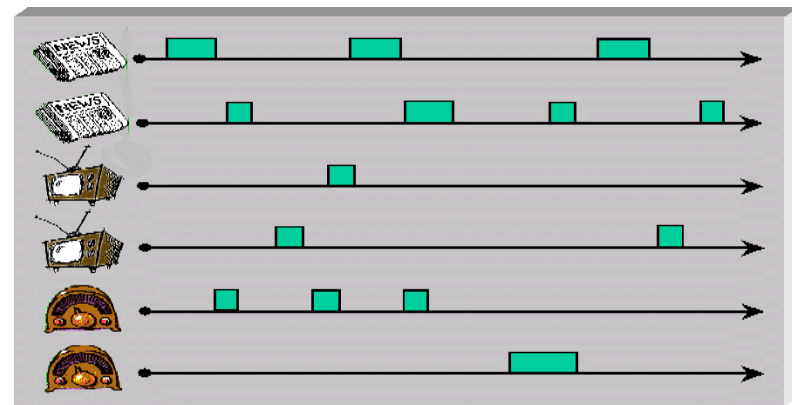


# Information Retrieval and Extraction

Berlin Chen 2005



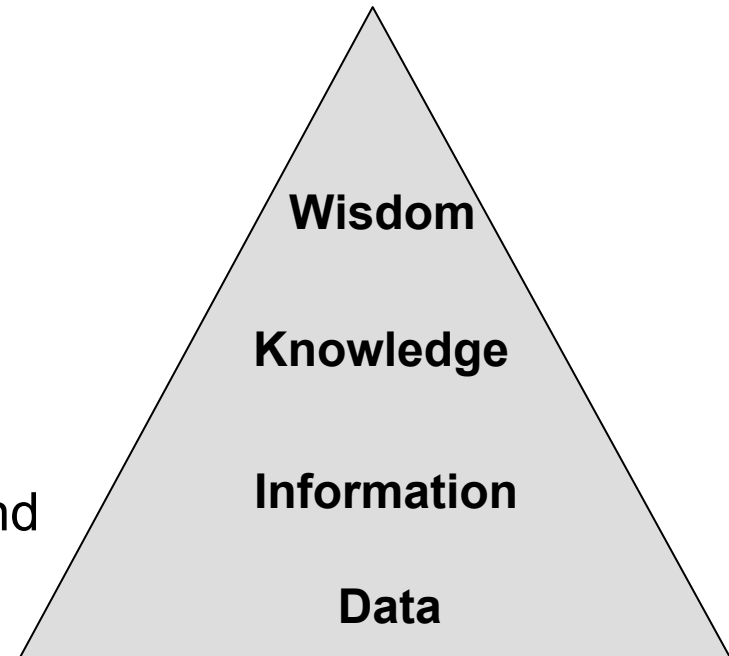
(Picture from the [TREC](#) web site)

# 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



Search engine/IR system

Emphasis is on the retrieval of information (not data)

# 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:
  - Modeling, 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

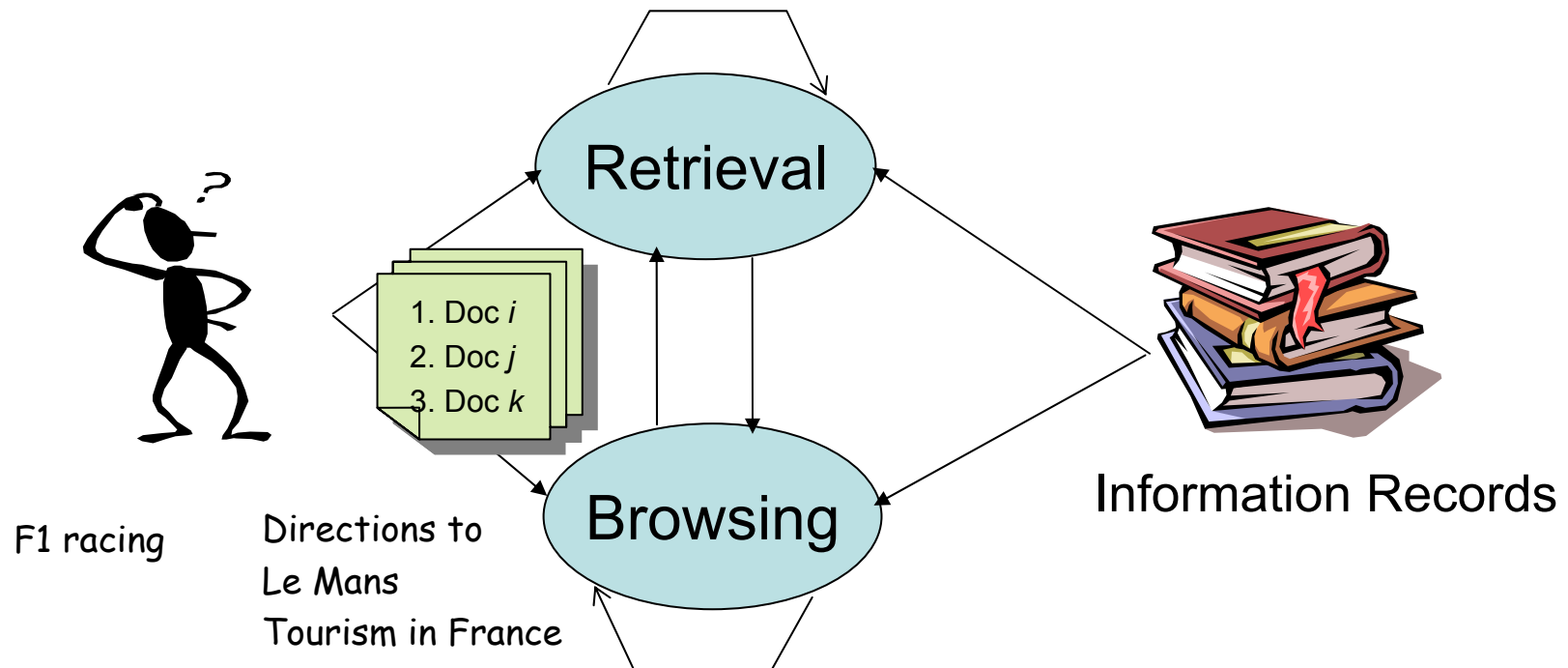


# IR Main Issues

- 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

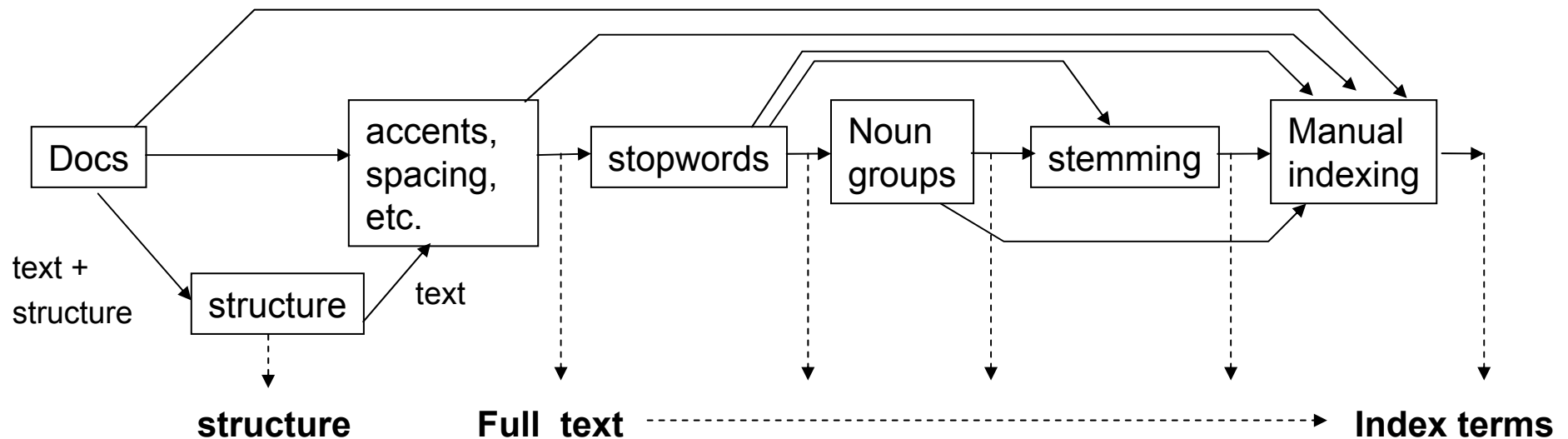


# Logical View of the Documents (1/2)

- 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

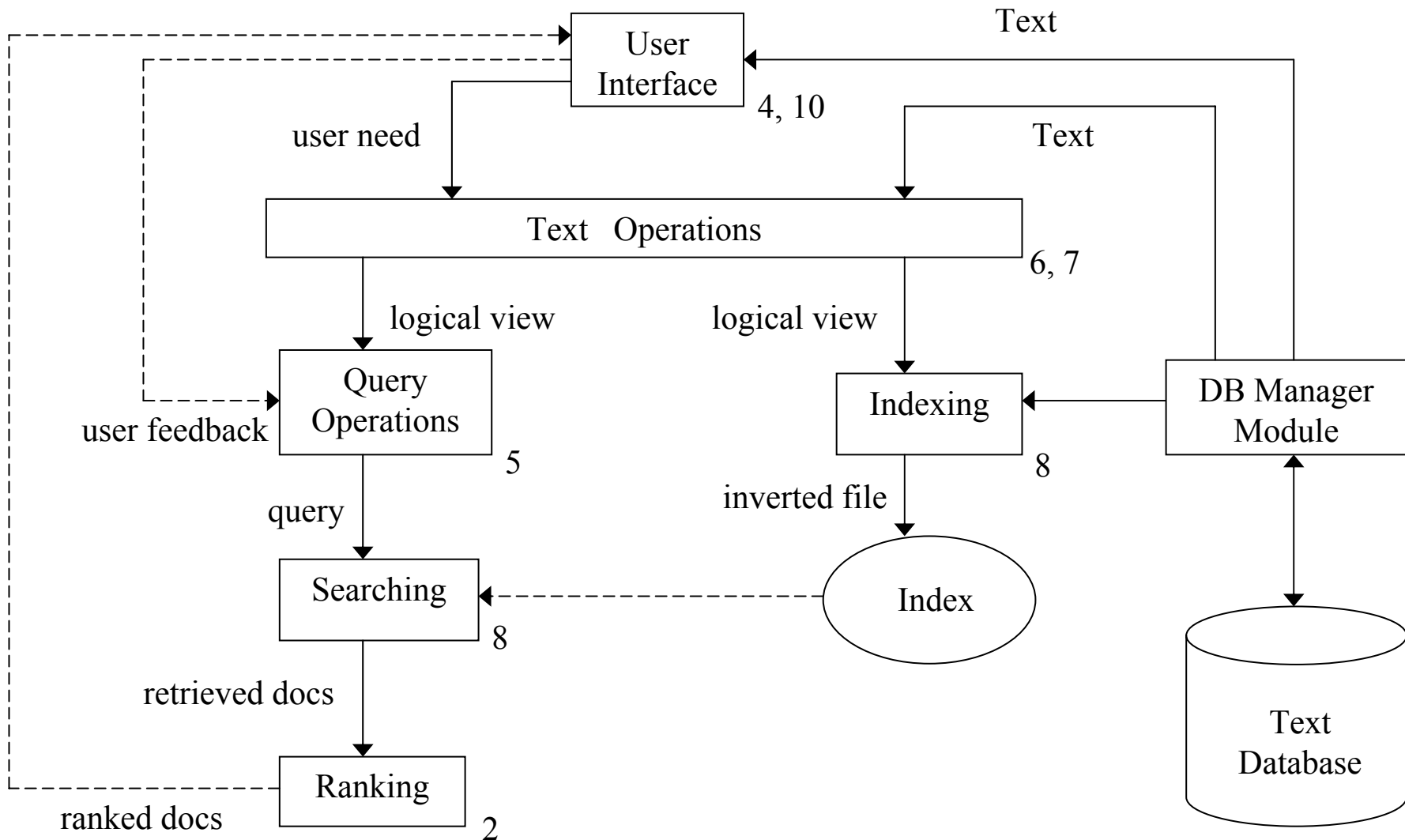
- Computer-centered (commercial perspective)
  - Efficient indexing approaches
  - High performance matching ranking algorithms
  
- Human-centered (academic perceptive)
  - Studies of user behaviors
  - Understanding of user needs

} 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)

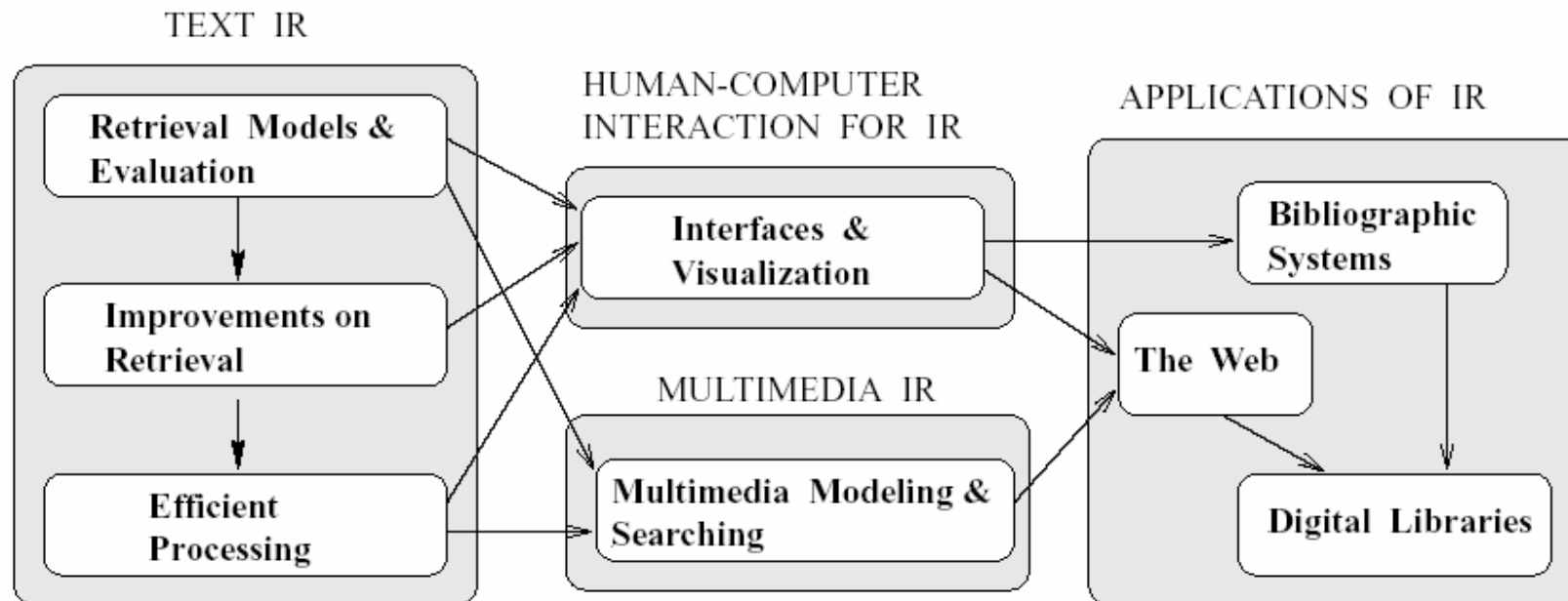
- In current retrieval systems
  - Users almost never declare his information need
    - Only a short queries composed few words (typically fewer than 4 words)
  - Users have no knowledge of the text or query operations

Poor formulated queries lead to poor retrieval !



# Major Topics (1/2)

- Four Main Topics

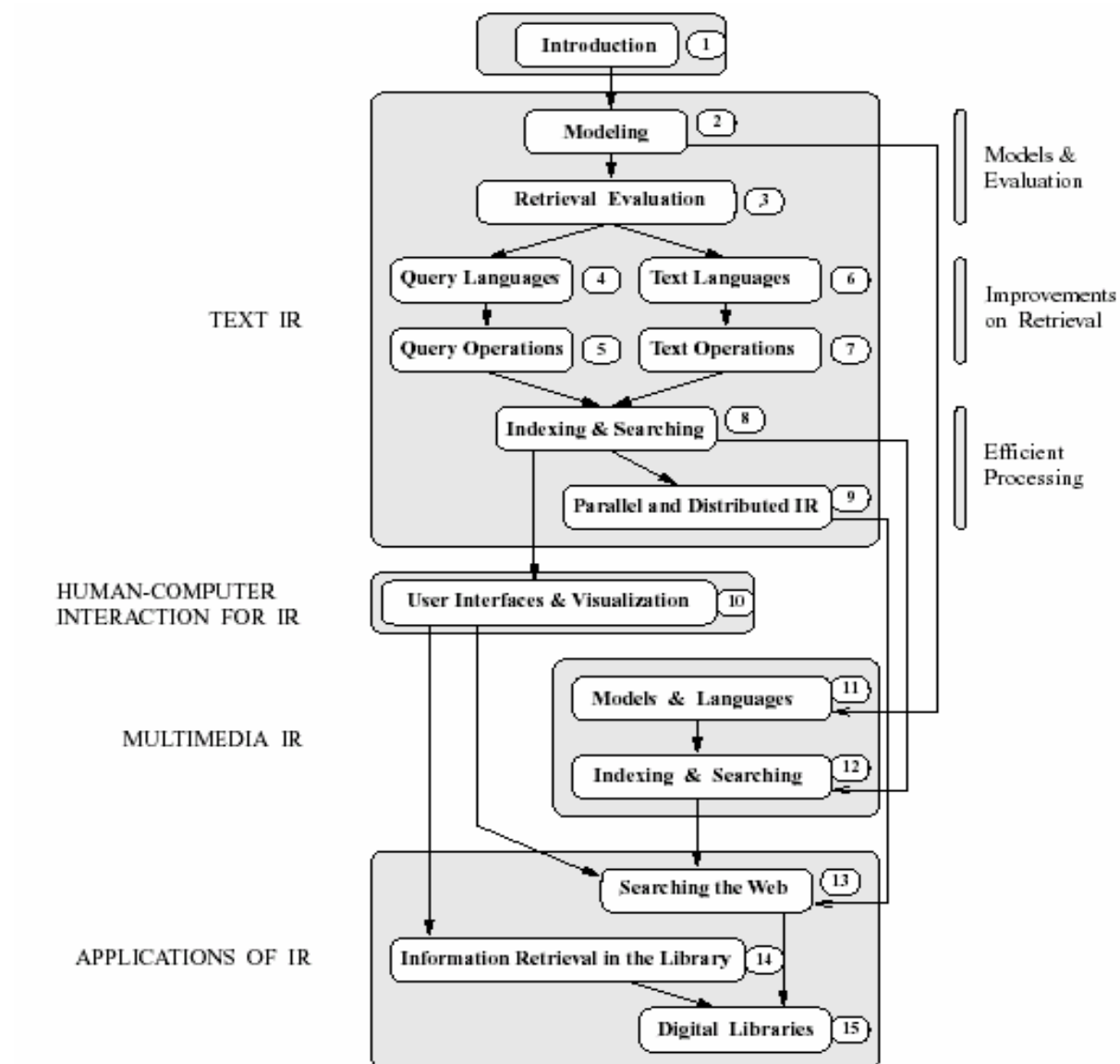


**Figure 1.4** Topics which compose the book and their relationships.

# Major Topics (2/2)

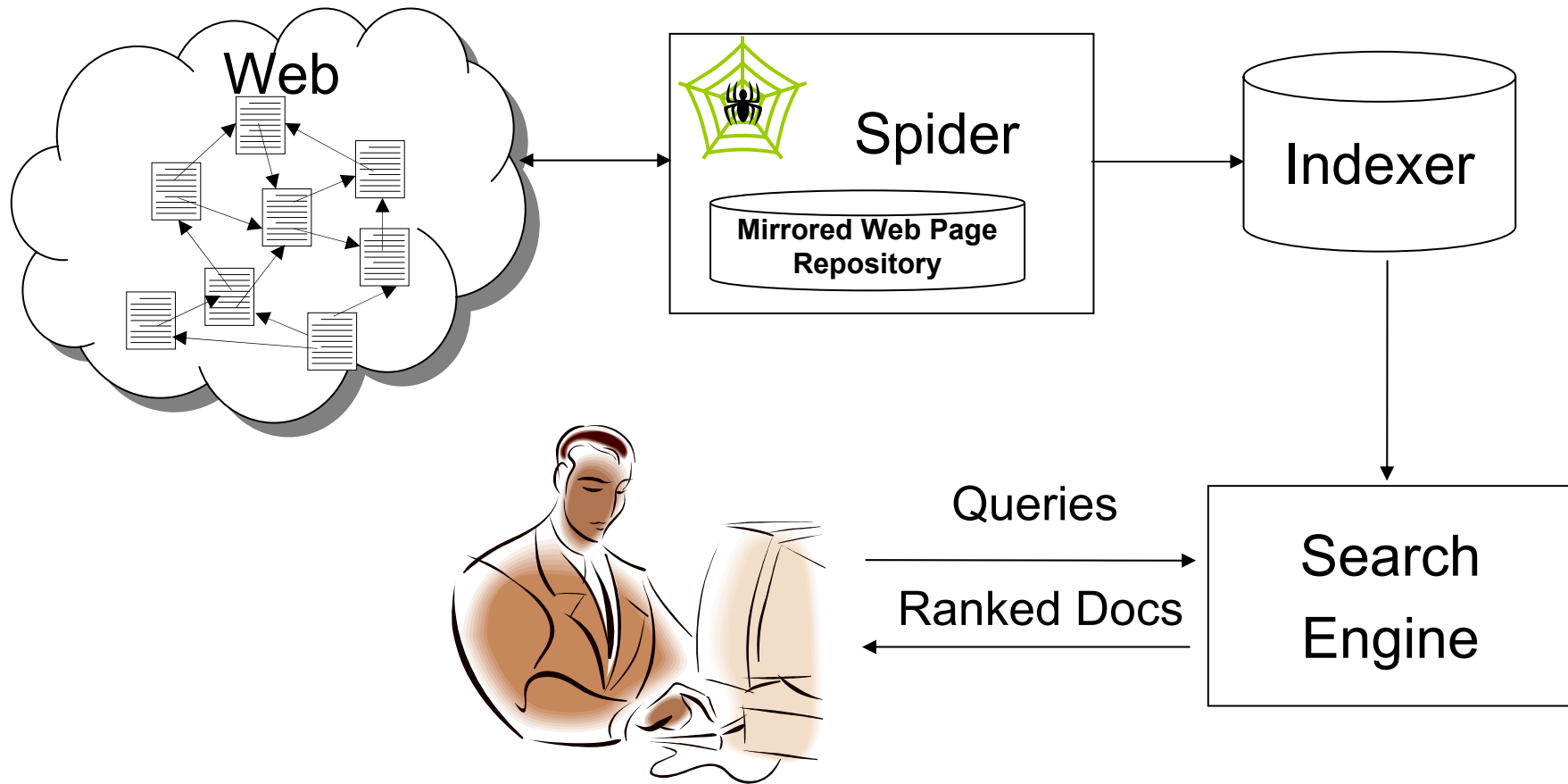
- Text IR
  - Retrieval models, evaluation methods, indexing
- Human-Computer Interaction (HCI)
  - Improved user interfaces and better data visualization tools
- Multimedia IR
  - Text, speech, audio and video contents
  - Multidisciplinary approaches
- Applications
  - Web, bibliographic systems, digital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

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(1) [image&Query=&QUERY=&query=%C6%5B%C3%FA&ServiceID=0](#) 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

觀霧 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](#)

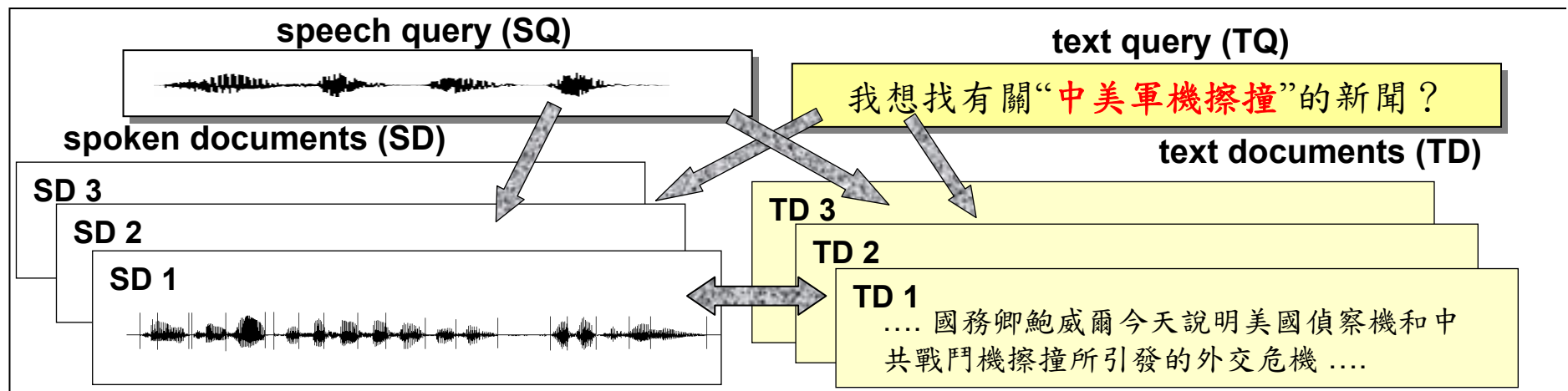
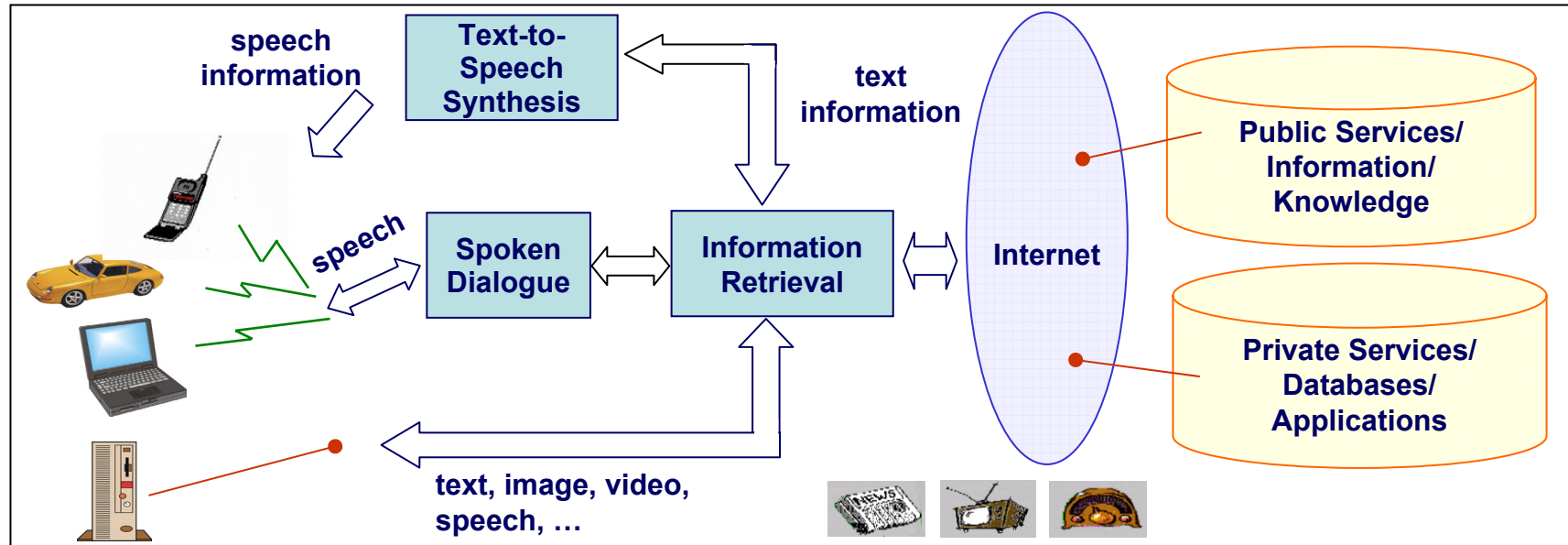
- 1. 觀霧農莊**  
介紹農莊風景及其服務項目、交通指南、住宿方式等。 公司名稱: ...  
<http://tree.2u.com.tw/> - 2002/12/11, 16k - [ 關鍵字 ] [ 更多結果 ]
- 2. 瀑布谷農場**  
自然休閒-擁抱山水-到雲海的舞台**觀霧** | 瀑布谷農場介紹 | | 交通路線圖 | | 旅遊注意事項  
| **觀霧**是雲的故鄉, 景色千變萬化, 體驗大自然、賞... 農場也準備卡拉OK讓您高歌一曲。  
注意事項×**觀霧**地區日夜溫差大請多加保暖衣物·請攜帶證件...  
簡介-介紹位在雪霸國家公園觀霧的瀑布谷農場, 經營民宿、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: 新竹縣五峰鄉掛山村巨石362號之1 公司電話:

# Text Information Retrieval (4/4)

- <http://www.baidu.com>

The screenshot shows a Baidu search interface. At the top, there is the Baidu logo and navigation links for '设百度为首页', '高级搜索', and '帮助'. A search box contains the text '陈柏琳', with buttons for '百度搜索' and '在结果中找'. Below the search box, there are tabs for '新闻', '网页', '贴吧', 'MP3', and '图片'. A status bar indicates '找到相关网页156篇, 用时0.158秒'. The main content area displays search results for '陈柏琳'. The first result is '陈柏琳 (Berlin Chen) 的网页', with a snippet: 'Welcome to Berlin's Homepage 2004 Berlin Chen, Assistant Professor, Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan, ROC Personal Information My...'. Below this is a link to 'www.csie.ntnu.edu.tw/~berlin/' and a '百度快照' link. The second result is 'Berlin Chen (陈柏琳) - Research', with a snippet: '邱炫盛、陈柏琳, "垃圾邮件过滤技术之初步研究," 投稿至「第十届人工智能与应用研讨会」, December 2...'. Below this is a link to '140.122.185.120/berlin\_research/research\_...' and a '百度快照' link. The third result is '百度\_choi吧 [[Charlene Choi相关电影资料]]', with a snippet: '的关机仪式, 该片导演刘伟信偕同主演谢霆锋、蔡卓妍、范冰冰、陈柏琳、BOYZ(关智斌、张致恒)、梁洛施、谭耀文、戴娇倩等人盛装出席。>> ... http://ent.tom.com/1636/1637/200517-115930.html 帖子相关图片: 作者: Angel\_...'. Below this is a link to 'post.baidu.com/?kz=8522392' and a '百度快照' link. The fourth result is '娱乐/中国宁波网', with a snippet: '陈柏琳在《...》中饰演...'. On the right side of the search results, there are several promotional boxes: '找陈柏琳商品在eBay易趣', '找陈柏琳创业项目在biz178', '访问通用网址陈柏琳', '找陈柏琳好项目到e26', 'DELL电脑低价直销3399起', '找陈柏琳创业项目在89178', '找陈柏琳项目在创业加盟网', '搜陈柏琳在阿里巴巴', '总有一人知道你问题的答案', '发表留言创建陈柏琳贴吧', '有许多话想对这个人说? 赶紧敲下来吧, 让她/他感受一种幸福和惊喜! 您的心意, 将在此一一传递.', and '给陈柏琳传情...'. At the bottom of the screenshot, there is a small text: '陈柏琳在《...》中饰演...'

# 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/>)

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: [input] [button]

» 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 [input] [button]

» Power Search » Help

Search for: [input] [button]

Topics: All Topics Dates: All 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 (3/4)

- 輸入聲音問句：“請幫我查總統府升旗典禮”

中文電視暨廣播新聞檢索系統 2002v1-Berlin Chen & Lin-shan Lee

辨識 I 等待輸入指令...

測靜音 放音 離開 載入新聞

語音辨識結果

總統府升旗典禮

← 聲音問句的語音辨識結果

Witerbi=>End\_Time= 100  
TotalFrame=362 1. (接受) 幫我找 8340.57 (時間) 28 100

文字檢索

語音辨識結果

FILE (Erroneous Transcription): FTV2002-004.txt

檢索到新聞的語音辨識結果

中華民國就是明年元旦總統府升旗典禮即將在下而星期二登場  
而今年首度社教有民間工商團體來舉辦  
新科立委金素梅將帶著實為原住民亦同高唱國歌  
展現多元文化的特性有以今年的元旦升旗典禮將打破傳統方式長  
經紀人龍門一千人到新竹美勞他擔任市為原住民

檢索到新聞的影音

可以選擇同時使用音節、字、詞等三種索引特徵

Rank	ID	Score
[ 1]	FTV2002-004	3.09164e-001
[ 2]	N200201211200-01	2.11802e-001
[ 3]	N200201091200-12	1.91467e-001
[ 4]	N200201091200-09	1.89940e-001
[ 5]	N200109061200-07	1.66562e-001
[ 6]	N200201211200-01	1.64992e-001
[ 7]	N200105071000-04	1.60819e-001
[ 8]	N200111131200-04	1.57109e-001
[ 9]	T200201211200-01	1.53650e-001
[ 10]	T200201211200-04	1.51319e-001
[ 11]	N200110031200-03	1.47177e-001
[ 12]	N200201171200-11	1.44006e-001
[ 13]	N200105071400-02	1.41382e-001
[ 14]	T200106191000-02	1.39268e-001
[ 15]	N200110291200-01	1.38799e-001
[ 16]	N200104301230-05	1.36488e-001
[ 17]	N200109051200-05	1.33595e-001
[ 18]	N200109141200-18	1.33158e-001
[ 19]	N200105142000-05	1.32321e-001
[ 20]	FTV2002-064	1.32147e-001
[ 21]	N200201181200-11	1.31223e-001

QueryByExemplars 檢索結果之排名

FILE (Erroneous Transcription): FTV2002-004.txt

24小時現場直播

元旦升旗 金素梅將帶原住民唱國歌

中二高龍升旗慶架倒樹 2工人重傷

# Speech Information Retrieval (4/4)

NTNU Broadcast News Retrieval System  
(本系統僅供內部語音辨識及語音資訊檢索實驗之用)

搜尋詞彙：賓拉登。  
 共找到 90 頁相關網頁。

① /Word/N200109141200-25.txt audio  
 包庇 賓拉登 的阿富汗塔利班因為美國九一一恐怖攻擊案而再度成為全球關注焦點根據阿富汗消息塔利班表示如果美國知名 賓拉登 的罪行將考慮把 賓拉登 已從審判目前阿富汗首都喀布爾籠罩在可能遭報復的緊張氣氛下並重已經陸續拋出可獲

② /Word/N200110081000-05.txt audio  
 而 賓拉登 今天也發表的公開談話阿富汗當地電視台實況提供一卷事先錄製好的 賓拉登 談話被西方媒體 賓拉登 表示阿富汗將奮戰到底並強調美國是罪魁禍首 賓拉登 讚揚就對美國發動的恐怖攻擊並且審慎那次攻擊是一群回教徒所謂 賓拉登 在這項錄影談話當中說美國在最軟弱的地方遭到真主的打擊毀毀的她最有名的建築物感謝真主他要求回教國家青年加入這場選戰

③ /Word/N200110081200-04.txt audio  
 美國 媒體 那 有名

④ /W  
 另  
 松  
 論  
 事  
 不  
 不  
 播

⑤ /Word/N200109191200-25.txt audio

Windows Media Player  
 檔案(E) 檢視(V) 播放(P) 工具(T) 說明(H)  
 現正播放  
 尋找專輯資訊  
 N200109141200-25  
 全部時間: 0:20  
 正在播放: 16 K 位元/秒  
 00:04

錄製好的 賓拉登 談話被西方  
 美國發動的恐怖攻擊並且審慎  
 到真主的打擊摧毀了他最有

佈分子 賓拉登 台灣目前金  
 觸表示 賓拉登 願意回應結  
 的問題包括了 賓拉登 在九一  
 幕及計畫如何使用的不過還  
 不確定 賓拉登 目前是生十四  
 礙國家安全為前提擷取片段

Browser 11:16

中文廣播新聞檢索系統 國立台灣師範大學資工所

◎錄音鍵 [Waveform]

辨識結果 美國總統大選 搜尋

摘要

040304-13.兩千年美國總統大選時  
 021216-24.二零零零年總統大選時高爾以此  
 040309-10.把總統到訪當成的將領希望帶  
 021210-23.因此如果國親兩黨有任何一個

全文

關心美國總統大選消息美國北卡羅來納  
 州參議員愛德華茲間  
 正式宣布退出民主黨總統候選人初選並  
 表示將全力協助麻州參議員凱  
 瑞期待美國總統布希而儘管美國十一月

--- 新聞影音播放 ---

File Settings

# Visual Information Retrieval (1/4)

- Content-based approach

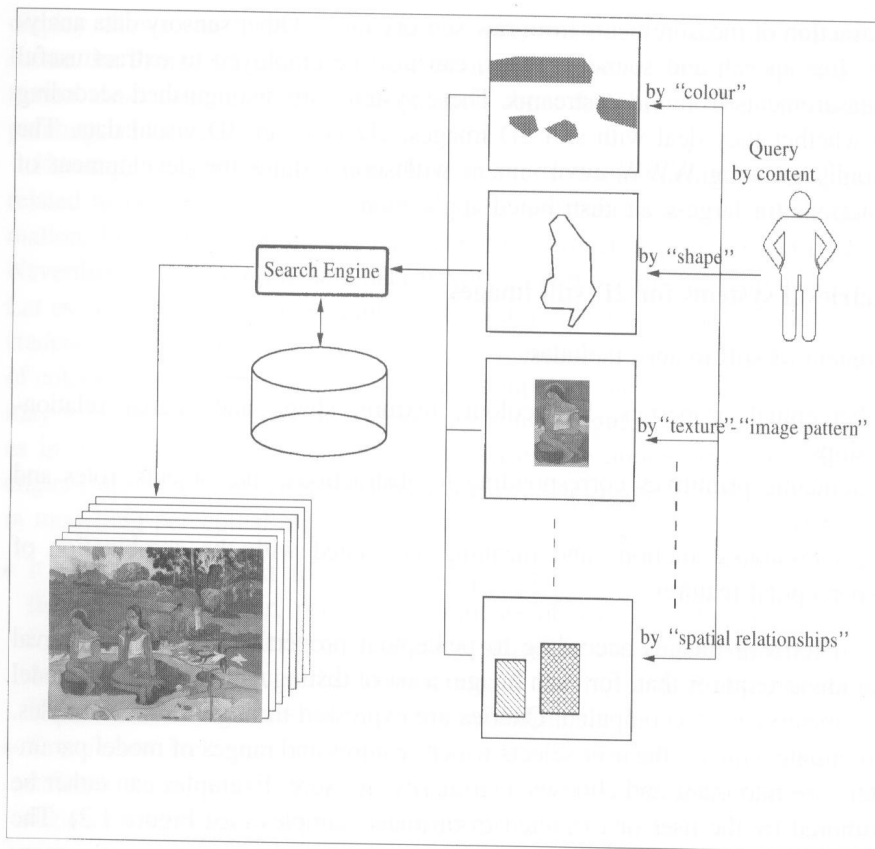


Figure 1.2 Different types of query by example.

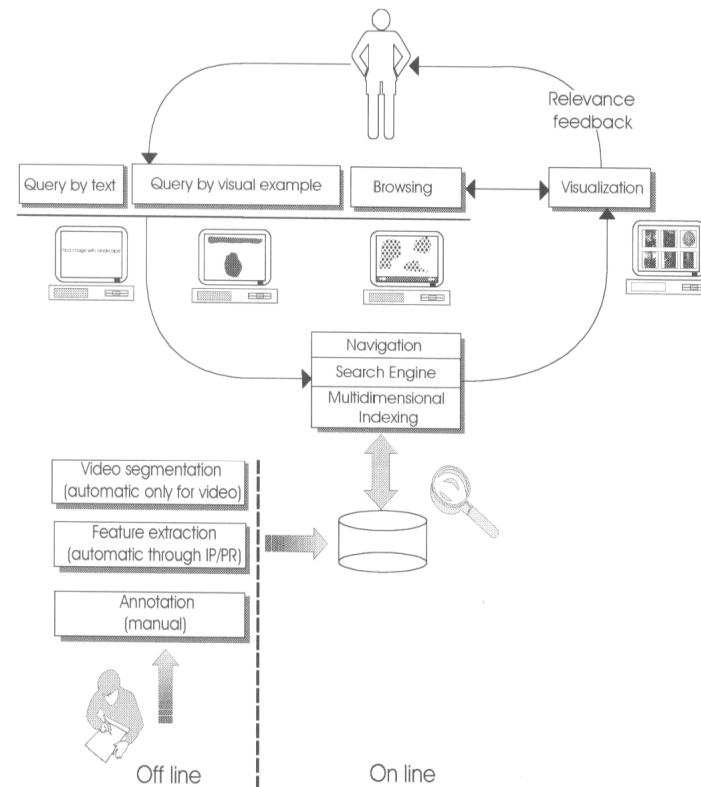
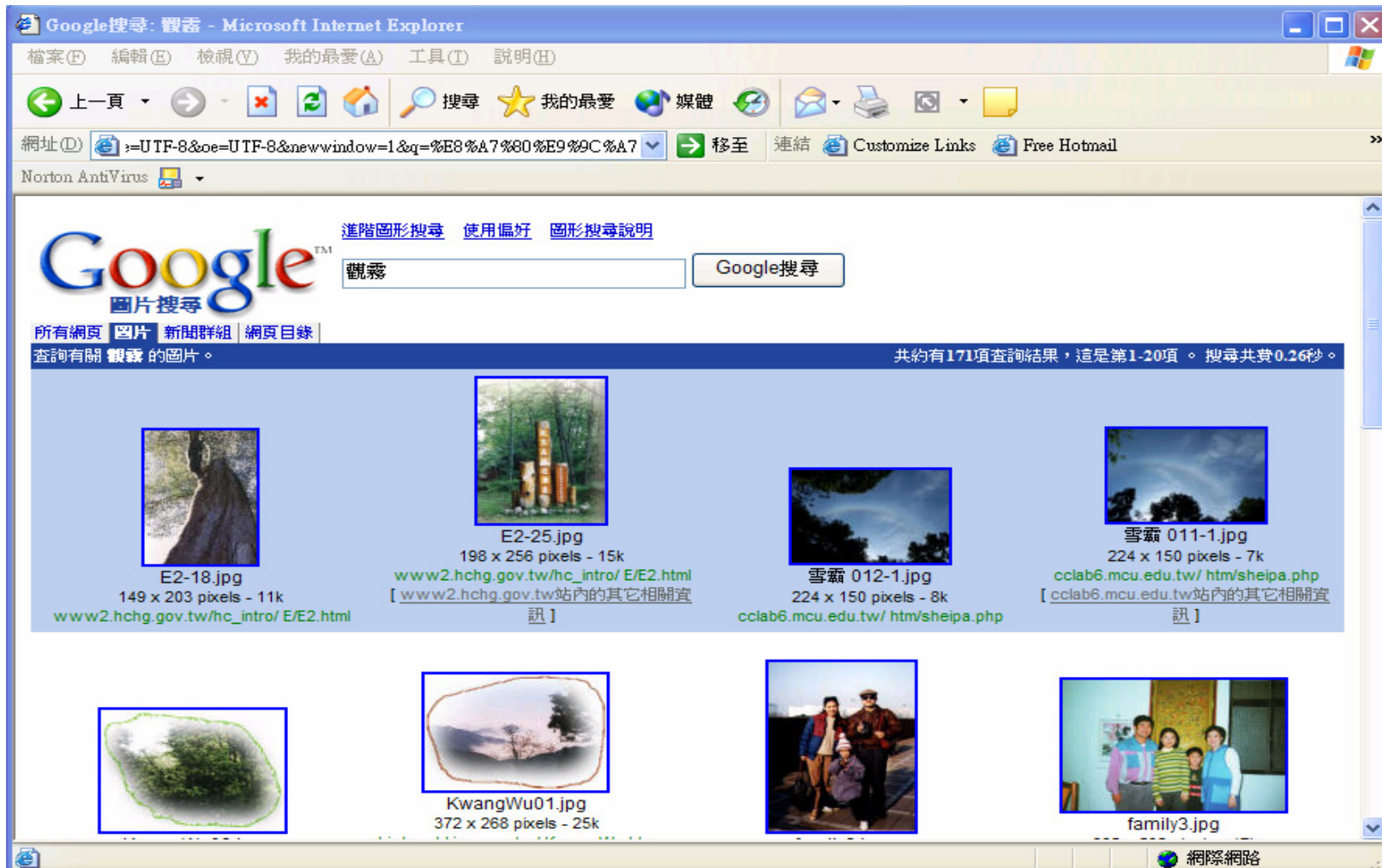


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

# Visual Information Retrieval (2/4)

- Images with Texts



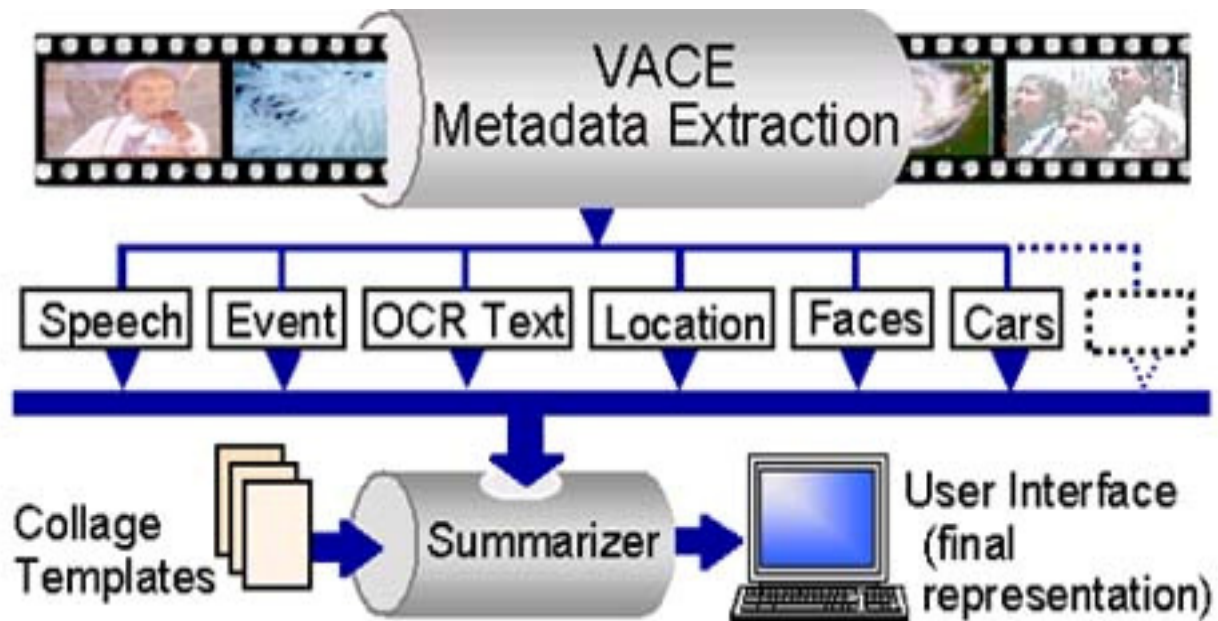
# Visual Information Retrieval (3/4)

- 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
- .....

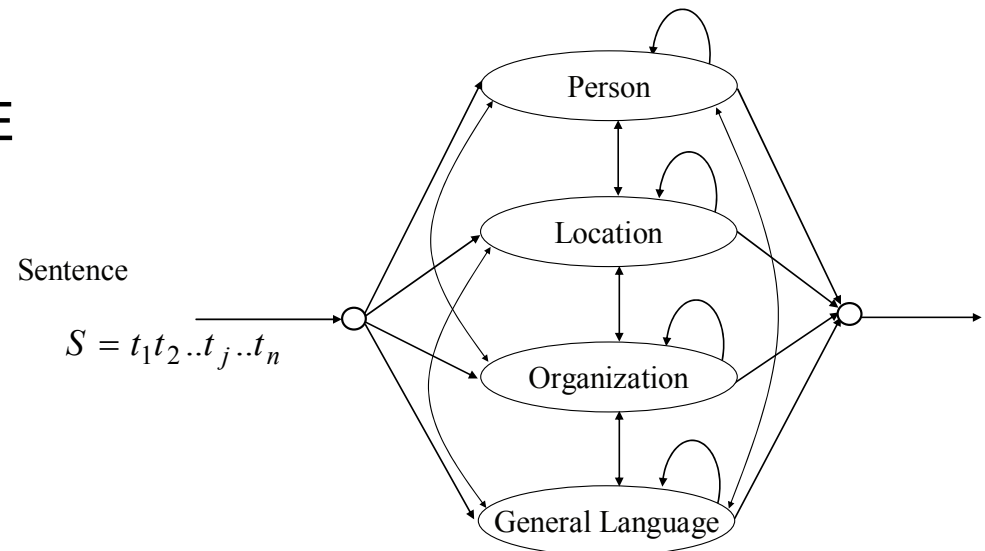


# Document Summarization

- 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 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



# 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

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

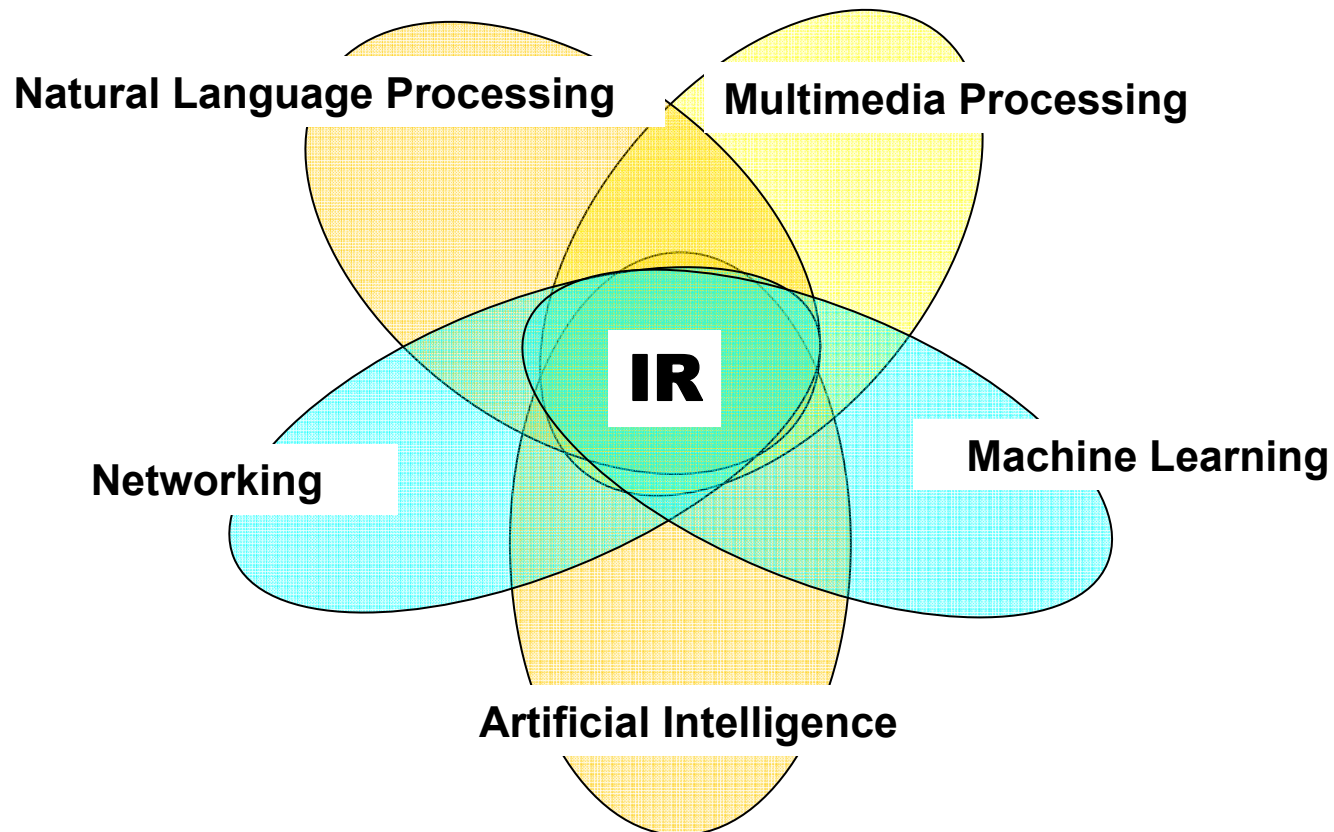
Query/Translation	Relevant Pages	Relevant Images
national palace museum	<ul style="list-style-type: none"> <li>* <a href="#">National Palace Museum</a> [Gloss translation: ]</li> <li>* <a href="#">TIT Museums: The National Palace Museum: 70 Years Young!</a> [Gloss translation: ]</li> <li>* <a href="#">Jades 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>	

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!

Machine-  
Extracted  
Translation

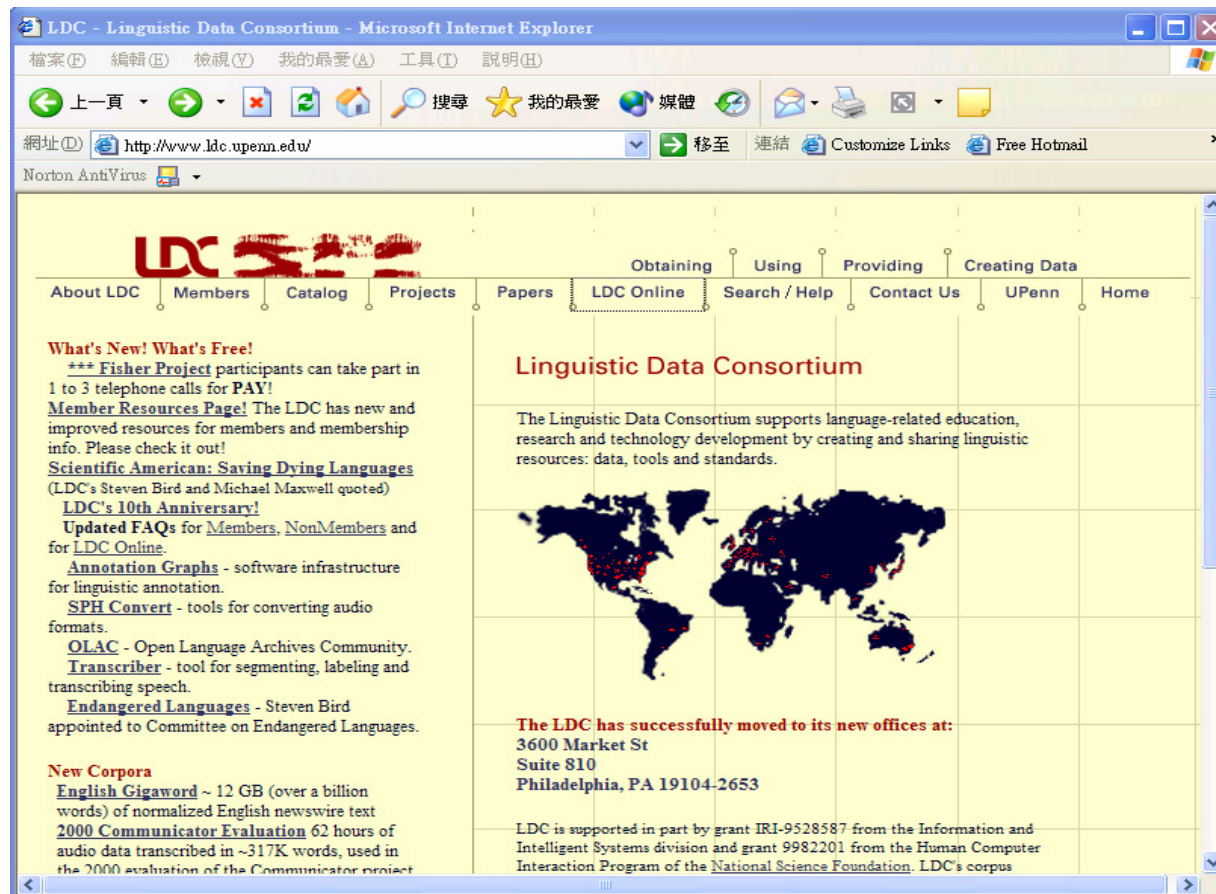


# Multidisciplinary Approaches



# Resources

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



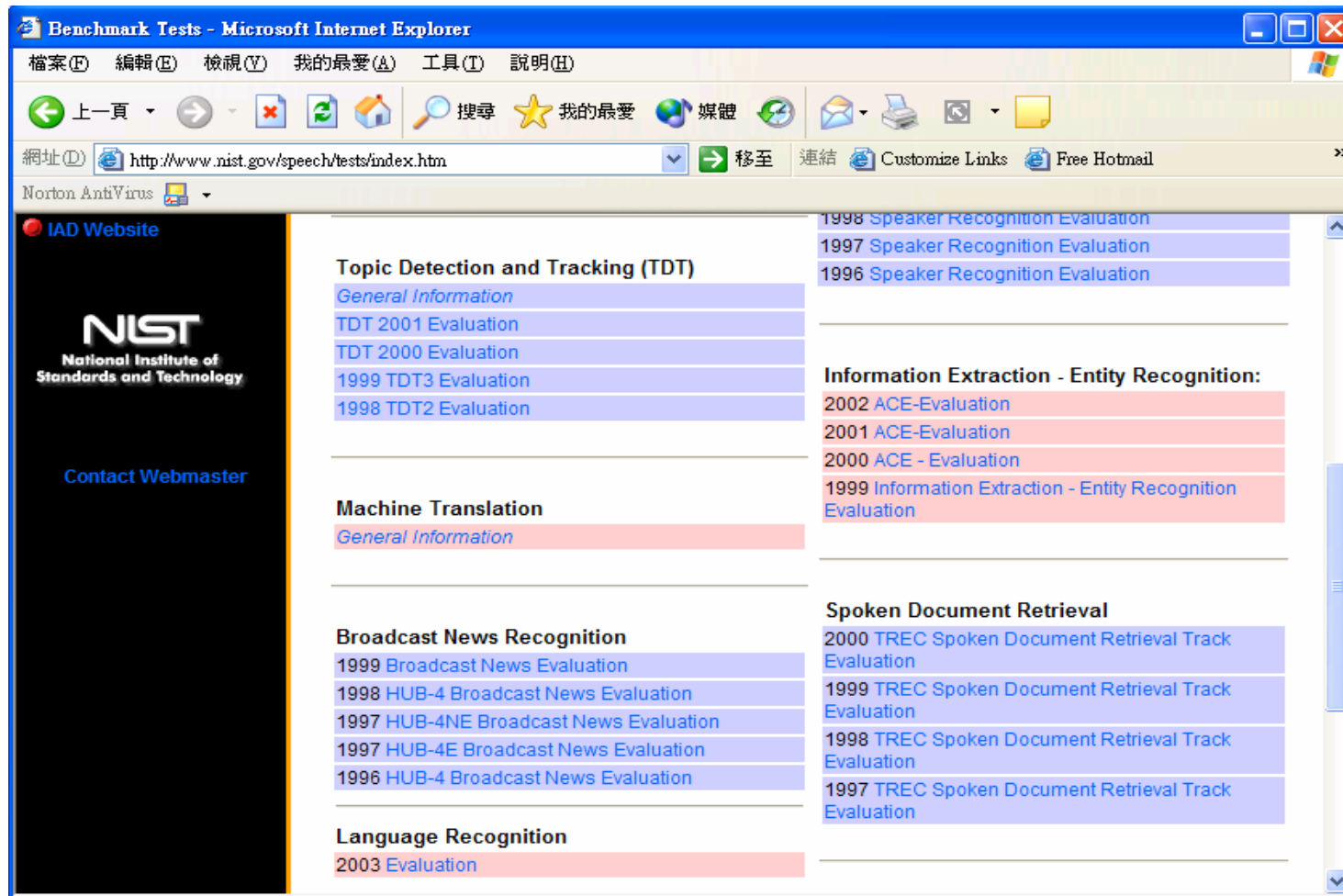
# Contests (1/2)

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



# Contests (2/2)

- US National Institute of Standards and Technology



# 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	<a href="#">Course Overview &amp; Introduction</a>
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 (HMMN-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	<b>Midterm</b>
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	<b>Break</b>
1/3	<b>Final</b>