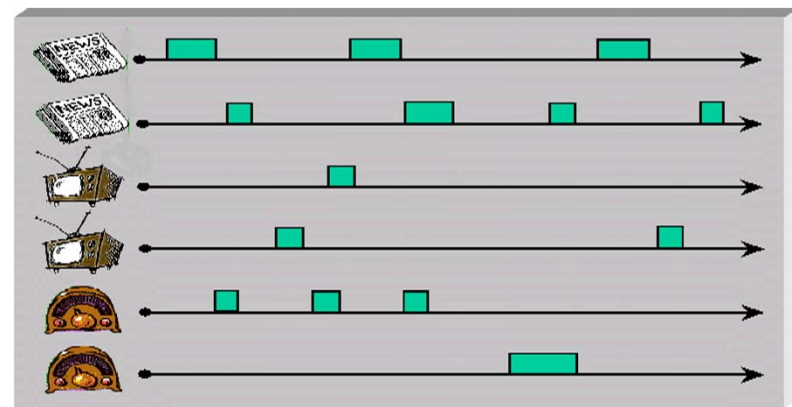


# Information Retrieval and Extraction

Berlin Chen



(Picture from the [TREC](http://www.trec.nist.gov/) web site)

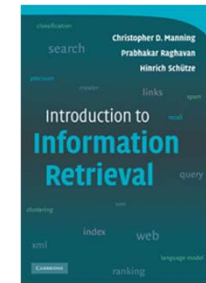
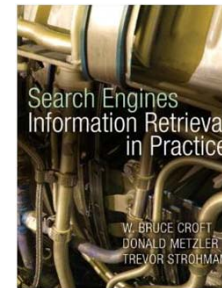
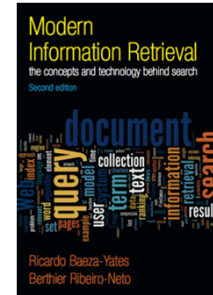
# Objectives of this Course

- Elaborate on the fundamentals of information retrieval (IR), a almost *sixty-year-old* discipline
  - Indexing, search, relevance, classification, organization, storage, browsing, visualization, etc.
- Focus on prominent *computer algorithms* and *techniques* used in IR systems from a computer scientist's perspective
  - How to provide users with easy access to information of interest
  - Rather than from a “librarian” perspective that put great emphasis on “*human-centered*” studies (e.g., user behaviors, psychology, etc. )
- Discuss ractical Issues on the Web
  - Crawling, retrieval, and ranking of Web documents
  - Electronic commerce; security, privacy, copy rights and pattern rights; multimedia and cross-language retrieval; digital libraries

# Textbook and References

- Textbooks

- R. Baeza-Yates and B. Ribeiro-Neto. ***Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)***, ACM Press, 2011
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, ***Introduction to Information Retrieval***, Cambridge University Press, 2008
- W. Bruce Croft, Donald Metzler, and Trevor Strohman, ***Search Engines: Information Retrieval in Practice***, Addison Wesley, 2009



- References

- C.X. Zhai, ***Statistical Language Models for Information Retrieval*** (Synthesis Lectures Series on Human Language Technologies), Morgan & Claypool Publishers, 2008
- W. B. Croft and J. Lafferty (Editors). ***Language Modeling for Information Retrieval***. Kluwer-Academic Publishers, July 2003
- D. A. Grossman, O. Frieder, ***Information Retrieval: Algorithms and Heuristics***, Springer. 2004
- 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. Schütze. ***Foundations of Statistical Natural Language Processing***. MIT Press, 1999

# Motivation (1/3)

- Information Overload: *Too much information kills information!*



•The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech 2009*



# Motivation (2/3)

- Information Hierarchy

- **Data**

- The raw material of information

- **Information**

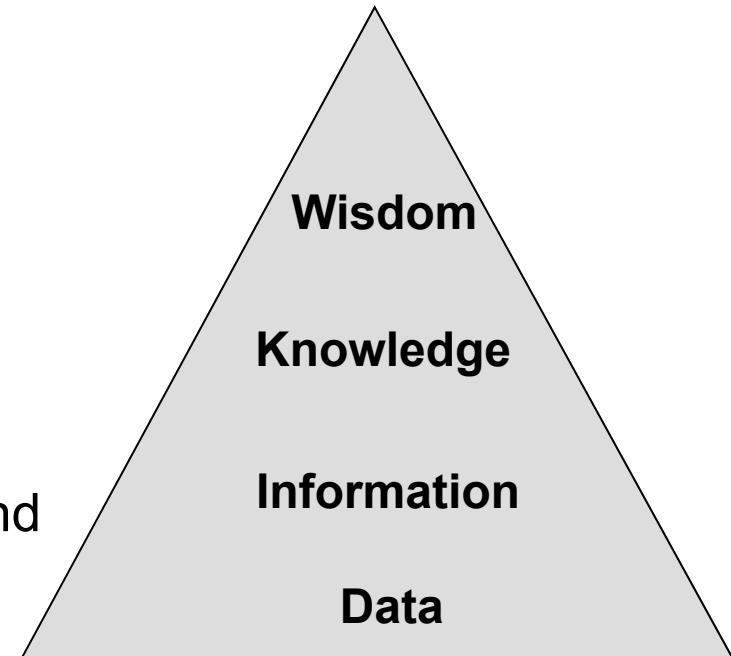
- Data organized and presented by someone

- **Knowledge**

- Information read, heard or seen and understood

- **Wisdom**

- Making appropriate use of distilled and integrated knowledge and understanding



- *Search* and *communication* (of information) are by far the most popular uses of the computer

# Motivation (3/3)

- How to satisfy users' information needs?
  - Find all docs containing information on college tennis teams which:
    - (1) are maintained by a USA university and
    - (2) participate in the NCAA (National Collegiate Athletic Association) tournament
    - (3) National ranking in last three years and contact information



Query



Search engine/IR system

Emphasis is on the retrieval of information (not data)

# Information Retrieval (1/2)

- Information retrieval (IR) is the field concerned with the structure, analysis, or organization, searching and retrieval of information items (documents, webpages, online catalogs, structured/unstructured records, multimedia objects)
  - Defined by Gerard Salton, a pioneer and leading figure in IR
- Early goals of the IR area: indexing text and searching for useful documents in a collection
- Nowadays, research in IR includes:
  - Modeling, Web search, text classification, systems architecture, user interfaces, data visualization, filtering and languages

# Information Retrieval (2/2)

- IR typically handles **natural language text** (or free text) which is not always well structured and could be semantically ambiguous
- Its focus is on the user information need
  - Information about a subject or topic
  - Semantics is frequently loose
  - Small errors are tolerated

A user of an IR system is willing to accept documents that contain synonyms of the query terms in the result set, even when those documents do not contain any query terms.



# Data Retrieval

- Determine which document of a collection contain the *keywords* in the user query
  - Such documents are regarded as database records, such as a bank account record or a flight reservation, consisting of structural elements such as fields or attributes (e.g., account number and current balance)
- 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 (attributes) in some specific fields?
  - Well defined semantics & structures
  - A single erroneous object implies (total) failure!

Data retrieval does not solve the problem of retrieving information about a **subject or topic**.

# Early Developments in IR (1/2)

- During the 50's, research efforts in IR were initiated by pioneers such as Hans Peter Luhn, Eugene Garfield, Philip Bagley, and Calvin Moores, who allegedly coined the term *Information Retrieval*
- In 1962, Cyril Cleverdon published the Cranfield studies on retrieval evaluation
- In 1963, Joseph Becker and Robert Hayes published the first book on IR
- In the late 60's, key research conducted by Karen Sparck Jones and Gerard Salton, among others, led to the definition of the *TF-IDF term weighting scheme*

## Early Developments in IR (2/2)

- In 1978, the first ACM SIGIR International Conference on Information Retrieval was held in Rochester
- In 1979, van Rijsbergen published a classic book entitled *Information Retrieval*, which focused on the Probabilistic Model
- In 1983, Salton and McGill published a classic book entitled *Introduction to Modern Information Retrieval*, which focused on the Vector Model

# IR at the Center of the Stage (1/2)

- Before 1990s
  - Until recently, IR was an area of interest restricted mainly to librarians and information experts
  - Such a tendentious vision prevailed for many years, despite the rapid dissemination, among users of modern personal computers, of IR tools for many applications
- After 1990s (WWW environment)
  - A single fact changed these perceptions—the introduction of the Web, which has become the largest repository of knowledge and culture in human history
    - Decentralized
    - Without frontiers: free universal access (*freedom to publish*)
    - Hypertext (HTTP protocol and browsers by Tim Berners-Lee)
    - Lack of well-defined data model



# IR at the Center of the Stage (2/2)

- Due to its enormous size, finding useful information on the Web usually requires running a search
- Searching on the Web is all about IR and its technologies
- Recall: typical tasks includes
  - Modeling, classification, clustering, filtering
  - User interfaces and visualization
  - Systems and languages

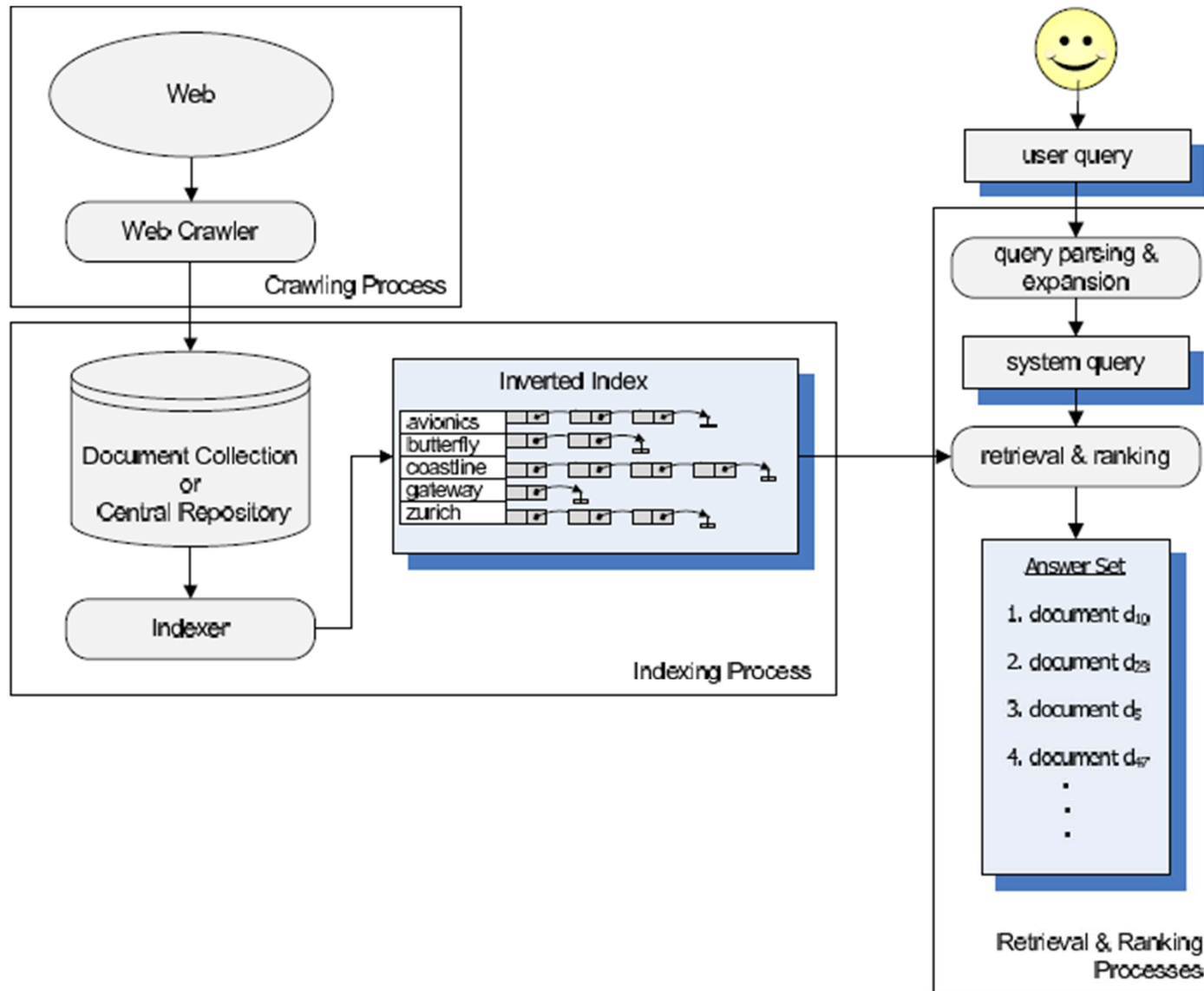
Restrictions imposed by mass communication media companies and by natural geographical barriers were almost entirely removed by the invention of the Web! (*e-Publishing Era*)

Thus, almost overnight, IR has gained a place with other technologies at the center of the stage

# Web Changed Search!

- Characteristics of document collection
  - Distributed natural => *crawling*
- The size of document collection
  - ~20 billion pages=> *performance* and *scalability* are big issues
- Relevance judgment in the face of the vast size of document collections
  - Hyperlinks and user clicks in documents => *clickthrough data*
- Going beyond seeking text information
  - E.g., price of a book, phone number of a hotel  
=> *effective answers* to various types of information needs  
(Question Answering -> *Apple's Siri!* )
- Web advertising and economic incentives
  - E-commerce, advertising <=> *Web spam*

# IR Systems: Schematic Depiction



# IR systems: Operations

- **Indexing**: assemble and interpret contents of information items (documents)
  - Most of the information in such documents is in the form of text which relatively unstructured
  - Efficient indexing is of much importance (**inverted indexes**)
- **Retrieval process**: generate a ranking that reflects relevance
  - A ranked list of documents returned according to a likelihood of relevance to the user
- Notion of **relevance** is most important
  - Relevance judgment  
(using **clickthrough data**? how to interpret **clickthrough data** as an indicative of relevance in an unsupervised manner?)
- The other important issues
  - Vocabulary mismatch problems
  - Evaluations of retrieval performance



# IR systems: Distinctions

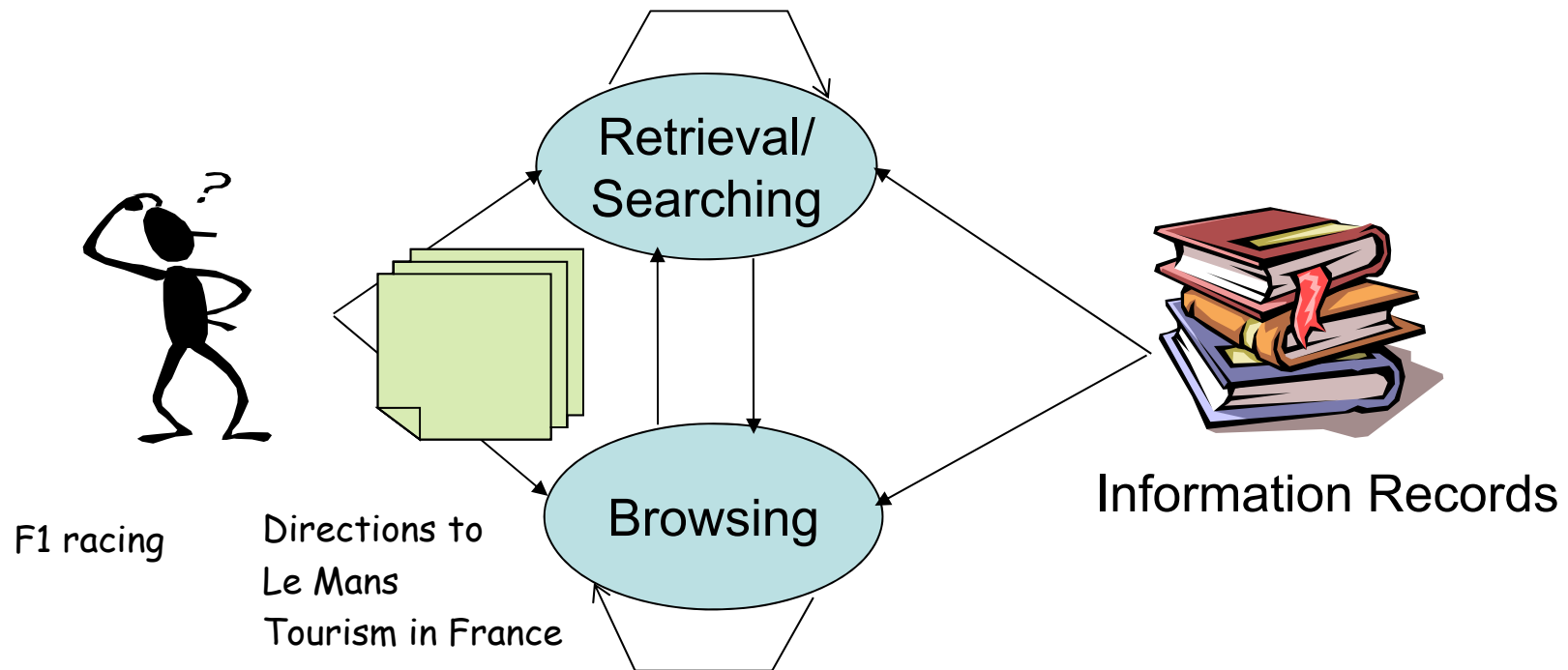
- IR systems can also be distinguished by the scale at which they operate
  - *Web search* (containing billions (or even trillions) of documents)
  - *Enterprise, institutional, and domain-specific search*
  - *Personal (desktop) search*
  - *Peer-to-peer (P2P) search*
  - *Forum search*
  - *Literature search*
  - *....*

# IR Main Issues

- The effective retrieval of relevant information affected by
  - The user task
    - Retrieval/searching and browsing
  - Logical view of the documents
    - Full-text/Keyword-based (text) operations; Indexing

# 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

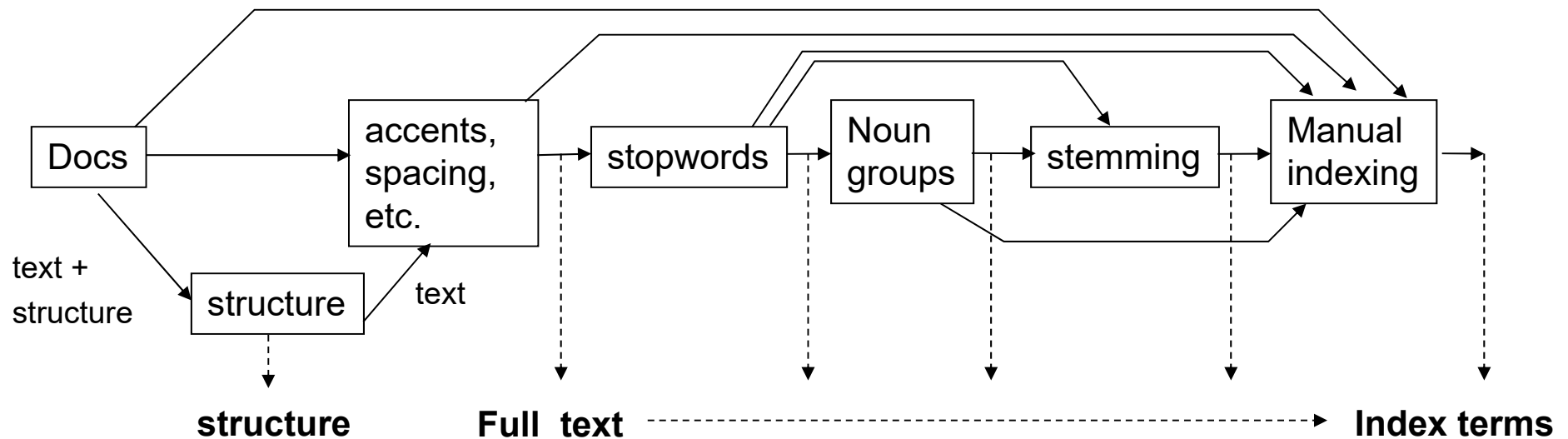


# Logical View of the Documents (1/2)

- 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

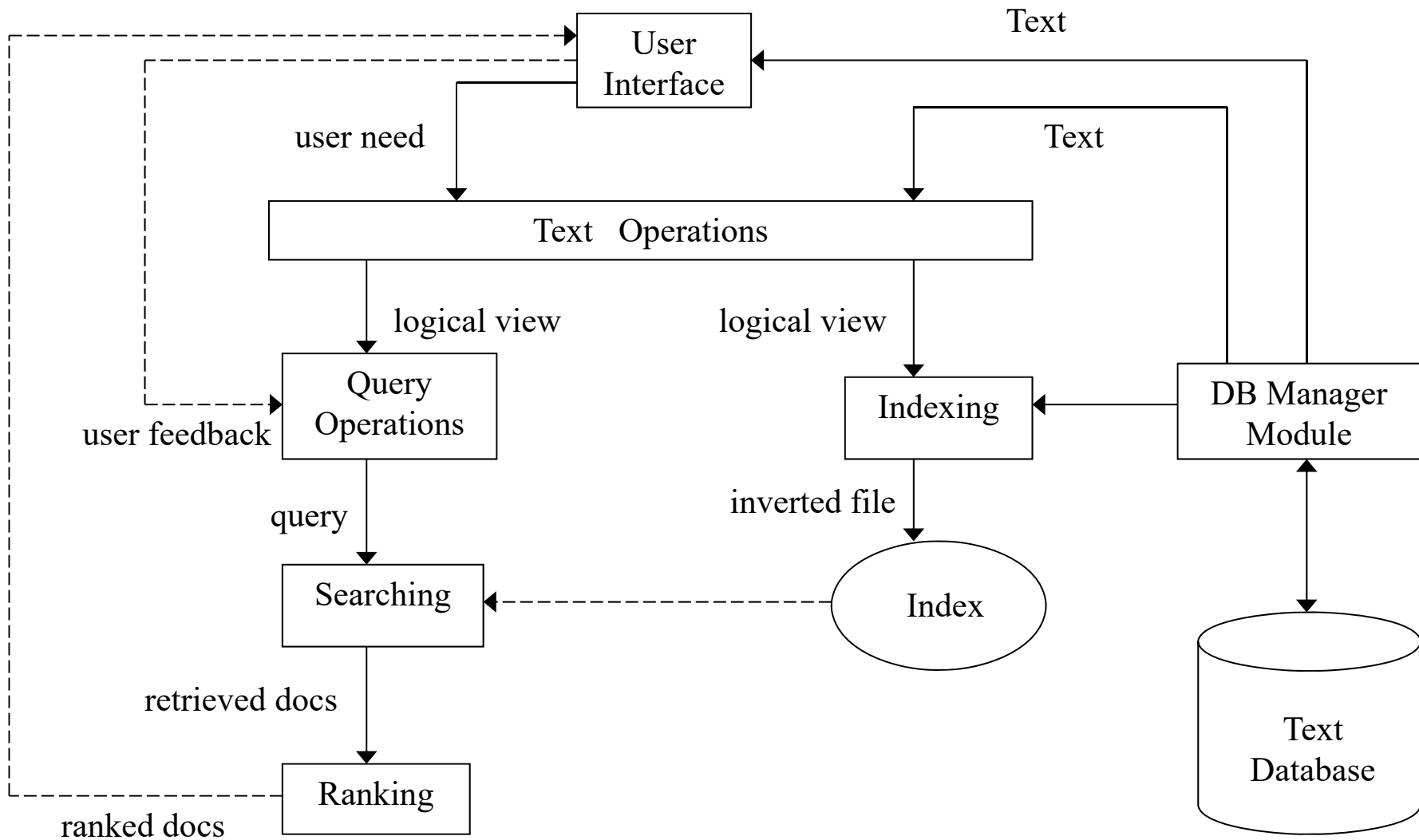
- Computer-centered (commercial perspective)
  - Efficient indexing approaches
  - High-performance ranking (matching) algorithms
- Human-centered (academic perceptive)
  - Studies of user behaviors
  - Understanding of user needs

} 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)

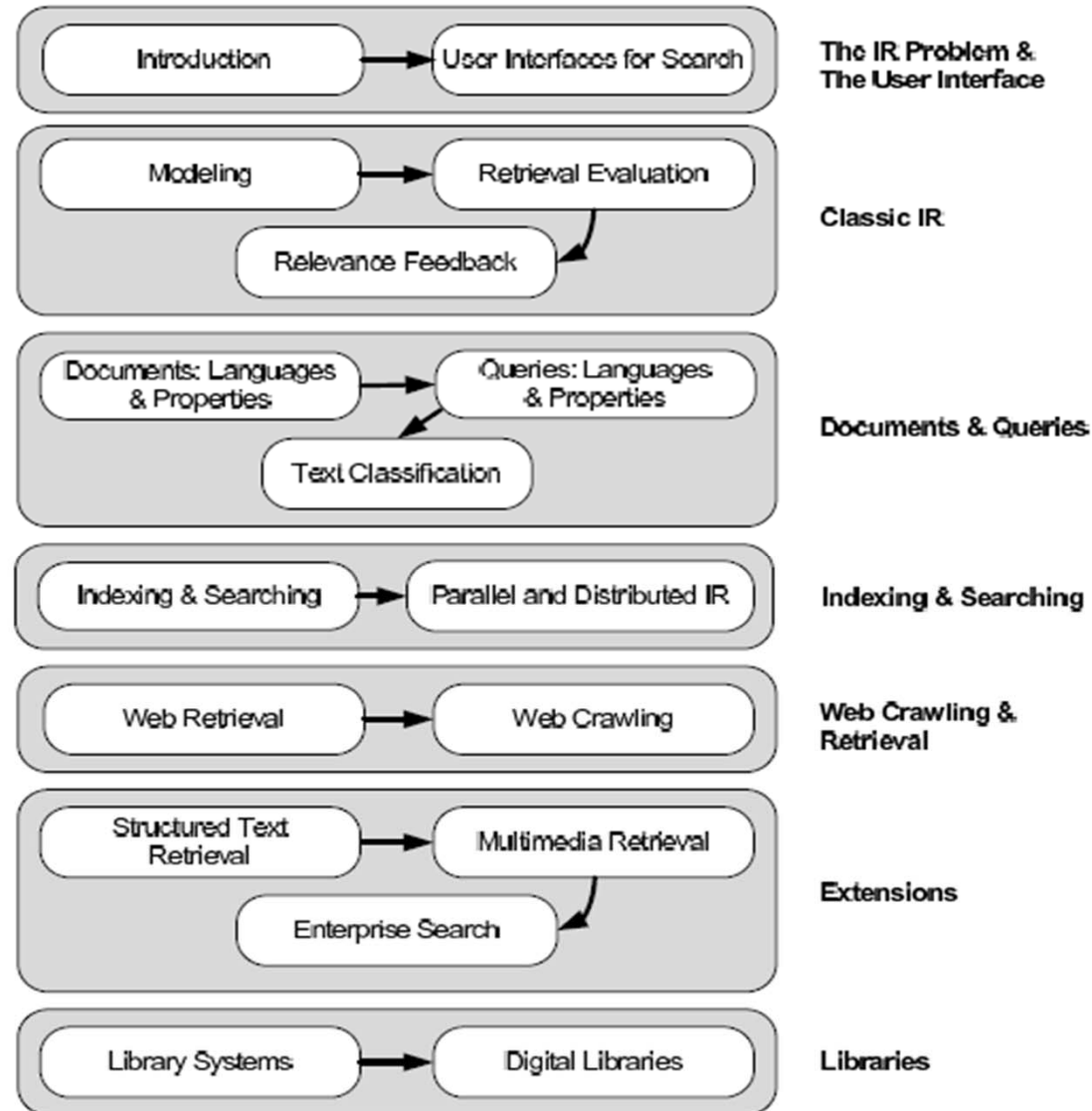
- In current retrieval systems
  - Users almost never declare his information need
    - Only a short queries composed few words (typically fewer than 4 words)
  - Users have no knowledge of the text or query operations

Poor formulated queries lead to poor retrieval !

# Major Topics (1/2)

- 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
  - Can multimedia be treated in a unified manner?
- Applications
  - Web, bibliographic systems, digital libraries, internet of things (IOT), among others

# Major Topics (2/2)



# Some Directions of Information Retrieval (1/2)

Example of Content	Example of Applications	Examples of Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned documents	(Personal) Desktop search	Question answering
Audio (Speech & Music)	Peer-to-peer search	

- In the past, most technology for searching non-text document relies on the descriptions of their content rather than the contents themselves
  - The need of “*content-based*” image/audio/music retrieval!
- In **vertical search** the domain of the search is restricted to particular topics
- **Enterprise search** is to find the required information in the huge variety of computer files scattered across a corporate intranet
- **Peer-to-peer search** involves finding information in networks of nodes or computers without any centralized control

# Some Directions of Information Retrieval (2/2)

- For **ad hoc retrieval**, retrieval is based on a user query, where the range of possible queries is huge and not prespecified
- **Filtering** involves detecting stories/documents of interest based on a person's interests (profile) and providing an alert using email or some other mechanism.
- **Classification** (categorization) uses a defined set of labels or classes and automatically assigns those labels to documents
- **Question answering** is similar to search (or ad hoc retrieval) but is aimed at more specific questions, such as “***What is the height of Mt. Everest?***”

# Core IR Issues and Search Engine Design

## Information Retrieval

### Relevance

*-Effective ranking*

### Evaluation

*-Testing and measuring*

### Information needs

*-User interaction*



## Search Engines

### Performance

*-Efficient search and indexing  
(response time throughput,  
indexing speed)*

### Incorporating new data

*-Coverage and freshness*

### Scalability

*-Growing with data and users*

### Adaptability

*-Tuning for applications  
(customizable)*

### Specific problems

*-e.g. Spam*

# More on Relevance and Retrieval Models

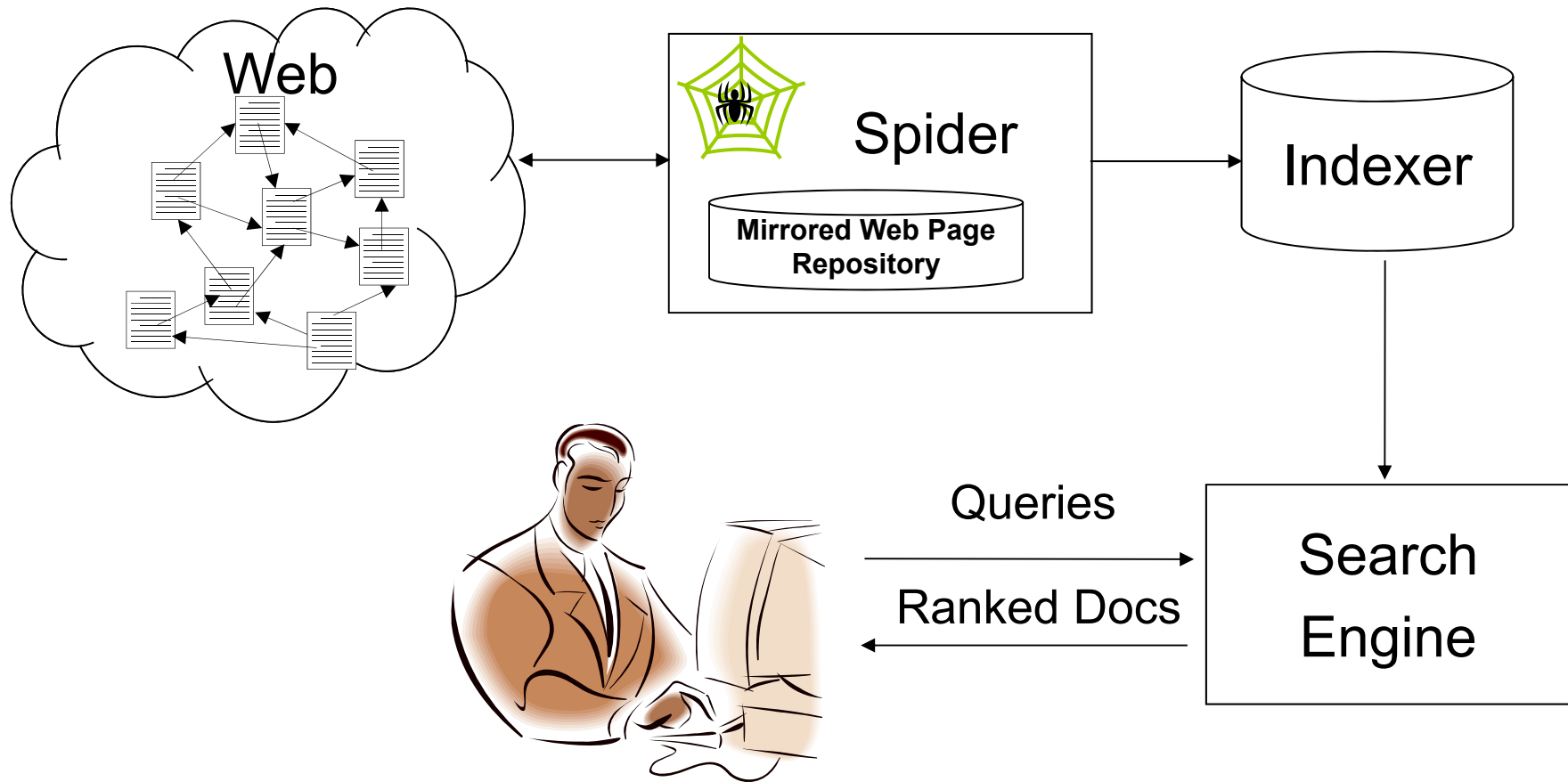
- Relevance
  - Loosely speaking, a relevant document contains the information that a person was looking for when he/she submitted a query to the search engine
  - Simply comparing the text of a query with the text of a document and looking for an exact match produces very poor results in terms of relevance
    - One obvious reason for this is that language can be used to express the same concepts in many different way, often with very different words

tornado vs. severe weather event

- Retrieval models
  - A retrieval model is a formal representation of the process of matching a query and a document
  - It is the basis of the **ranking algorithm** that is used in a search engine to produce the ranked list of documents

# 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> (Service is No Longer Available)

Openfind Taiwan Webpage Search: 觀霧 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

← 上一頁 → 搜尋 ★ 我的最愛 媒體

網址(D)  移至 連結 Customize Links Free Hotmail

Norton AntiVirus

**Openfind** 免費撥接服務 電話號碼: 40508888 使用名稱: openfind 密碼: openfind

網頁 BBS文章 新聞 分類 圖片 音樂 軟體 文件

不限日期 查詢 進階 - 喜好 - 說明

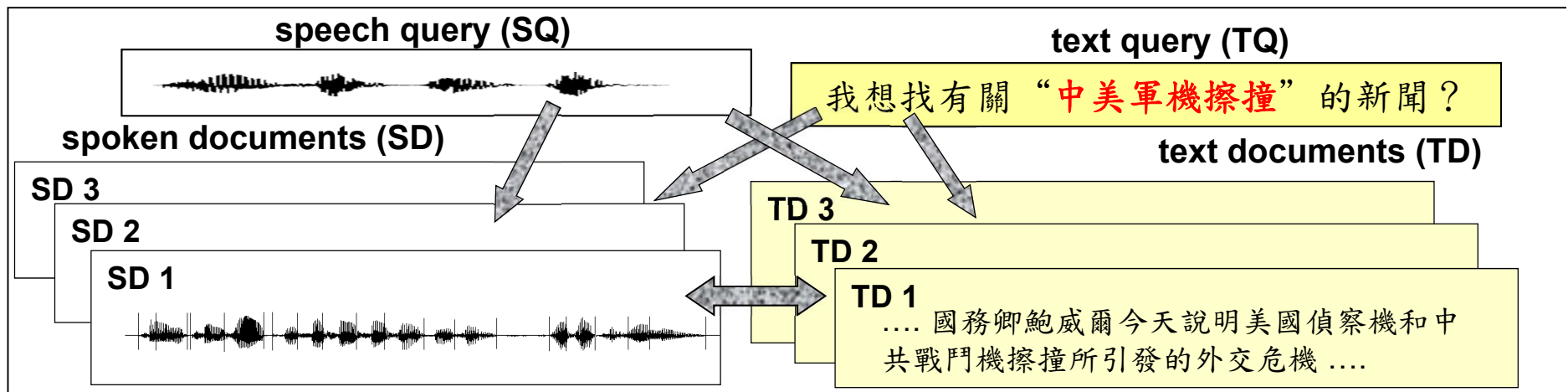
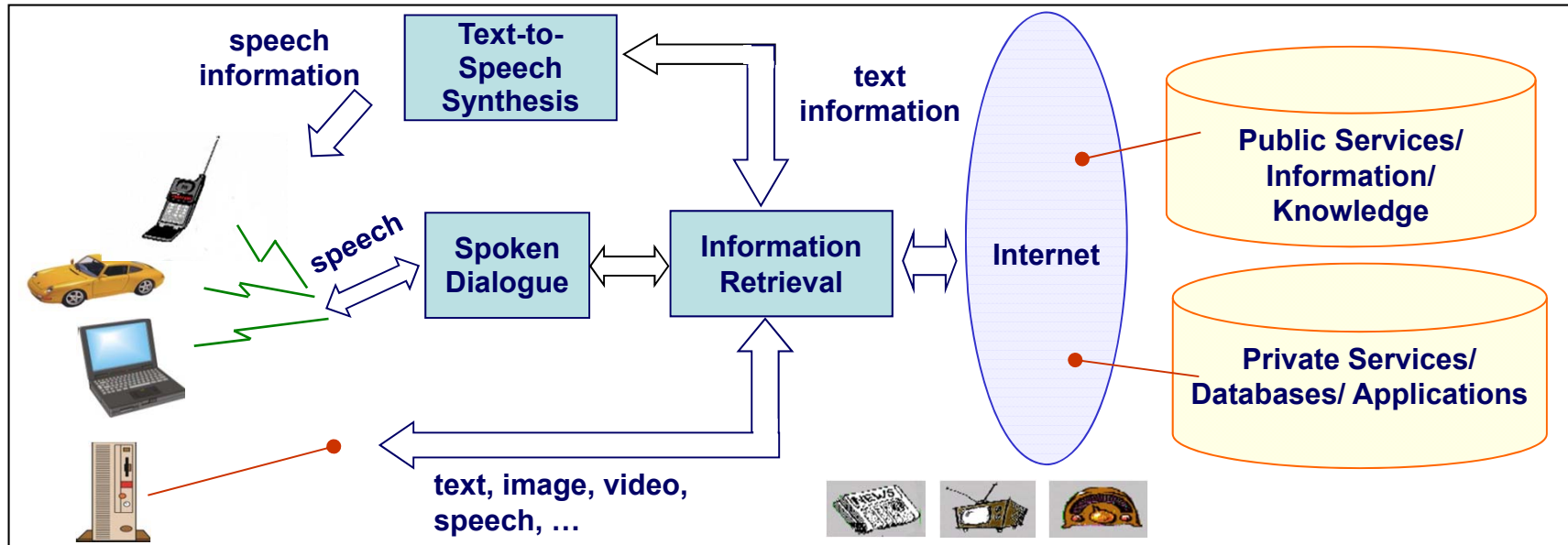
相關查詢 8 筆 · [雪霸](#) · [雪霸國家公園](#) · [大霸尖山](#) · [林道](#) · [竹東](#) · [觀霧山莊](#) · [觀霧之旅](#) · [觀霧農場](#)

Openfind 找到 5,594 篇相關網頁 [有效增加網站曝光](#)

- 1. 觀霧農莊**  
介紹農莊風景及其服務項目、交通指南、住宿方式等。公司名稱: ...  
<http://tree.2u.com.tw/> - 2002/12/11, 16k - [ [關鍵字](#) ] [ [更多結果](#) ]
- 2. 瀑布谷農場**  
自然休閒-擁抱山水-到雲海的舞台觀霧 | 瀑布谷農場介紹 | | 交通路線圖 | | 旅遊注意事項  
| 觀霧是雲的故鄉, 景色千變萬化, 體驗大自然、賞... 農場也準備卡拉OK讓您高歌一曲。  
注意事項×觀霧地區日夜溫差大請多加保暖衣物、請攜帶證件...  
簡介 - 介紹位在雪霸國家公園觀霧的瀑布谷農場, 經營民宿、餐飲、水密...  
<http://ppg.2u.com.tw/> - 2002/06/04, 2k - [ [庫存頁面](#) ] [ [關鍵字](#) ]
- 3. 觀霧雲山農場**  
觀霧雲山農場位在雪霸國家公園內, 提供遊客餐飲及住宿服務。公司名稱: 觀霧雲山農場  
公司地址: 新竹縣五峰鄉觀霧山莊村石362號 → 1 公司電話:



# Speech Information Retrieval (1/4)





# Speech Information Retrieval (2/5)

- HP Research Group – Speechbot System  
(Service is No Longer Available)
  - 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/>)

HP SpeechBot - Microsoft Internet Explorer

http://speechbot.research.compaq.com

United States-English

» HP Home » Products & Services » Support & Drivers » Solutions » How to Buy

» Contact HP

Search:

HP Labs  All of HP US

**hp**  
invent

**SpeechBot™**  
audio search using speech recognition

» HP Labs

» Research

» News and events

» Technical reports

» About HP Labs

» Careers @ HP Labs

» People

» Worldwide sites

» Cambridge Research Lab

» Downloads

**Search** » Power Search » Help

Search for:

Topics:  Dates:

**Tip:** An asterisk \* at the end of a partial word will match all words starting with the partial word (e.g. "surf\*" matches "surfers", "surfs" etc.)

SpeechBot is a search engine for audio & video content that is hosted and played from other websites (listed below). **Note:** Transcripts of the content based on speech recognition are not exact.

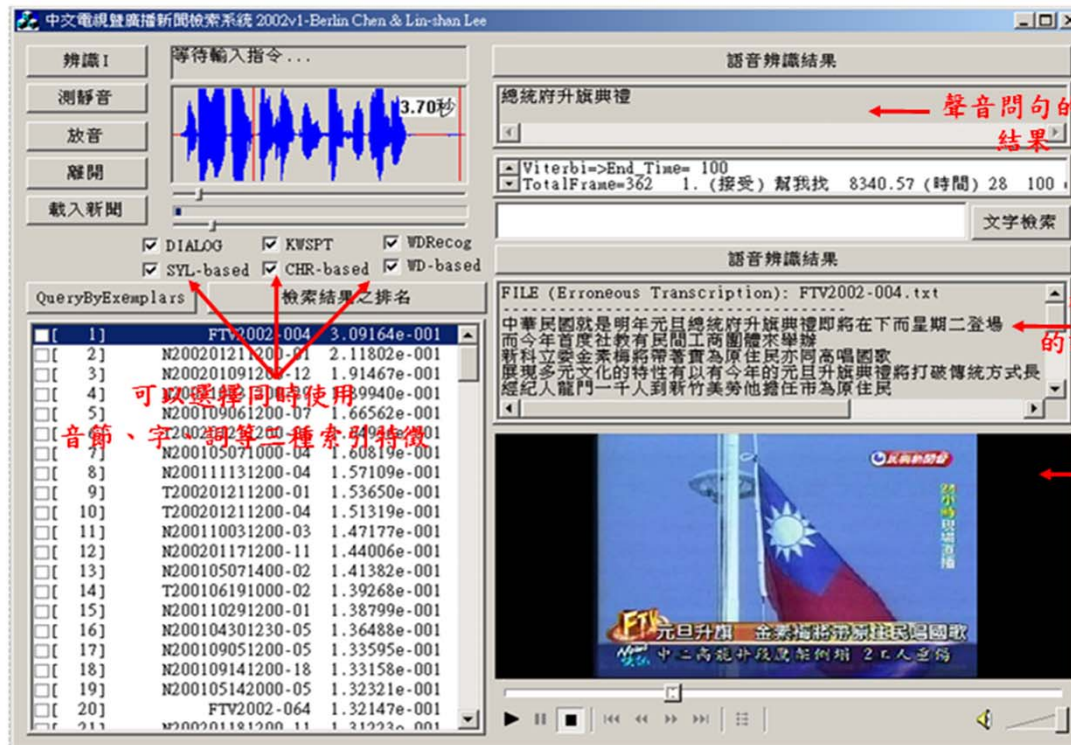
SpeechBot currently indexes **14791 hours of content** from the following websites:

<b>Arts &amp; Entertainment</b> <ul style="list-style-type: none"><li>» Fresh Air</li></ul>	<b>Government &amp; Military</b> <ul style="list-style-type: none"><li>» AFRTS Radio News</li><li>» The White House</li><li>» U.S. Department of Defense Briefings</li></ul>	<b>Sports</b> <ul style="list-style-type: none"><li>» Only A Game</li><li>» Scuba Radio</li></ul>
<b>Current Events</b> <ul style="list-style-type: none"><li>» American RadioWorks</li><li>» Here and Now</li><li>» On Point</li><li>» PBS Online NewsHour</li></ul>	<b>Music</b> <ul style="list-style-type: none"><li>» Soundcheck</li></ul>	<b>Talk</b> <ul style="list-style-type: none"><li>» Car Talk Radio Show</li><li>» Public Interest</li><li>» The Brian Lehrer Show</li><li>» The Charlie Rose Show</li><li>» The Connection</li><li>» The Diane Rehm Show</li></ul>
	<b>Personal Investment</b> <ul style="list-style-type: none"><li>» Marketplace Radio</li></ul>	

# Speech Information Retrieval (3/5)

- Speech Summarization and Retrieval

輸入聲音問句：“請幫我查總統府升旗典禮”



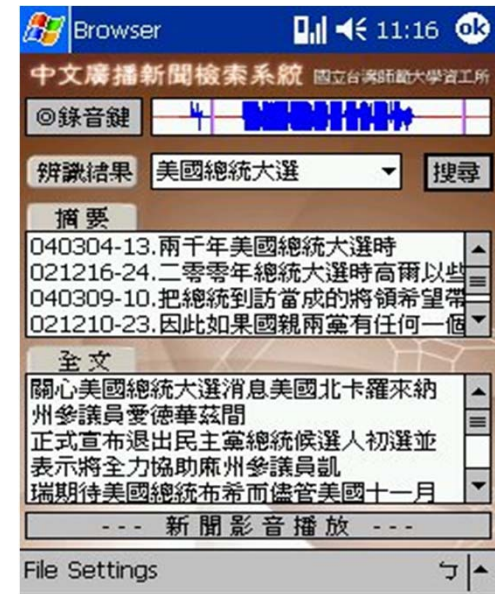
可以同時使用音節、字詞等三種索引特徵

聲音問句的語音辨識結果

檢索到新聞的語音辨識結果

檢索到新聞的影音

中文影音多媒體資訊檢索雛形展示系統。



- B. Chen, H.-M. Wang, L.-S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," IEEE Transactions on Speech and Audio Processing, July 2002.
- B. Chen, Yi.T.Chen, C.-H. Chang, H.-B. Chen, "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," Interspeech 2005

# Speech Information Retrieval (4/5)

- Speech Organization

**廣播新聞搜尋瀏覽系統**  
Broadcast News Retrieval/Browsing System

**(a)** 國外政治 [International Political News] Topic Map  
 國內政治 [Local Political News] Topic Map  
 國外財經 [International Business] Topic Map  
 國內財經 [Local Business] Topic Map  
 國外影劇 [International Entertainment] Topic Map  
 國內影劇 [Local Entertainment] Topic Map  
 國外體育 [International Sports] Topic Map  
 國內體育 [Local Sports] Topic Map

**(b)** 伊拉克 巴格達  
 美軍 陸戰隊  
 以色列 阿拉法特  
 巴勒斯坦 迦薩市

**(c)** 國土安全部 民航機  
 蓋達組織 中情局  
 聯合國 安理會  
 武檢人員 武器

**(d)** 阿拉法特 阿巴斯  
 雷馬拉 任命  
 以色列 夏隆  
 約旦河 美國  
 中東 鮑爾  
 和平 路線  
 巴格達 炸彈  
 自殺 巴士

**(e)** [ ] 1 以色列結束對阿拉法特總部的包圍 [sum.] 02.09.21  
 [x] 2 阿拉法特反對以色列保所提結束包圍條件 [sum.] 02.09.21  
 [ ] 3 以色列部隊進攻阿拉法特總部後撤軍 [sum.] 02.10.22  
 [ ] 4 以色列結束對阿拉法特總部的包圍 [sum.] 02.10.01  
 [ ] 5 以色列坦克撤出阿拉法特辦公室 [sum.] 02.09.21  
 [ ] 6 以色列與巴勒斯坦展開安全問題會議 [sum.] 02.11.23  
 [ ] 7 以色列在加薩擊斃一名回教聖戰組織領袖 [sum.] 02.06.05  
 [ ] 8 以色列巴勒斯坦就伯利恆撤軍達成協議 [sum.] 02.02.12  
 [ ] 9 以色列坦克闖入加薩難民營 兩人喪生 [sum.] 02.04.21

阿拉法特原則接受歐盟所提中東和平計畫 [summary]  
 (May 03/02/12:00)  
 英美就解決阿拉法特所受包圍與巴方展開談判 [summary]  
 (May 06/02/12:00)  
 阿拉法特反對以色列保所提結束包圍條件 [summary]  
 (Sep 20/02/12:00)  
 阿拉法特宣布新內閣引發巴勒斯坦國會激辯 [summary]  
 (Oct 30/02/12:00)  
 阿拉伯人支持阿拉法特及巴勒斯坦人正當抵抗 [summary]  
 (Nov 02/02/12:00)

go to Level-1  
 go to Level-2

- L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 2005.

# Speech Information Retrieval (5/5)

- Google, Apple and Microsoft's Deployed Services



Google Voice Search

<http://www.google.com/mobile/voice-search/>



Apple Siri

<http://www.apple.com/iphone/features/siri.html>



Microsoft Cortana

[http://zh.wikipedia.org/wiki/Microsoft\\_Cortana](http://zh.wikipedia.org/wiki/Microsoft_Cortana)



# Visual Information Retrieval (1/4)

- Content-based approach

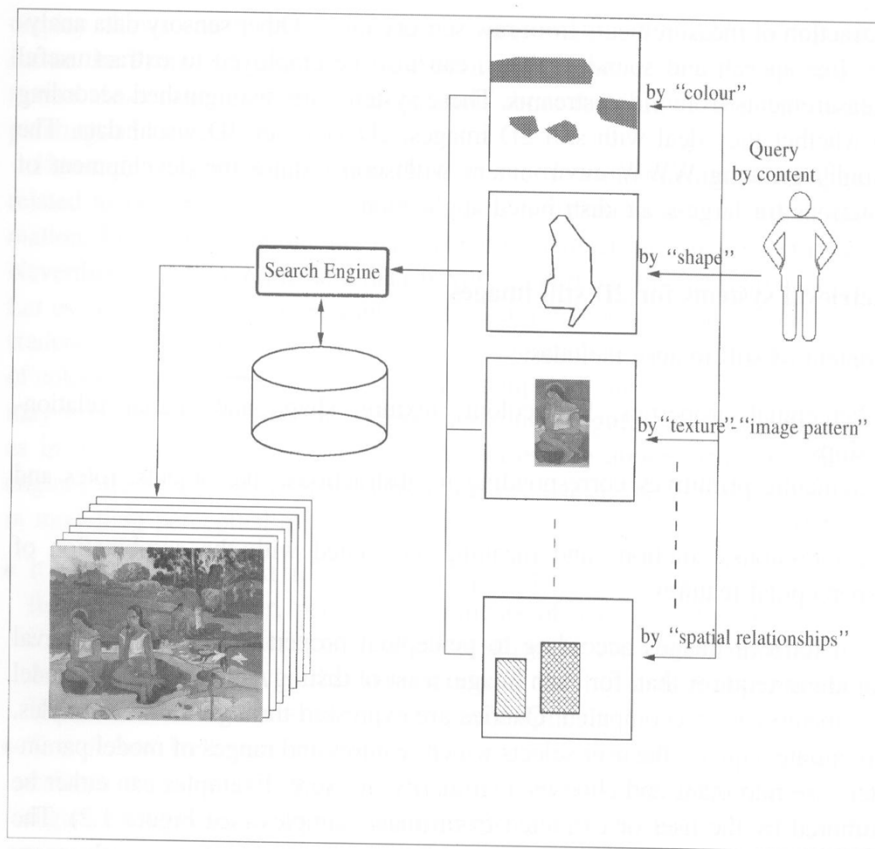


Figure 1.2 Different types of query by example.

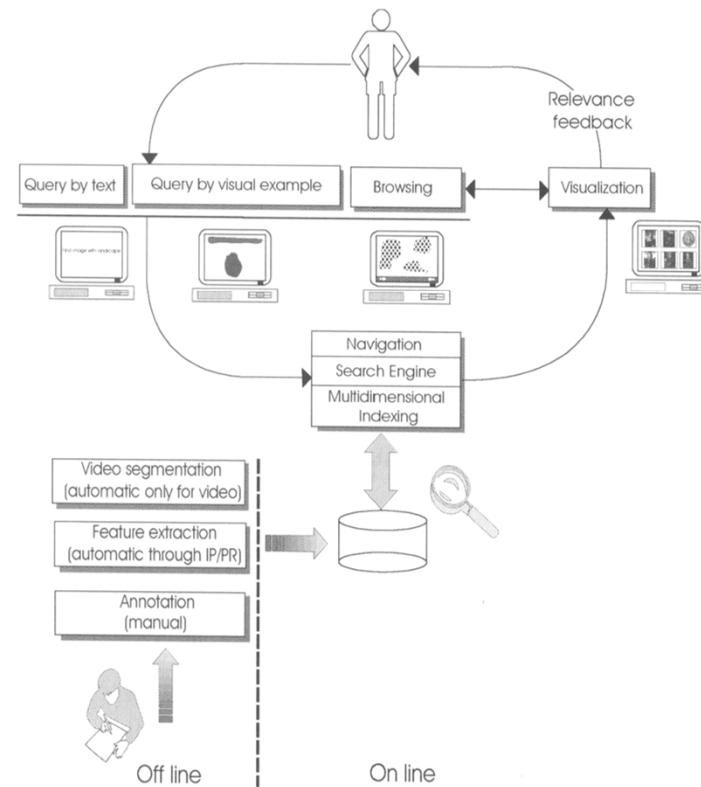
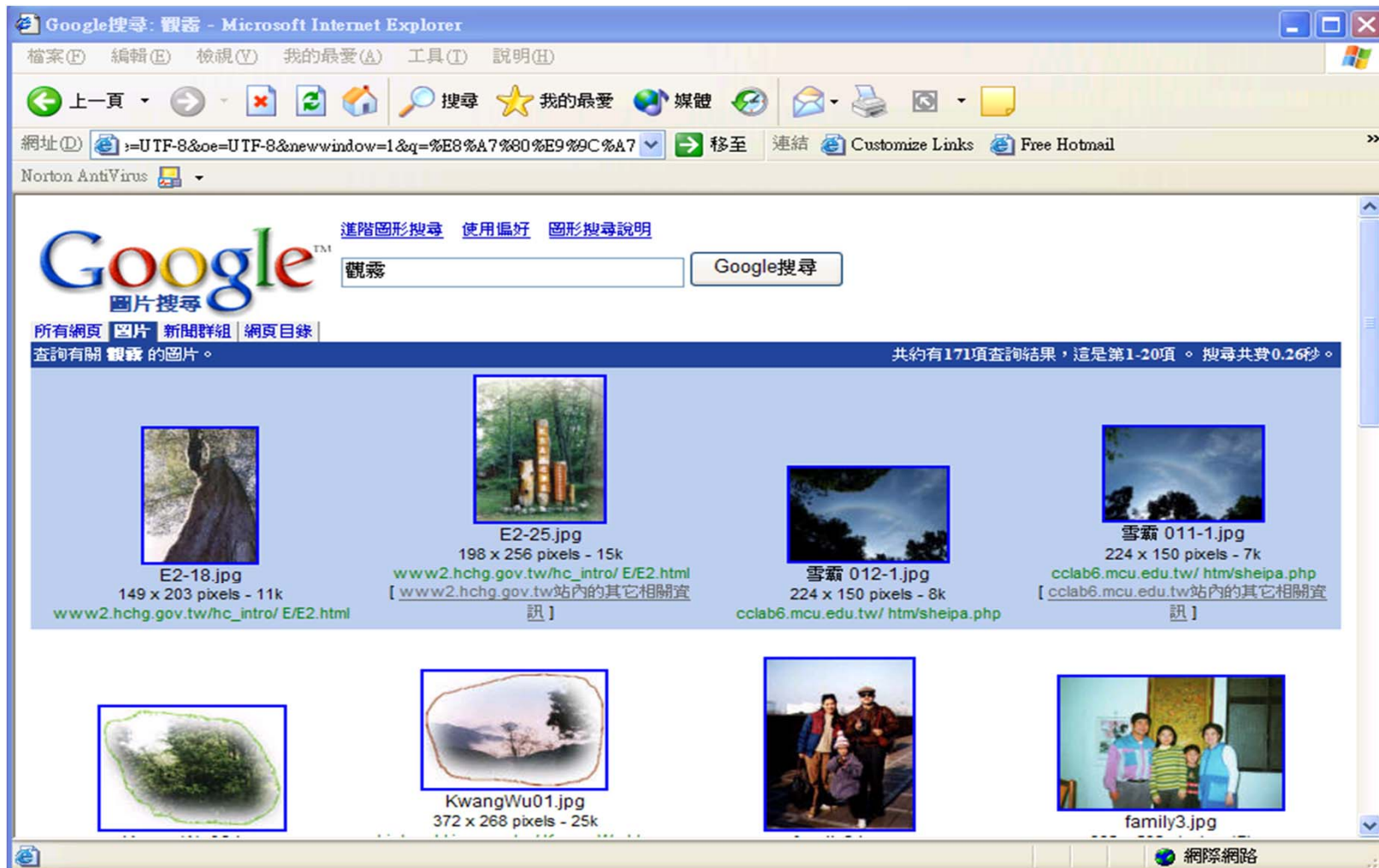


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

# Visual Information Retrieval (2/4)

- Images with Texts (Metadata)



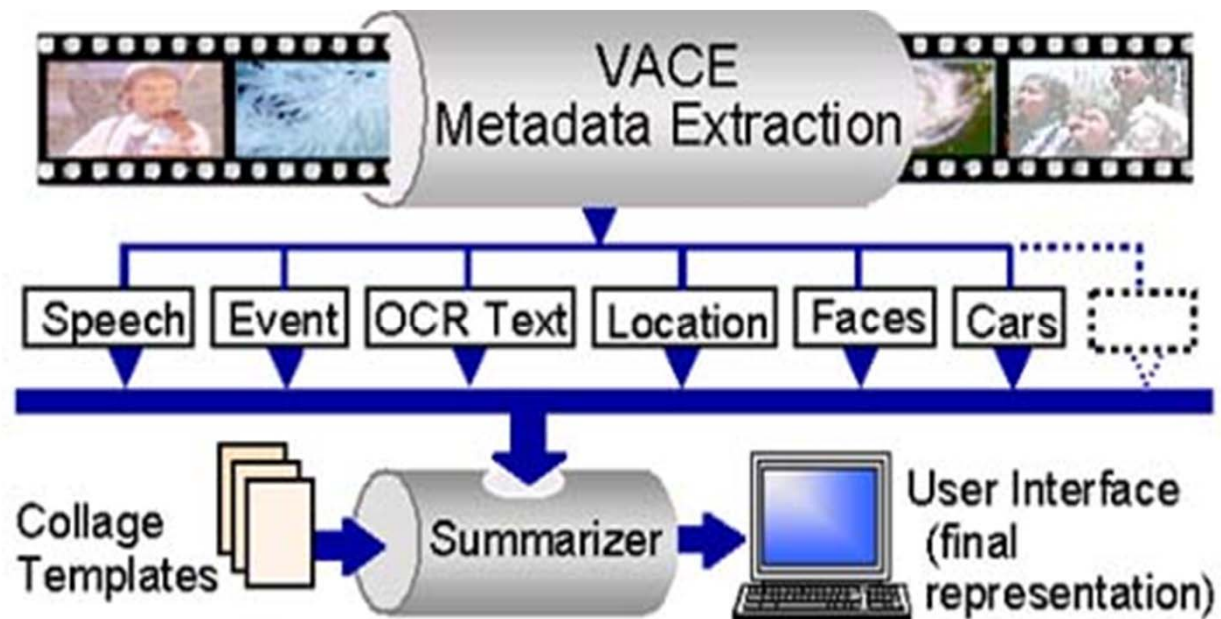
# Visual Information Retrieval (3/4)

- Content-based Image Retrieval

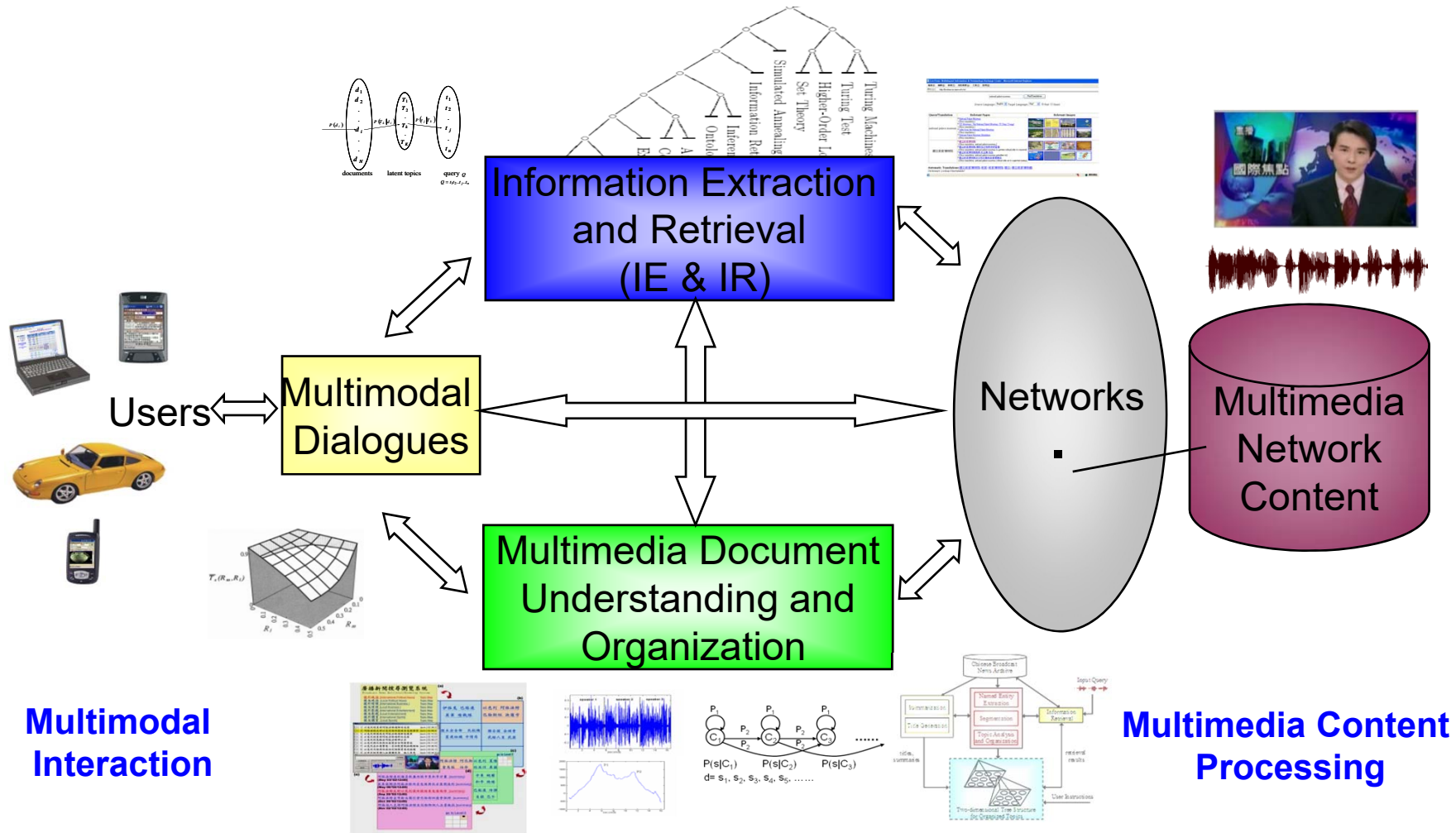


# Visual Information Retrieval (4/4)

## Video Analysis and Content Extraction



# Scenario for Multimedia information access

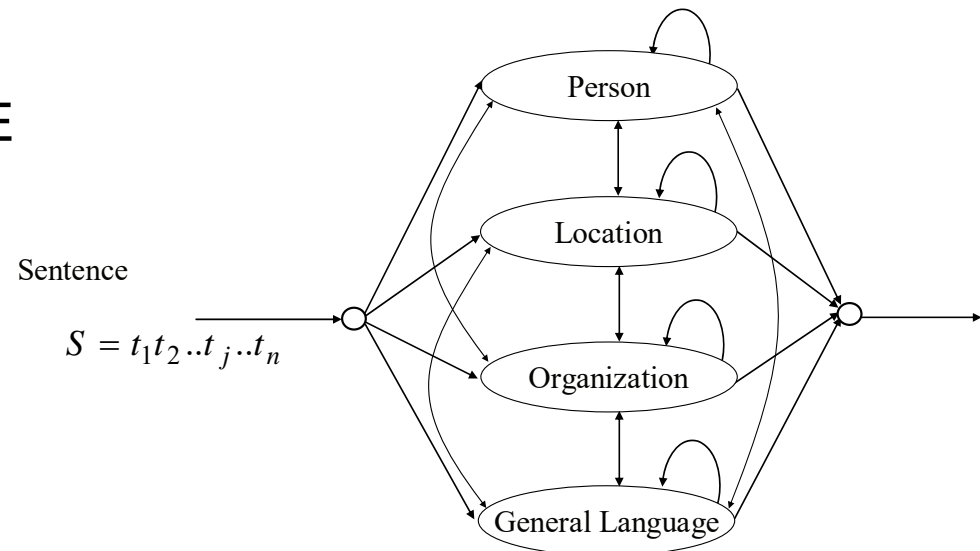


# Other IR-Related Tasks

- Information filtering and routing
- Term/Document categorization
- Term/Document clustering
- Crosslingual information retrieval
- Information extraction
- Document summarization
- Question answering
  - “*What is the height of Mt. Everest?*”
- .....

# Information Extraction

- E.g., Named-Entity Extraction
  - NE has its 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





# Cross-lingual 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

LiveTrans: Multilingual Information & Terminology Exchange Center - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

網址(D) http://livetrans.iis.sinica.edu.tw/

national palace museum FindTranslations

Source Language: English Target Language: Big5  Fast  Smart

Query/Translation	Relevant Pages	Relevant Images
national palace museum	<ul style="list-style-type: none"> <li>* <a href="#">National Palace Museum</a> [Gloss translation:]</li> <li>* <a href="#">TIT Museums: The National Palace Museum: 70 Years Young!</a> [Gloss translation:]</li> <li>* <a href="#">Jades from the National Palace Museum</a> [Gloss translation:]</li> <li>* <a href="#">National Palace Museum Exhibition</a> [Gloss translation:]</li> </ul>	
國立故宮博物院	<ul style="list-style-type: none"> <li>* <a href="#">國立故宮博物院</a> [Gloss translation: national palace museum,]</li> <li>* <a href="#">國立故宮博物院 預防性文物保存研習會</a> [Gloss translation: national palace museum to prevent cultural relic to conserve]</li> <li>* <a href="#">國立故宮博物院院長 杜正勝 先生</a> [Gloss translation: national palace museum president sir]</li> <li>* <a href="#">國立故宮博物院古文物及藝術品管理辦法</a> [Gloss translation: national palace museum cultural relic art to supervise means]</li> </ul>	

Machine-Extracted Translation

Automatic Translations: [國立故宮博物院](#); [故宮](#); [故宮博物院](#); [國立](#); [國立故宮博物館](#);  
Dictionary Lookup: Unavailable!

網際網路

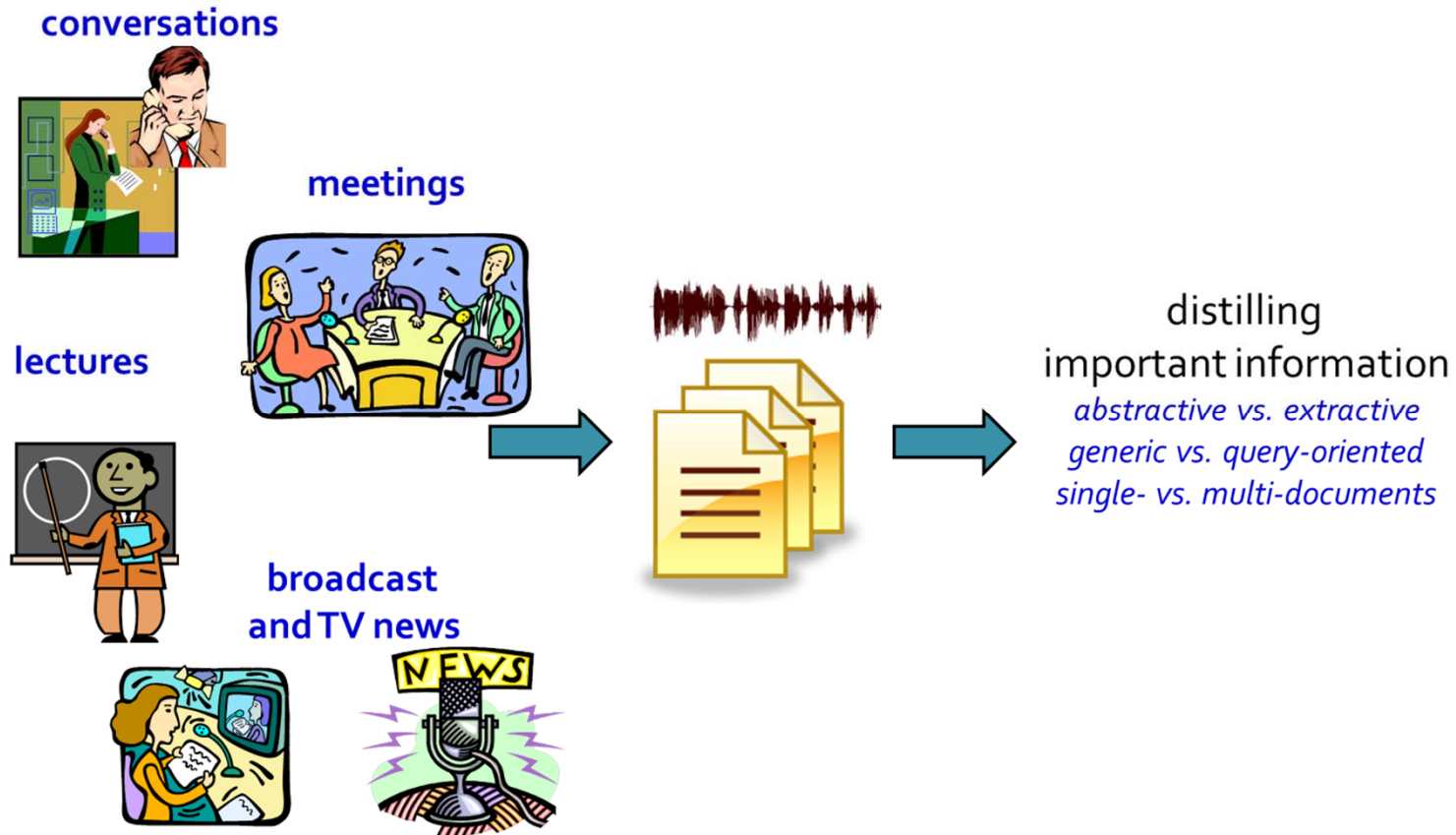


# Document Summarization (1/2)

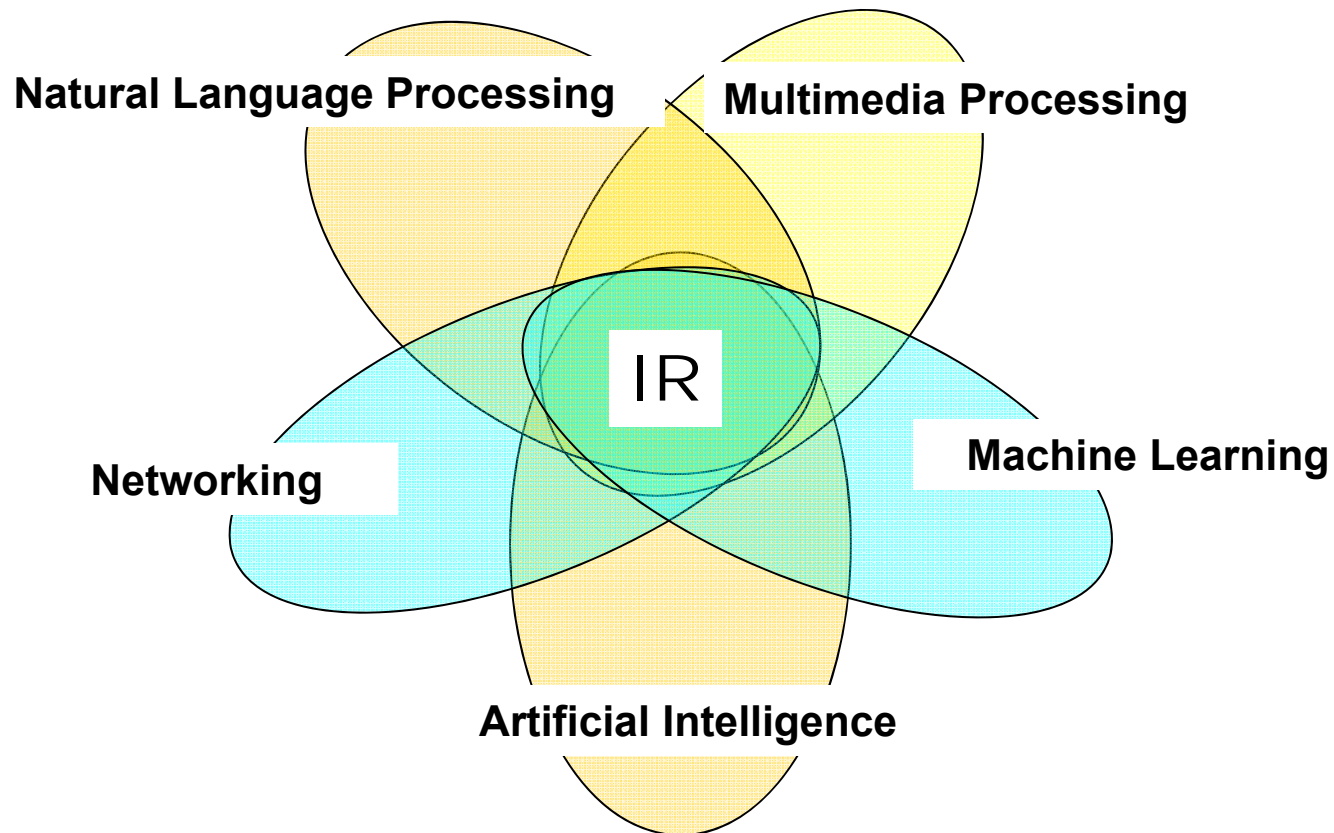
- Audience
  - Generic summarization
  - User-focused summarization
    - Query-focused summarization
    - Topic-focused summarization
- Function
  - Indicative summarization
  - Informative summarization
- Extracts vs. abstracts
  - Extract: consists wholly of portions from the source
  - Abstract: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

# Document Summarization (2/2)

- Speech Summarization

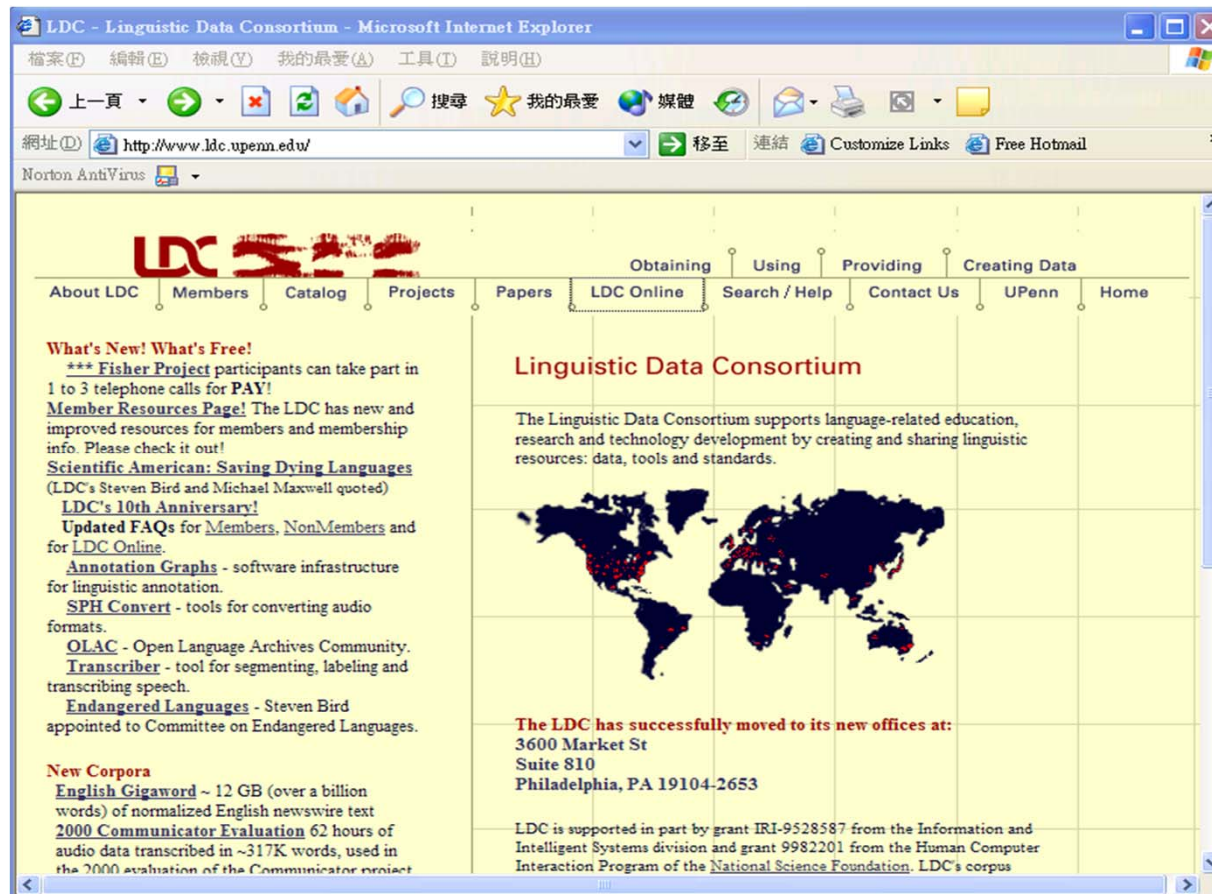


# Multidisciplinary Approaches



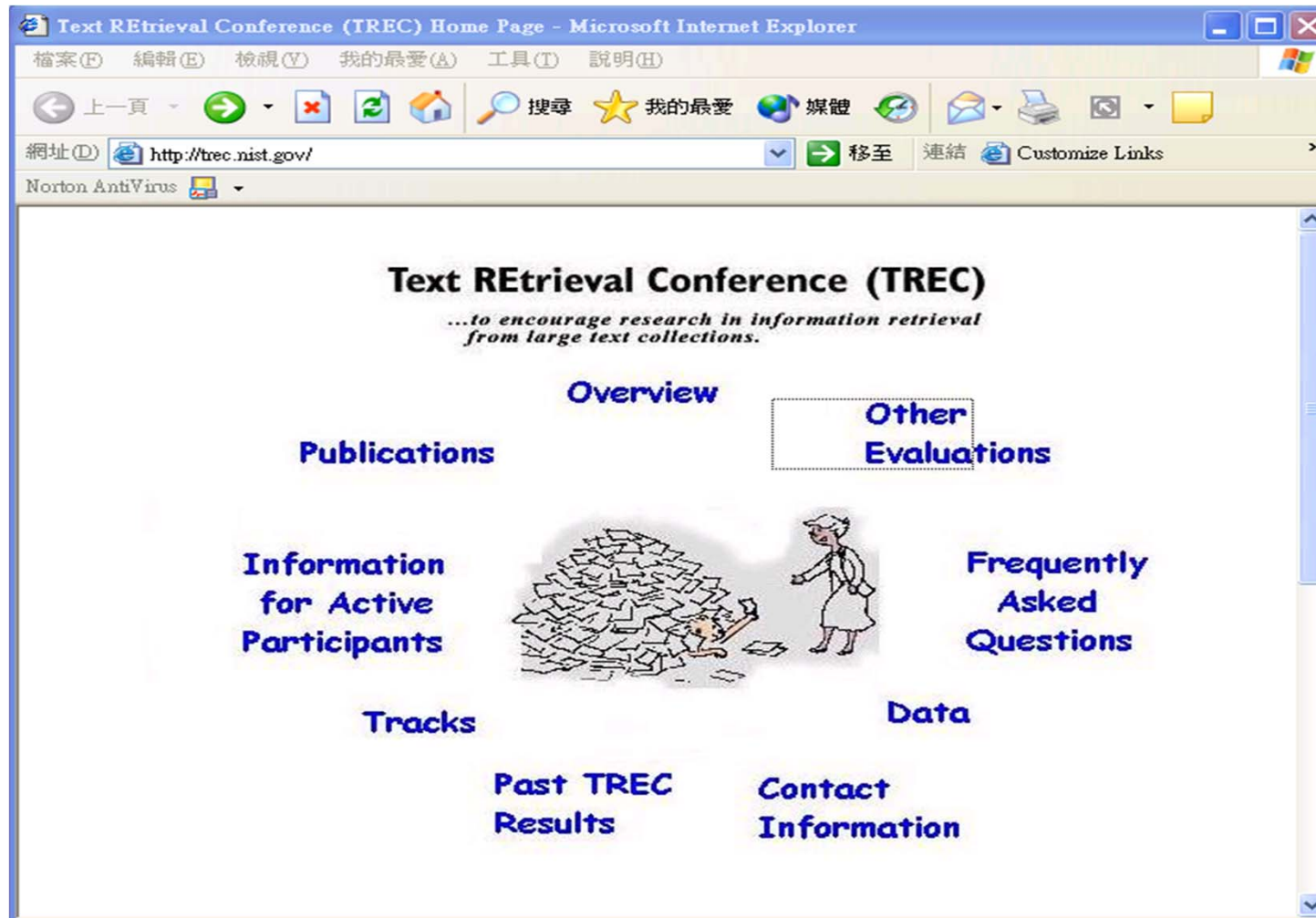
# Resources

- Corpora (Speech/Language resources)
  - Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them
    - [LDC - Linguistic Data Consortium](http://www ldc.upenn.edu/)



# Contests (1/2)

- [Text REtrieval Conference \(TREC\)](http://trec.nist.gov/)





# Contests (2/2)

- US National Institute of Standards and Technology

The screenshot shows a Microsoft Internet Explorer browser window displaying the NIST Benchmark Tests website. The browser's address bar shows the URL <http://www.nist.gov/speech/tests/index.htm>. The website content is organized into several sections:

- Conversational Telephone Recognition**
  - 2001 HUB-5 Evaluation Plan, multiple languages
  - 2000 HUB-5 Evaluation Plan, multiple languages
  - 1998 HUB-5 English Evaluation
  - 1997 HUB-5NE Evaluation
  - 1997 HUB-5E Evaluation
- Topic Detection and Tracking (TDT)**
  - General Information
  - TDT 2004 Evaluation
  - TDT 2003 Evaluation
  - TDT 2002 Evaluation
  - TDT 2001 Evaluation
  - TDT 2000 Evaluation
  - 1999 TDT3 Evaluation
  - 1998 TDT2 Evaluation
- Machine Translation**
  - General Information
- Information Extraction - Entity Recognition:**
  - 2002 ACE-Evaluation
  - 2001 ACE-Evaluation
  - 2000 ACE - Evaluation
  - 1999 Information Extraction - Entity Recognition Evaluation
- Spoken Document Retrieval**
  - 2000 TREC Spoken Document Retrieval Track Evaluation
  - 1999 TREC Spoken Document Retrieval Track Evaluation
  - 1998 TREC Spoken Document Retrieval Track Evaluation
  - 1997 TREC Spoken Document Retrieval Track

The NIST logo and "National Institute of Standards and Technology" are visible in the top left corner of the page. A "Contact Webmaster" link is also present. The browser's taskbar at the bottom shows the system clock and network status.

# Conferences/Journals

- Conferences

- ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR )
- ACM Conference on Information Knowledge Management (CIKM)
- ...

- Journals

- Journal of the American Society for Information Science (JASIS)
- ACM Transactions on Information Systems (TOIS)
- Information Processing and Management (IP&M)
- ACM Transactions on Asian Language Information Processing (TALIP)
- ...

# Tentative Topic List

Course Overview & Introduction
Retrieval Models (I) - Classic Retrieval Models (Boolean, Vector Space and Probabilistic Models)
Retrieval Performance Evaluation - Measures
Retrieval Performance Evaluation - Collections
Retrieval Models (II) - Improved Approaches (Fuzzy Set, Extended Boolean, Generalized Vector Space Models)
Query Operations (Query Expansion and Term Re-weighting)
Retrieval Models (III) - Latent Semantic Analysis (LSA)
Retrieval Models (IV) - Language Models
Retrieval Models (V) - Learning to Rank
Clustering for Information Retrieval
Classification for Information Retrieval
Efficient Indexing and Searching
Web Search Basics
Cross-lingual Information Retrieval
Spoken Document Recognition, Retrieval and Summarization



# Grading (Tentative)

- Midterm (or Final): 30%
- Homework/Projects: 40%
- Presentation: 15%
- Attendance/Other: 15%