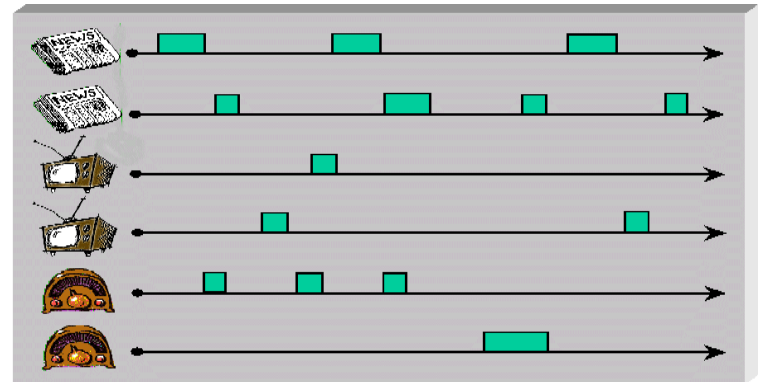


Information Retrieval and Extraction

Berlin Chen 2003



Textbook and References

- Textbook

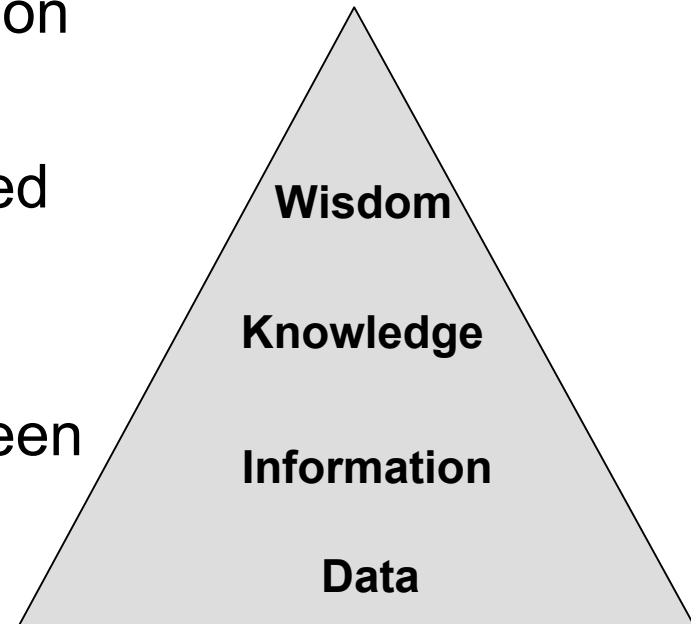
- R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman, 1999.

- References

- W. B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures & Algorithms, Prentice-Hall, 1992.
- A. D. Bimbo, "Visual Information Retrieval", Morgan Kaufmann, 1999.
- 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.
- D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice-Hall, 2000.

Motivation

- **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

- **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

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

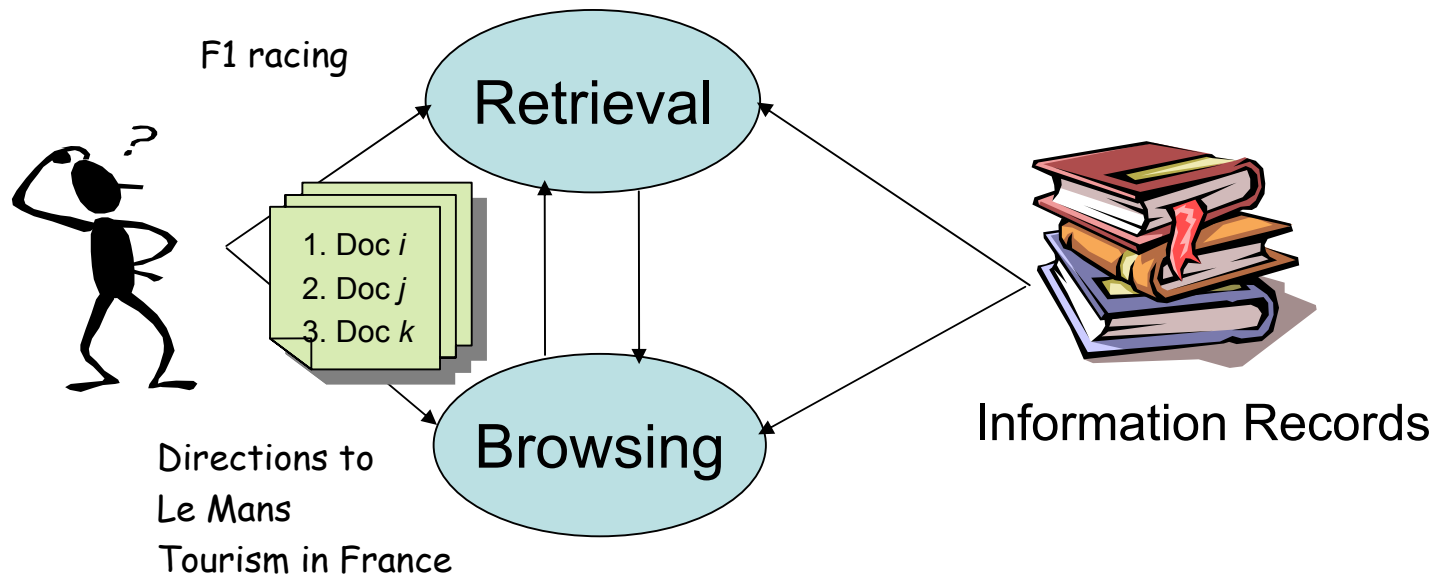
- Determine which document of a collection contain the keywords in the use query
- Retrieve all objects (attributes) which satisfy clearly defined conditions in a regular expression or a relational algebra expression
 - Which documents contain a set of keywords?
 - Well defined semantics
 - A single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

Motivation

- **IR system**
 - Interpret contents of information items
 - Generate a ranking which reflects relevance
 - Notion of relevance is most important

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

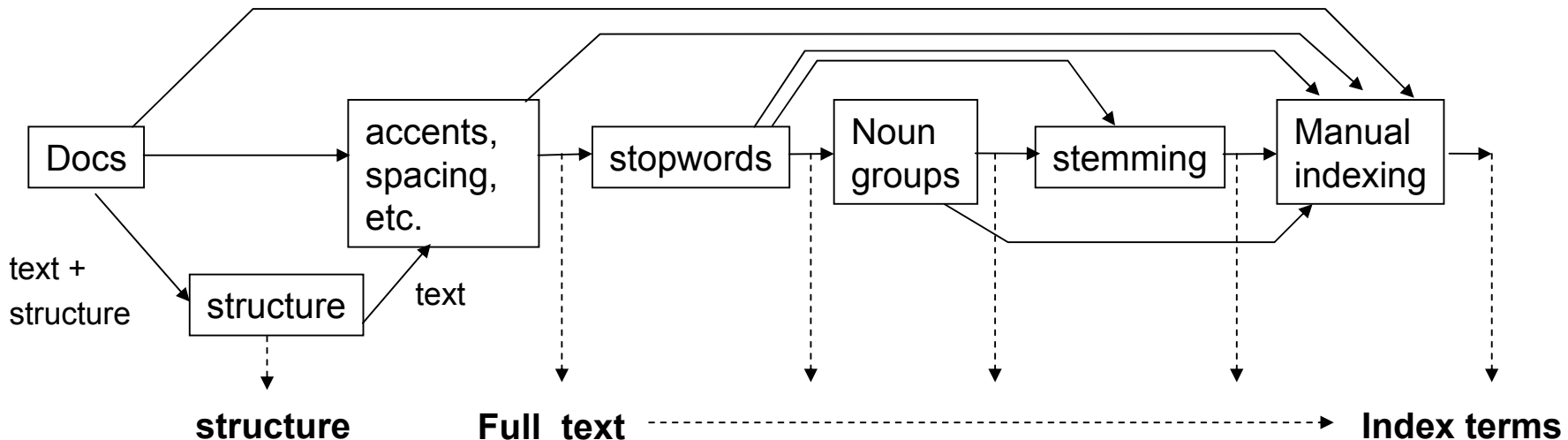


Logical view of the documents

- A full text view (representation)
 - Represent document by its whole set of words
- 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

- 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, ...)



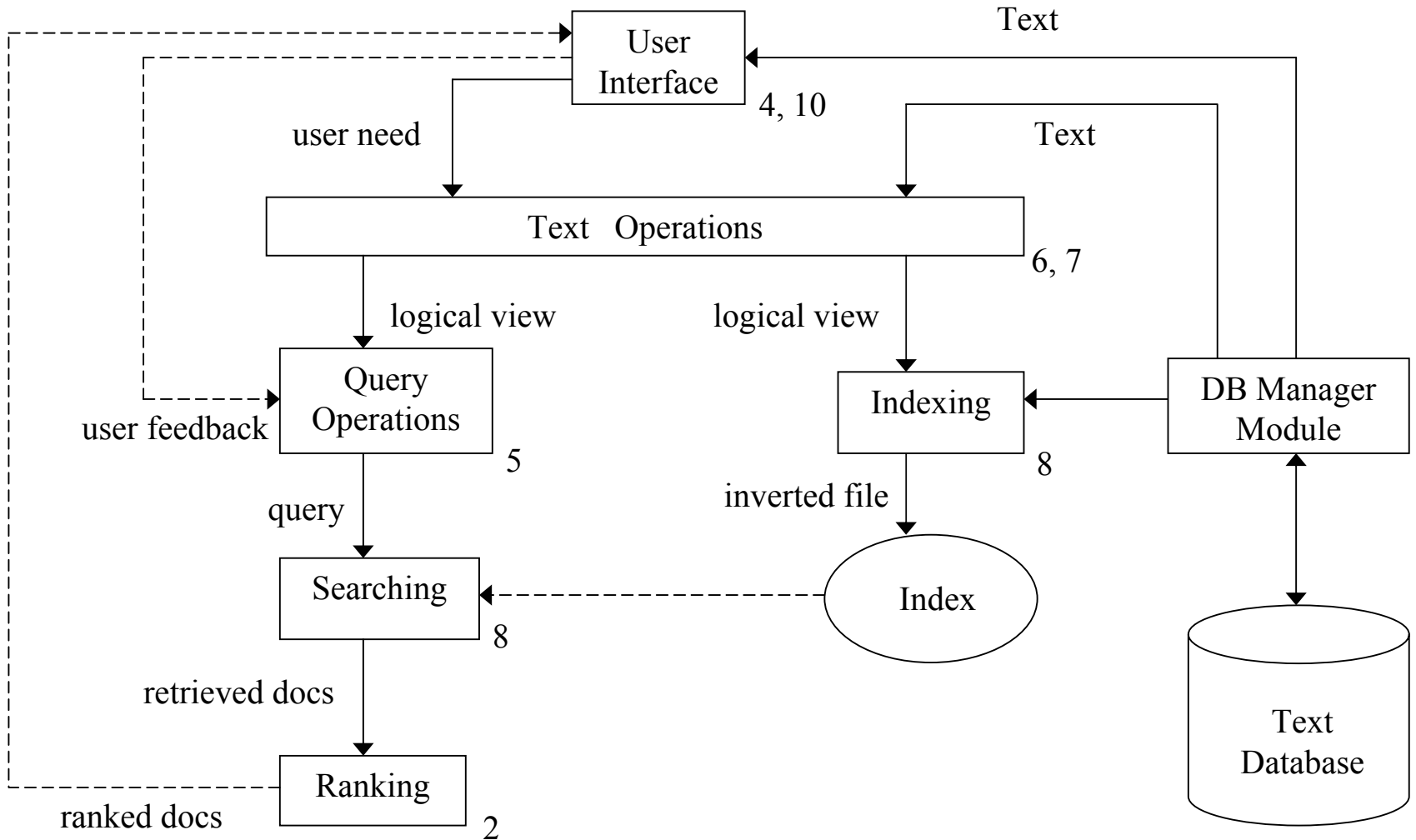
Different Views of the IR Problem

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

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
 - The user interaction with the system (HCI, Human Computer Interaction)
- Other concerns
 - Security and privacy
 - Copyright and patent

The Retrieval Process



The Retrieval Process

- 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

- Four Main Topics

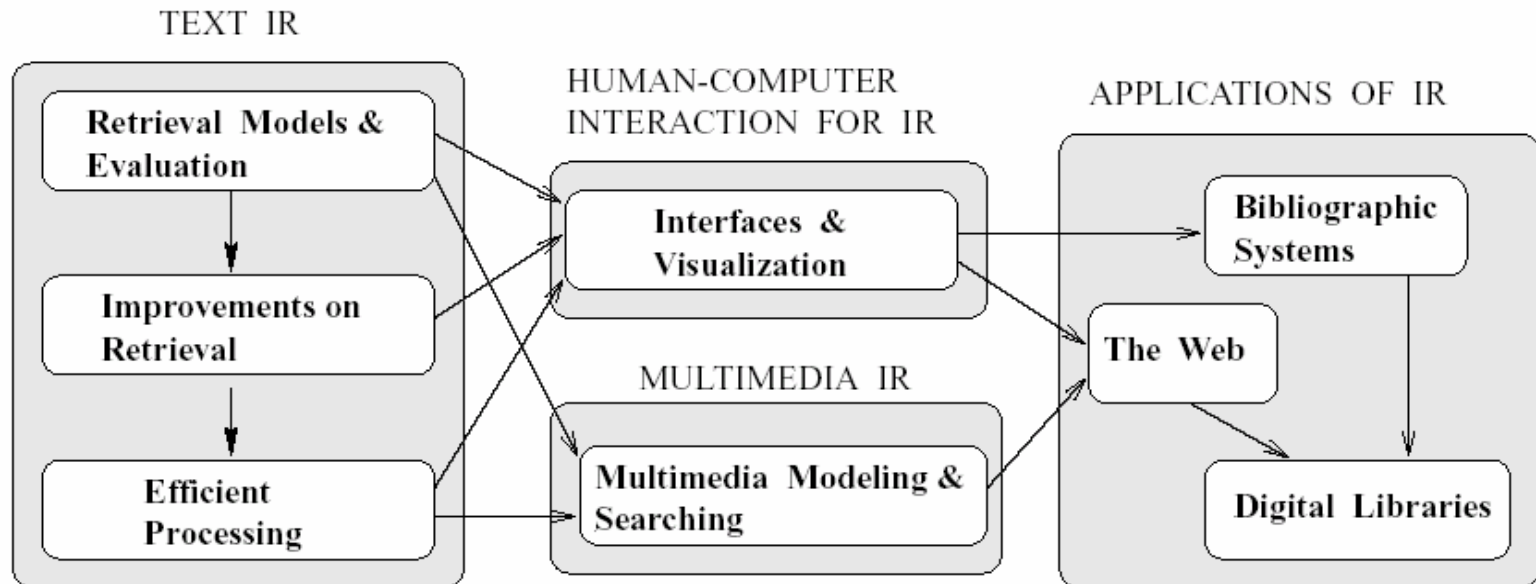
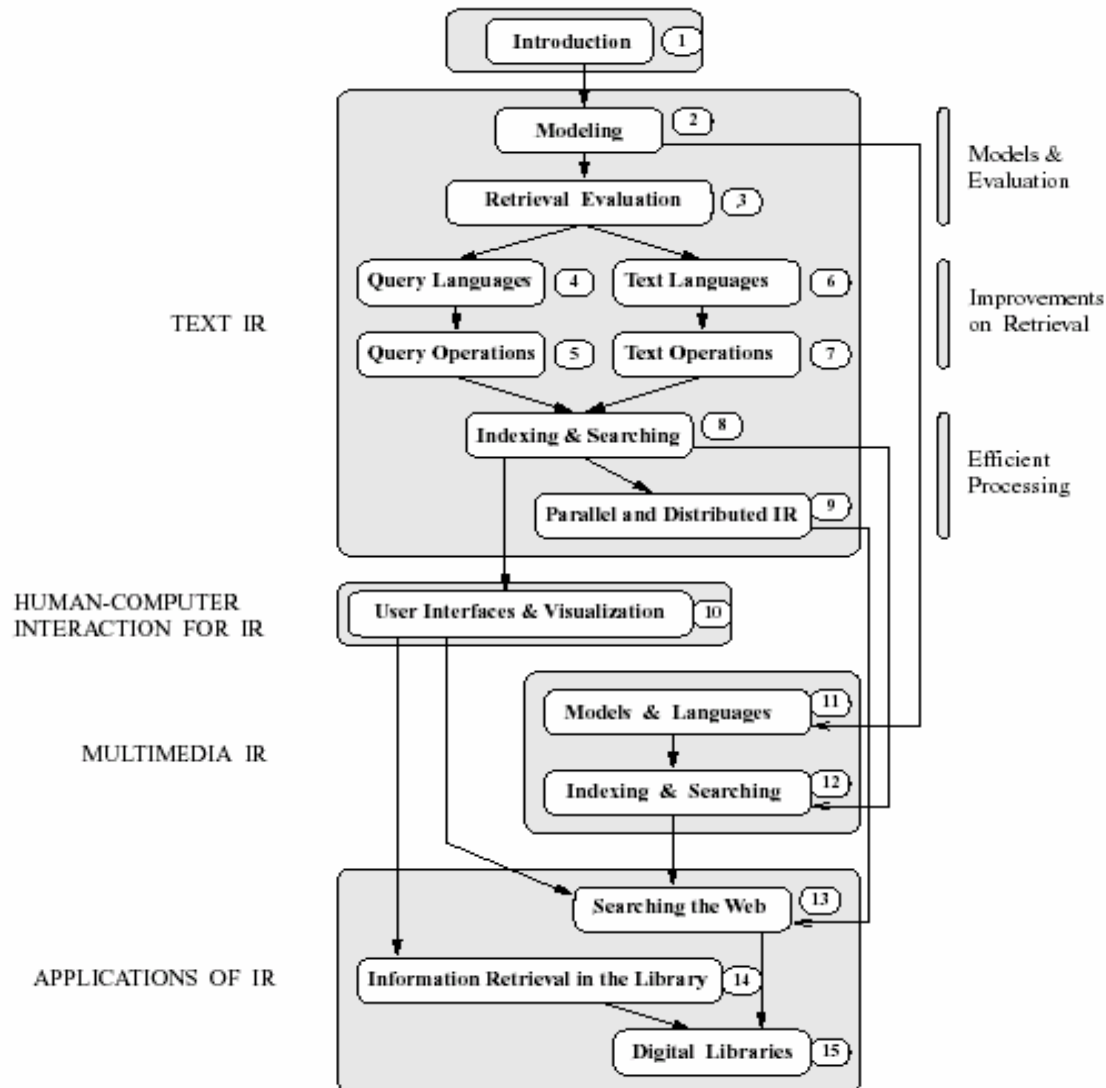


Figure 1.4 Topics which compose the book and their relationships.

Major Topics

- 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

Google 搜尋: 觀霧 - Microsoft Internet Explorer

網址 (D) TF-8&newwindow=1&q=%E8%A7%80%E9%9C%A7&sa=N&tab=iv

Google 搜尋

觀霧

搜尋所有網站 搜尋所有中文網頁 搜尋中文(繁體)網頁

已向所有網站搜尋觀霧。 共有 5,660 頁查詢結果, 這是第 1-10 頁。 搜尋共費 0.11 秒。

EbioTW-觀霧
觀霧位於新竹五峰鄉與苗栗泰安鄉交界, 雪霸國家公園範圍內, 為攀登大霸尖山必經之路, 終年雲霧繚繞, 是台灣觀賞雲霧風景的最佳景點之一。年平均溫度約 14-15 度, 冬季偶而飄雪, 夏季涼爽, 空氣清新無污染, 景色優美 ...
www.ebio2.com/ebiotw/leisure/shinjux-KuanWu.htm - 31k - 頁庫存檔 - 類似網頁

觀霧農莊
tree.2u.com.tw/ - 17k - 頁庫存檔 - 類似網頁

搜主網網複合式書店
· 觀霧 · 觀霧森林遊樂區 觀霧森林遊樂區位於新竹和苗栗交界處, 海拔約 2000 公尺, 區內林木茂密, 視野遼闊, 可遠眺雪山山脈的縱谷, 同時也是攀登大霸尖山的必經之地, 為近年來國內熱門的森林浴場所之一 ...
www.soidea.com.tw/soidea_model_index.cfm?CONSULTATENO=45 - 41k - 頁庫存檔 - 類似網頁

雪霸國家公園-觀霧遊憩區
地形特色: 在觀霧山莊、橫山步道、樂山林道等處可眺望蜿蜒曲折、岩巒高聳的聖稜線景觀。植物景觀: 橫山步道的四、五月可見高山杜鵑的綻放, 如霧氏杜鵑、台灣杜鵑等, 檜山步道沿線陰濕林下或林緣可見黃花鳳凰花 ...
www.spnp.gov.tw/chinese/information/kuanwuc.htm - 7k - 頁庫存檔 - 類似網頁

Openfind Taiwan Webpage Search: 觀霧 - Microsoft Internet Explorer

網址 (D) image&Query=&QUERY=query=%C6%5B%C3%FA&ServiceID=0

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
----	-------	----	----	----	----	----	----

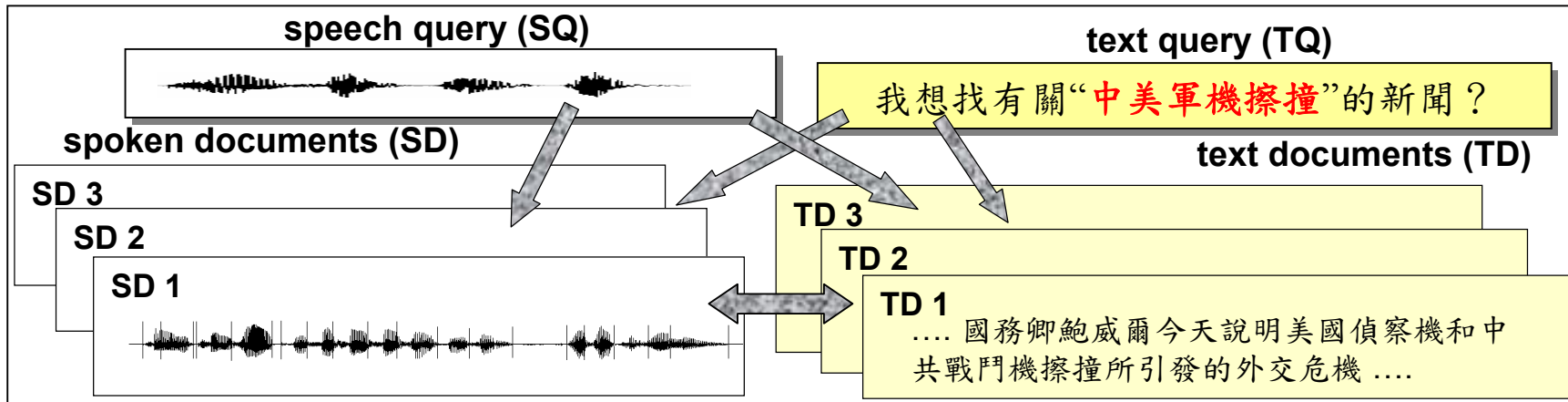
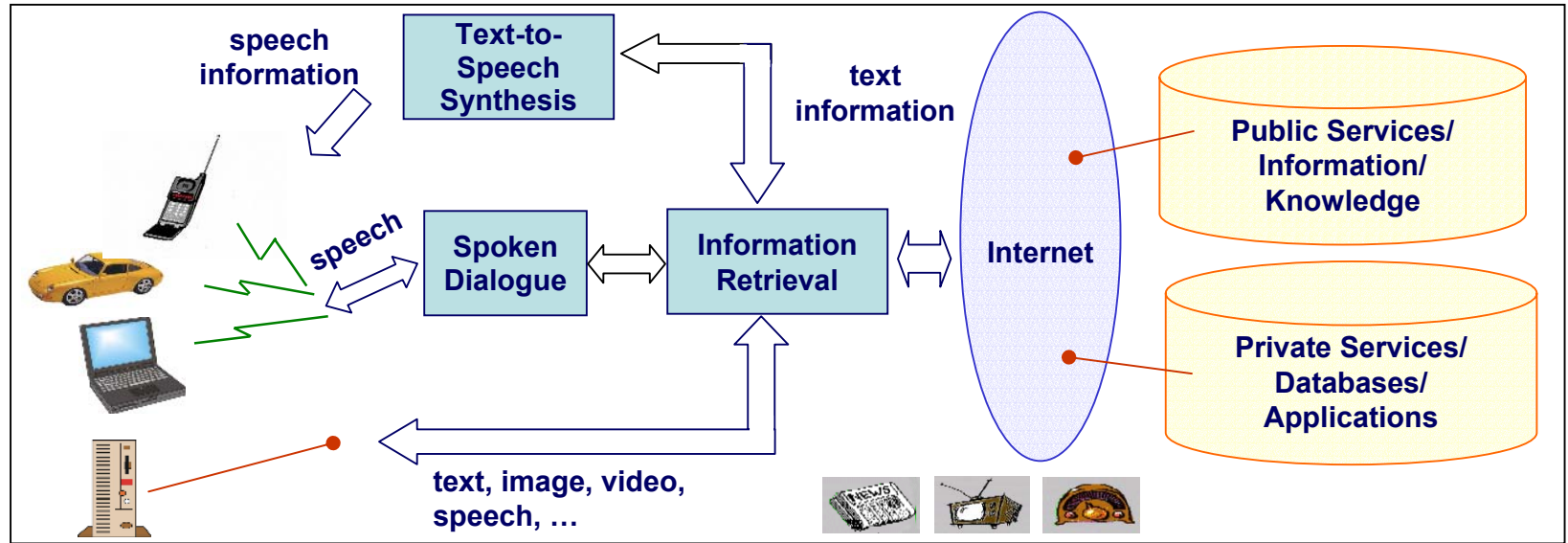
觀霧 不限日期 查詢 進階 - 喜好 - 說明

相關查詢 8 筆 · 雲霧 · 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

- 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]
- 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://ppg2u.com.tw/ - 2002/06/04, 2k - [庫存頁面] [關鍵字]
- 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: 新竹縣五峰鄉橫山村尾362號之1 公司電話:

Speech Information Retrieval



Speech Information Retrieval

- Compaq Research Group – Speechbot System
 - Broadcast news speech recognition, Information retrieval, and topic segmentation (SIGIR2001)
 - Currently indexes **15,588 hours of content** (2003/02/21, <http://speechbot.research.compaq.com/>)

hp.com - SpeechBot - 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) <http://speechbot.research.compaq.com/> 移至 連結 Customize Links Free Hotmail

Norton AntiVirus

hp **SpeechBot™**
audio search using speech recognition

invent

Simple Search **Power Search** **Help** > [FAQ](#) > [About SpeechBot](#) > [Feedback](#)

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 **15588 hours of content** from the following websites:

Advice Dr. Toni Grant	Government and Military AFRTS Radio News The White House U.S. Department of Defense Briefings	Science & Technology The Rapidly Changing Face of Computing
Arts & Entertainment Fresh Air	Internet Geeks in Space InternetNews Radio Streaming Media Talk	Sport Only A Game Resort Sports Network Scuba Radio The Golf Power Show
Current Events American RadioWorks Here and Now On Point PBS Online NewsHour	Paranormal Sightings on the Radio with Jeff Rense	Talk Car Talk Radio Show One Union Station Public Interest The Charlie Rose Show The Connection The Diane Rehm Show
Gardening On the Garden Line	Personal Investment Informed Investors, Inc. Radio Forum Marketplace Radio Business News Motley Fool Radio Show	

Speech Information Retrieval

- 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
測靜音
放音
離開
載入新聞

等待輸入指令...

3.70秒

語音辨識結果

總統府升旗典禮

Viterbi=>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

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

檢索到新聞的語音辨識結果

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	File ID	Score
1	FTV2002_004	3.09164e-001
2	N200201211200-01	2.11802e-001
3	N200201091200-12	1.91467e-001
4	N200109061200-07	1.66562e-001
5	T200201211200-04	1.57109e-001
6	N200105071000-04	1.53650e-001
7	N200111131200-04	1.51319e-001
8	T200201211200-01	1.47177e-001
9	N200201171200-11	1.44006e-001
10	N200105071400-02	1.41382e-001
11	T200106191000-02	1.38799e-001
12	N200110291200-01	1.36488e-001
13	N200104301230-05	1.33595e-001
14	N200109051200-05	1.33158e-001
15	N200109141200-18	1.32321e-001
16	N200105142000-05	1.32147e-001
17	N200201181200-11	1.31223e-001
18	N200105071000-04	1.29949e-001
19	N200105071400-02	1.29949e-001
20	N200105071400-02	1.29949e-001
21	N200105071400-02	1.29949e-001

檢索到新聞的影音

元旦升旗 金素梅將帶原住民唱國歌
中二高龍升段鷹架倒塌 2工人重傷

Visual Information Retrieval

- Content-based approach

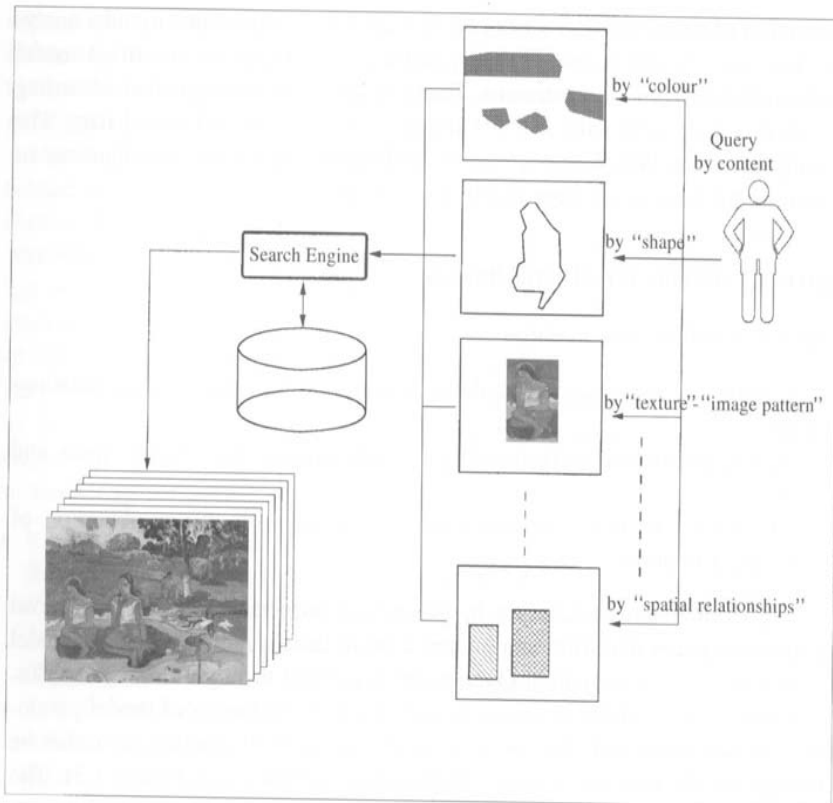


Figure 1.2 Different types of query by example.

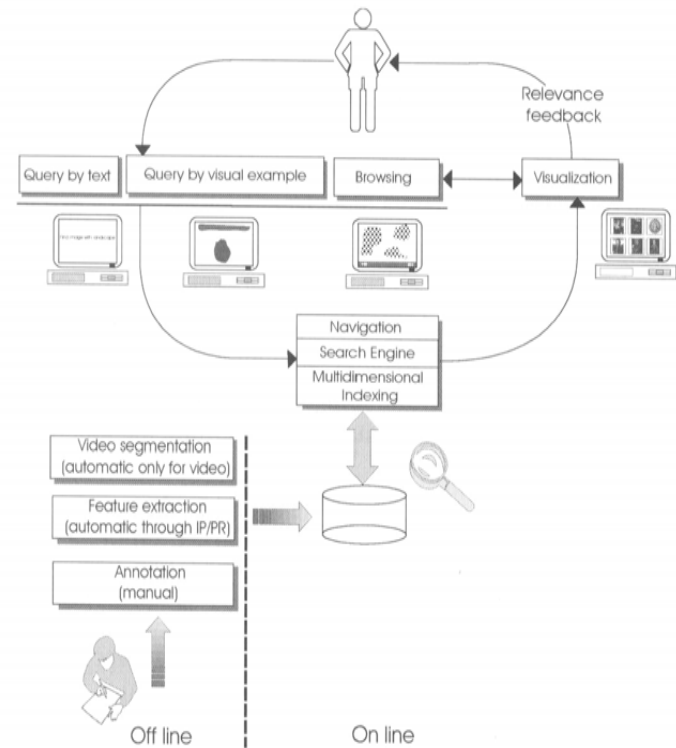
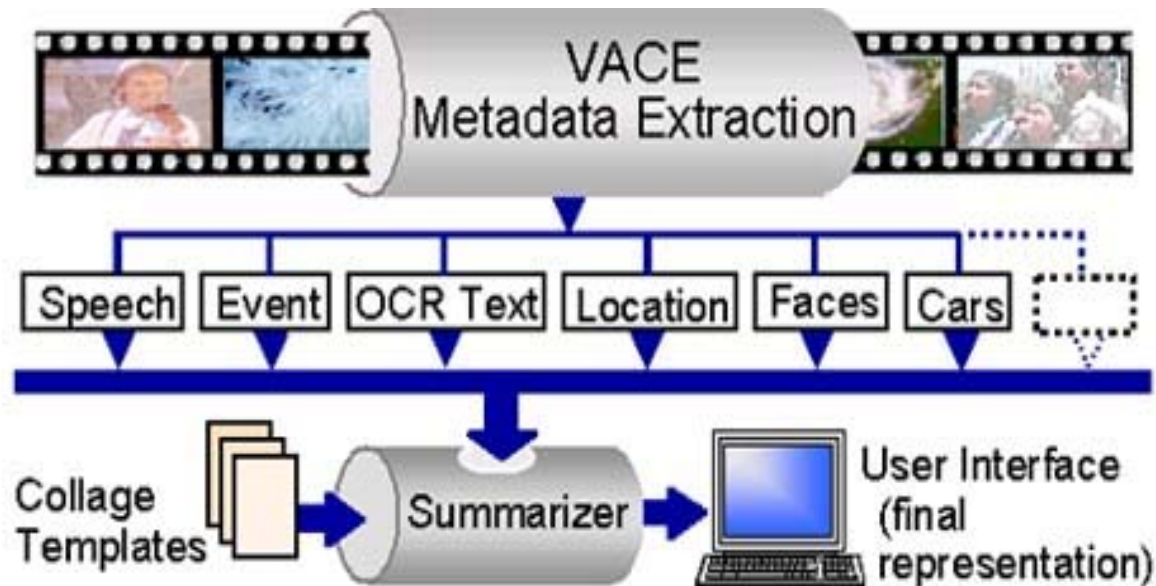


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

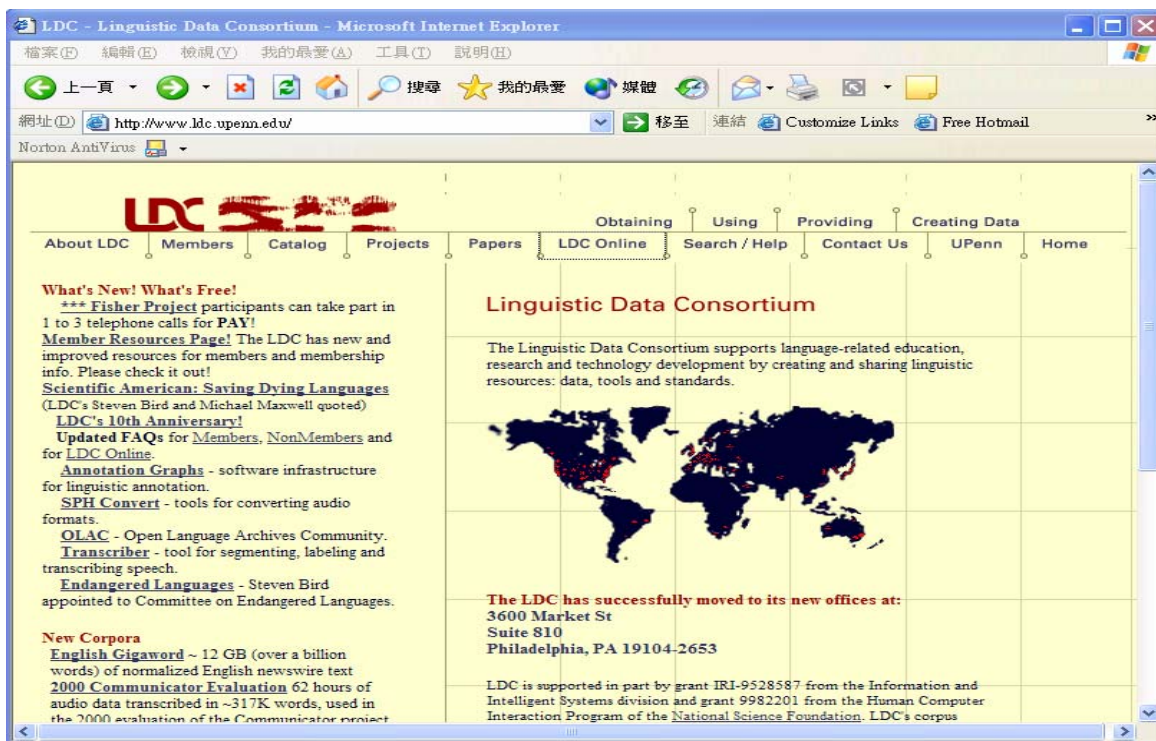
Visual Information Retrieval

Video Analysis and Content Extraction



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/)



The screenshot shows the LDC website in a Microsoft Internet Explorer browser window. The address bar displays <http://www ldc.upenn.edu/>. The website features a navigation menu with links for About LDC, Members, Catalog, Projects, Papers, LDC Online, Search / Help, Contact Us, UPenn, and Home. The main content area is divided into two columns. The left column contains several news items under the heading "What's New! What's Free!", including announcements about the Fisher Project, member resources, a Scientific American article on saving dying languages, the LDC's 10th anniversary, updated FAQs, annotation graphs, SPH Convert tools, OLAC, a transcriber tool, and endangered languages. The right column features the "Linguistic Data Consortium" logo, a description of the consortium's mission, a world map with red dots indicating office locations, and a notice that the LDC has moved to its new offices at 3600 Market St, Suite 810, Philadelphia, PA 19104-2653. At the bottom, it mentions support from the Information and Intelligent Systems division and the Human Computer Interaction Program of the National Science Foundation.

Institutes/Researchers

- Taiwan

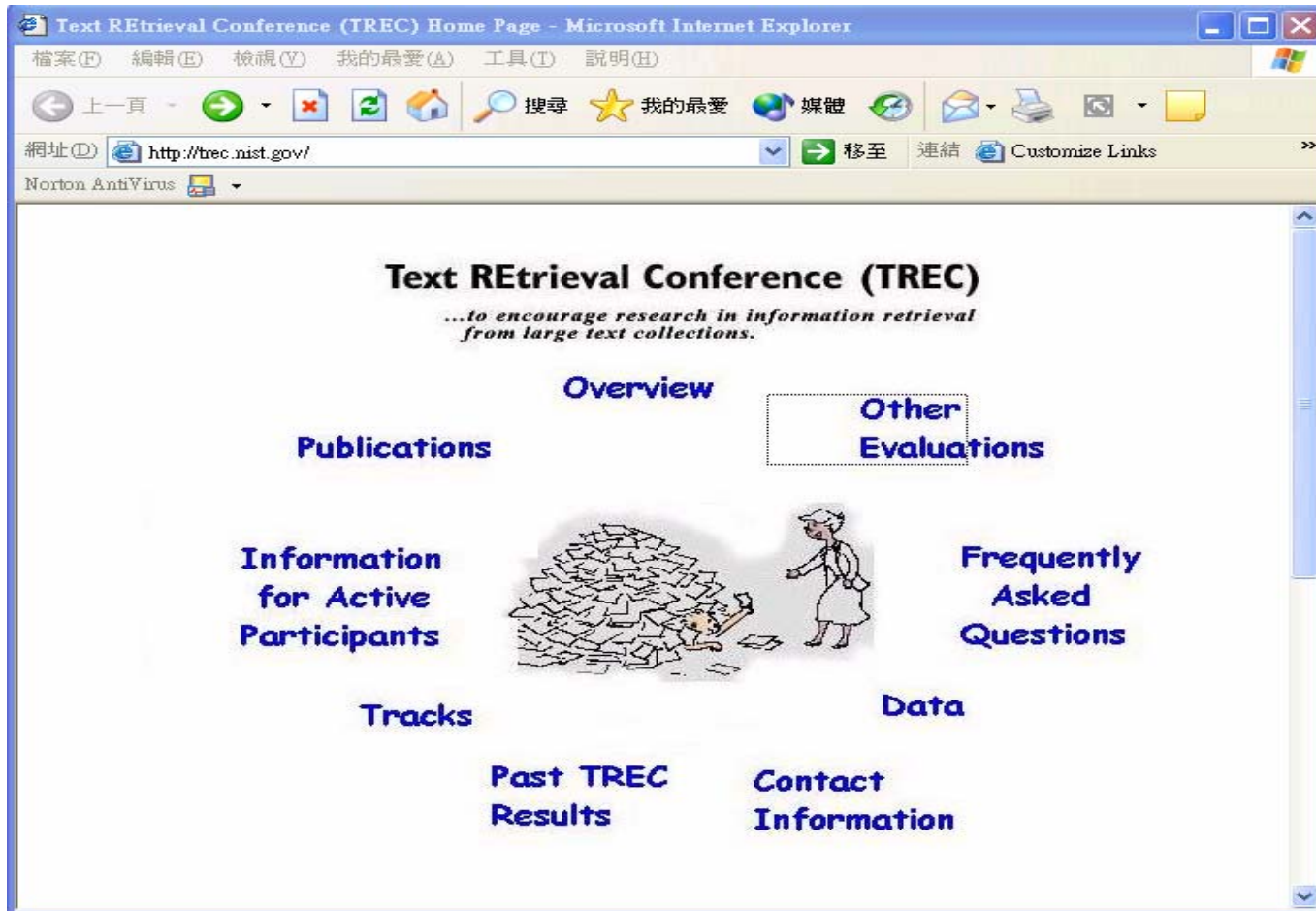
- 中研院：簡立峰(Text) 、王新民(Speech)
- 台大：陳信希(Text) 、陳光華(Text) ；李琳山(Speech)
- 成大：簡仁宗(Speech)
- 清大：張智星(Audio)
- 中央：楊接期(Text) ；張嘉惠(Text)
- 暨南：張景新(Text) 、林宣華(Text)
- 政大：劉昭麟(Text)

Institutes/Researchers

- Foreign
 - MIT (Speech)
 - **CMU** (Video/Speech)
 - **UMass** (Text/Speech)
 - **Cambridge** (Text/Speech)
 - Microsoft (Text/Speech)
 - IBM (Text/Speech)
 - MITRE (Text/Speech)
 - BBN (Speech)
 - HP (Speech/Text)
 -

Contests

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



Contests

- US National Institute of Standards and Technology

Benchmark Tests - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

地址(D) <http://www.nist.gov/speech/tests/index.htm>

IAD Website

NIST
National Institute of
Standards and Technology

[Contact Webmaster](#)

Topic Detection and Tracking (TDT)

- [General Information](#)
- [TDT 2001 Evaluation](#)
- [TDT 2000 Evaluation](#)
- [1999 TDT3 Evaluation](#)
- [1998 TDT2 Evaluation](#)

Machine Translation

- [General Information](#)

Broadcast News Recognition

- [1999 Broadcast News Evaluation](#)
- [1998 HUB-4 Broadcast News Evaluation](#)
- [1997 HUB-4NE Broadcast News Evaluation](#)
- [1997 HUB-4E Broadcast News Evaluation](#)
- [1996 HUB-4 Broadcast News Evaluation](#)

Language Recognition

- [2003 Evaluation](#)

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 Evaluation](#)

1998 Speaker Recognition Evaluation

1997 Speaker Recognition Evaluation

1996 Speaker Recognition Evaluation

Conferences/Journals

- Conferences
 - ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR)
 - ACM Conference on Information Knowledge Management (CIKM)
 - ...
- Journals
 - Information Processing and Management
 - Journal of the American Society for Information Science
 - ACM Transactions on Asian Language Information Processing
 - ...