Processing

- ** With the Vector Space Model, retrieval can be based on a query-by-example paradigm.
 - ☐ The user can present a text document and present the query as "find document like this one".
- # Relevance ranking: documents are ranked by ascending order of relevance.
- # Then, we can use a cut-off point to measure recall and precision, e.g., the first twenty returned.

Query	Database	Scores	Sorted Scores
$ec{q} = [w_{q_1}, w_{q_2}, w_{q_3},, w_{q_t}]$	$\vec{d}_{1} = [w_{11}, w_{12}, w_{13},, w_{1t}]$ $\vec{d}_{2} = [w_{21}, w_{22}, w_{23},, w_{2t}]$ $\vec{d}_{3} = [w_{31}, w_{32}, w_{33},, w_{3t}]$ $\vec{d}_{4} = [w_{41}, w_{42}, w_{43},, w_{4t}]$ $\vec{d}_{5} = [w_{51}, w_{52}, w_{53},, w_{5t}]$ $\vec{d}_{6} = [w_{61}, w_{62}, w_{63},, w_{6t}]$ $\vec{d}_{N} = [w_{N1}, w_{N2}, w_{N3},, w_{Nt}]$	$egin{array}{c} S_1 \ S_2 \ S_3 \ S_4 \ S_5 \ \end{array}$	$egin{array}{c} S_8 \ S_{3^0} \ S_3 \ S_9 \ S_1 \ S_7 \ \dots \ \dots \ S_5 \ \end{array}$

Similarity Measurement

- # In a ranking process, query's vector is compared for similarity or dissimilarity to vectors corresponding to documents in a given database.
- # Similarity is computed based on methods such as Cosine measure:

Similaritý
$$D_q, D_j$$
) = $S(\vec{q}, \vec{d}_j) = \frac{\vec{q} \cdot \vec{d}_j}{|\vec{q}| \times |\vec{d}_j|} = \frac{\sum_{k=1}^t w_{qk} w_{jk}}{\sqrt{\sum_{k=1}^t w_{qk}^2 \sum_{k=1}^t w_{jk}^2}}$

Where

 $ec{q}$ is the term vector of a given query

 $ec{d}_{j}$ is the term vector of the j-th document in the database.

User Interaction In IR

- #User interaction method in IR is used to improve retrieval effectiveness, through query expansion process.
- #In practice, most users find it difficult to formulate queries which are well designed for retrieval purposes.
- #In IR, query is started by a tentative query and repeated by relevance feedback.

Query Formulation Process

In a relevance feedback cycle, the user is presented with a list of the retrieved documents and, after examining them, marks those which are relevant.

Retrieved Documents

- = Relevance and Non-relevance Items
- # Which are then used to reweight the query's terms: Query Formulation

Query Formulation Process

Definitions:

 D_r set of relevant documents defined by the user, among the retrieved documents;

 D_n set of non-relevant documents;

 α, β, γ constants;

The modified query is calculate as:

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Original Query: $\alpha \vec{q} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \end{bmatrix}$

$$\alpha \vec{q} = [1]$$

Relevant Terms: $\frac{\beta}{|D_r|} \sum_{\forall \vec{d}_i \in D_r} \vec{d}_j = \begin{bmatrix} 2 & 0 & 3 & 1 & 0 & 0 \end{bmatrix}$

Non-Relevant $\frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_i \in D_n} \vec{d}_j = \begin{bmatrix} 1 & 1 & 0 & 0 & 2 & 2 \end{bmatrix}$ Terms:

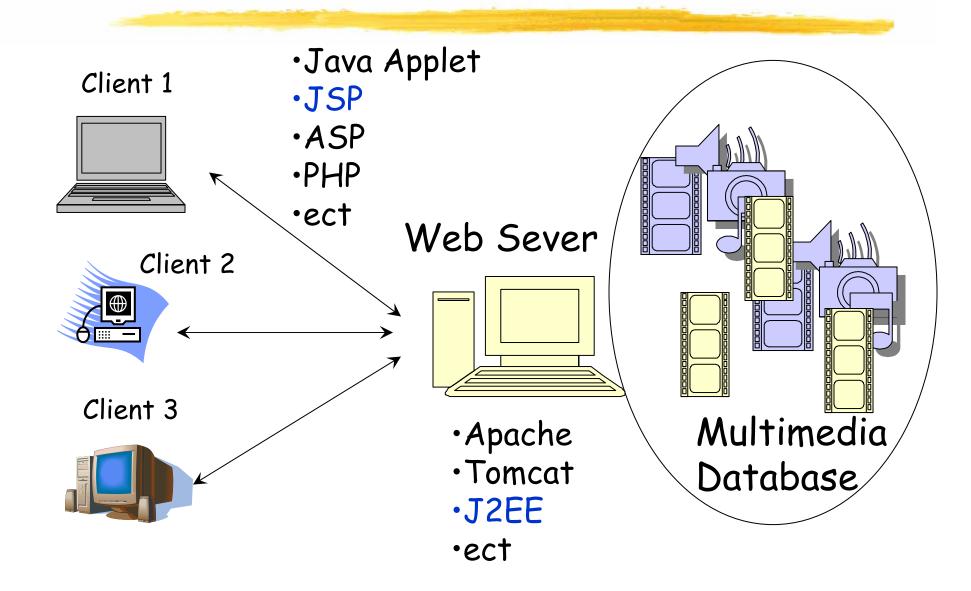
Modified Query: $\vec{q}_m = \begin{bmatrix} 2 & -1 & 4 & 1 & -2 & -2 \end{bmatrix}$

Summary

- # There is an argent need for automatic indexing and retrieval, following the explosion of multimedia data over Internet.
- # It is difficult to address semantic meaning in multimedia representation.
- # Thus, many search engines always have relevance feedback.
- # In text retrieval, Term Vector Model and relevance feedback are the basic techniques.

Setting Up J2EE Server and JDBC

Simple Client-Sever Architecture



Setting J2EE Server

Install Java JDK: j2sdk-1 3 1 01-win # Install J2EE: j2sdkee-1 3 01-win **# Configuration Your System** Set variable JAVA HOME=c:\jdk1.3.1 01 Set variable J2EE HOME=C:\j2sdkee1.3 △ Set PATH=%JAVA HOME%\BIN;%J2EE_HOME%\BIN Set CLASSPATH=.;%J2EE HOME%\lib\j2ee.jar; %J2EE HOME%\lib\sound.jar; %J2EE HOME%\lib\jmf.jar; %J2EE HOME%\LIB\SYSTEM\cloudscape.jar; %J2EE HOME%\LIB\SYSTEM\cloudutil.jar; %J2EE HOME%\LIB\cloudscape\RmiJdbc.jar; %J2EE HOME%\LIB\cloudscape\cloudclient.jar; %J2EE HOME%\LIB\cloudscape\cloudview.jar;

%J2EE HOME%\LIB\cloudscape\jh.jar;

Running J2EE Server

- **Start the Server:>> j2ee -verbose**
- #Stop the Server:>> j2ee -stop
- # Deploy applications:>> deploytool
- #J2EE Server can be access at port: 8000, http://localhost:8000
- # Try to deploy your applications

Setting JDBC and Cloudscape Database

- # Copy the files "cloudview.jar" and "jh.jar" to directory C:\j2sdkee1.3\lib\cloudscape
- # Running Cloudscape:>> cloudscape -start
- # Stop Cloudscape:>> cloudscape -stop
- # Graphic User Interface:>> java COM.cloudscape.tools.cview
- # Try to import and export data into database