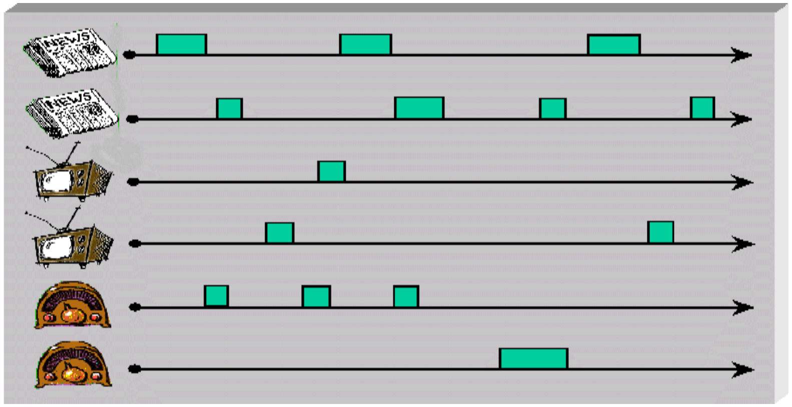
Information Retrieval and Extraction

Berlin Chen 2004



(Picture from the TREC web site)

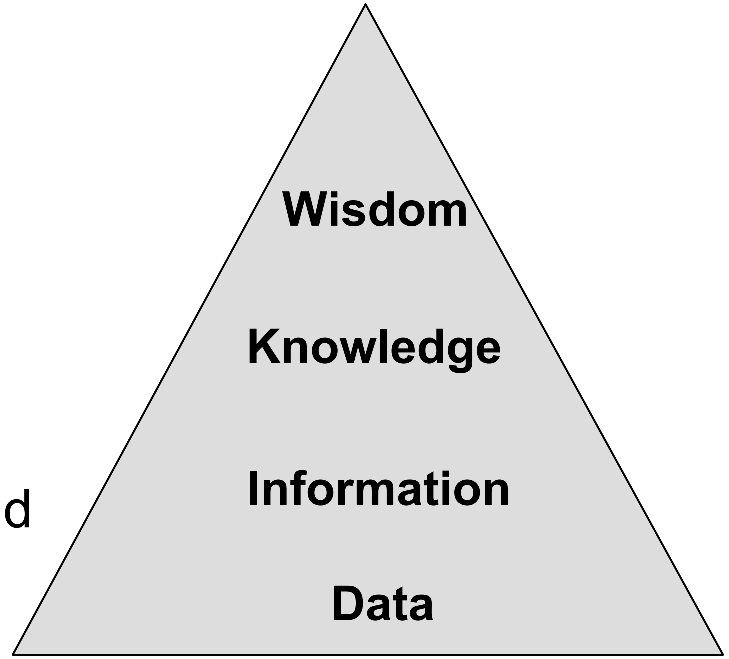
Textbook and References

- 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

- 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



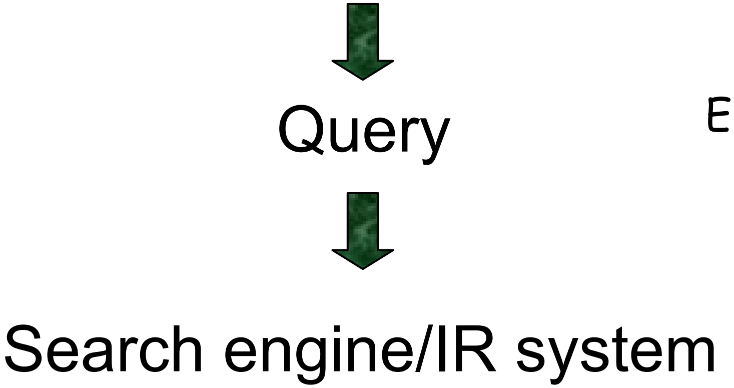
Motivation (cont.)

- 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!

Motivation (cont.)

- 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

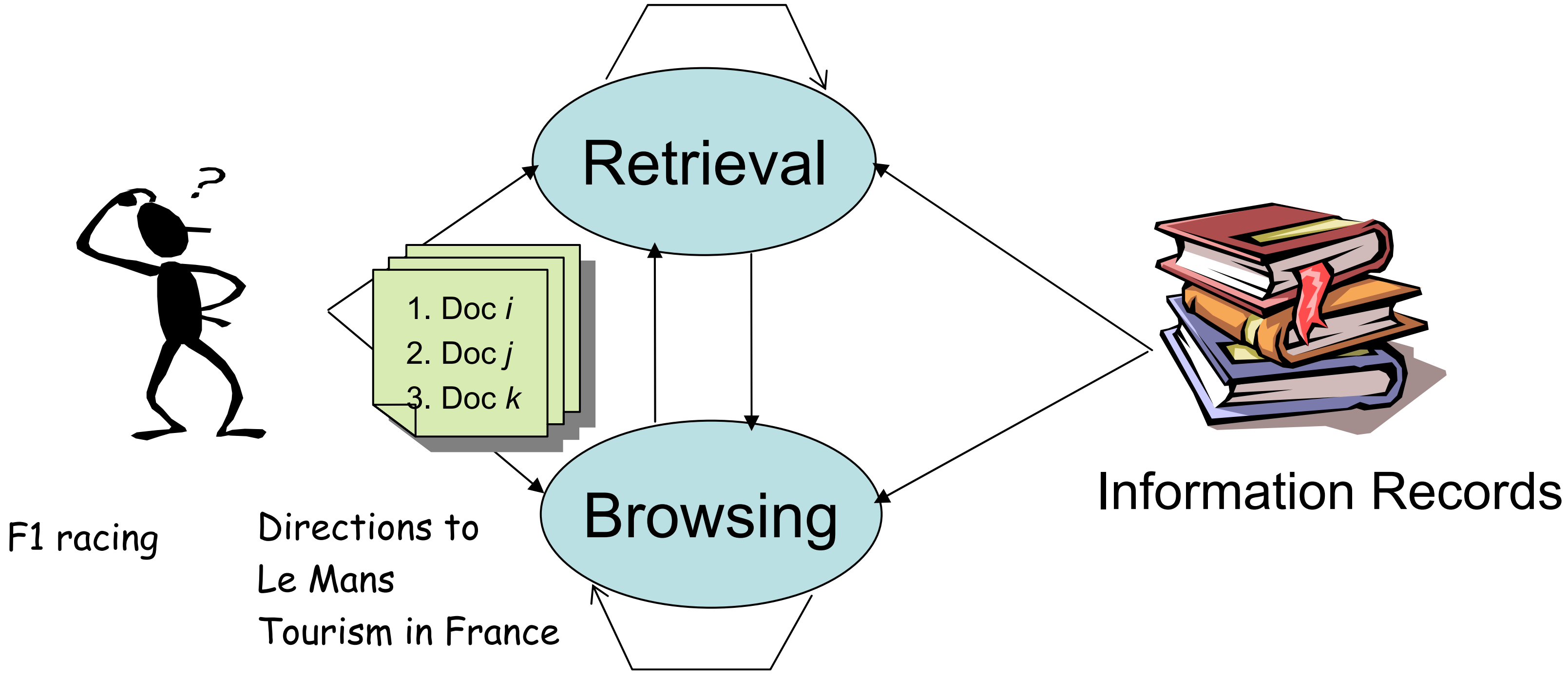
IR Main Issues

- 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

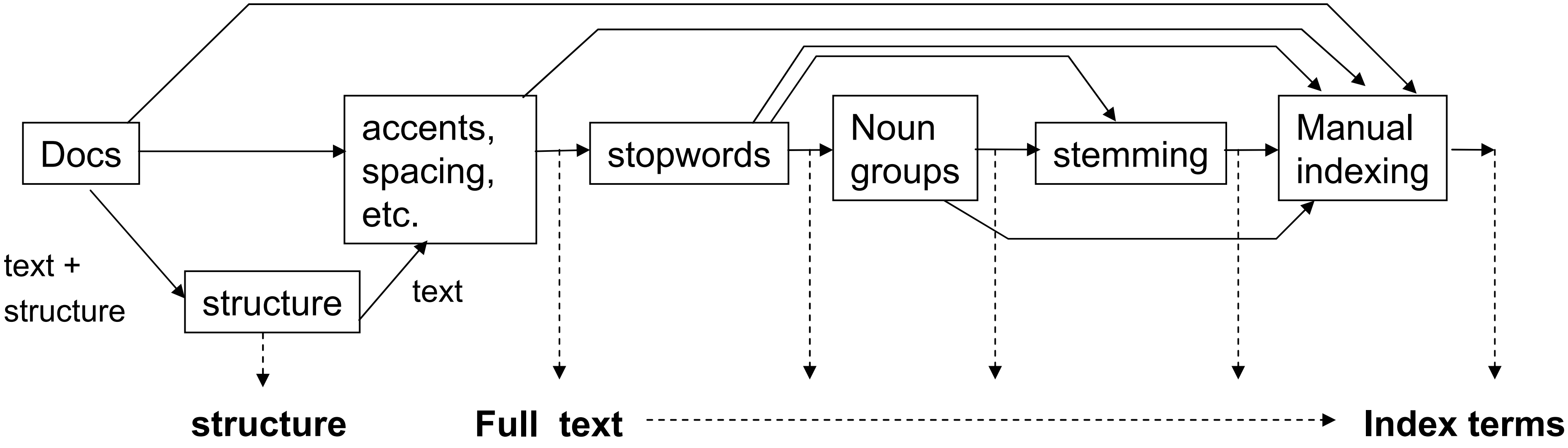
- A full text view (representation)
  - Represent document by its whole set of words
    - Complete but higher computational cost

- 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 (cont.)

- 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

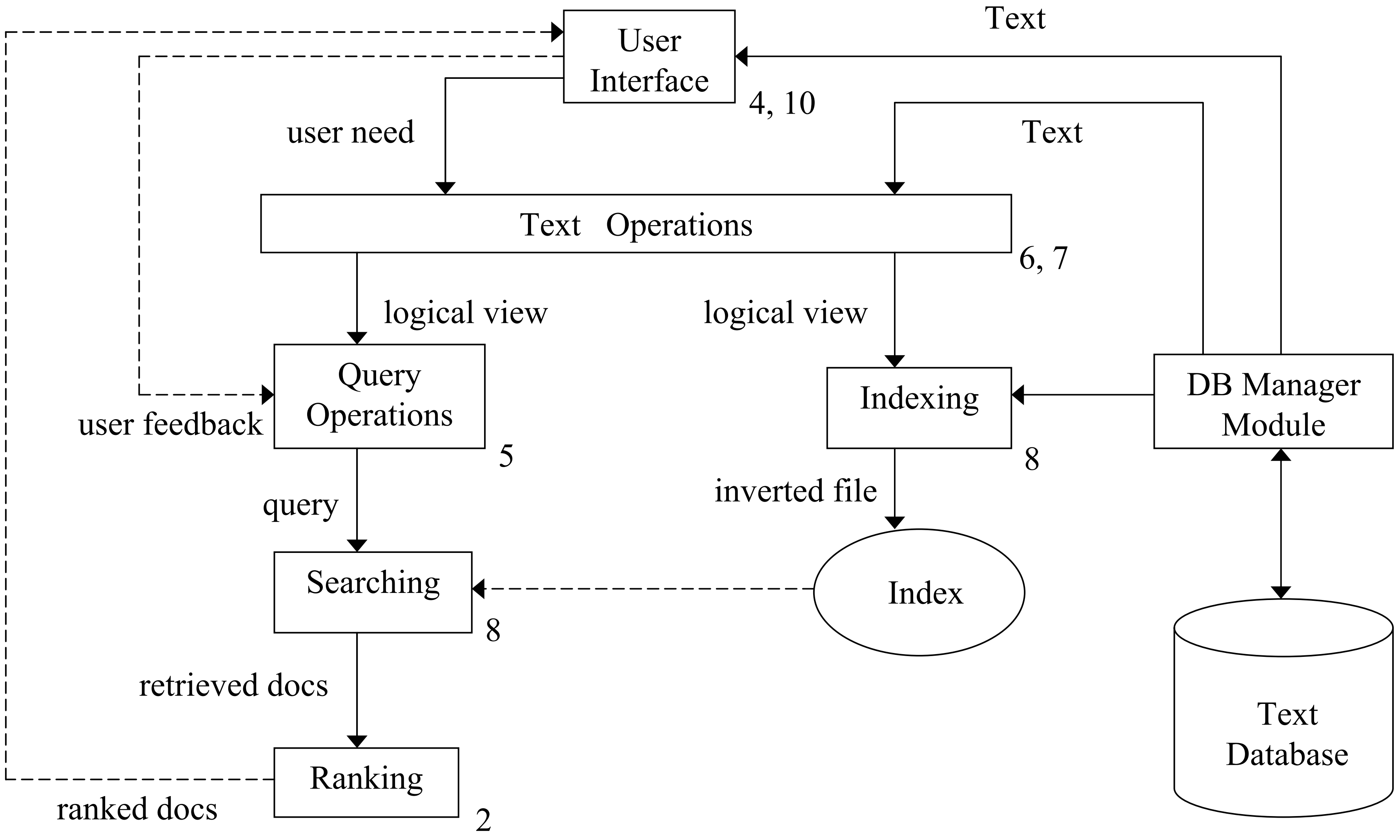
….

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



The Retrieval Process (cont.)

- 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

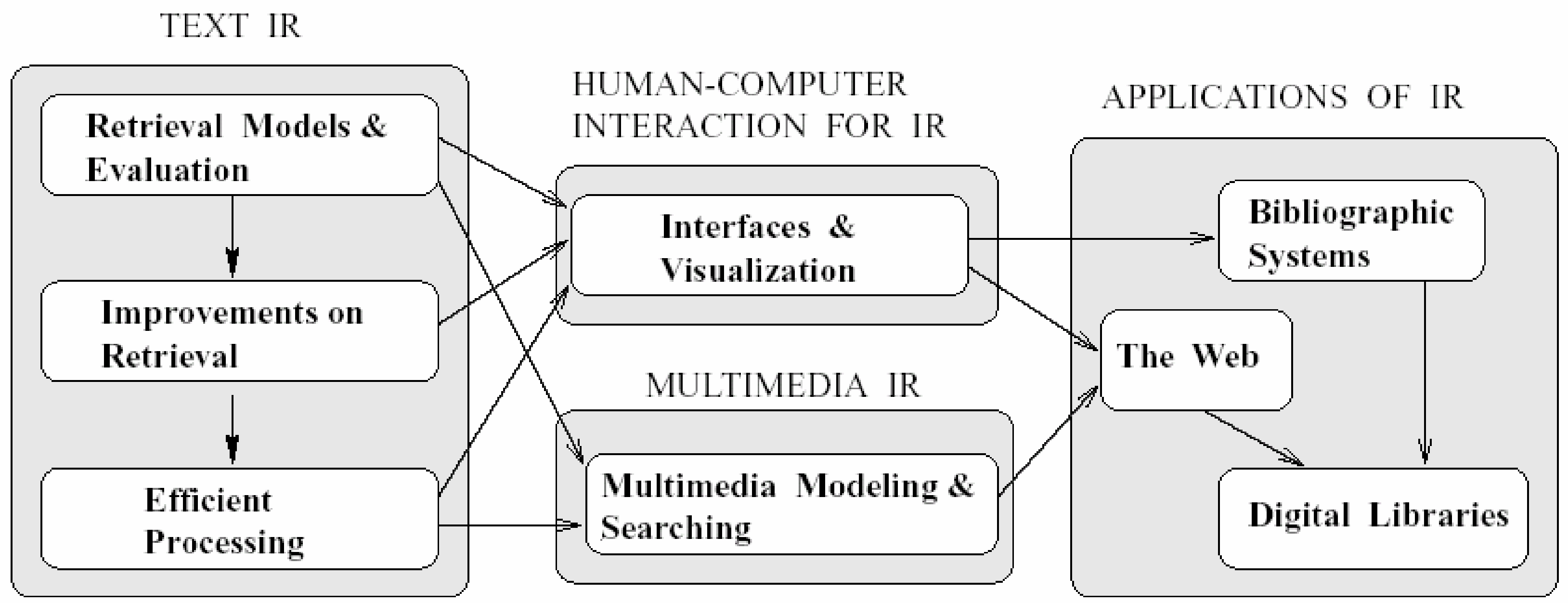
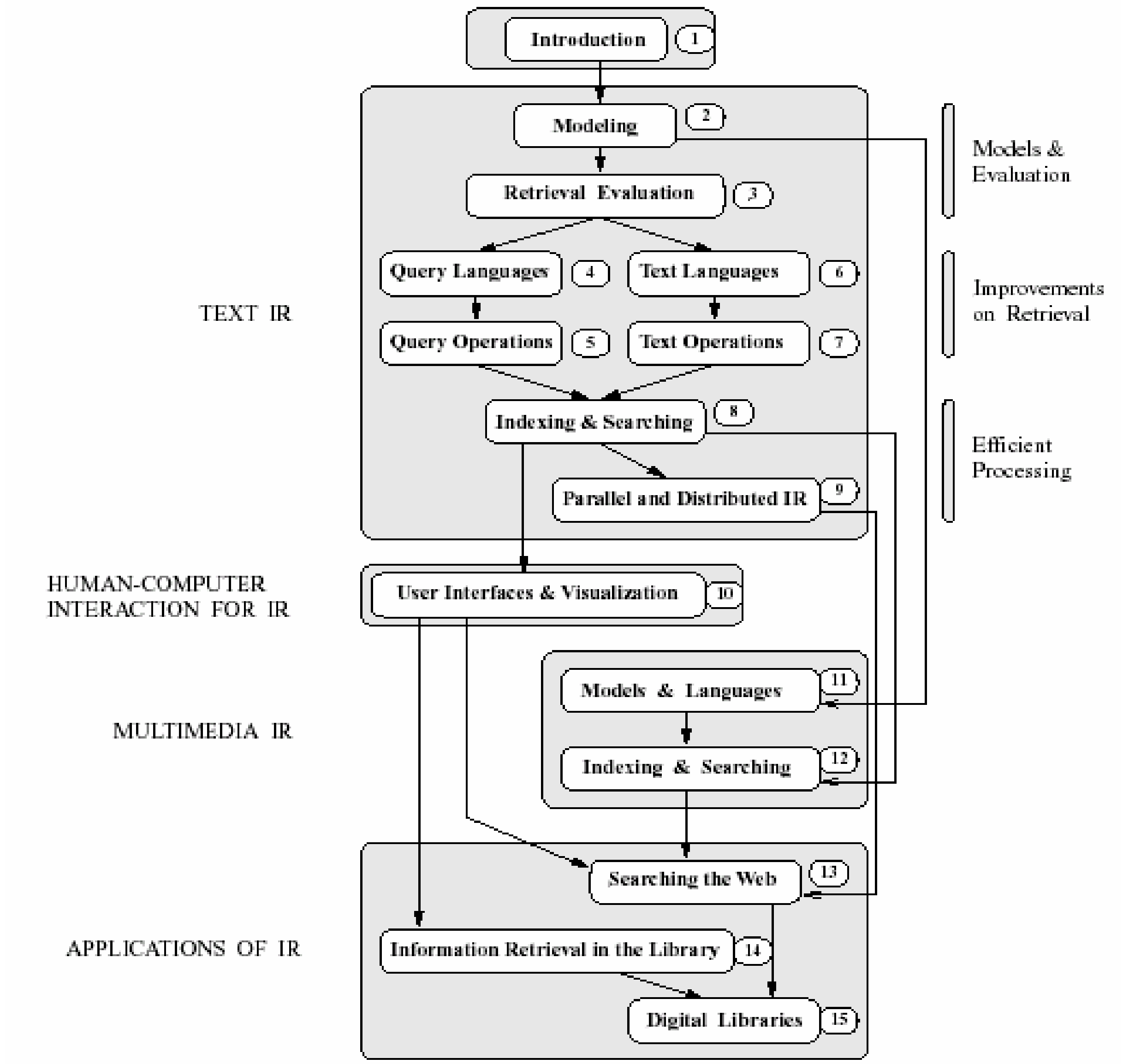

TEXT IR
Retrieval Models & Evaluation
Improvements on Retrieval
Efficient Processing
HUMAN-COMPUTER INTERACTION FOR IR
Interfaces & Visualization
MULTIMEDIA IR
Multimedia Modeling & Searching
APPLICATIONS OF IR
Bibliographic Systems
The Web
Digital Libraries

Figure 1.4 Topics which compose the book and their relationships.

Major Topics (cont.)

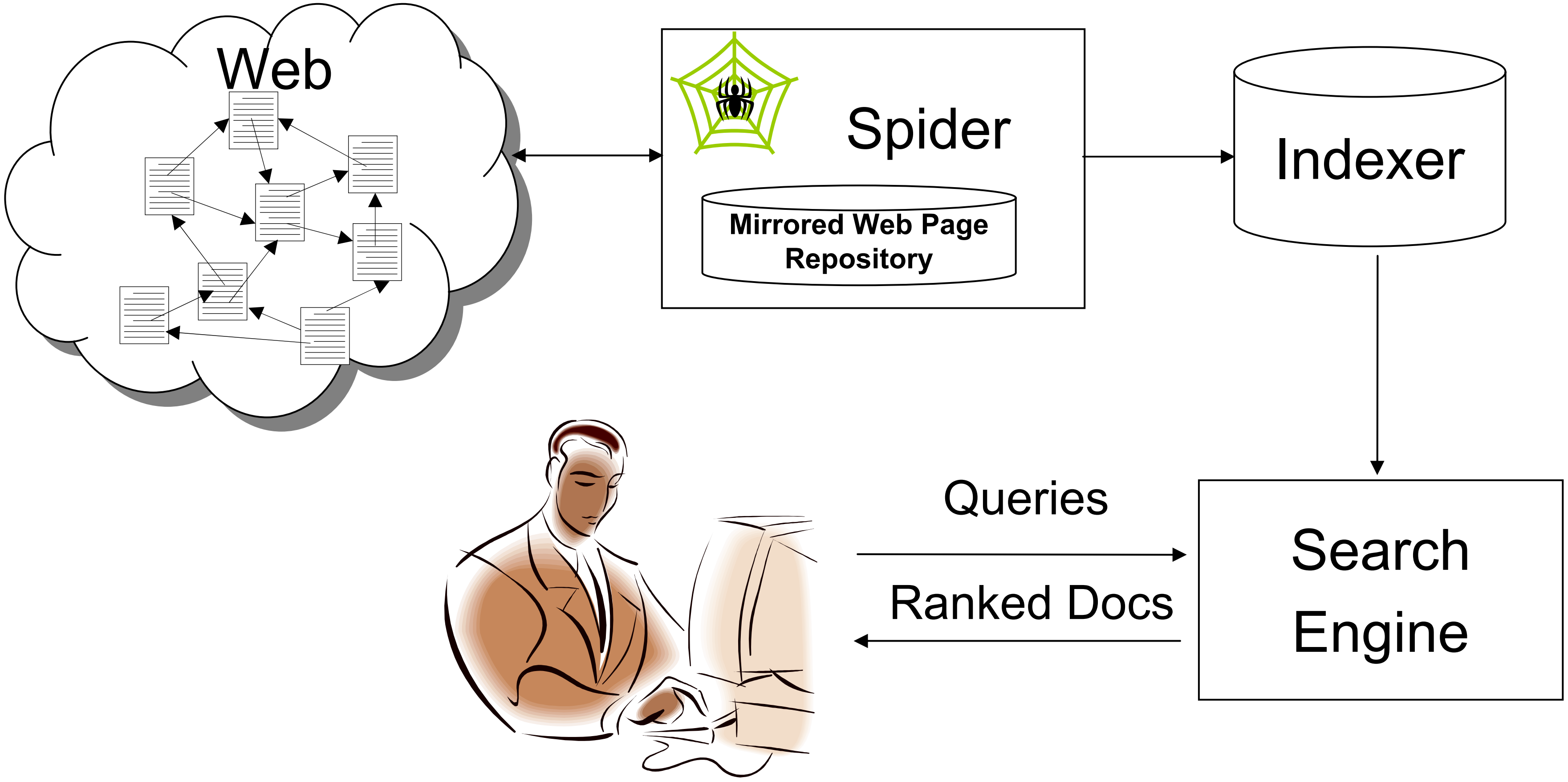
- Text IR
  - Retrieval models, evaluation methods, indexing
- Human-Computer Interaction (HCI)
  - Improved user interfaces and better data visualization tools
- Multimedia IR
  - Text, speech, audio and video contents
  - Multidisciplinary approaches
- Applications
  - Web, bibliographic systems, digital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

Textbook Topics

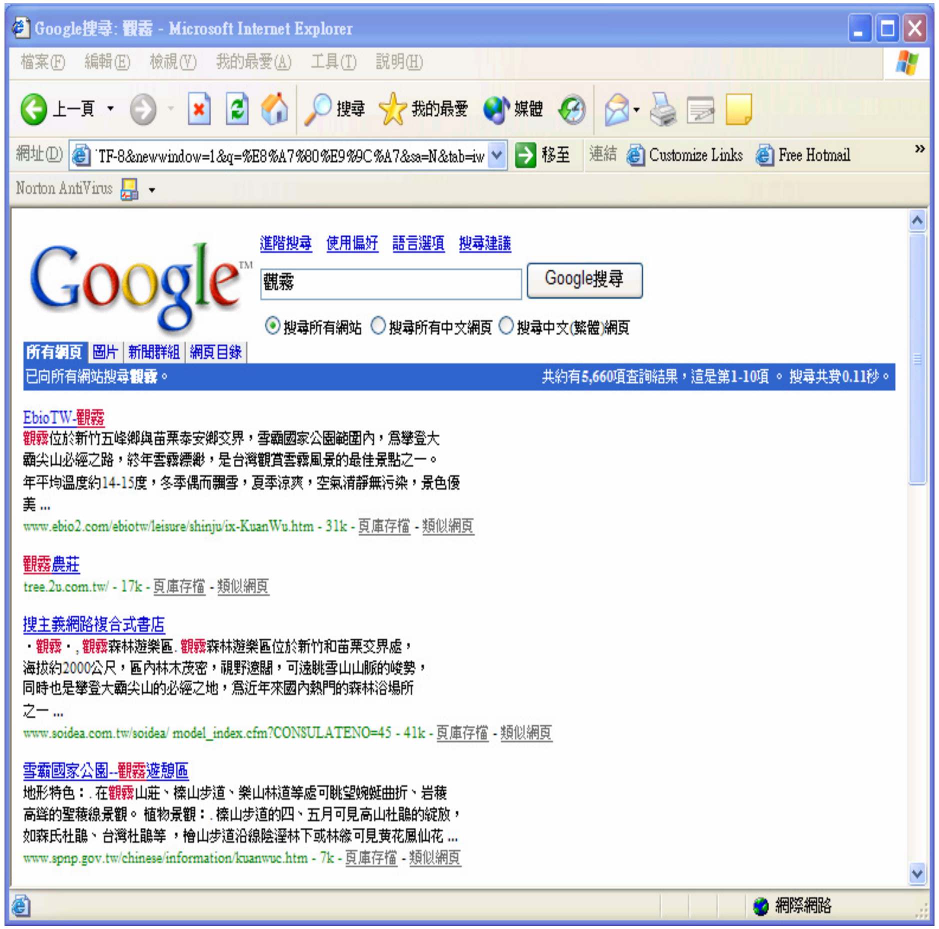


Text Information Retrieval

- Internet searching engine



Text Information Retrieval (cont.)


Google搜尋: 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) TF-8&newwindow=1&q=%E8%A7%80%E9%9C%A7&sa=N&tab=iw 移至 連結 Customize Links Free Hotmail
Norton AntiVirus
Google
進階搜尋 使用偏好 語言選項 搜尋建議
Google搜尋
所有網頁 圖片 新聞群組 網頁目錄
網際網路

觀霧

● 搜尋所有網站 ○ 搜尋所有中文網頁 ○ 搜尋中文(繁體)網頁

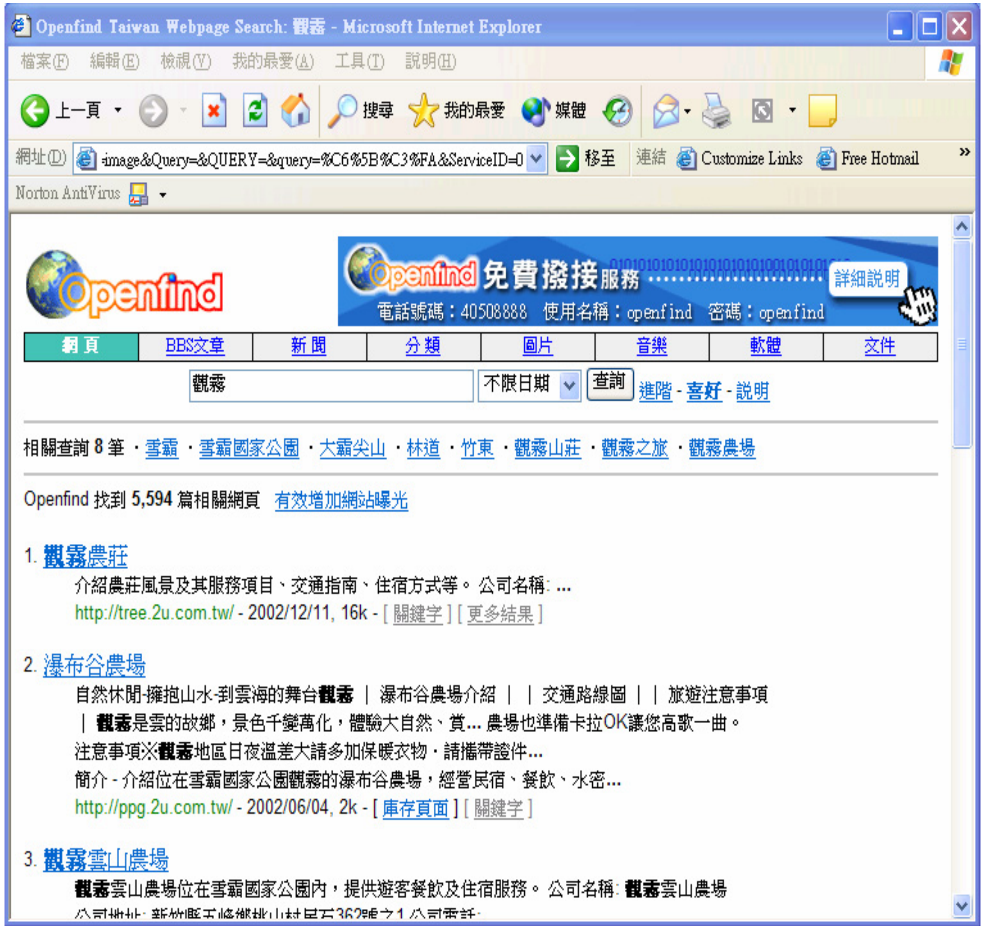
已向所有網站搜尋**觀霧**。 共約有**5,660**項查詢結果，這是第**1**-**10**項 。 搜尋共費**0.11**秒。

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，爲攀登大
霸尖山必經之路，終年雲霧縹緲，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，空氣清靜無污染，景色優
美 ...
www.ebio2.com/ebiotw/leisure/shinju/ix-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

搜主義網路複合式書店
・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，可遠眺雪山山脈的峻勢，
同時也是攀登大霸尖山的必經之地，爲近年來國內熱門的森林浴場所
之一 ...
www.soidea.com.tw/soidea/ model_index.cfm?CONSULATENO=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、榛山步道、樂山林道等處可眺望婉蜒曲折、岩稜
高聳的聖稜線景觀。 植物景觀：. 榛山步道的四、五月可見高山杜鵑的綻放，
如森氏杜鵑、台灣杜鵑等 ，檜山步道沿線陰溼林下或林緣可見黃花鳳仙花 ...
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
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) -image&Query=&QUERY=&query=%C6%5B%C3%FA&ServiceID=0 移至 連結 Customize Links Free Hotmail
Norton AntiVirus
Openfind
Openfind 免費撥接服務
詳細說明
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

觀霧

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

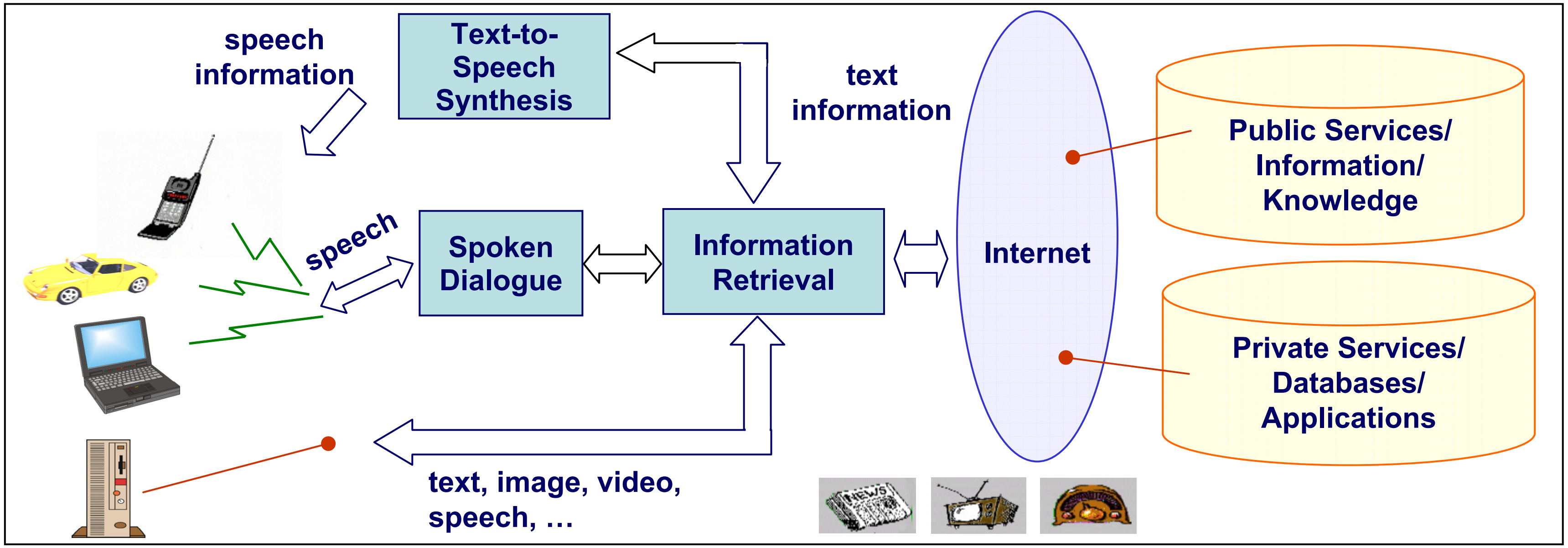
Openfind 找到 **5,594** 篇相關網頁 有效增加網站曝光

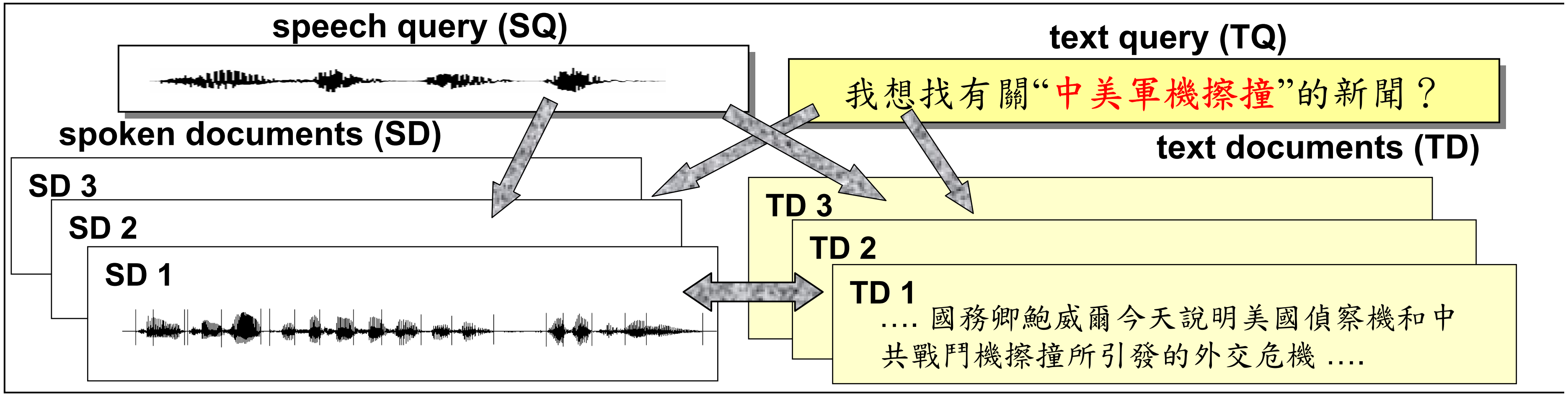
1. **觀霧**農莊
介紹農莊風景及其服務項目、交通指南、住宿方式等。 公司名稱: ...
http://tree.2u.com.tw/ - 2002/12/11, 16k - [關鍵字] [更多結果]

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: 新竹縣五峰鄉桃山村民石362號之1 公司電話:

Speech Information Retrieval


speech information
Text-to-Speech Synthesis
text information
speech
Spoken Dialogue
Information Retrieval
Internet
Public Services/ Information/ Knowledge
Private Services/ Databases/ Applications
text, image, video, speech, …

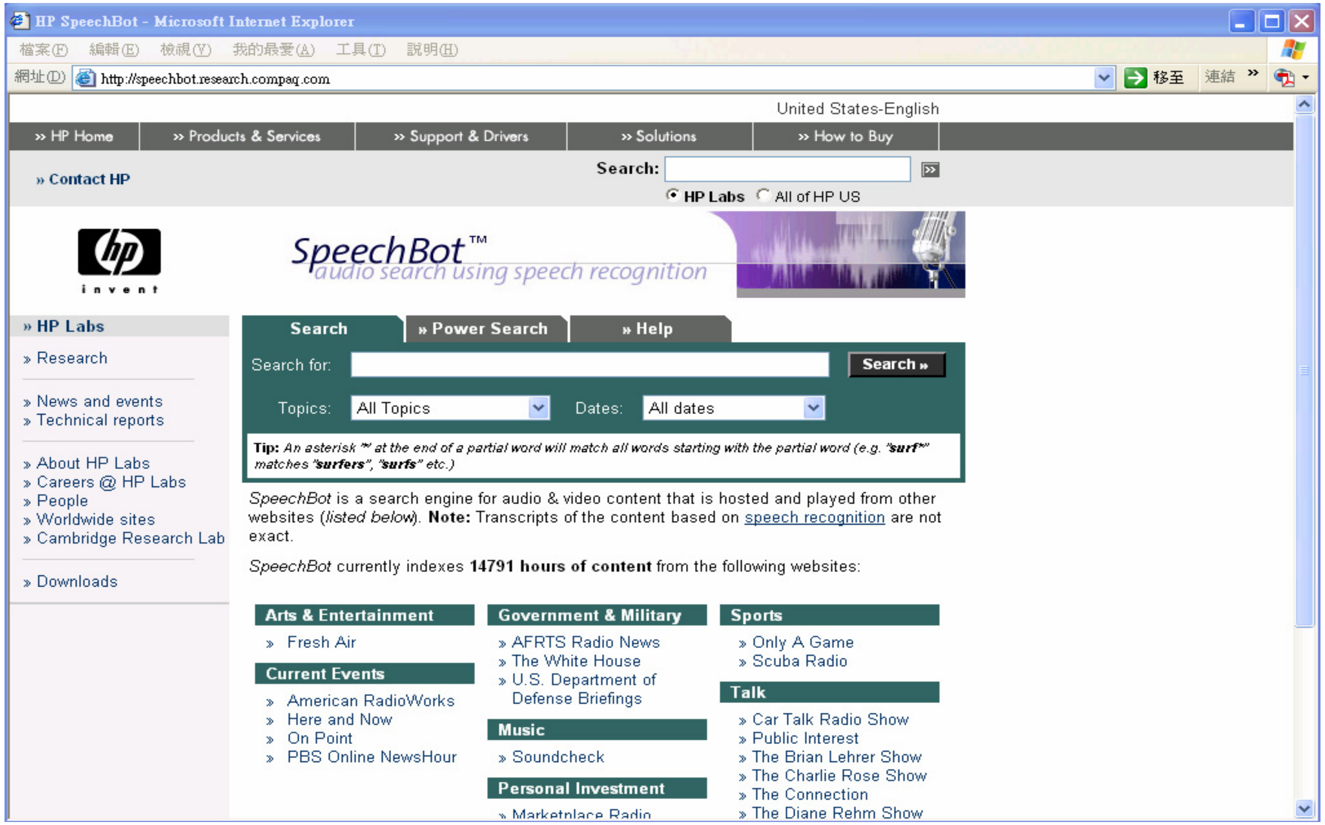


我想找有關“中美軍機擦撞”的新聞？

…. 國務卿鮑威爾今天說明美國偵察機和中共戰鬥機擦撞所引發的外交危機 ….

Speech Information Retrieval (cont.)

- 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/)



Search for:

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:

Arts & Entertainment
- Fresh Air

Current Events
- American RadioWorks
- Here and Now
- On Point
- PBS Online NewsHour

Government & Military
- AFRTS Radio News
- The White House
- U.S. Department of Defense Briefings

Music
- Soundcheck

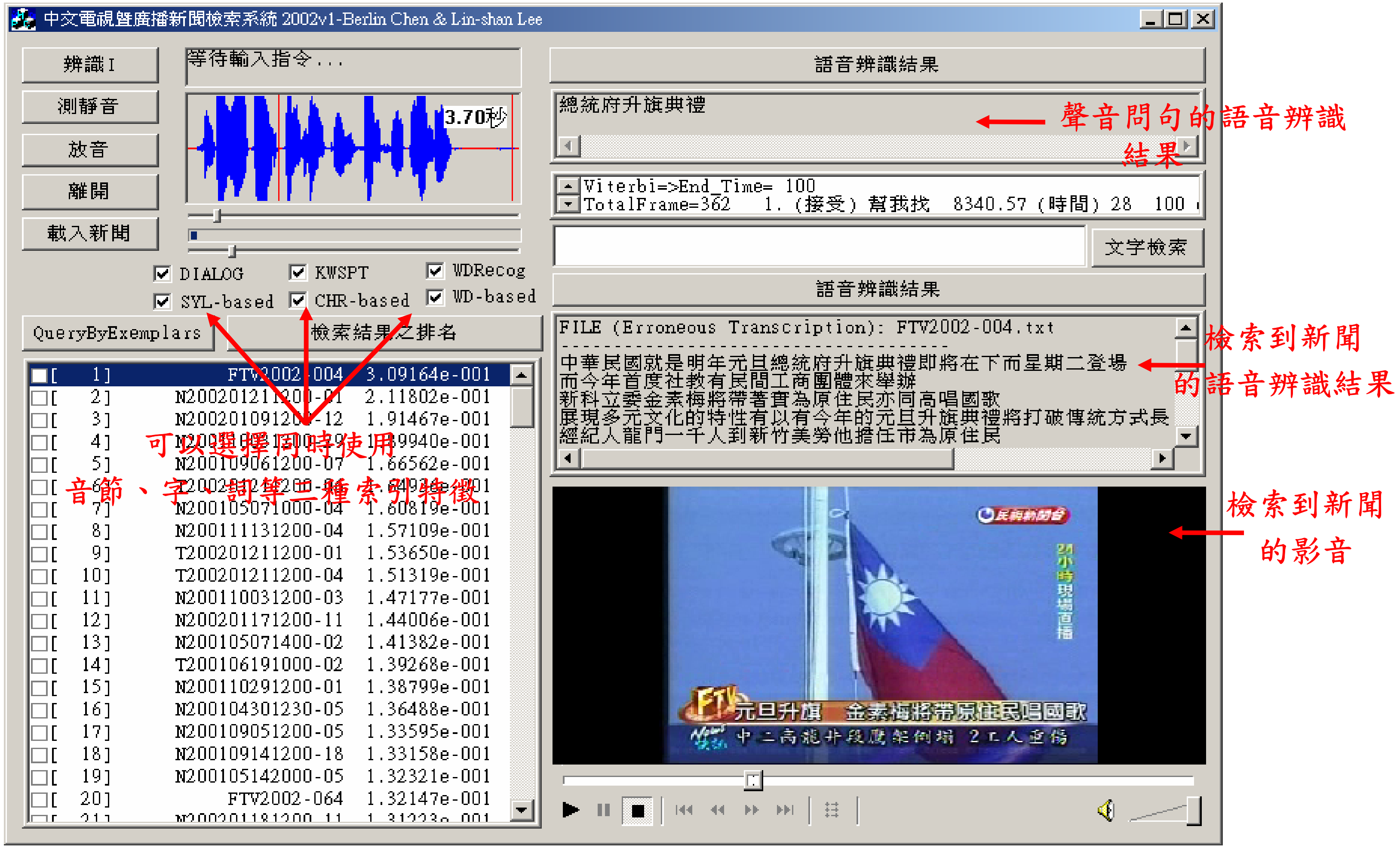
Personal Investment
- Marketplace Radio

Sports
- Only A Game
- Scuba Radio

Talk
- Car Talk Radio Show
- Public Interest
- The Brian Lehrer Show
- The Charlie Rose Show
- The Connection
- The Diane Rehm Show

Speech Information Retrieval (cont.)

- 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

等待輸入指令...

3.70秒

☑ DIALOG ☑ KWSPT ☑ WDRecog
☑ SYL-based ☑ CHR-based ☑ WD-based

檢索結果之排名

```
□[  1]      FTV2002-004  3.09164e-001
□[  2] N200201211200-01  2.11802e-001
□[  3] N200201091200-12  1.91467e-001
□[  4] N2002010[illegible]1200-09  1.[illegible]9940e-001
□[  5] N200109061200-07  1.66562e-001
□[  6] T2002[illegible]1200-[illegible]6  1.6[illegible]e-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
```

可以選擇同時使用

音節、字、詞等三種索引特徵

語音辨識結果

總統府升旗典禮

聲音問句的語音辨識結果

```
Viterbi=>End_Time= 100
TotalFrame=362    1. (接受) 幫我找   8340.57 (時間) 28   100
```

語音辨識結果

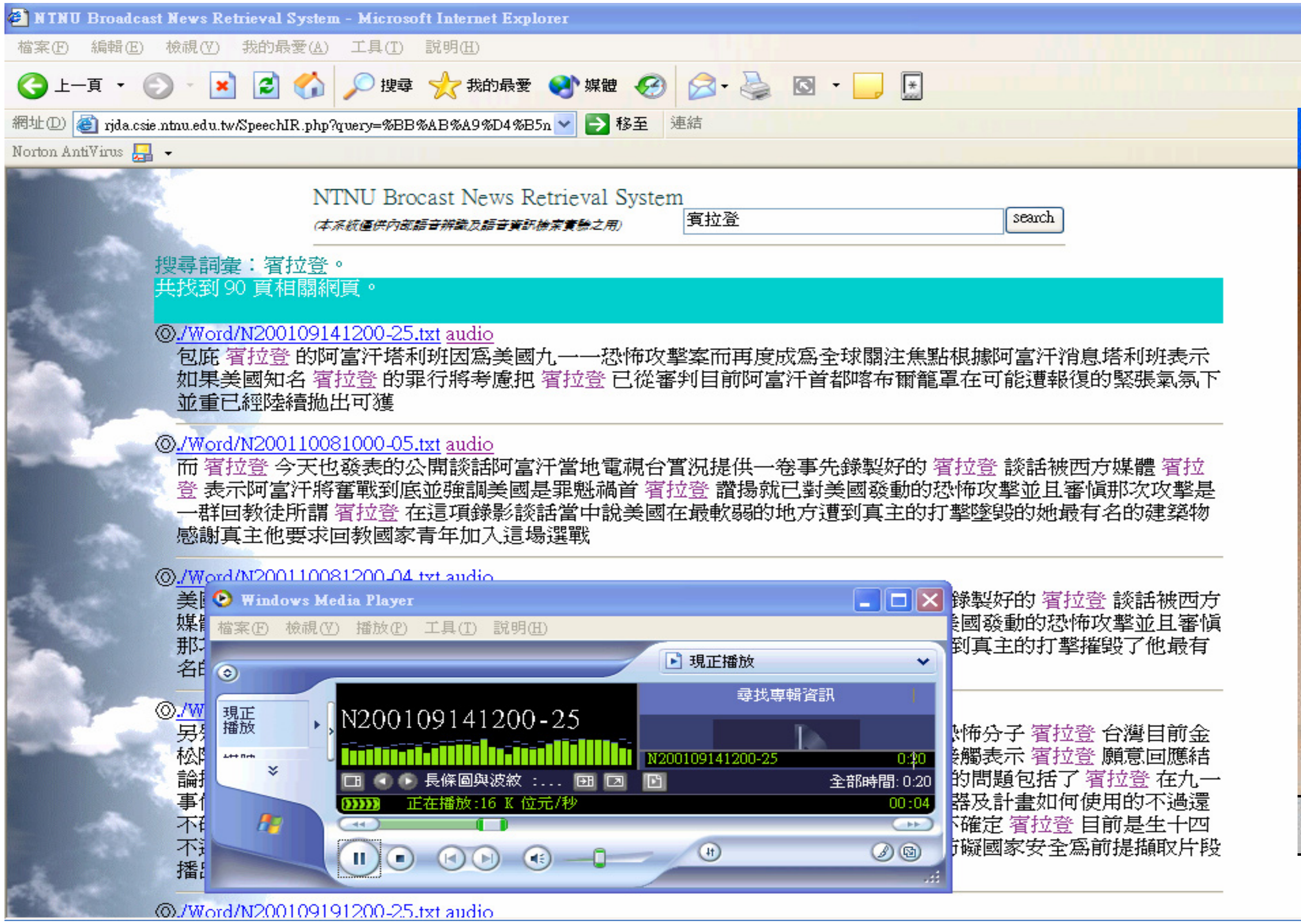
```
FILE (Erroneous Transcription): FTV2002-004.txt
-----------------------------------------
中華民國就是明年元旦總統府升旗典禮即將在下而星期二登場
而今年首度社教有民間工商團體來舉辦
新科立委金素梅將帶著實為原住民亦同高唱國歌
展現多元文化的特性有以有今年的元旦升旗典禮將打破傳統方式長
經紀人龍門一千人到新竹美勞他擔任市為原住民
```

檢索到新聞的語音辨識結果

檢索到新聞的影音

中文影音多媒體資訊檢索雛形展示系統。

Speech Information Retrieval (cont.)


NTNU Broadcast News Retrieval System - 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)
rjda.csie.ntnu.edu.tw/SpeechIR.php?query=%BB%AB%A9%D4%B5n
移至
連結
Norton AntiVirus
search
Windows Media Player
檔案(F) 檢視(V) 播放(P) 工具(T) 說明(H)
現正播放
尋找專輯資訊
長條圖與波紋 :...
全部時間: 0:20
正在播放:16 K 位元/秒
00:04
0:20

NTNU Brocast 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)

賓拉登

搜尋詞彙：賓拉登。

共找到 90 頁相關網頁。

◎./Word/N200109141200-25.txt audio

包庇 賓拉登 的阿富汗塔利班因爲美國九一一恐怖攻擊案而再度成爲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

◎./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 讚揚就已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墜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

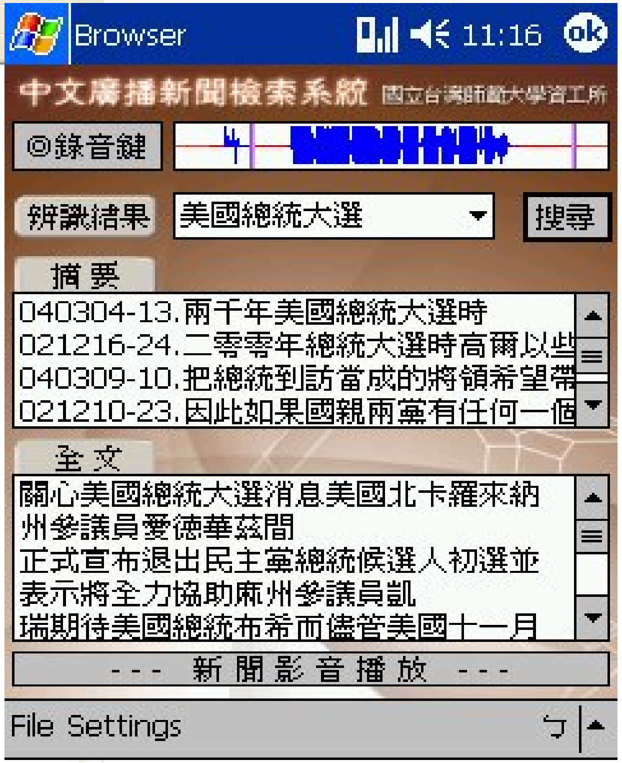
◎./Word/N200110081200-04.txt audio

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

◎./W
另 … 恐怖分子 賓拉登 台灣目前金
松 … 接觸表示 賓拉登 願意回應結
論 … 的問題包括了 賓拉登 在九一
事 … 器及計畫如何使用的不過還
不 … 不確定 賓拉登 目前是生十四
不 … 防礙國家安全爲前提擷取片段
播

◎./Word/N200109191200-25.txt audio

N200109141200-25


Browser
11:16
ok
中文廣播新聞檢索系統 國立台灣師範大學資工所
◎錄音鍵
辨識結果
搜尋
摘 要
全 文
--- 新 聞 影 音 播 放 ---
File Settings

美國總統大選

040304-13.兩千年美國總統大選時
021216-24.二零零年總統大選時高爾以些
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

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

Visual Information Retrieval

- Content-based approach

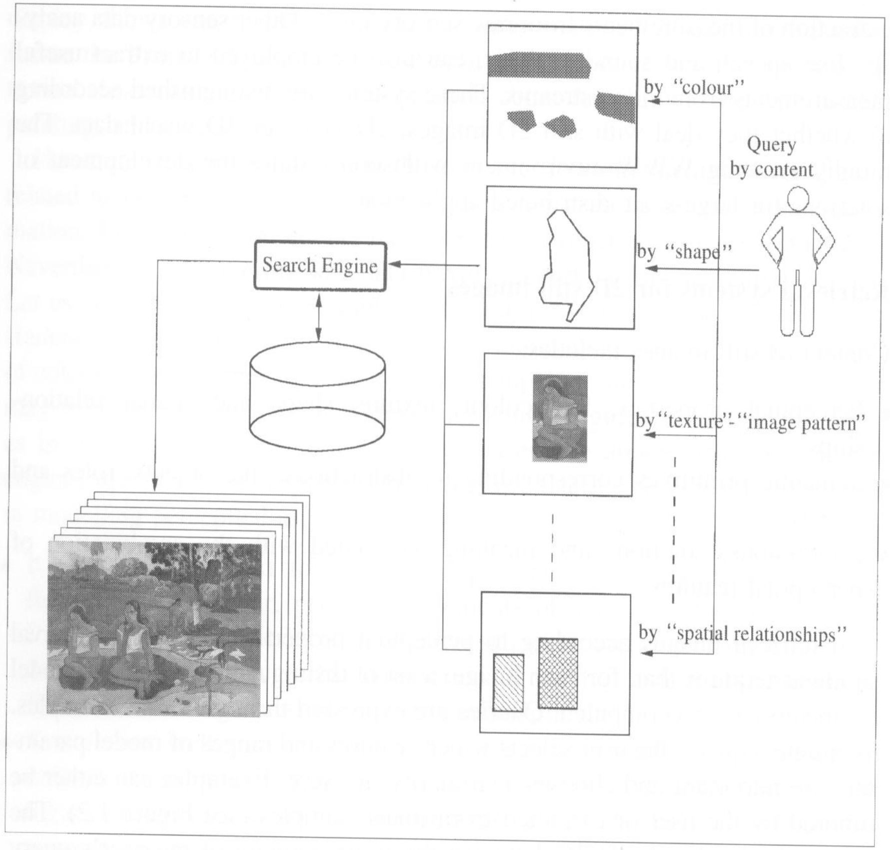


Figure 1.2 Different types of query by example.

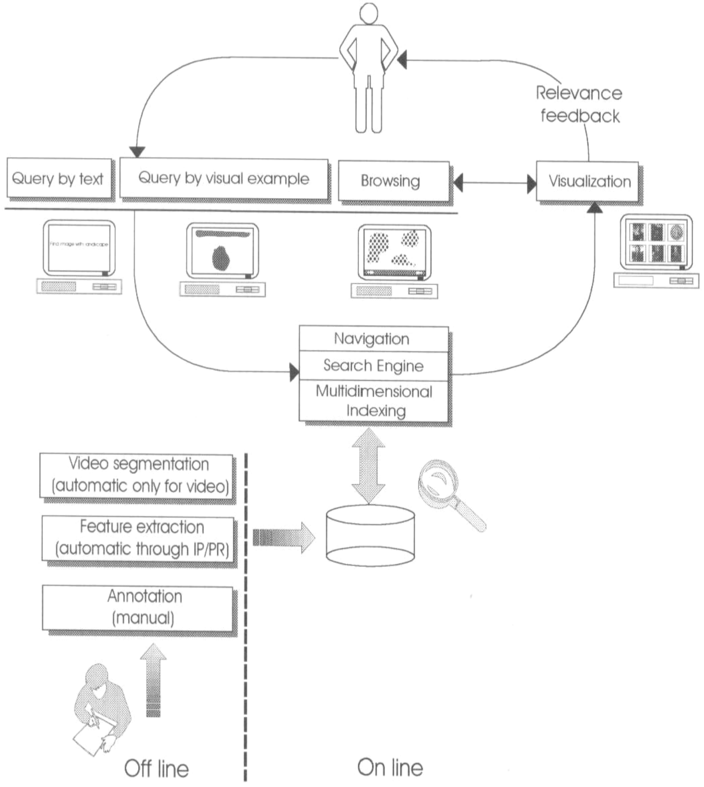
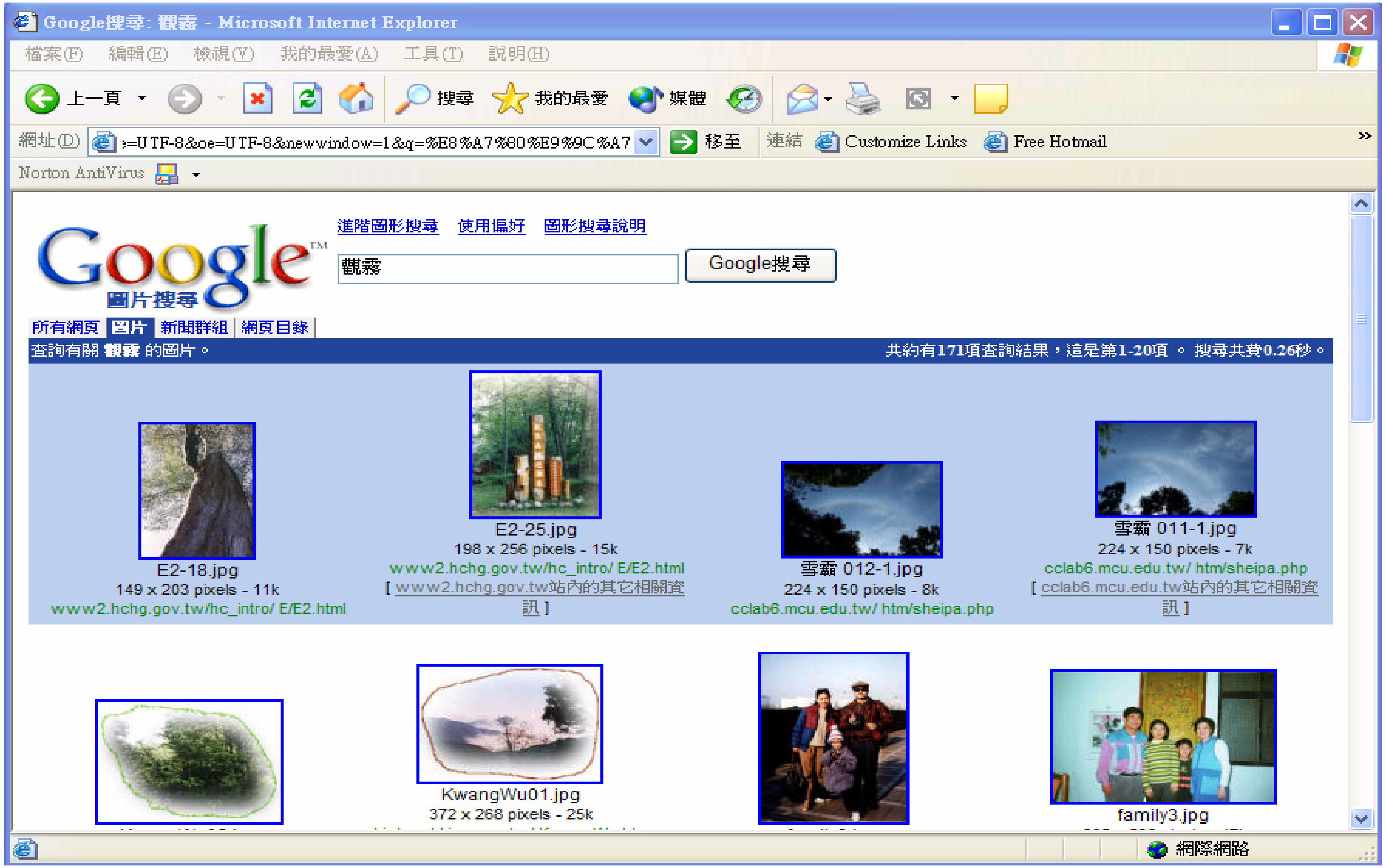


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

Visual Information Retrieval (cont.)

- Images with Texts


Google搜尋: 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)
=UTF-8&oe=UTF-8&newwindow=1&q=%E8%A7%80%E9%9C%A7
移至
連結
Customize Links
Free Hotmail
Norton AntiVirus
Google
圖片搜尋
進階圖形搜尋 使用偏好 圖形搜尋說明
觀霧
Google搜尋
所有網頁 圖片 新聞群組 網頁目錄
網際網路

查詢有關 **觀霧** 的圖片。 共約有171項查詢結果，這是第1-20項 。 搜尋共費0.26秒。

E2-18.jpg
149 x 203 pixels - 11k
www2.hchg.gov.tw/hc_intro/ E/E2.html

E2-25.jpg
198 x 256 pixels - 15k
www2.hchg.gov.tw/hc_intro/ E/E2.html
[www2.hchg.gov.tw站內的其它相關資訊]

雪霸 012-1.jpg
224 x 150 pixels - 8k
cclab6.mcu.edu.tw/ htm/sheipa.php

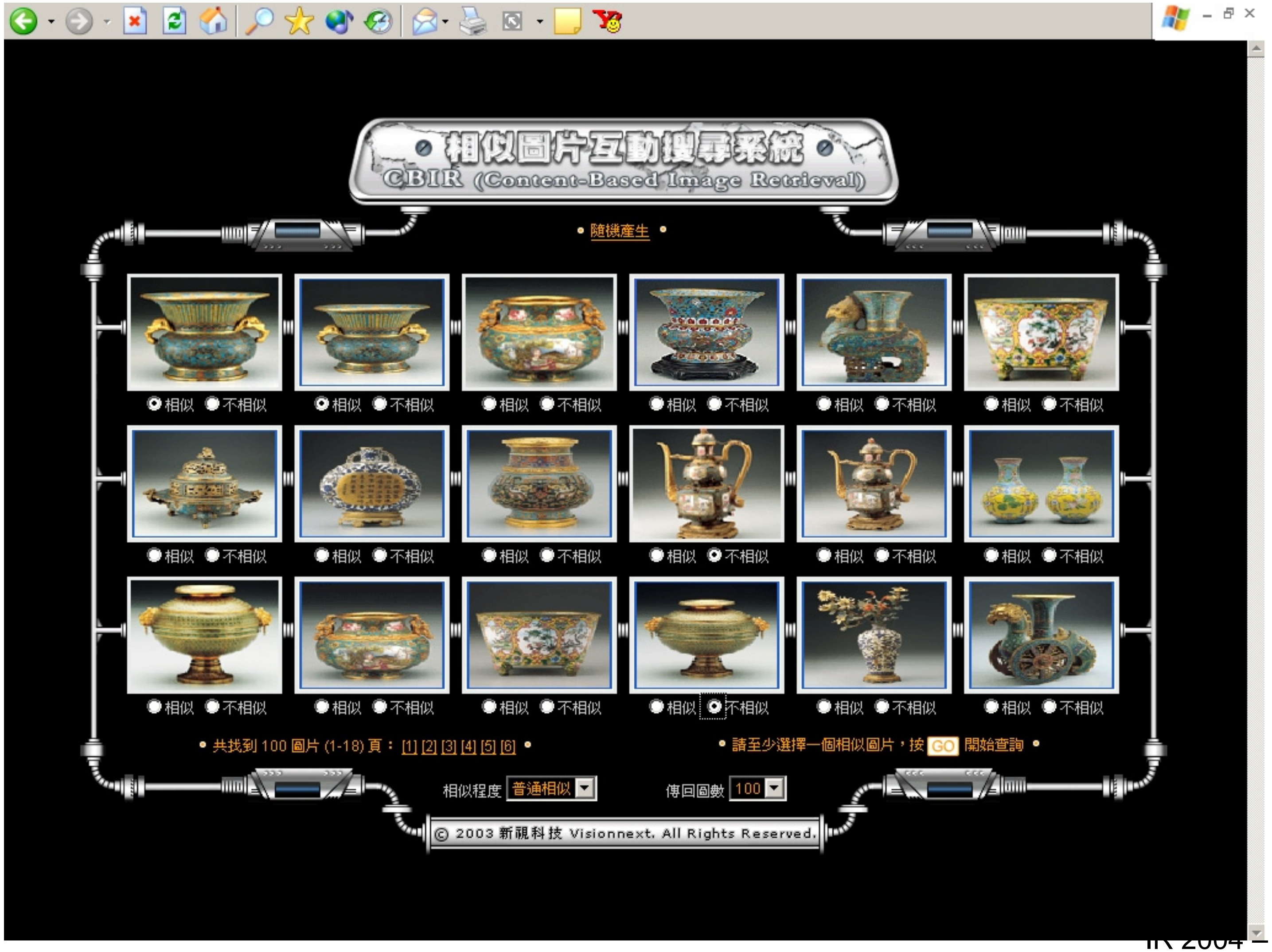
雪霸 011-1.jpg
224 x 150 pixels - 7k
cclab6.mcu.edu.tw/ htm/sheipa.php
[cclab6.mcu.edu.tw站內的其它相關資訊]

KwangWu01.jpg
372 x 268 pixels - 25k

family3.jpg

Visual Information Retrieval (cont.)

- Content-based Image Retrieval



相似圖片互動搜尋系統

CBIR (Content-Based Image Retrieval)

隨機產生

● 相似 ○ 不相似	● 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似
○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ● 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似
○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似	○ 相似 ● 不相似	○ 相似 ○ 不相似	○ 相似 ○ 不相似

共找到 100 圖片 (1-18) 頁：[1] [2] [3] [4] [5] [6]

請至少選擇一個相似圖片，按 GO 開始查詢

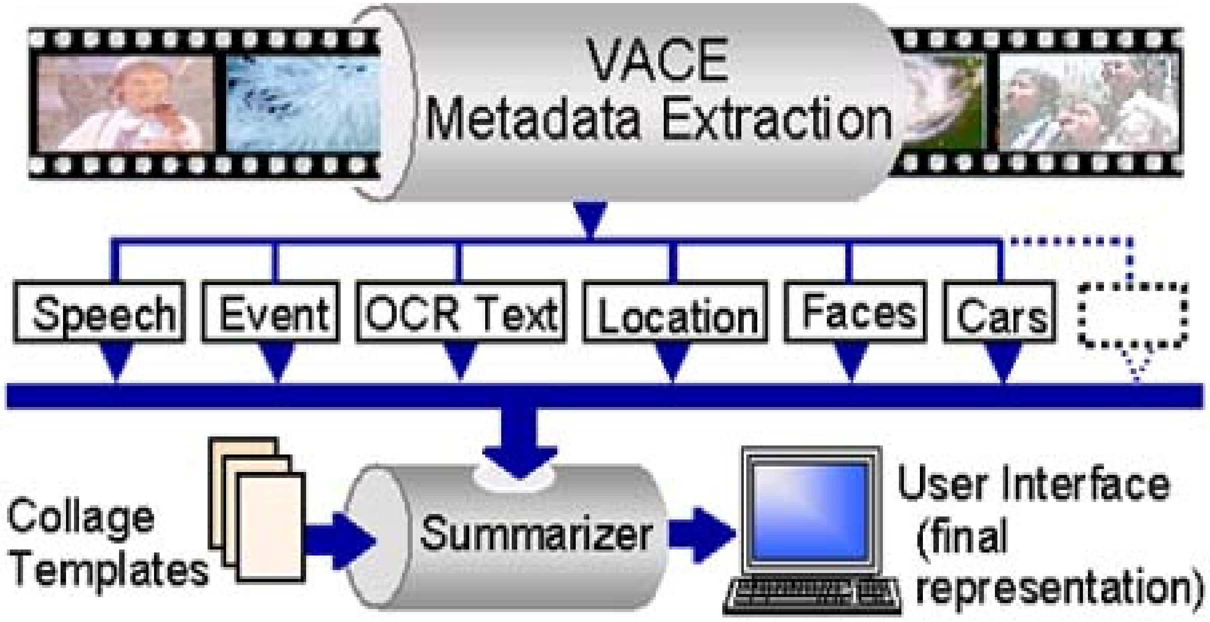
相似程度 普通相似

傳回圖數 100

© 2003 新視科技 Visionnext. All Rights Reserved.

Visual Information Retrieval (cont.)

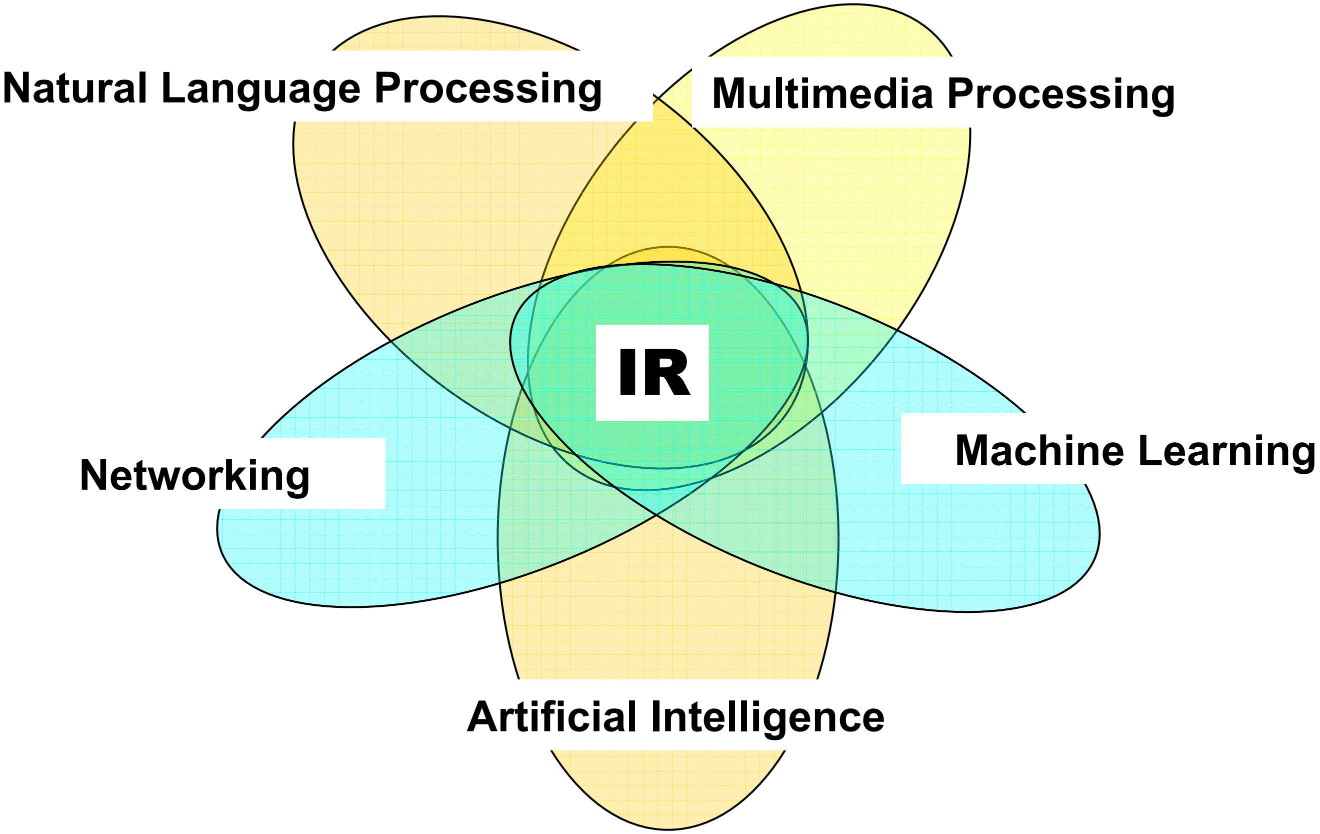
Video Analysis and Content Extraction



Other IR-Related Tasks

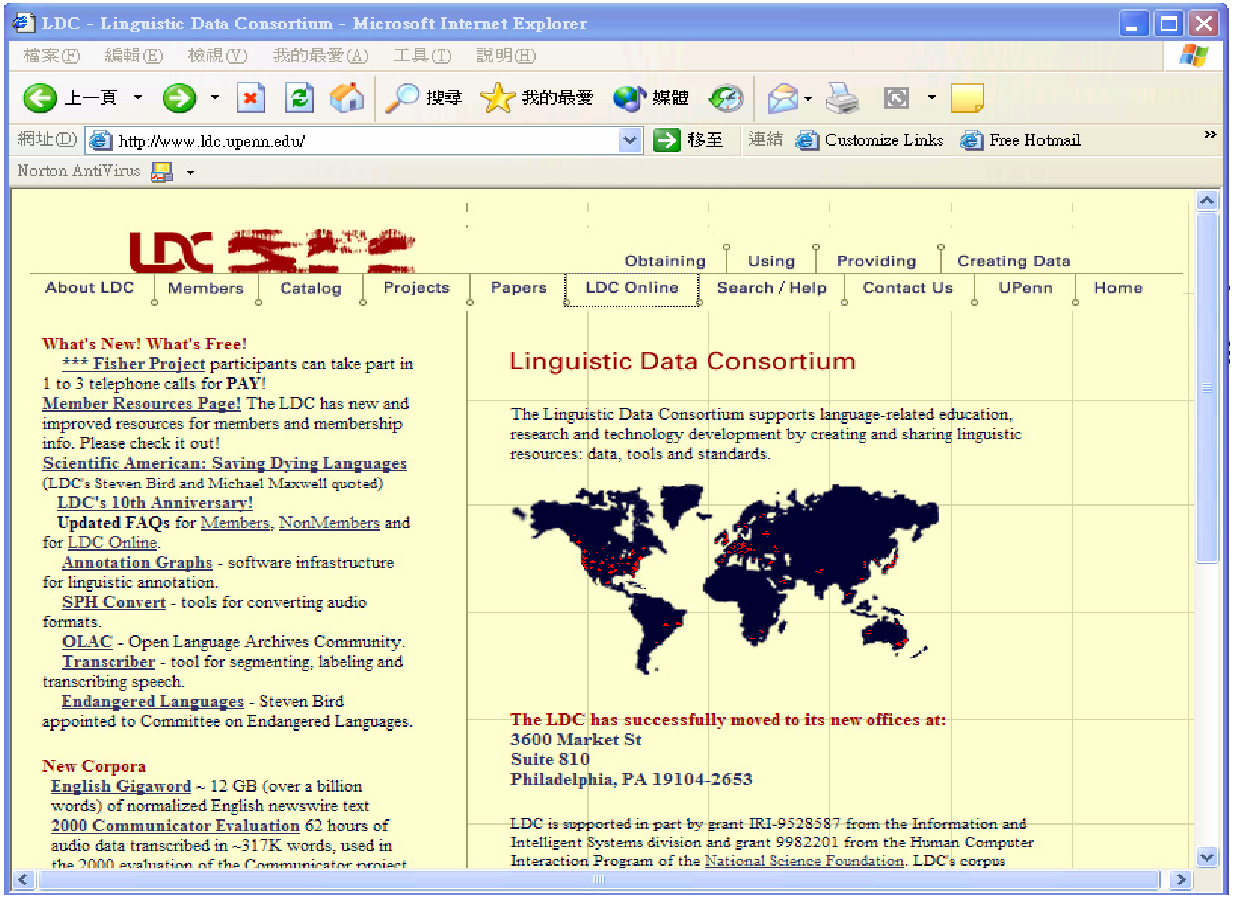
- Information filtering and routing
- **Document categorization**
- **Document clustering**
- **Document summarization**
- Information extraction
- Question answering
- Crosslingual information retrieval
- …..

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



What's New! What's Free!

*** Fisher Project participants can take part in 1 to 3 telephone calls for **PAY**!

Member Resources Page! The LDC has new and improved resources for members and membership info. Please check it out!

Scientific American: Saving Dying Languages (LDC's Steven Bird and Michael Maxwell quoted)

LDC's 10th Anniversary!

Updated FAQs for Members, NonMembers and for LDC Online.

Annotation Graphs - software infrastructure for linguistic annotation.

SPH Convert - tools for converting audio formats.

OLAC - Open Language Archives Community.

Transcriber - tool for segmenting, labeling and transcribing speech.

Endangered Languages - Steven Bird appointed to Committee on Endangered Languages.

New Corpora

English Gigaword ~ 12 GB (over a billion words) of normalized English newswire text

2000 Communicator Evaluation 62 hours of audio data transcribed in ~317K words, used in the 2000 evaluation of the Communicator project

Linguistic Data Consortium

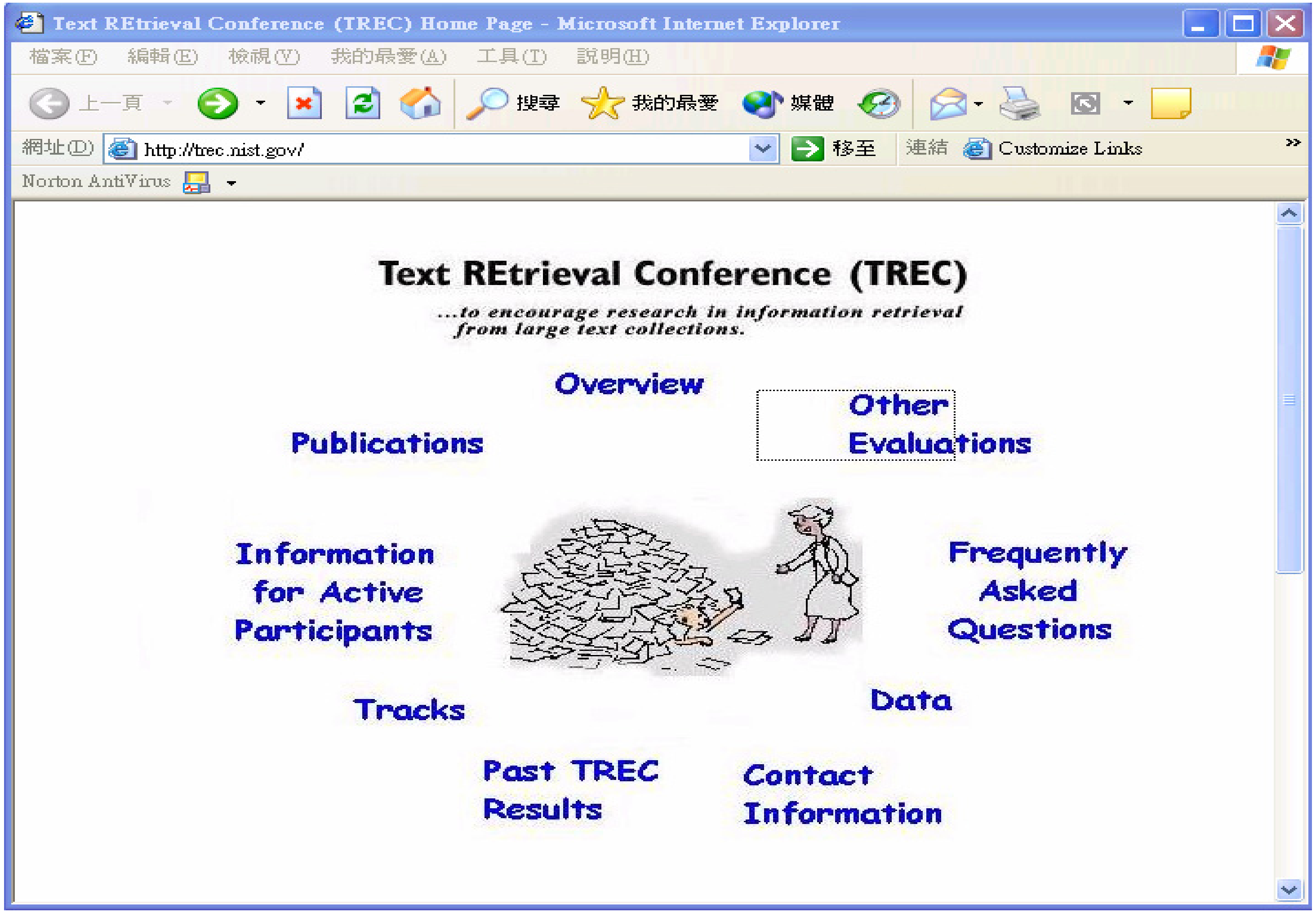
The Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards.

The LDC has successfully moved to its new offices at:
3600 Market St
Suite 810
Philadelphia, PA 19104-2653

LDC is supported in part by grant IRI-9528587 from the Information and Intelligent Systems division and grant 9982201 from the Human Computer Interaction Program of the National Science Foundation. LDC's corpus

Contests

- Text REtrieval Conference (TREC)


Text REtrieval Conference (TREC) Home Page - 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://trec.nist.gov/
移至
連結
Customize Links
Norton AntiVirus

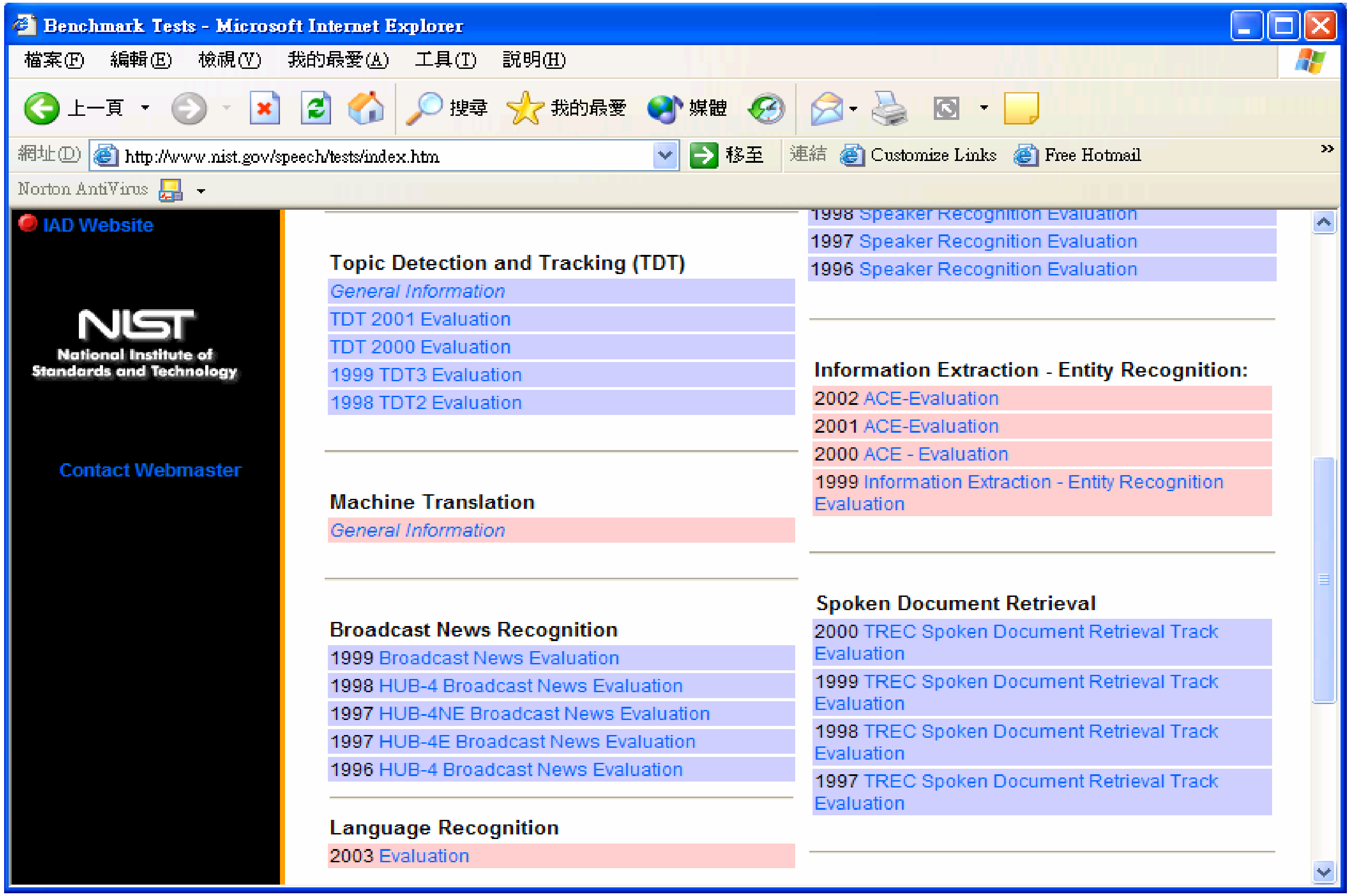
Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview
Other Evaluations
Publications
Information for Active Participants
Frequently Asked Questions
Tracks
Data
Past TREC Results
Contact Information

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)
上一頁
搜尋
我的最愛
媒體
網址(D)
http://www.nist.gov/speech/tests/index.htm
移至
連結
Customize Links
Free Hotmail
Norton AntiVirus
IAD Website
NIST
National Institute of Standards and Technology
Contact Webmaster

1998 Speaker Recognition Evaluation
1997 Speaker Recognition Evaluation
1996 Speaker Recognition Evaluation

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

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)
  - …