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>	Sex:<s>	Age:<s>	NRIC:<s>	…..

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

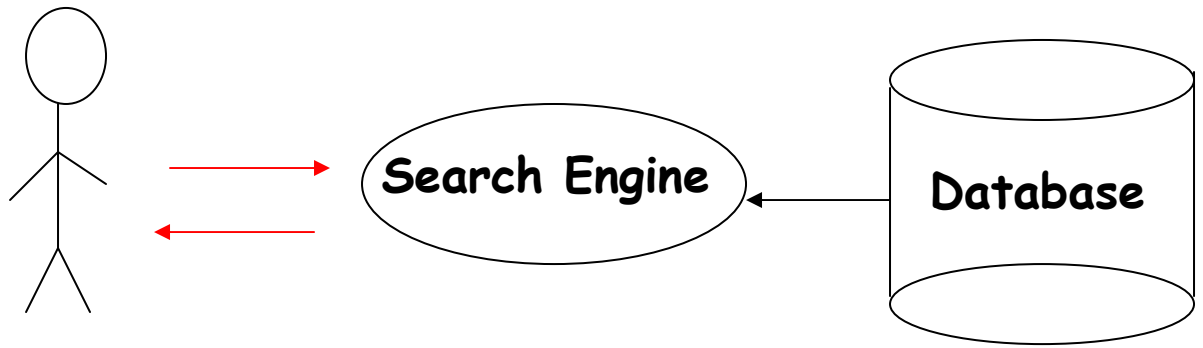
SEARCH LINC'S WEB SITE

Retrieval Procedure

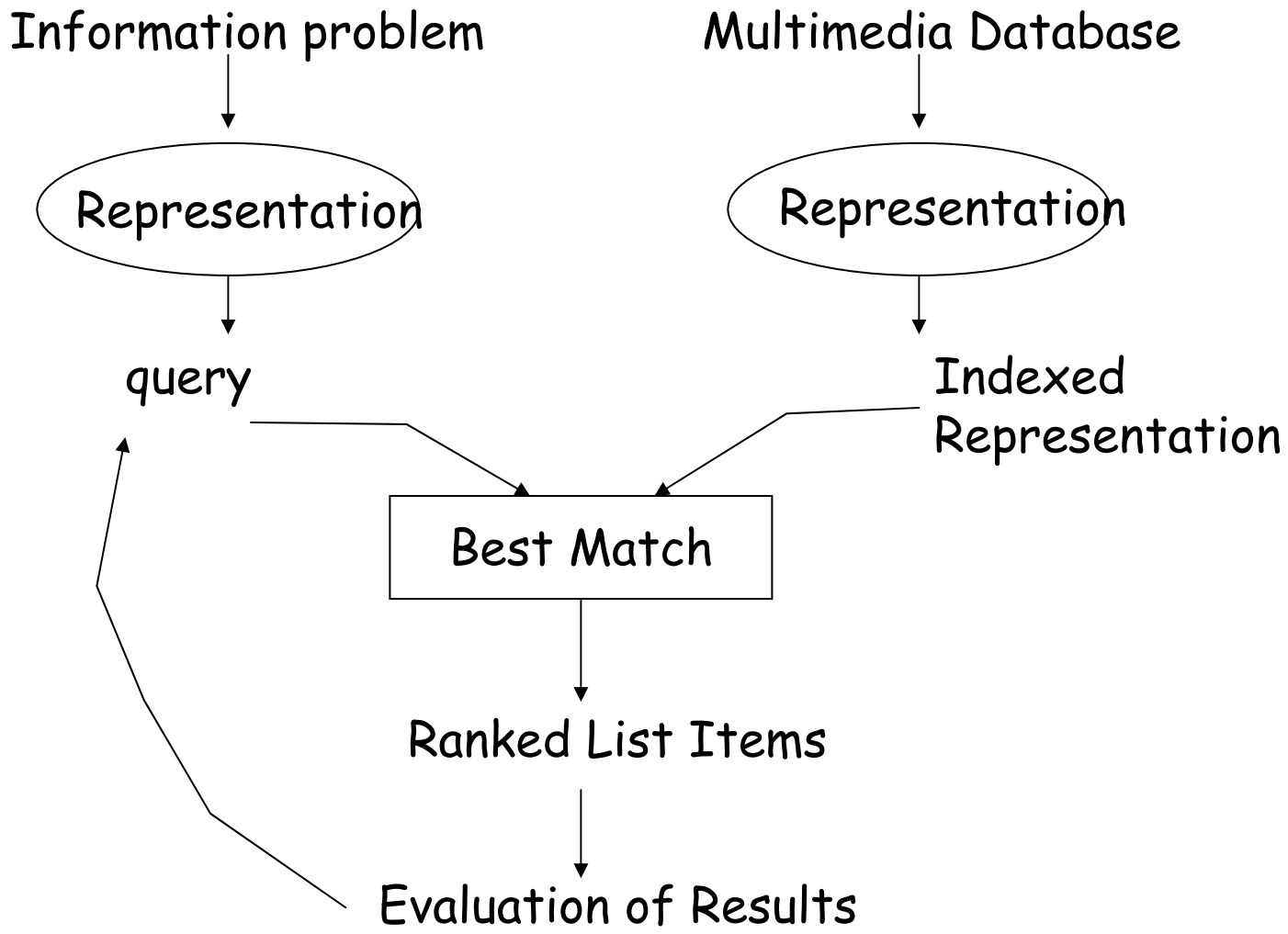
- The purpose of an automatic retrieval strategy is to retrieve all the relevant documents whilst at the same time retrieving as few of the non-relevant ones as possible.
- Simple retrieval procedure:
  - **Step I**: Query
  - **Step II**: Similarity Evaluation and Ranking
  - **Step III**: Show the 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?"

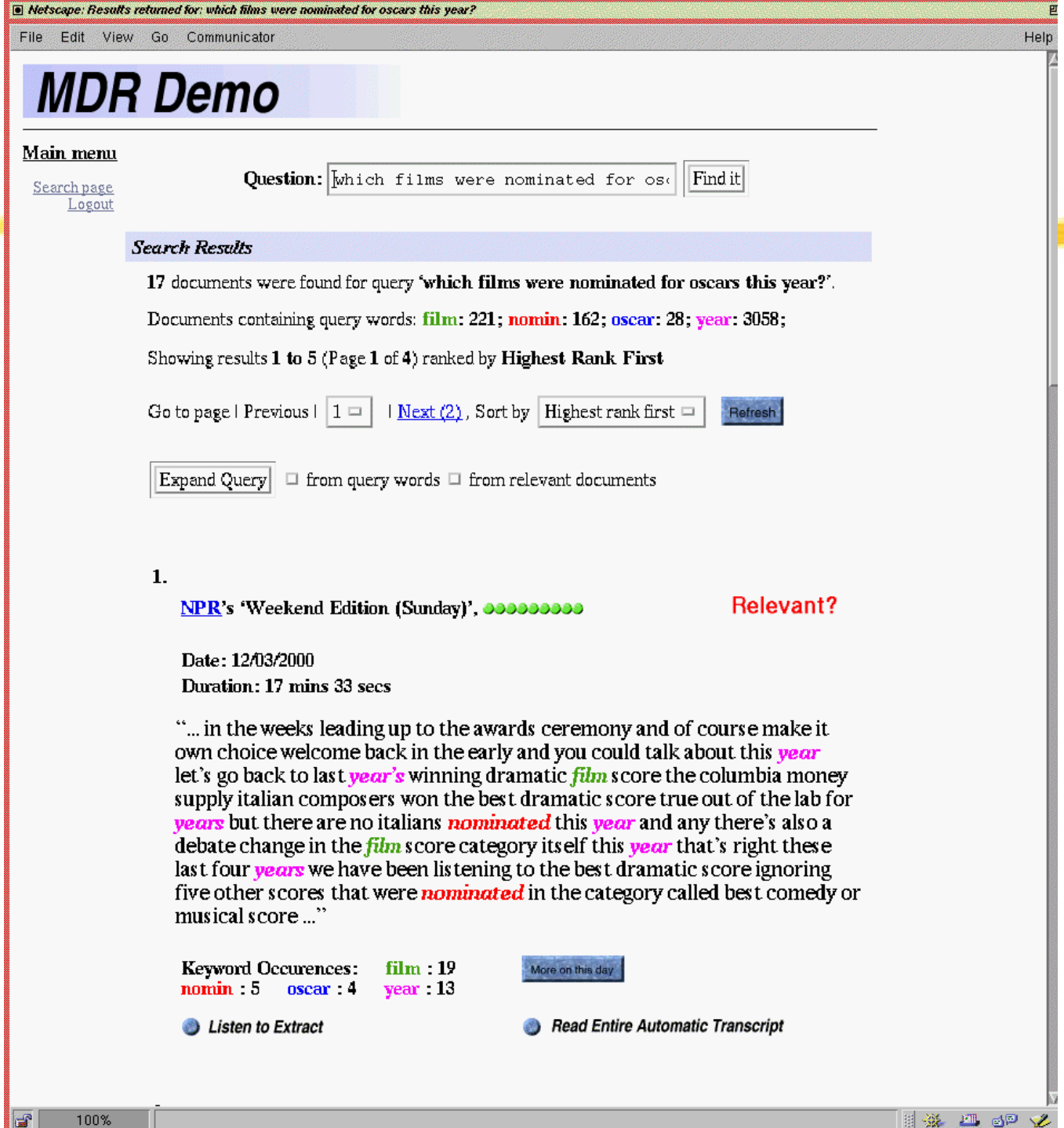


Figure1: Results Page

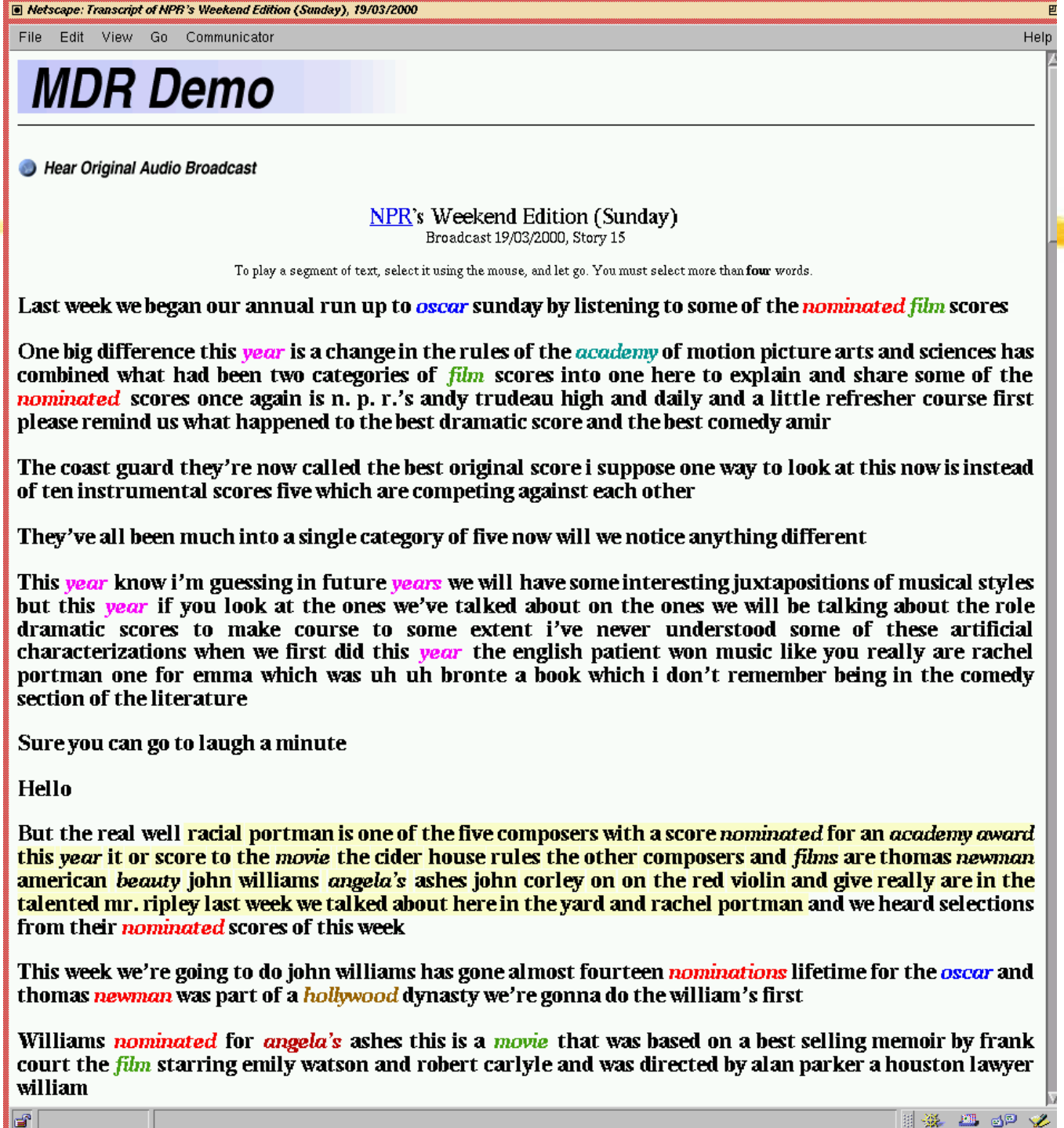


Figure2: Complete Transcription Page

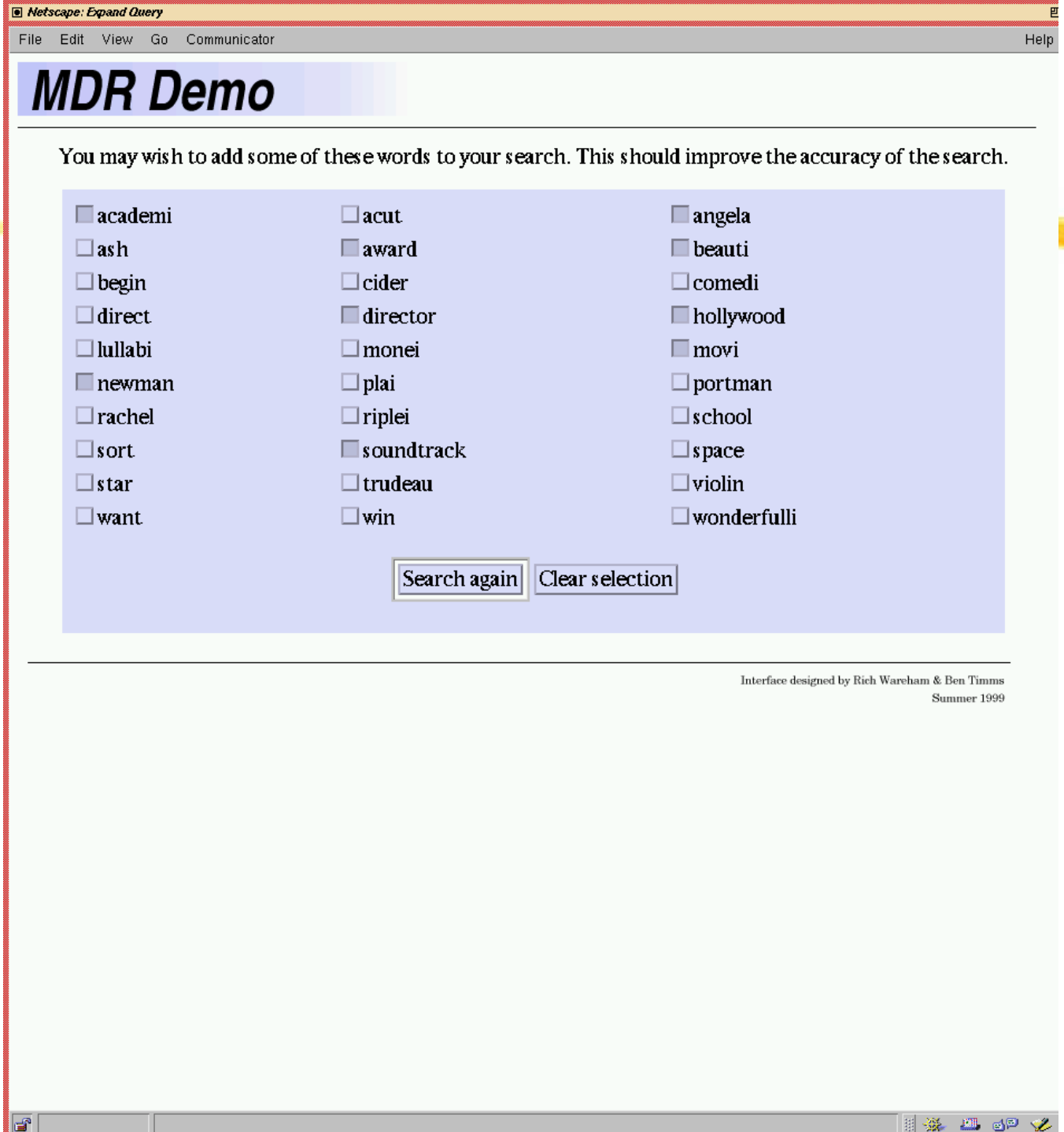


Figure3: Query Expansion

Measures of Effectiveness

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

$$R = \frac{\text{No.of relevant documents retrived}}{\text{No.of relevant documents in the database}}$$

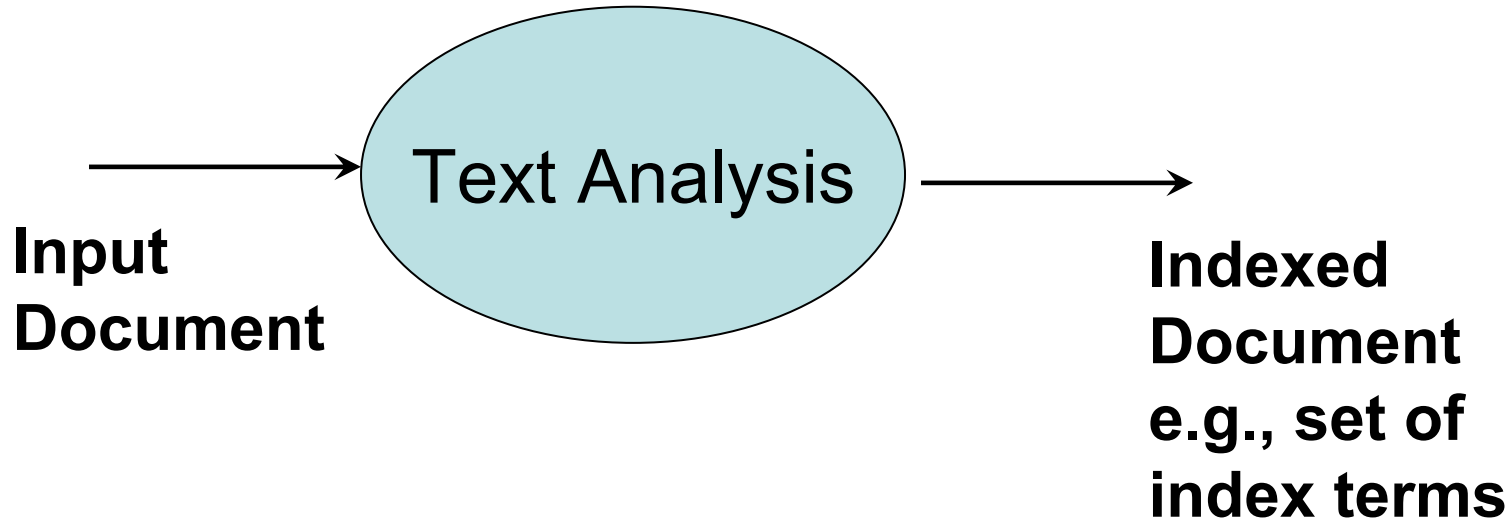
  - Precision,

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

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

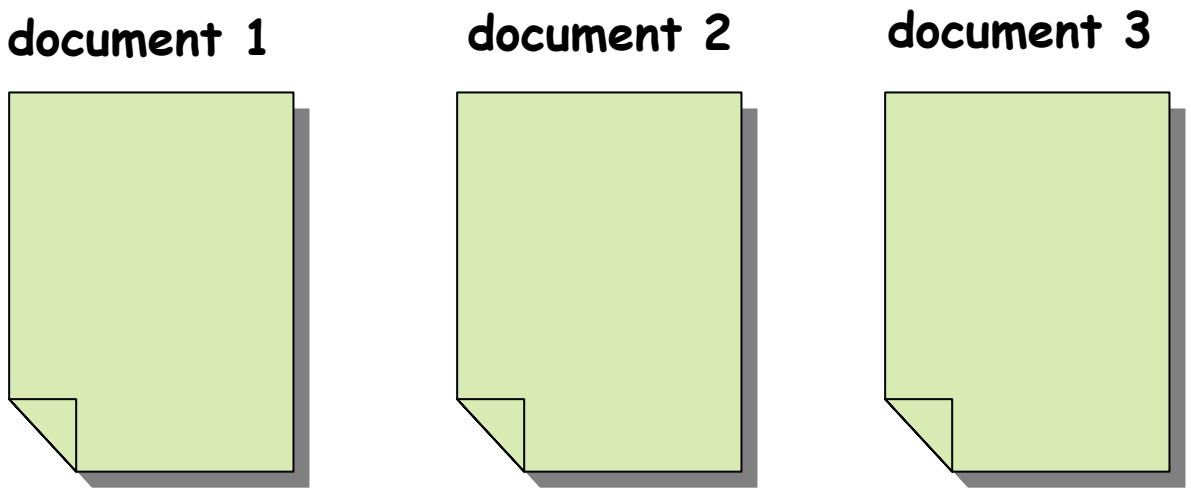


- Text Analysis Methods:
  - Single Document Processing
  - A collection of Documents

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:

$$D = (T_1, w_1; T_2, w_2; \ldots\ldots;; T_n, w_n)$$

- Where $w_j = tf_j$

w_j indicates the importance of term j in the document,

tf_j gives the no. of occurrences of term j in the document.

Algorithm

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

Doc2 $\vec{d}_2 = [w_{21}, w_{22}, w_{23}, \ldots, w_{2t}]$

Doc3 $\vec{d}_3 = [w_{31}, w_{32}, w_{33}, \ldots, w_{3t}]$

Doc4 $\vec{d}_4 = [w_{41}, w_{42}, w_{43}, \ldots, w_{4t}]$

Doc5 $\vec{d}_5 = [w_{51}, w_{52}, w_{53}, \ldots, w_{5t}]$

Doc6 $\vec{d}_6 = [w_{61}, w_{62}, w_{63}, \ldots, w_{6t}]$

...

...

...

...

...

DocN $\vec{d}_N = [w_{N1}, w_{N2}, w_{N3}, \ldots, w_{Nt}]$

TFxIDF Model

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

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

- Where

w_{ij} indicates the importance of term j in document i

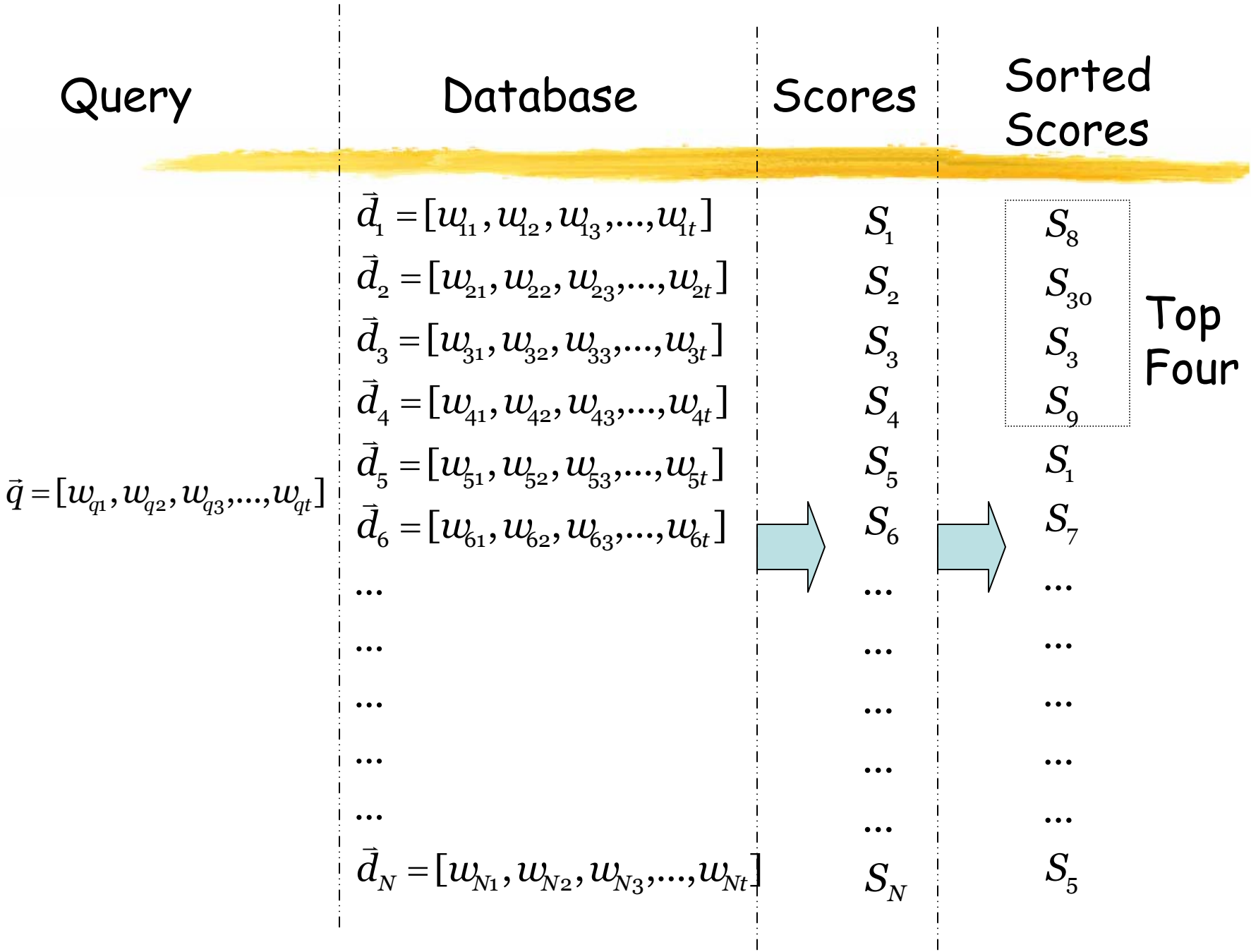
tf_{ij} gives the no. of occurrences of terms j in document i

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
$\vec{d}_1 = [w_{11}, w_{12}, w_{13}, ..., w_{1t}]$
$\vec{d}_2 = [w_{21}, w_{22}, w_{23}, ..., w_{2t}]$
$\vec{d}_3 = [w_{31}, w_{32}, w_{33}, ..., w_{3t}]$
$\vec{d}_4 = [w_{41}, w_{42}, w_{43}, ..., w_{4t}]$
$\vec{d}_5 = [w_{51}, w_{52}, w_{53}, ..., w_{5t}]$
$\vec{d}_6 = [w_{61}, w_{62}, w_{63}, ..., w_{6t}]$
...
...
...
...
...
$\vec{d}_N = [w_{N1}, w_{N2}, w_{N3}, ..., w_{Nt}]$
$\vec{q} = [w_{q1}, w_{q2}, w_{q3}, ..., w_{qt}]$
S_1
S_2
S_3
S_4
S_5
S_6
...
...
...
...
...
S_N
S_8
S_{30}
S_3
S_9
S_1
S_7
...
...
...
...
...
S_5
Top Four

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:

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

- Where

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

  $\vec{d}_j$ 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: $\alpha\vec{q} = [1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0]$

Relevant Terms: $\frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j = [2 \quad 0 \quad 3 \quad 1 \quad 0 \quad 0]$

Non-Relevant Terms: $\frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j = [1 \quad 1 \quad 0 \quad 0 \quad 2 \quad 2]$

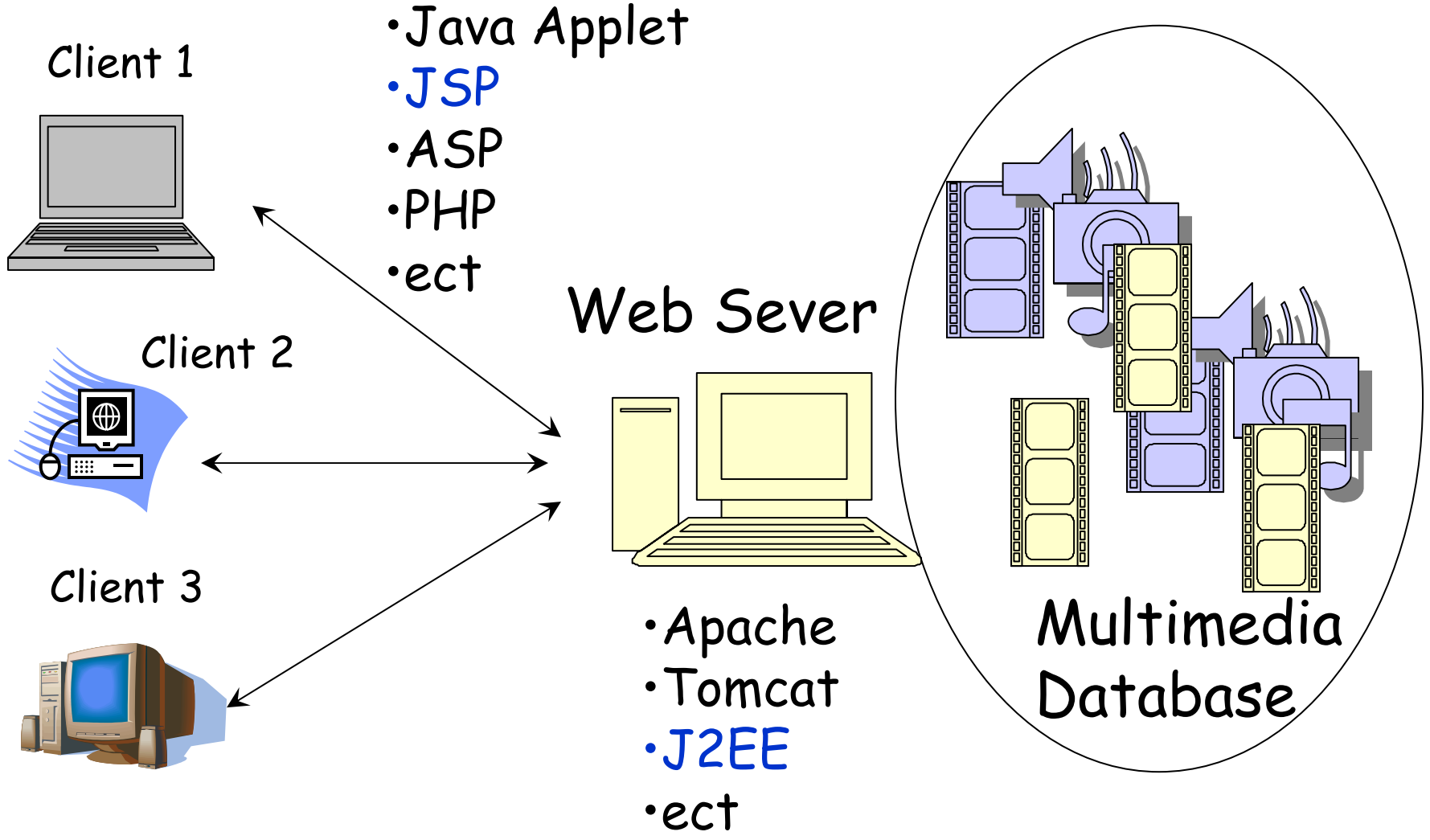
Modified Query: $\vec{q}_m = [2 \quad -1 \quad 4 \quad 1 \quad -2 \quad -2]$

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