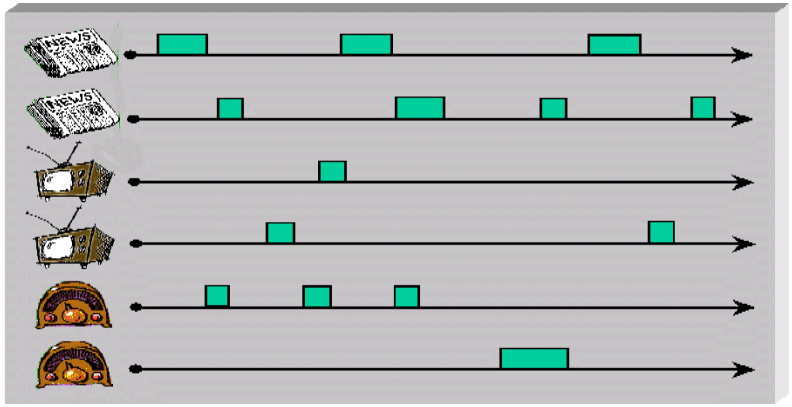
Information Retrieval and Extraction

Berlin Chen 2005



(Picture from the TREC web site)

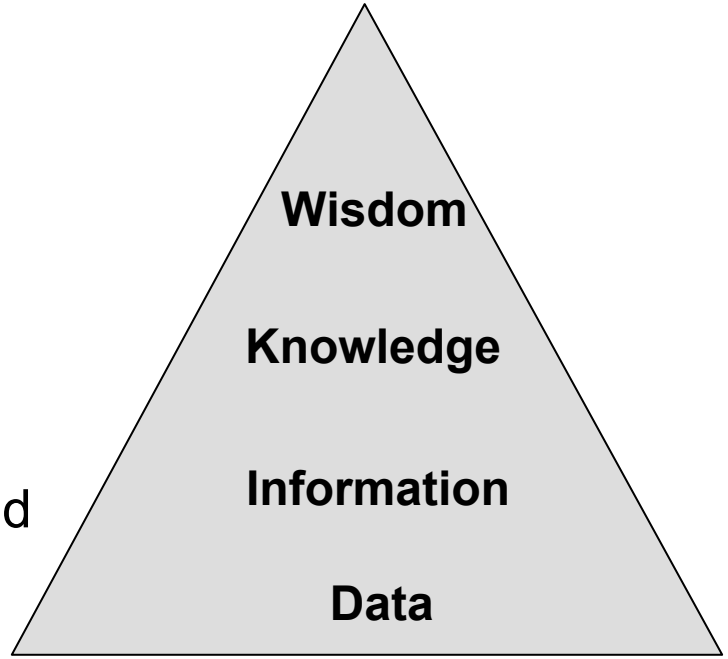
Textbook and References

- Textbook
  - R. Baeza-Yates and B. Ribeiro-Neto. ***Modern Information Retrieval.*** Addison Wesley Longman, 1999

- References
  - W. B. Croft and J. Lafferty (Editors). ***Language Modeling for Information Retrieval***. Kluwer-Academic Publishers, July 2003
  - W. B. Frakes and R. Baeza-Yates. ***Information Retrieval: Data Structures & Algorithms***. Prentice-Hall, 1992
  - I. H. Witten, A. Moffat, and T. C. Bell. ***Managing Gigabytes: Compressing and Indexing Documents and Images***. Morgan Kaufmann Publishing, 1999
  - C. Manning and H. Schutze. ***Foundations of Statistical Natural Language Processing***. MIT Press, 1999
  - A. D. Bimbo. ***Visual Information Retrieval***. Morgan Kaufmann, 1999

Motivation (1/2)

- Information Hierarchy
  - Data
    - The raw material of information
  - Information
    - Data organized and presented by someone
  - Knowledge
    - Information read, heard or seen and understood
  - Wisdom
    - Distilled and integrated knowledge and understanding



Motivation (2/2)

- User information need
  - Find all docs containing information on college tennis teams which:
    (1) are maintained by a USA university and
    (2) participate in the NCAA tournament
    (3) National ranking in last three years and contact information

Query

Emphasis is on the retrieval of information (not data)

Search engine/IR system

Information Retrieval

- Deal with the representation, storage, organization of, and access to information items

- Focus is on the user information need
  - Information about a subject or topic
  - Semantics is frequently loose
  - Small errors are tolerated

- Handle natural language text which is not always well structured and could be semantically ambiguous

Data Retrieval

- Determine which document of a collection contain the *keywords* in the user query
- Retrieve all objects (attributes) which satisfy clearly defined conditions in a regular expression or a relational algebra expression
  - Which documents contain a set of keywords?
  - Well defined semantics
  - A single erroneous object implies failure!

IR system

- Interpret contents of information items (docs)
- Generate a ranking which reflects relevance
- Notion of *relevance* is most important

IR at the Center of the Stage

- IR in the last 20 years:
  - Modelng, classification, clustering, filtering
  - User interfaces and visualization
  - Systems and languages

- WWW environment (90~)
  - Universal repository of knowledge and culture
  - Without frontiers: free universal access
  - Lack of well-defined data model

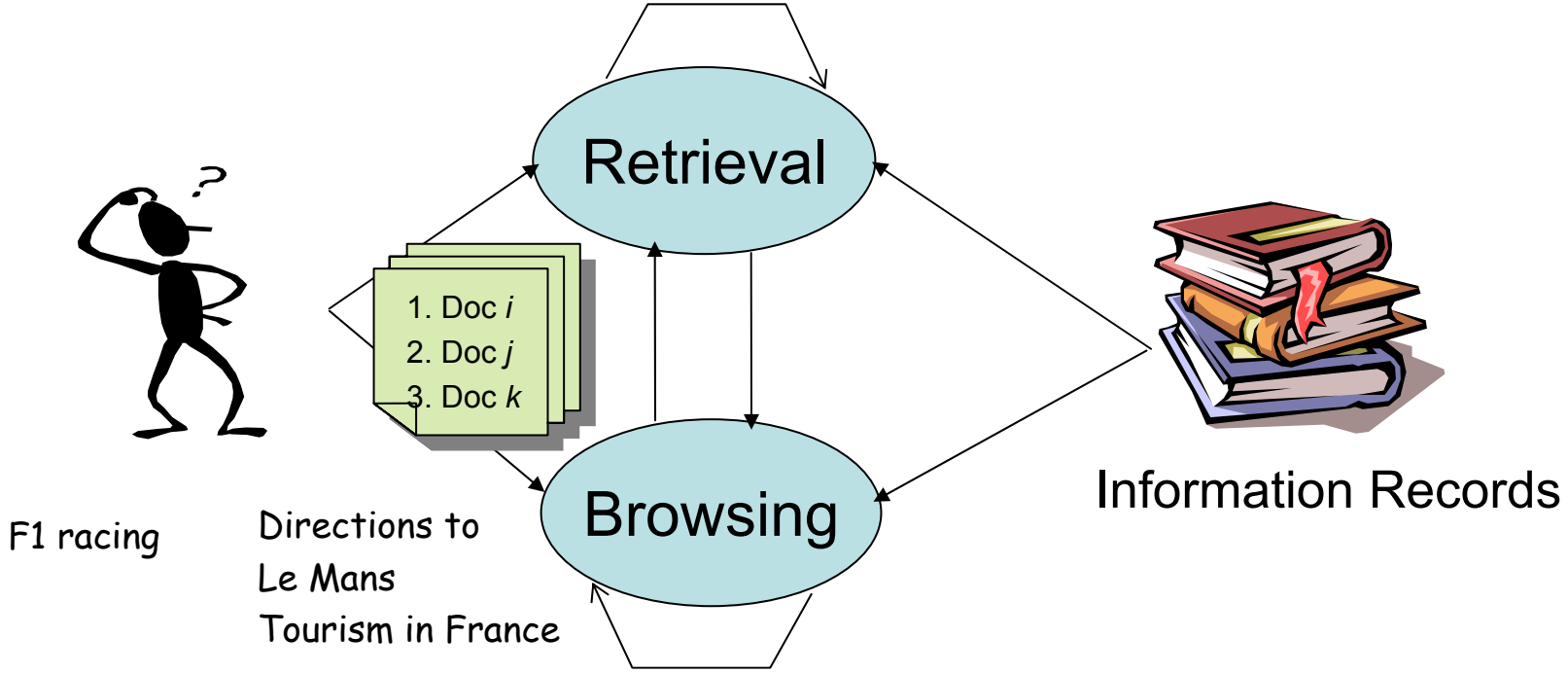
IR Main Issues

- The effective retrieval of relevant information affected by
  - The user task
  - Logical view of the documents

The User Task

- Translate the information need into a query in the language provided by the system
  - A set of words conveying the semantics of the information need

- Browse the retrieved documents

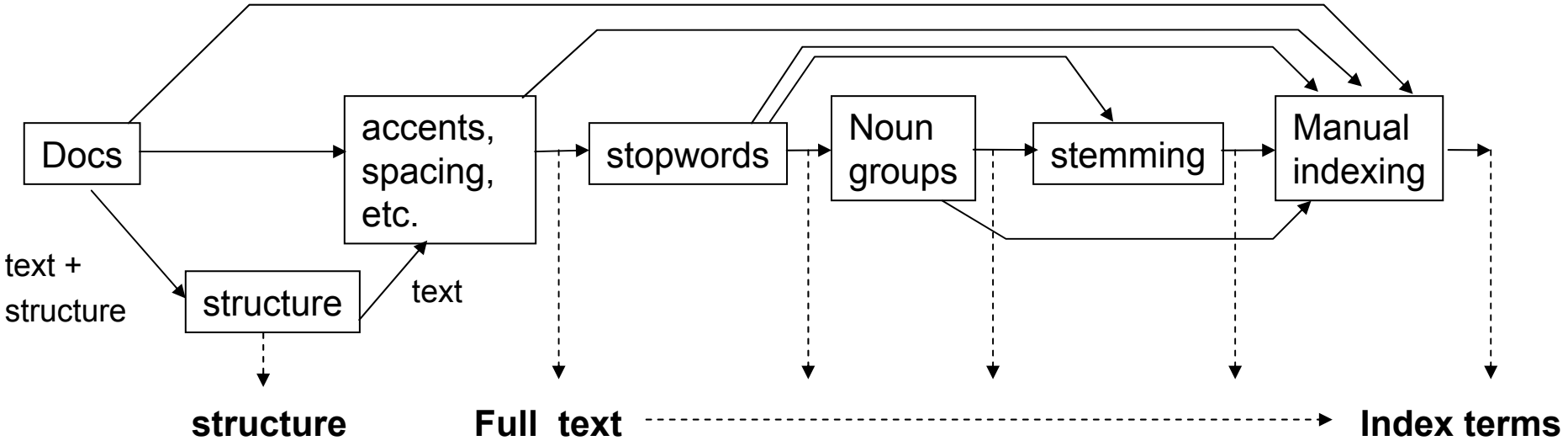


Logical View of the Documents (1/2)

- A full text view (representation)
  - Represent document by its whole set of words
    - Complete but higher computational cost
- A set of index terms by a human subject
  - Derived automatically or generated by a specialist
    - Concise but may poor
- An intermediate representation with feasible *text operations*

Logical View of the Documents (2/2)

- Text operations
  - Elimination of stop-words (e.g. articles, connectives, …)
  - The use of stemming (e.g. tense, …)
  - The identification of noun groups
  - Compression ….
- Text structure (chapters, sections, …)



Different Views of the IR Problem

- Computer-centered (commercial perspective)
  - Efficient indexing approaches
  - High performance matching ranking algorithms

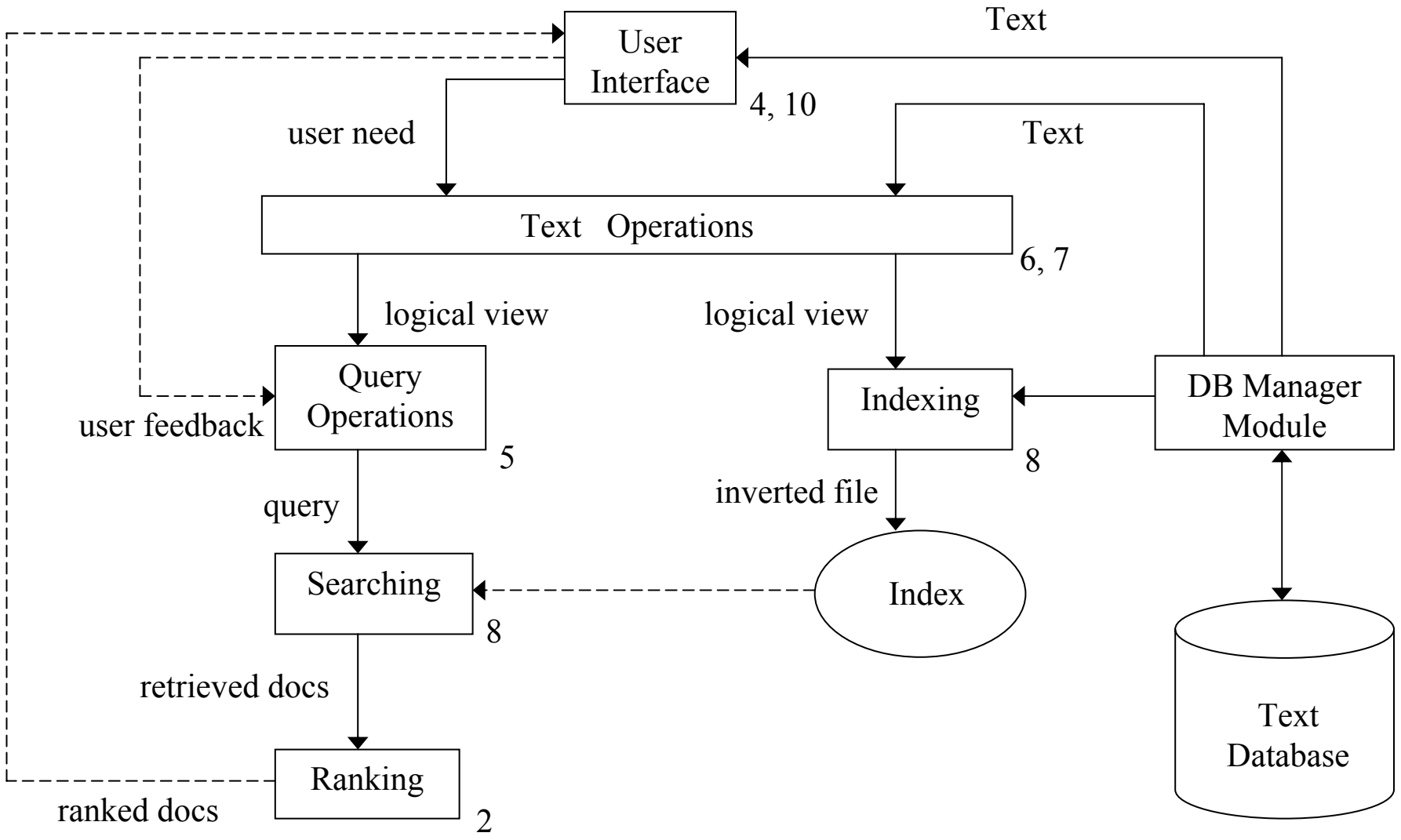
- Human-centered (academic perceptive)
  - Studies of user behaviors
  - Understanding of user needs

Library science
psychology
....

IR for Web and Digital Libraries

- Questions should be addressed
  - Still difficult to retrieve information relevant to user needs
  - Quick response is becoming more and more a pressing factor (Precision vs. Recall)
  - The user interaction with the system (HCI, Human Computer Interaction)
- Other concerns
  - Security and privacy
  - Copyright and patent

The Retrieval Process (1/2)



The Retrieval Process (2/2)

- In current retrieval systems
  - Users almost never declare his information need
    - Only a short queries composed few words (typically fewer than 4 words)
  - Users have no knowledge of the text or query operations

Poor formulated queries lead to poor retrieval !

Major Topics (1/2)

- Four Main Topics

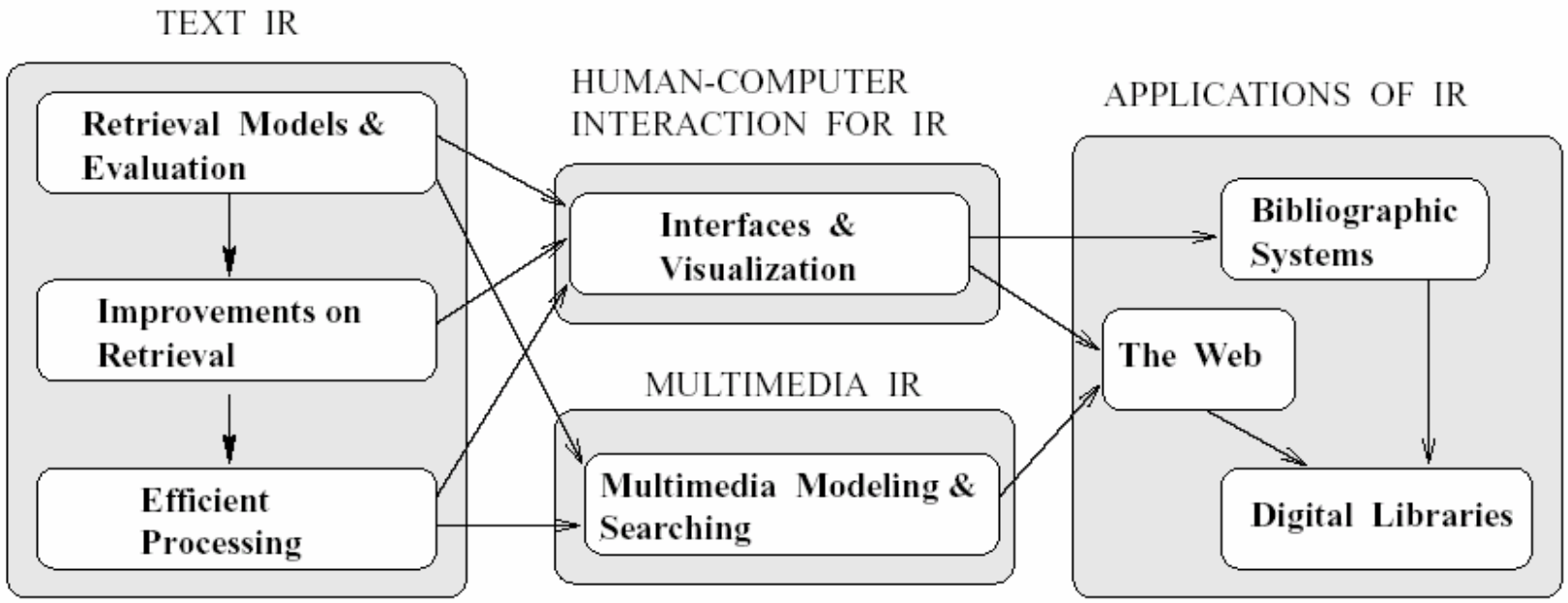
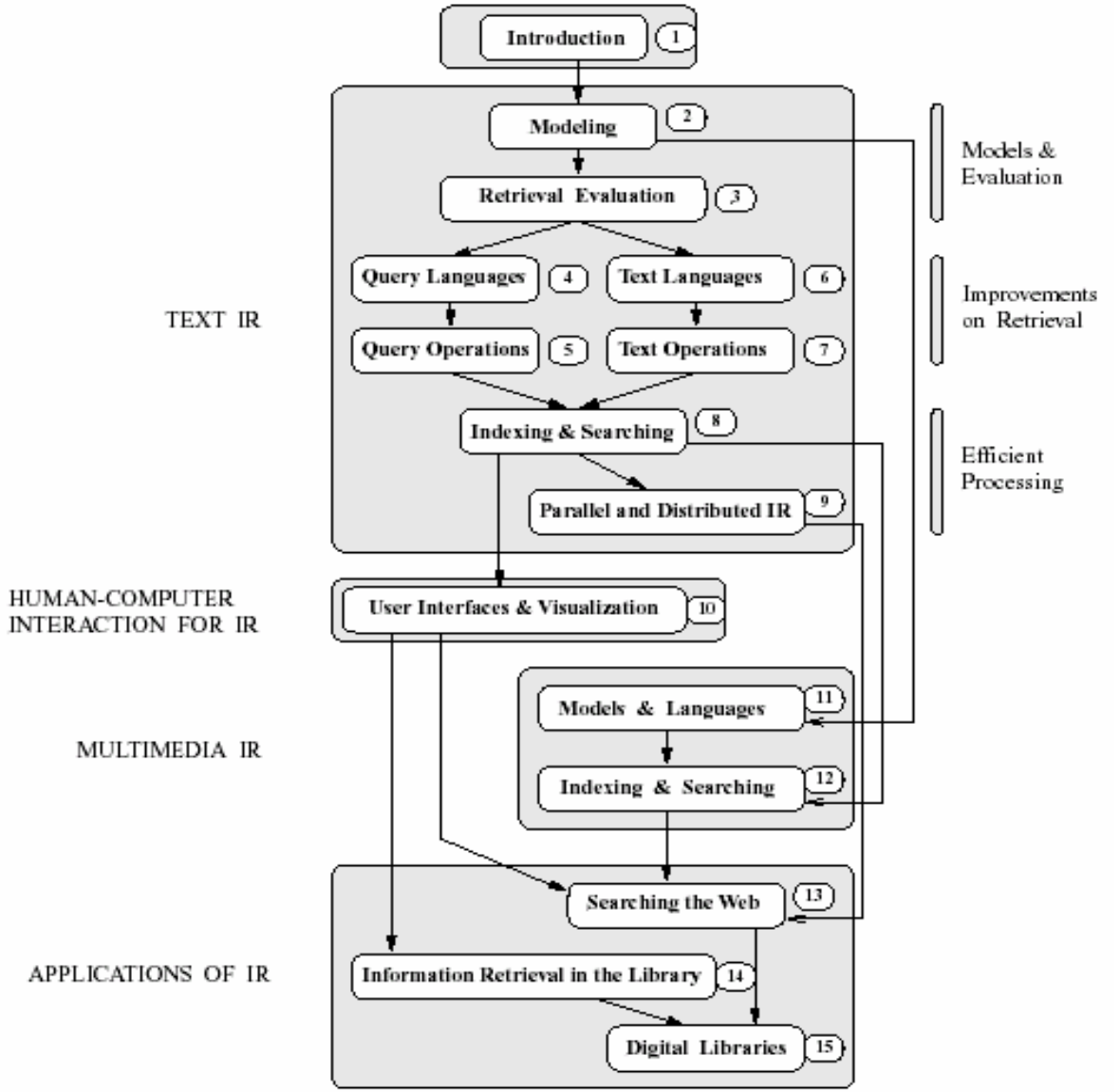


Figure 1.4 Topics which compose the book and their relationships.

Major Topics (2/2)

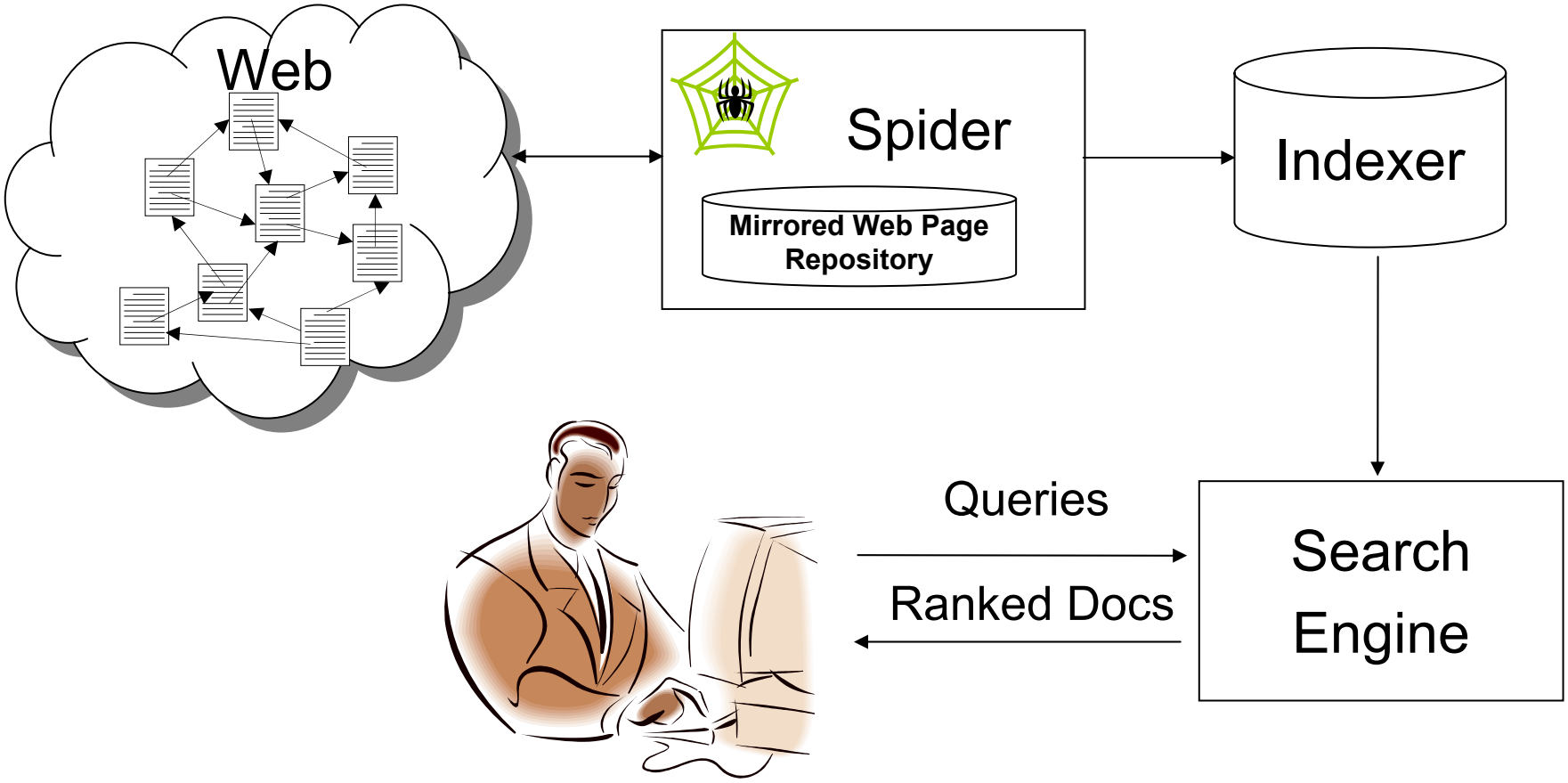
- Text IR
  - Retrieval models, evaluation methods, indexing
- Human-Computer Interaction (HCI)
  - Improved user interfaces and better data visualization tools
- Multimedia IR
  - Text, speech, audio and video contents
  - Multidisciplinary approaches
- Applications
  - Web, bibliographic systems, digital libraries

Textbook Topics



Text Information Retrieval (1/4)

- Internet searching engine



Text Information Retrieval (2/4)

- http://www.google.com

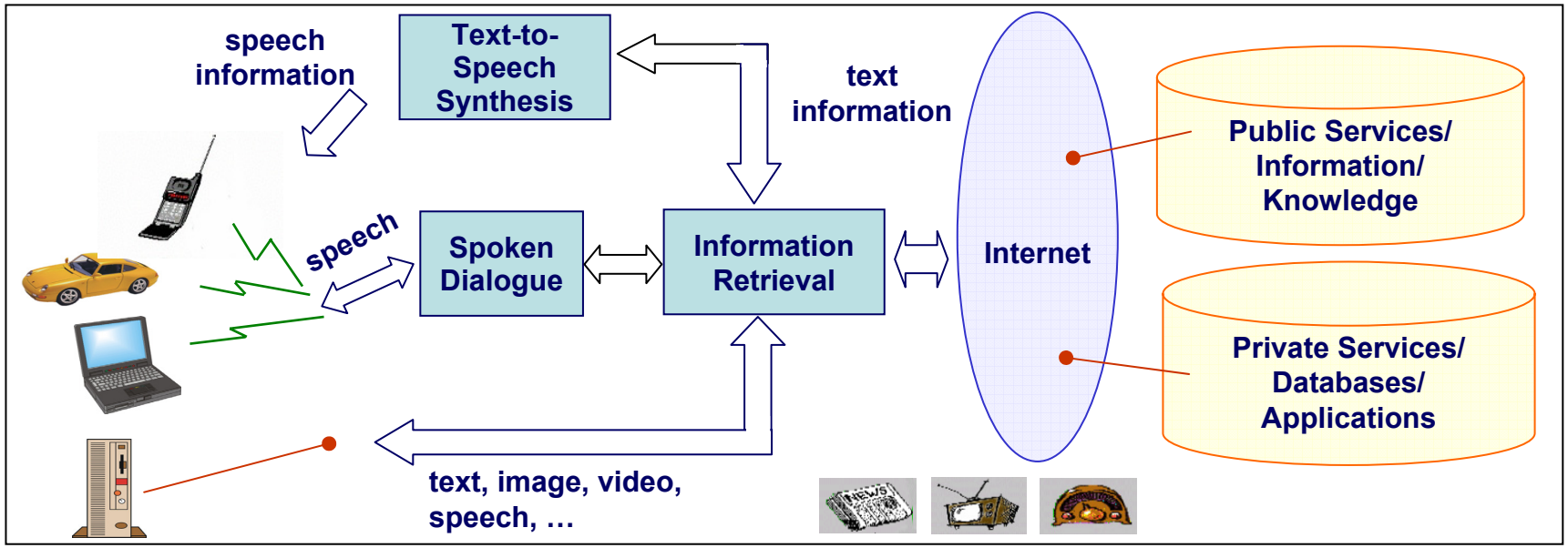
Text Information Retrieval (3/4)

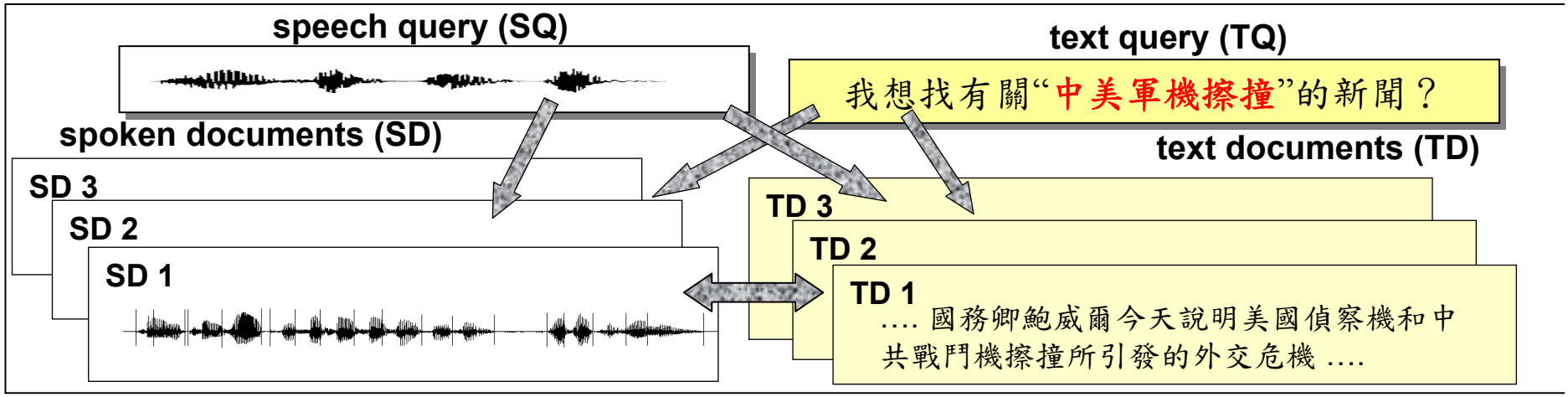
- http://www.openfind.com.tw

Text Information Retrieval (4/4)

- http://www.baidu.com

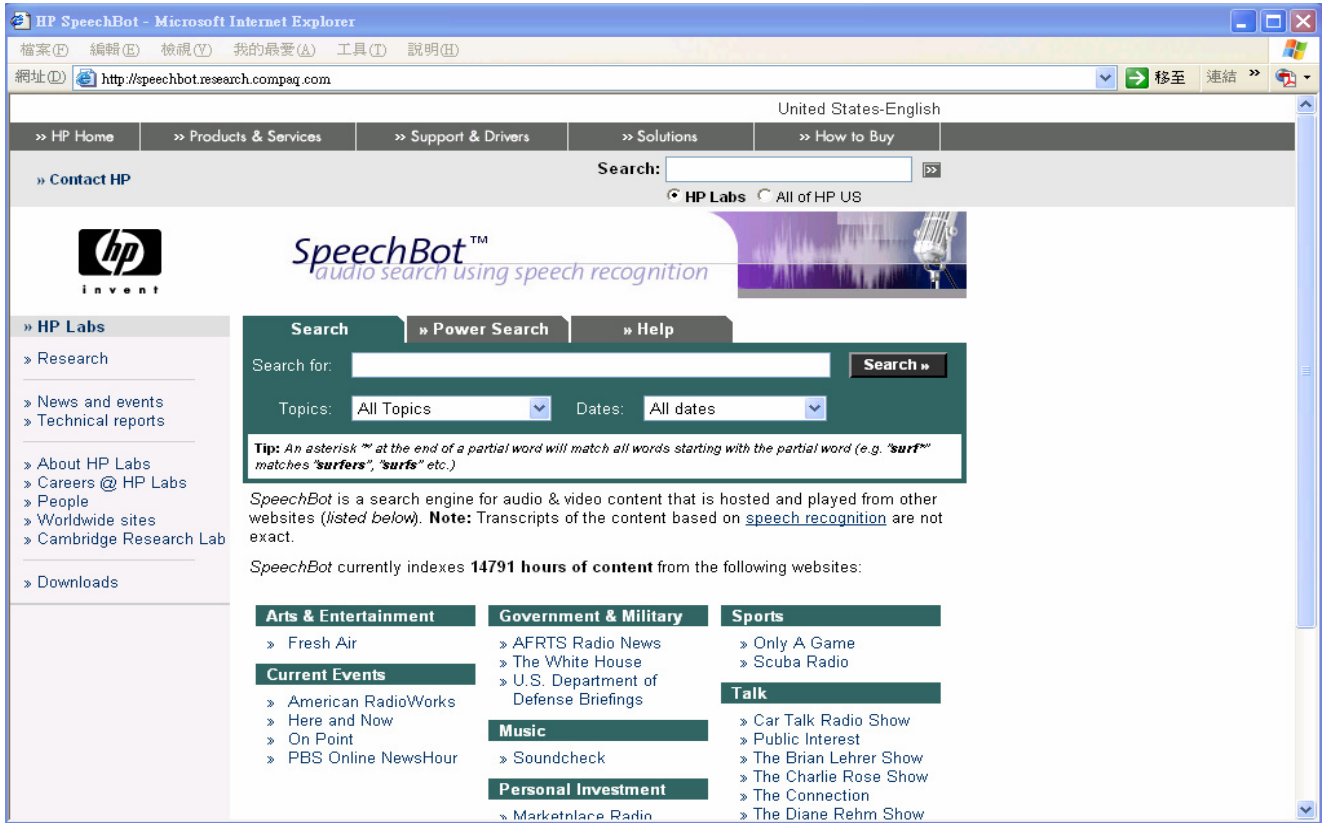
Speech Information Retrieval (1/4)





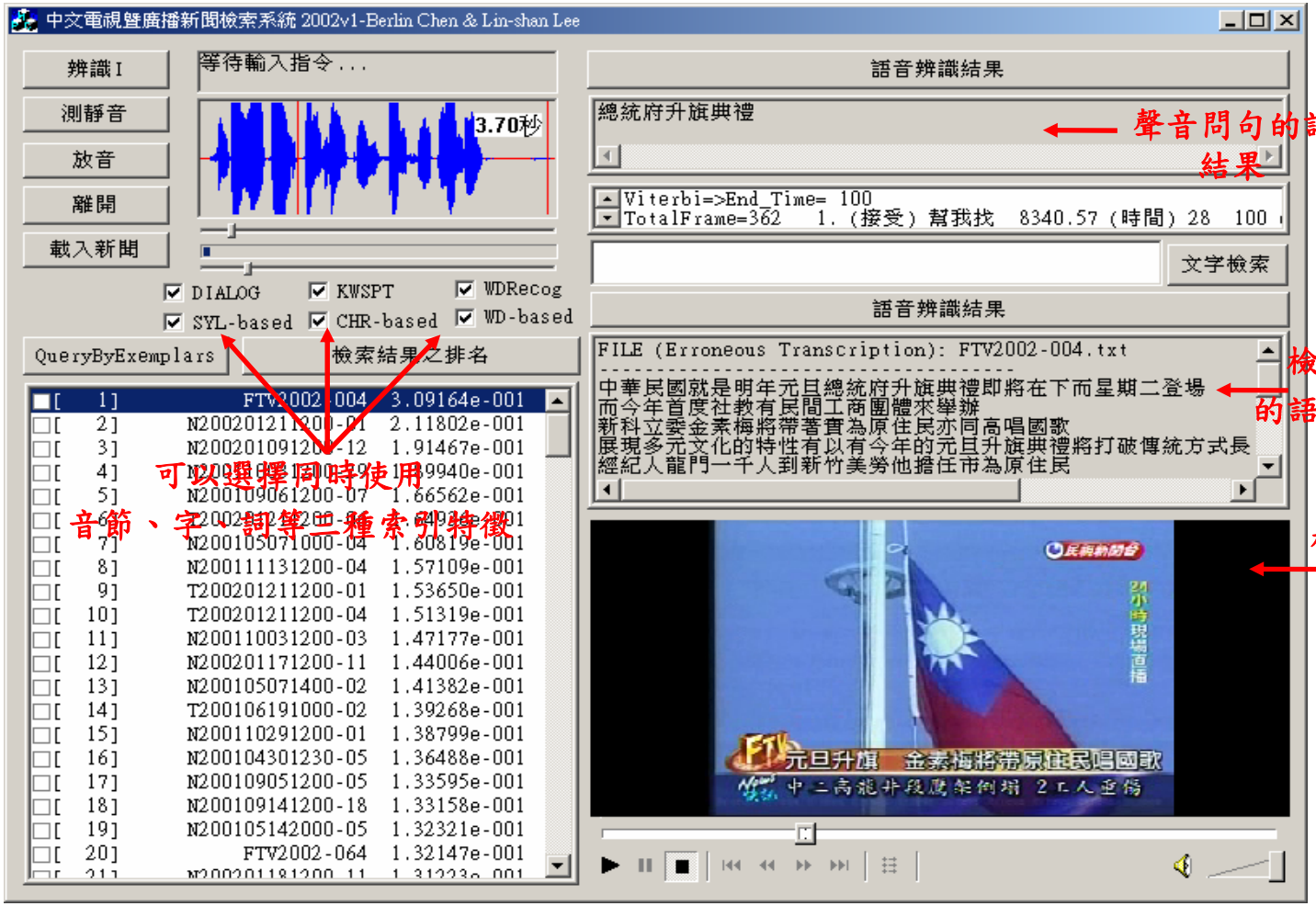
Speech Information Retrieval (2/4)

- Compaq Research Group – Speechbot System
  - Broadcast news speech recognition, Information retrieval, and topic segmentation (SIGIR2001)
  - Currently indexes **14,791 hours of content** (2004/09/22, http://speechbot.research.compaq.com/)



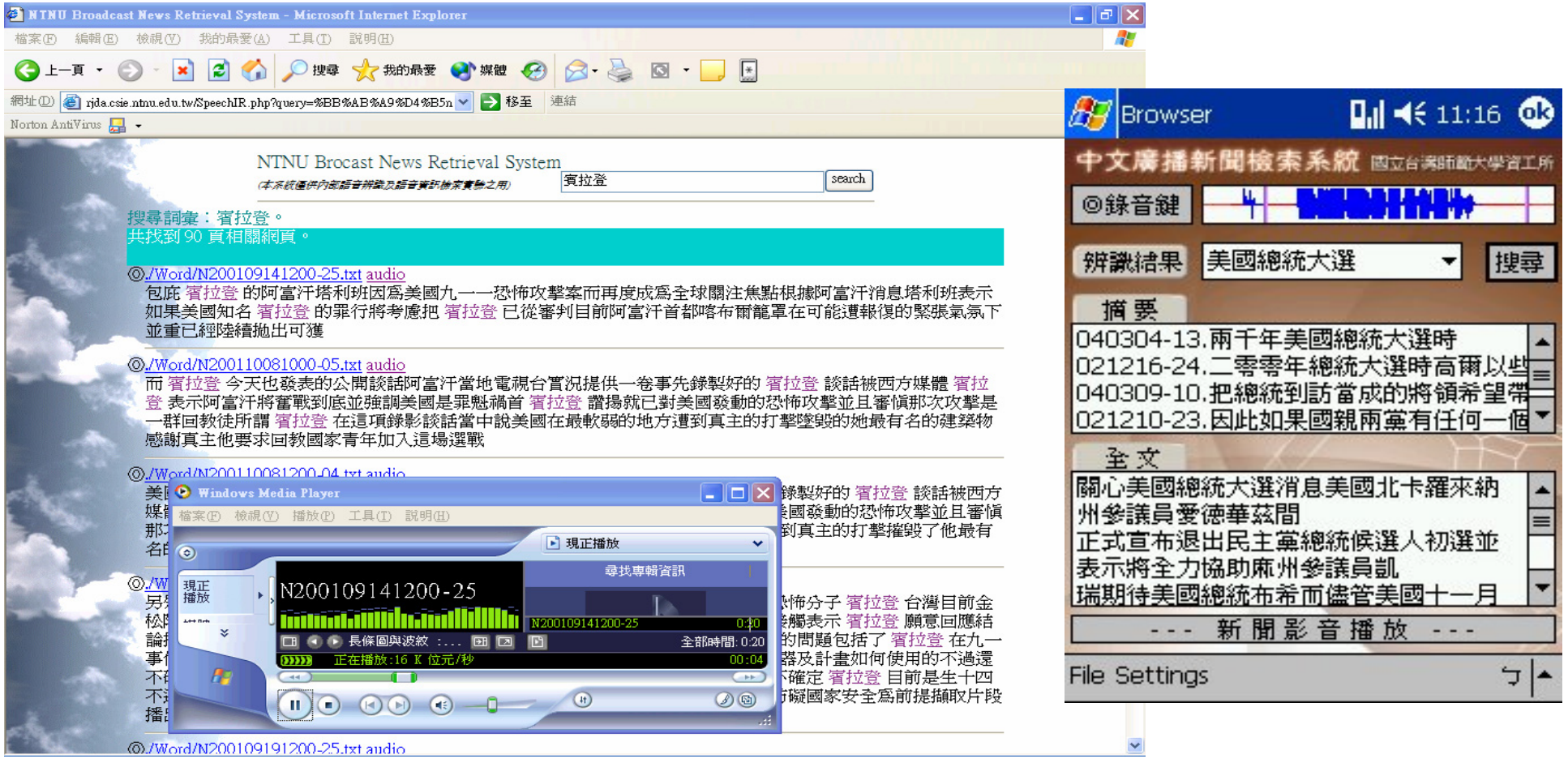
Speech Information Retrieval (3/4)

- 輸入聲音問句：“請幫我查總統府升旗典禮”



中文影音多媒體資訊檢索雛形展示系統。

Speech Information Retrieval (4/4)



Visual Information Retrieval (1/4)

- Content-based approach

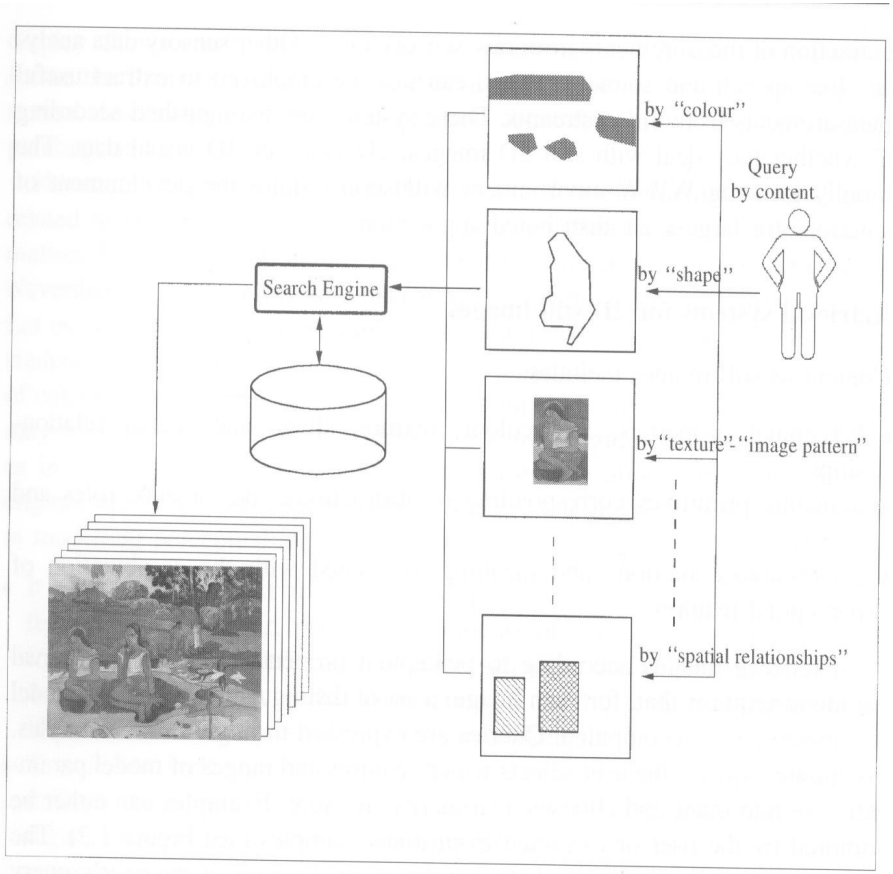


Figure 1.2 Different types of query by example.

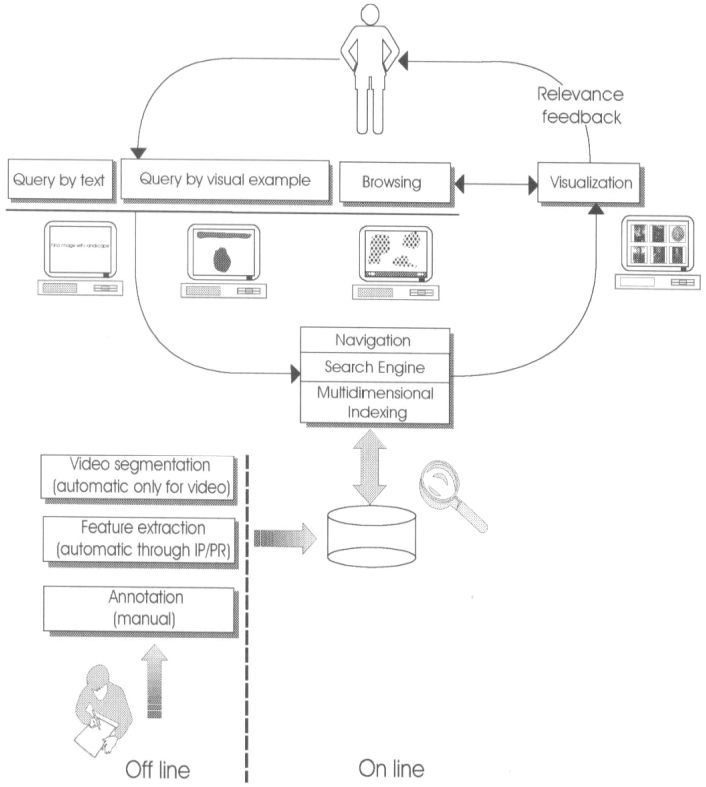


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

Visual Information Retrieval (2/4)

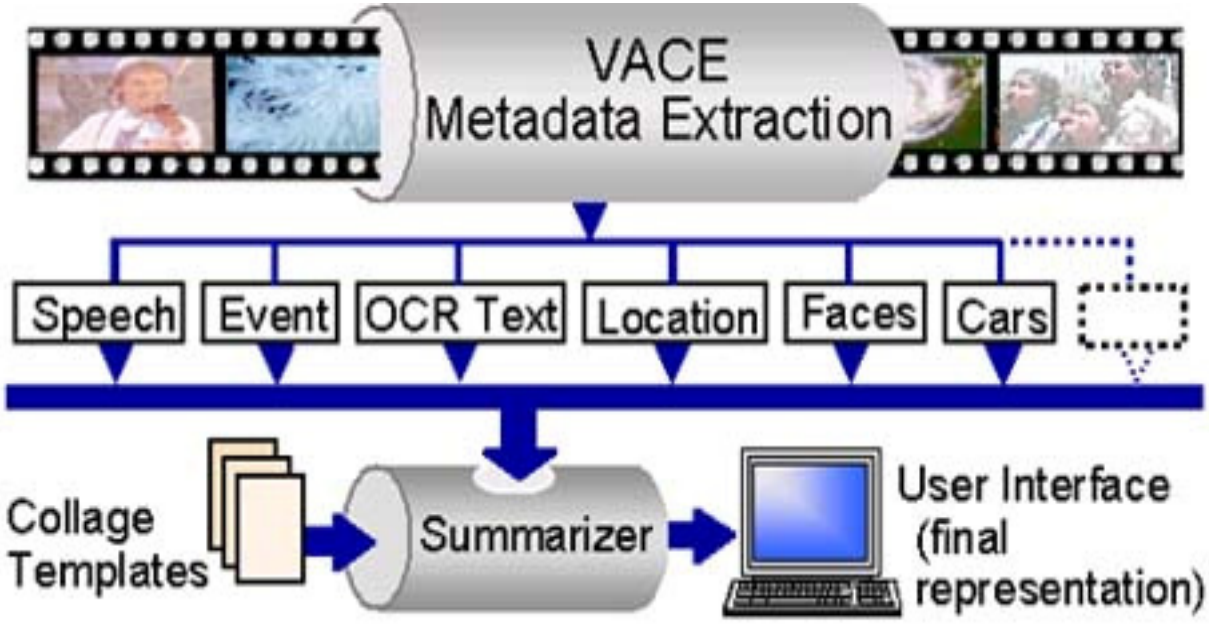
- Images with Texts

Visual Information Retrieval (3/4)

- Content-based Image Retrieval

Visual Information Retrieval (4/4)

Video Analysis and Content Extraction



Other IR-Related Tasks

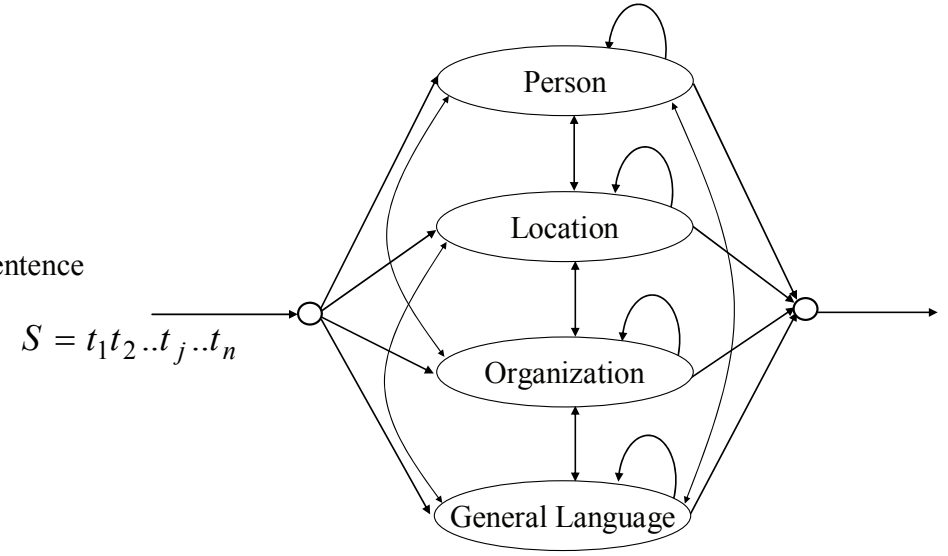
- Information filtering and routing
- **Term/Document categorization**
- **Term/Document clustering**
- **Document summarization**
- **Information extraction**
- Question answering
- Crosslingual information retrieval
- …..

Document Summarization

- Audience
  - Generic summarization
  - User-focused summarization
    - Query-focused summarization
    - Topic-focused summarization
- Function
  - Indicative summarization
  - Informative summarization
- Extracts vs. abstracts
  - Extract: consists wholly of portions from the source
  - Abstract: contains material which is not present in the source
- Output modality
  - Speech-to-text summarization
  - Speech-to-speech summarization
- Single vs. multiple documents

Information Extraction

- E.g., Named-Entity Extraction
  - NE has it origin from the Message Understanding Conferences (MUC) sponsored by U.S. DARPA program
    - Began in the 1990's
    - Aimed at extraction of information from text documents
    - Extended to many other languages and spoken documents (mainly broadcast news)
  - Common approaches to NE
    - Rule-based approach
    - Model-based approach
    - Combined approach

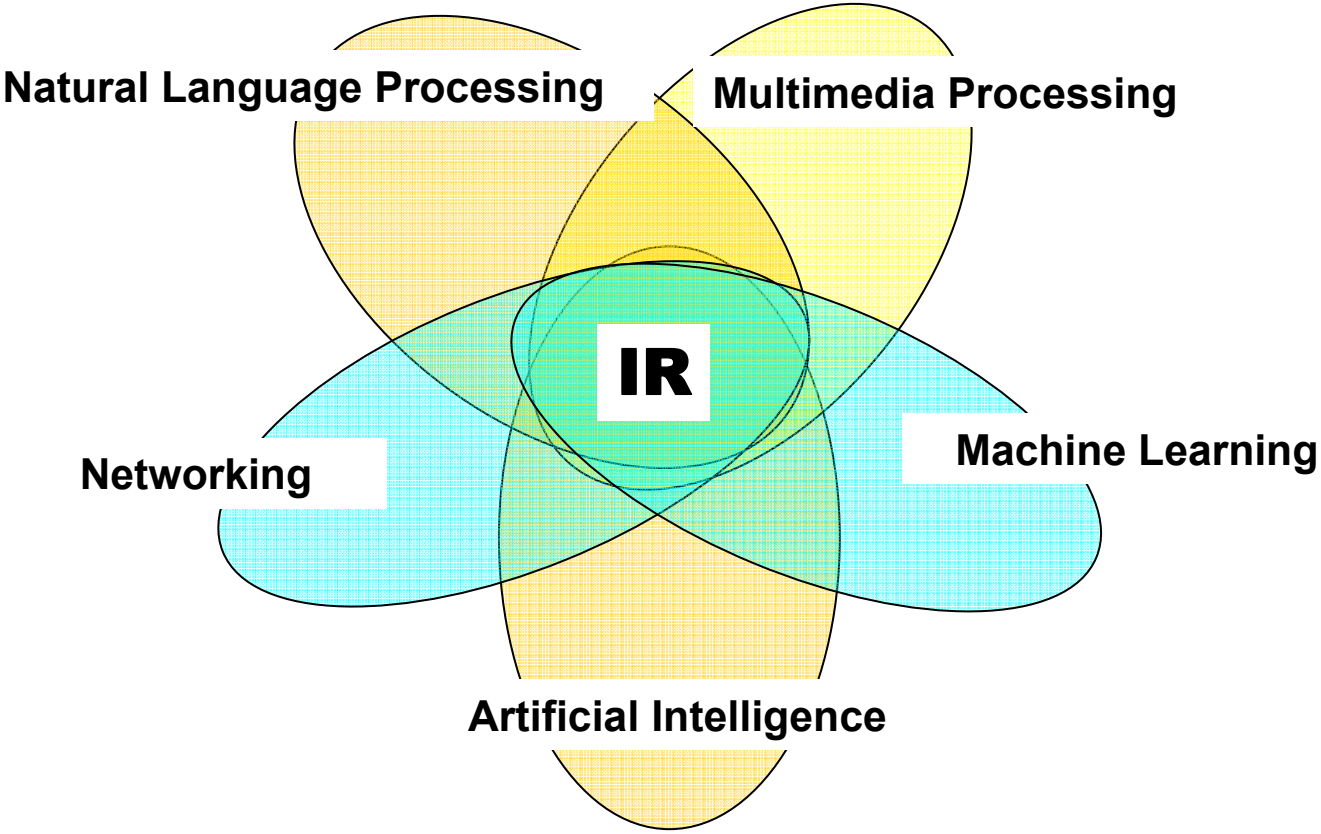


Crosslingual Information Retrieval

- E.g., Automatic Term Translation
  - Discovering translations of unknown query terms in different languages
  - E.g., The Live Query Term Translation System (LiveTrans) developed at Academia Sinica/by Dr. Chien Lee-Feng

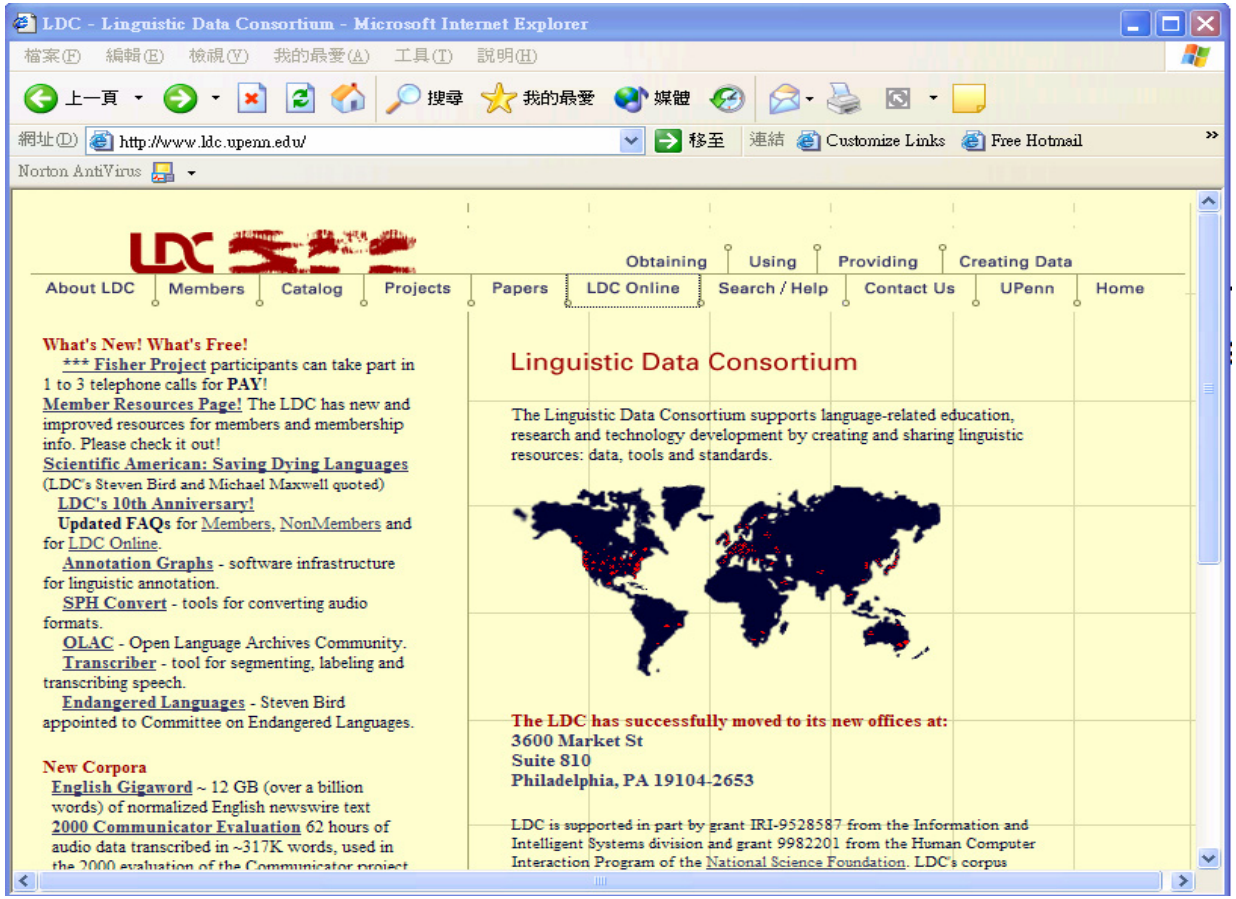
Machine-Extracted Translation

Multidisciplinary Approaches



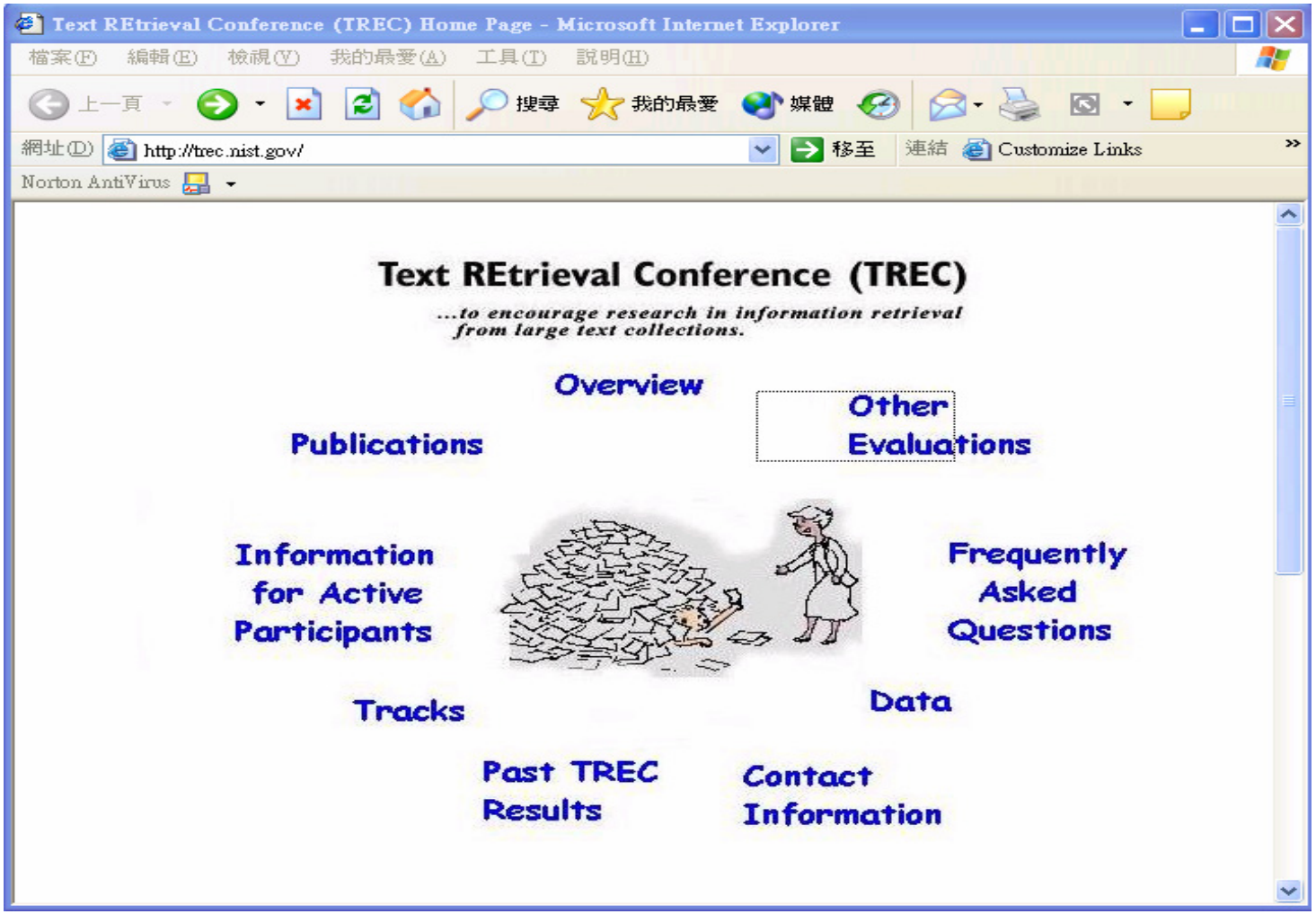
Resources

- Corpora (Speech/Language resources)
  - Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them
    - LDC - Linguistic Data Consortium



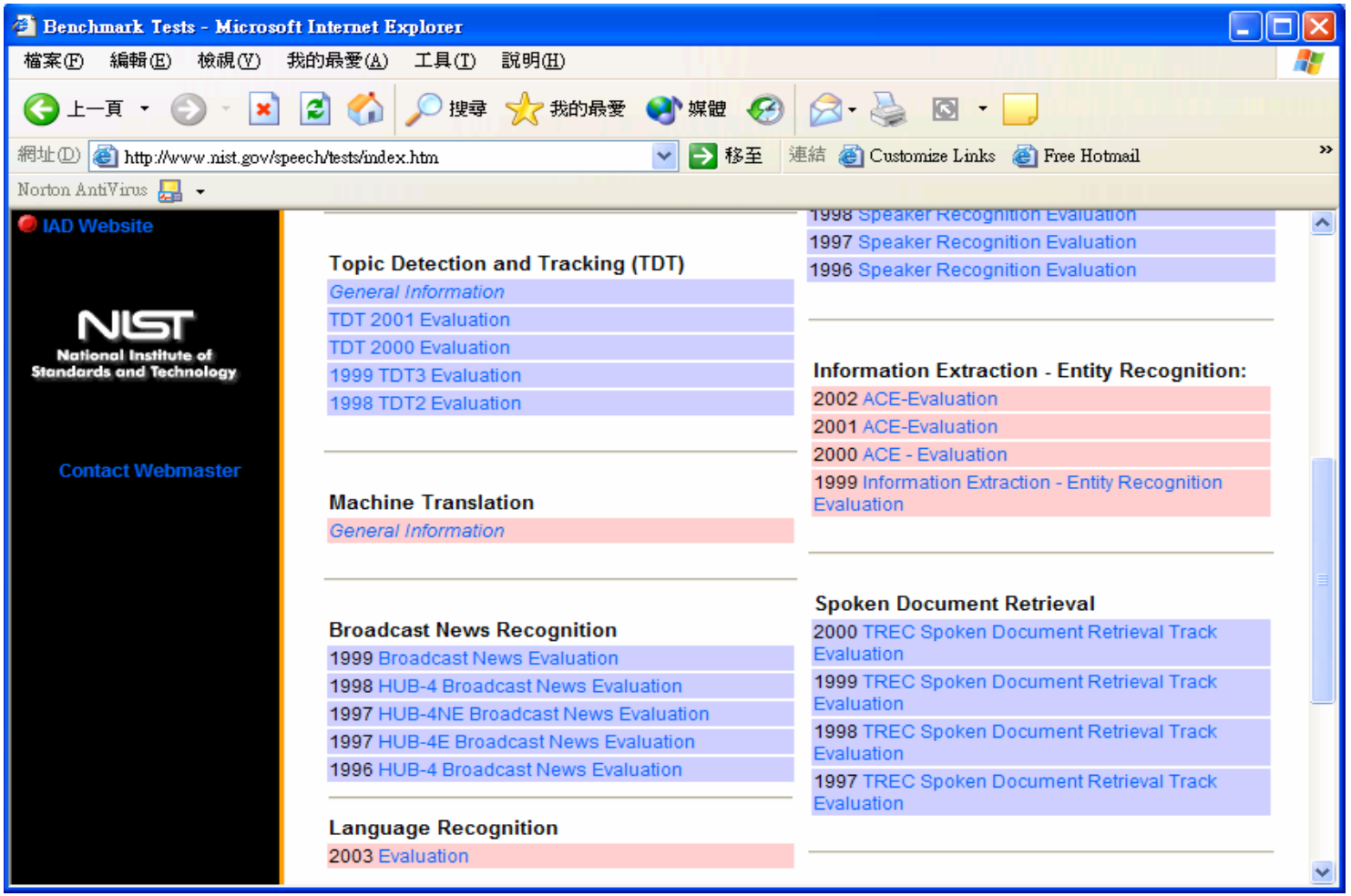
Contests (1/2)

- Text REtrieval Conference (TREC)



Contests (2/2)

- US National Institute of Standards and Technology



Conferences/Journals

- Conferences
  - ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR)
  - ACM Conference on Information Knowledge Management (CIKM)
  - …
- Journals
  - ACM Transactions on Information Systems (TOIS)
  - ACM Transactions on Asian Language Information Processing (TALIP)
  - Information Processing and Management (IP&M)
  - Journal of the American Society for Information Science (JASIS)
  - …

Tentative Schedule

9/13	Course Overview & Introduction
9/20	Retrieval Models (I) - Classic Retrieval Models (Boolean, Vector Space and Probabilistic Models)
9/27	Retrieval Performance Evaluation (I) - Measures
10/4	Retrieval Performance Evaluation (II) - Reference Collections
10/11	Retrieval Models (II) - Improved Approaches (Fuzzy Set, Extended Boolean, Generalized Vector Space Models)
10/18	Query Operations (Query Expansion and Term Re-weighting)
10/25	Retrieval Models (III) - Statistical Modeling Approaches (HMM/N-Gram: Language Model Approach)
11/1	Retrieval Models (III) - Statistical Modeling Approaches (TMM: Topical Mixture Model)
11/8	Retrieval Models (III) - Statistical Modeling Approaches (LSA, PLSA)
11/15	**Midterm**
11/22	Text Clustering & LSA Toolkit
11/29	Retrieval Models (IV) - Structural Retrieval Models and Browsing Models
12/6	Query Languages, Text Languages and Text Statistics
12/13	Text Preprocessing, Indexing and Searching
12/20	Text Categorization, Text Summarization, Information Extraction
12/27	Internet Search Engine
1/3	Paper Survey
1/10	**Break**
1/3	**Final**