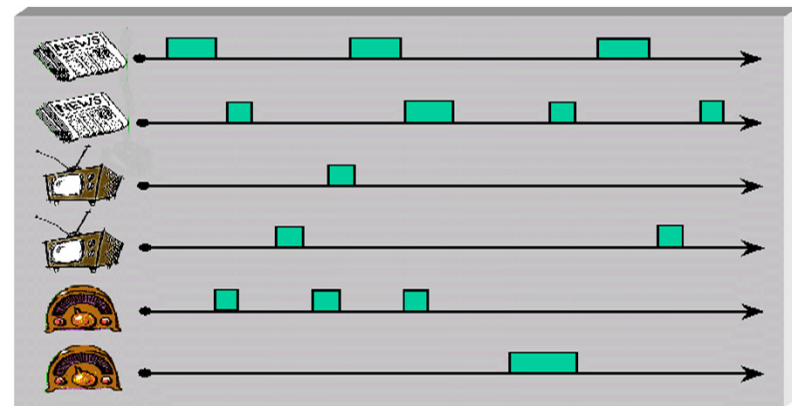


Information Retrieval and Extraction

Berlin Chen



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

Objectives of this Course

- Elaborate on the fundamentals of information retrieval (IR), a almost *fifty-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.)
- Practical 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 (1/2)

- Information Hierarchy

- **Data**

- The raw material of information

- **Information**

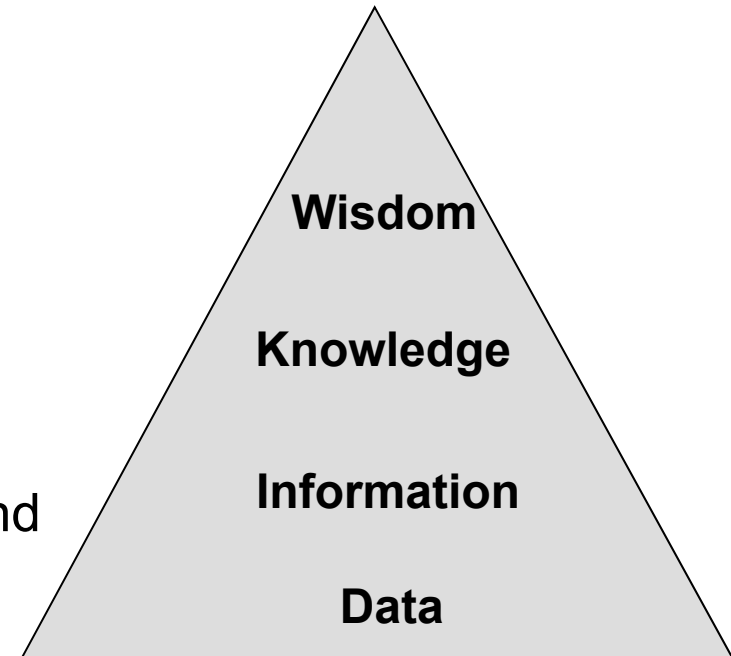
- Data organized and presented by someone

- **Knowledge**

- Information read, heard or seen and understood

- **Wisdom**

- 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 (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

- Information retrieval (IR) is the field concerned with the structure, analysis, or organization, searching and retrieval of information
 - Defined by Gerard Salton, a pioneer and leading figure in IR
- Handle **natural language text** (or free text) which is not always well structured and could be semantically ambiguous
- 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 failure!

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

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 of documents)
 - *Enterprise, institutional, and domain-specific search*
 - *Personal (desktop) search*
 -

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
 - Decentralized
 - Without frontiers: free universal access (*freedom to publish*)
 - Hypertext (HTTP protocol and browsers by Tim Berners-Lee)
 - Lack of well-defined data model

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*)

Web Changed Search!

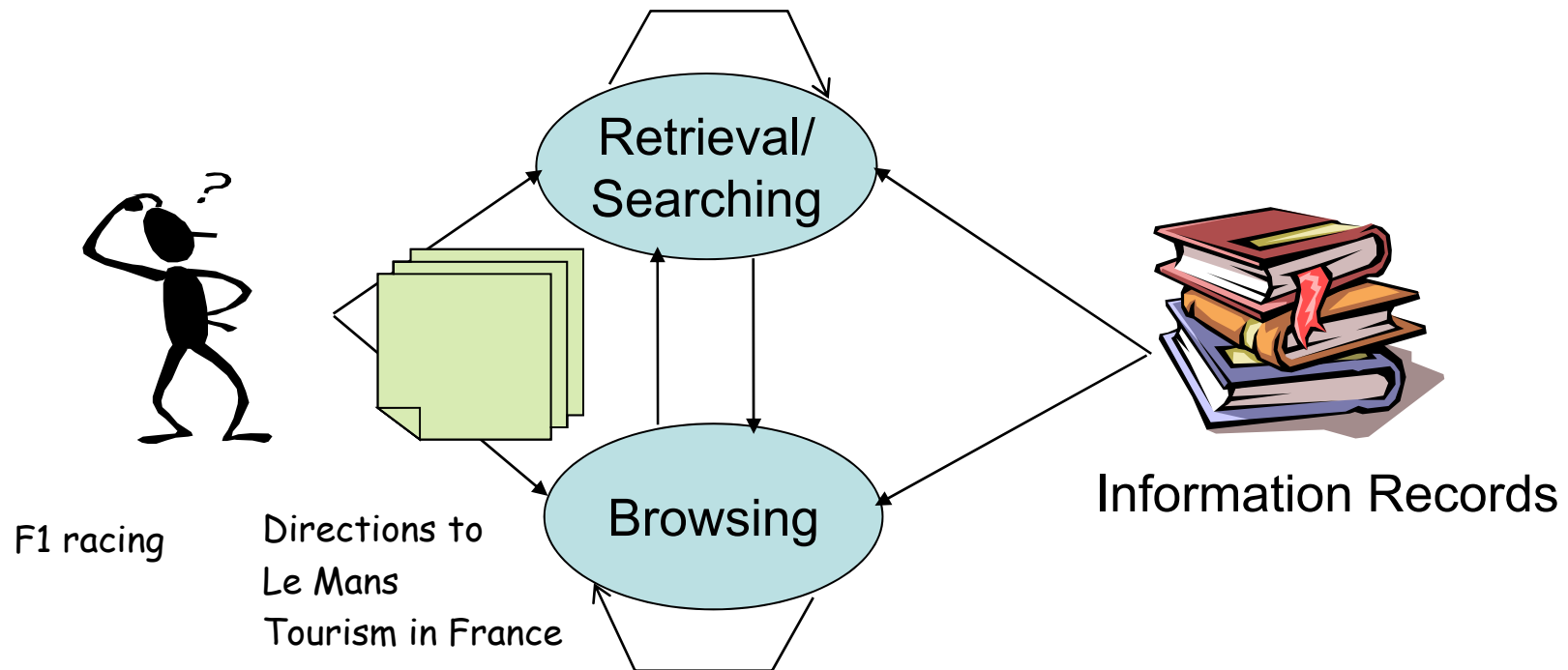
- Characteristics of document collection
 - Distributed natural => *crawling*
- The size of document collection
 - ~20 billion pages=> *performance* and *scalability* are big issues
- Relevance judgment in the face of the vast size of document collections
 - Hyperlinks and user clicks in documents => *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
- Web advertising and economic incentives
 - E-commerce, advertising <=> *Web spam*

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

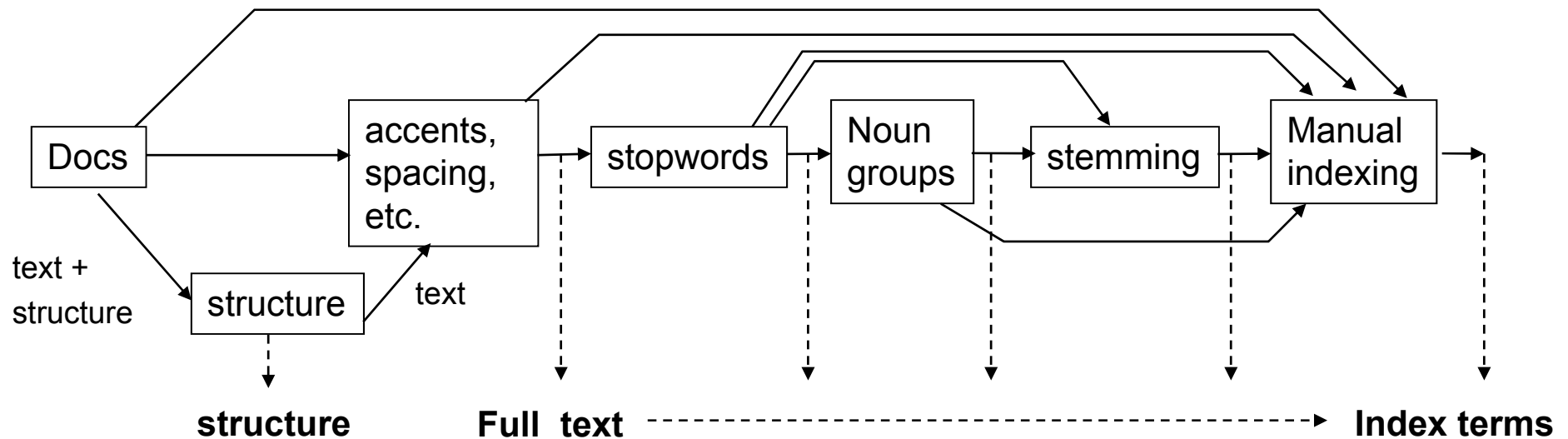


Logical View of the Documents (1/2)

- 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 matching ranking algorithms

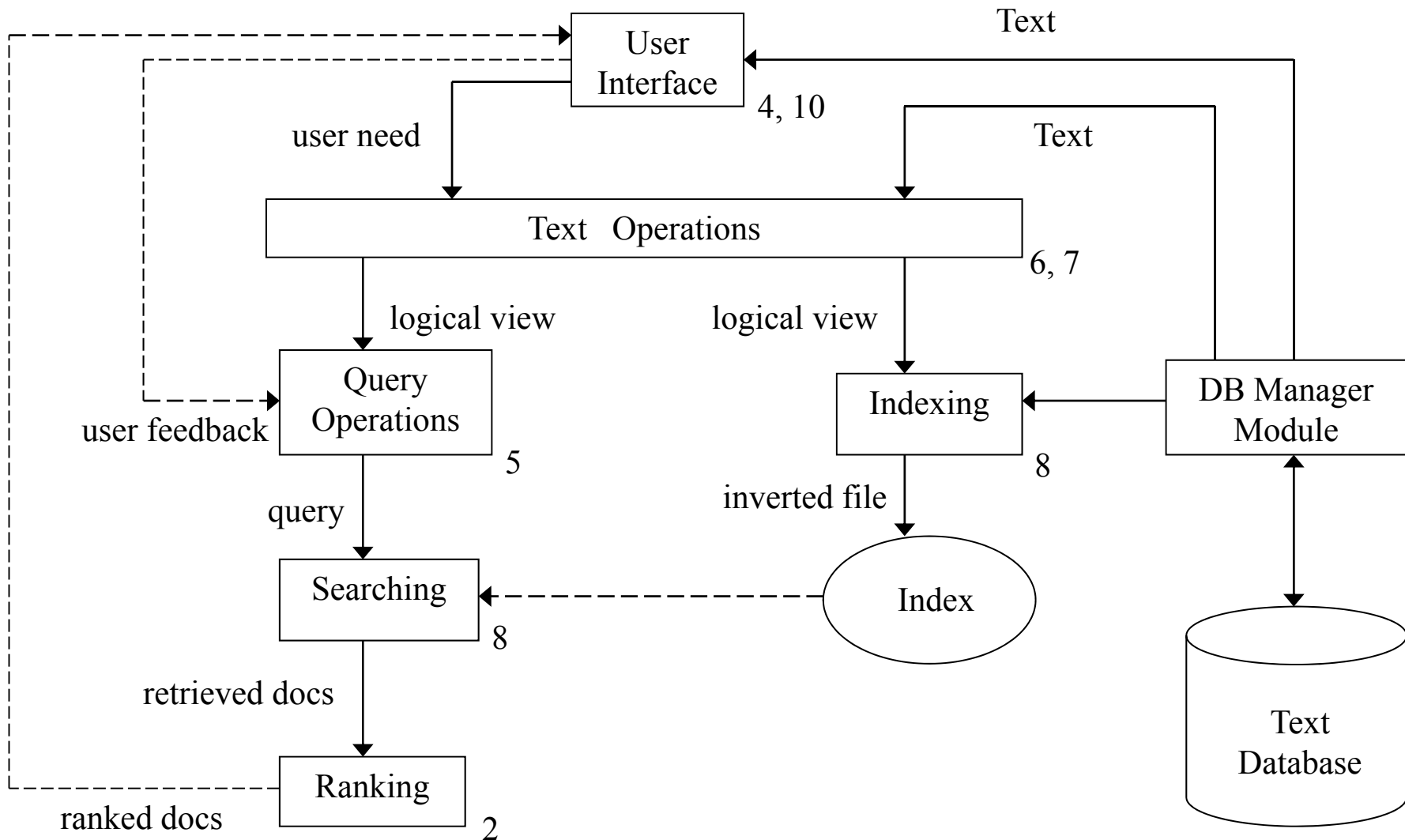
- Human-centered (academic perspective)
 - Studies of user behaviors
 - Understanding of user needs

} 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)

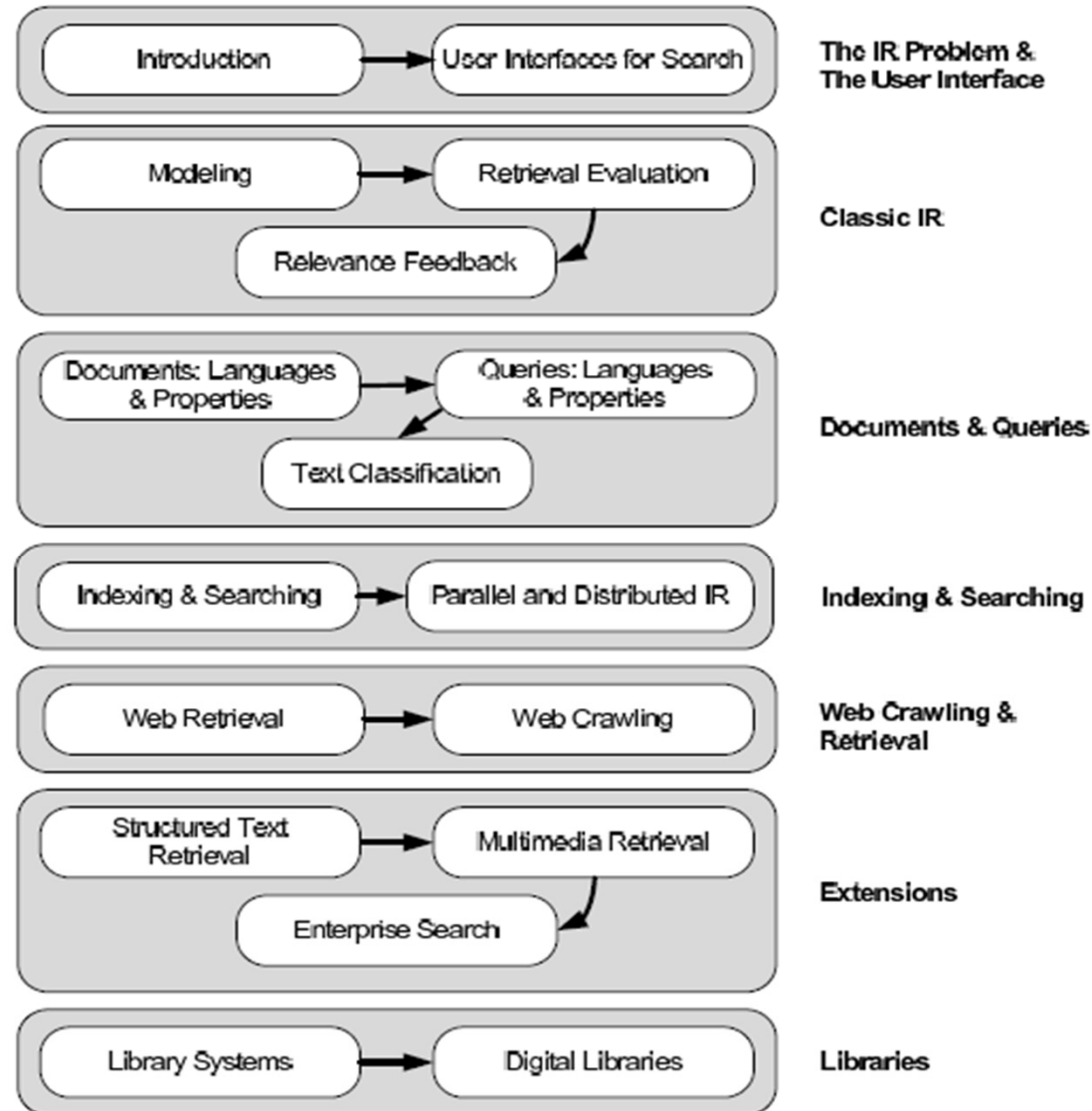
- In current retrieval systems
 - Users almost never declare his information need
 - Only a short queries composed few words (typically fewer than 4 words)
 - Users have no knowledge of the text or query operations

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

Major Topics (2/2)



Some Directions of Information Retrieval

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)	Peer-to-peer search	
Music		

- 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 !
- Peer-to-peer search involves finding information in networks of nodes or computers without any centralized control

IR and Search Engines

Information Retrieval

Relevance

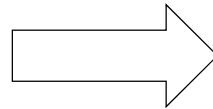
-Effective ranking

Evaluation

-Testing and measuring

Information needs

-User interaction



Search Engines

Performance

-Efficient search and indexing

Incorporating new data

-Coverage and freshness

Scalability

-Growing with data and users

Adaptability

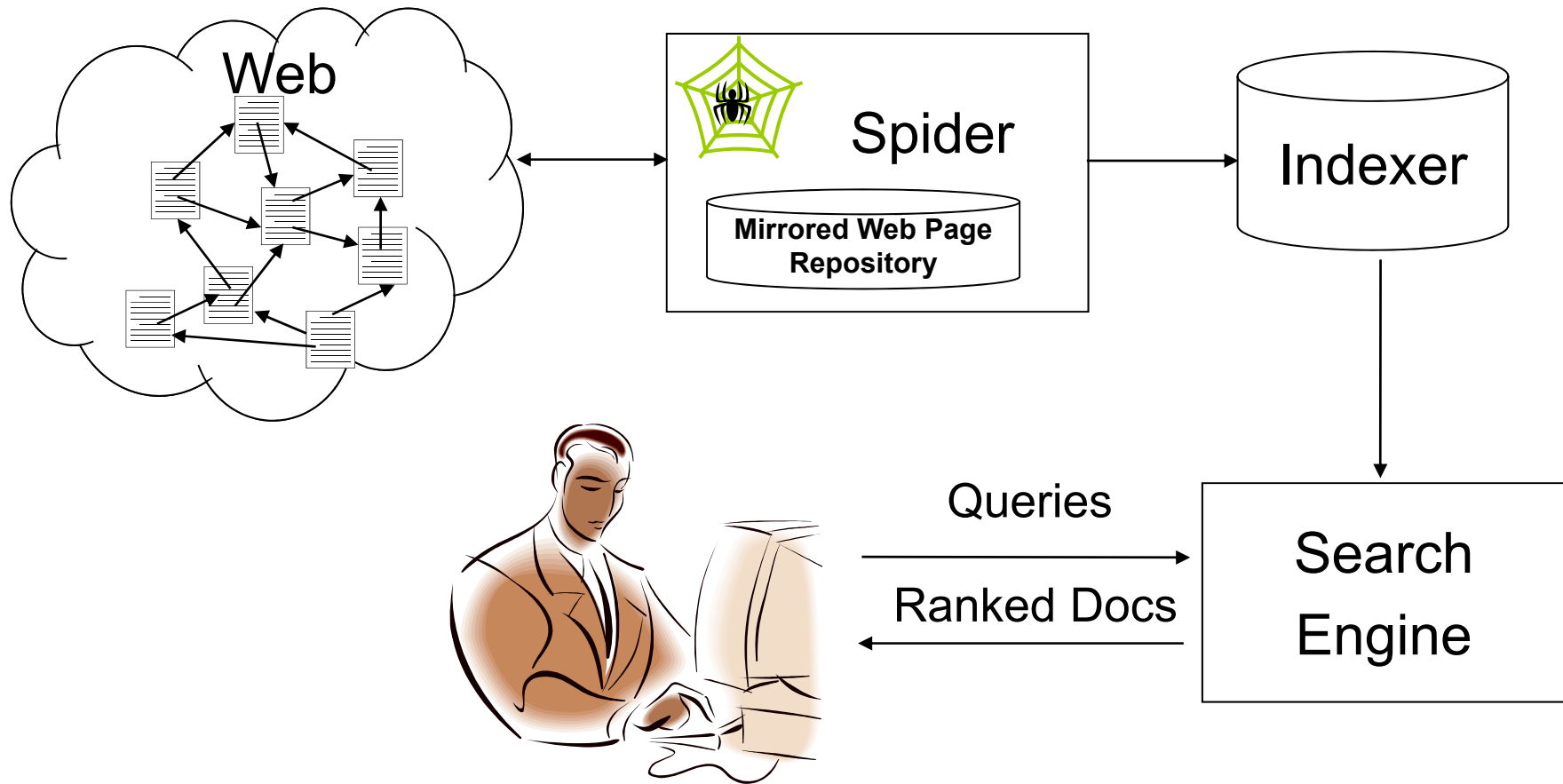
-Tuning for applications

Specific problems

-e.g. Spam

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 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

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](#)

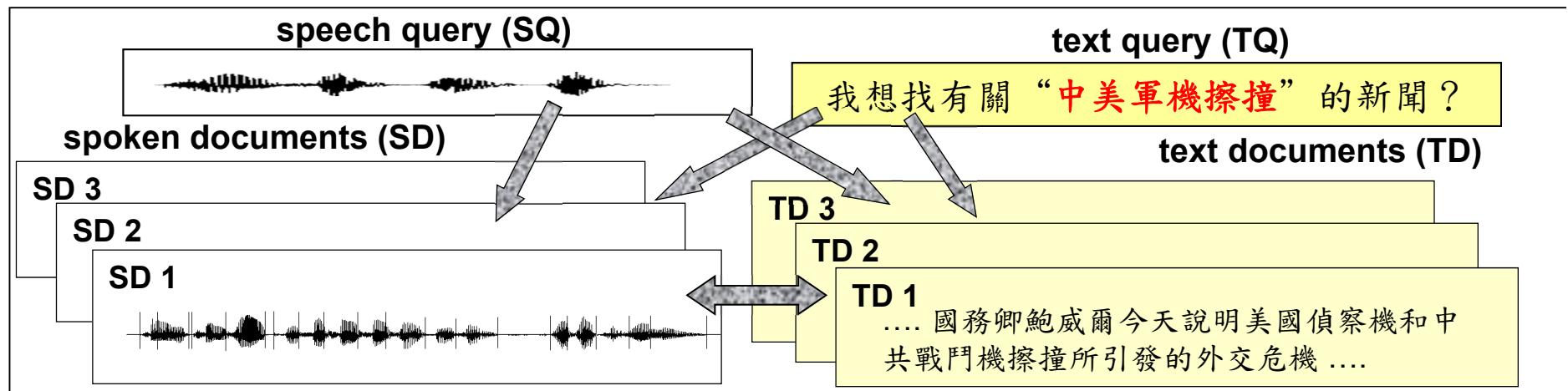
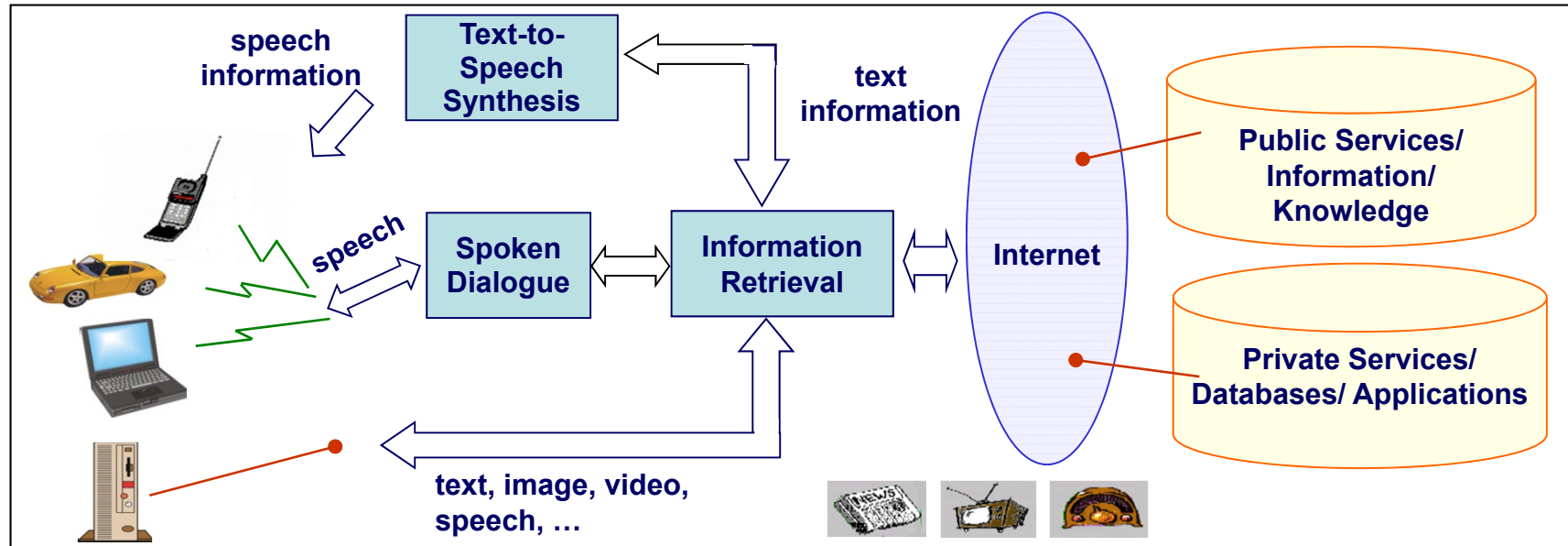
- 1. 觀霧農莊**
介紹農莊風景及其服務項目、交通指南、住宿方式等。 公司名稱: ...
<http://tree.2u.com.tw/> - 2002/12/11, 16k - [關鍵字] [更多結果]
- 2. 瀑布谷農場**
自然休閒-擁抱山水-到雲海的舞台**觀霧** | 瀑布谷農場介紹 | | 交通路線圖 | | 旅遊注意事項
| **觀霧**是雲的故鄉, 景色千變萬化, 體驗大自然、賞... 農場也準備卡拉OK讓您高歌一曲。
注意事項×**觀霧**地區日夜溫差大請多加保暖衣物·請攜帶證件...
簡介-介紹位在雪霸國家公園觀霧的瀑布谷農場, 經營民宿、餐飲、水密...
<http://ppg.2u.com.tw/>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

Text Information Retrieval (4/4)

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

The screenshot shows the Baidu search engine 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 bar contains the text '陈柏琳', with buttons for '百度搜索' and '在结果中找'. Below the search bar, 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 a homepage link: '陈柏琳 (Berlin Chen) 的网页', with a description: '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...'. It includes a URL 'www.csie.ntnu.edu.tw/~berlin/' and a date '12K 2004-9-21'. The second result is 'Berlin Chen (陈柏琳) - Research', with a description: '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-...'. It includes a URL '140.122.185.120/berlin_research/research_...' and a date '38K 2005-8-15'. The third result is '百度_choi吧 [[Charlene Choi相关电影资料]]', with a description: '的关机仪式, 该片导演刘伟强偕同主演谢霆锋、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。>>...'. It includes a URL 'http://ent.tom.com/1636/1637/200517-115930.html' and a date '125K 2005-8-6'. On the right side, there are several utility links: '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', and '搜陈柏琳在阿里巴巴'. Below these are buttons for '总有一人知道你问题的答案' and '发表留言创建陈柏琳贴吧'. At the bottom right, there is a section titled '有许多话想对这个人说?' with the text: '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..'. At the very bottom, there is a link '给陈柏琳传情...'. The bottom of the page shows the start of a footer: '娱乐/中国宁波网'.

Speech Information Retrieval (1/4)



Speech Information Retrieval (2/4)

- 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/>)



Speech Information Retrieval (3/4)

- Speech Summarization and Retrieval

輸入聲音問句：“請幫我查總統府升旗典禮”

辨識 I 等待輸入指令...

測靜音 收音 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

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貴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

可以選擇同時使用音節、字詞等二種索引特徵

QueryByExemplars	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]	N200109061200-07 1.66562e-001
[5]	N200105071000-04 1.60819e-001
[6]	N200111131200-04 1.57109e-001
[7]	T200201211200-01 1.53650e-001
[8]	T200201211200-04 1.51319e-001
[9]	N200110031200-03 1.47177e-001
[10]	N200201171200-11 1.44006e-001
[11]	N200105071400-02 1.41382e-001
[12]	T200106191000-02 1.39268e-001
[13]	N200110291200-01 1.38799e-001
[14]	N200104301230-05 1.36488e-001
[15]	N200109051200-05 1.33595e-001
[16]	N200109141200-18 1.33158e-001
[17]	N200105142000-05 1.32321e-001
[18]	FTV2002-064 1.32147e-001
[19]	N200201181200-11 1.31222e-001

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離形展示系統。

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

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

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

Speech Information Retrieval (4/4)

- Speech Organization

(a) 廣播新聞搜尋瀏覽系統
Broadcast News Retrieval/Browsing System

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
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

(c)

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

(d)

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)

(e)

Video player showing a speech waveform and a video frame of a man speaking.

- L.-S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine* 22(5), pp. 42-60, Sept. 2005

Visual Information Retrieval (1/4)

- Content-based approach

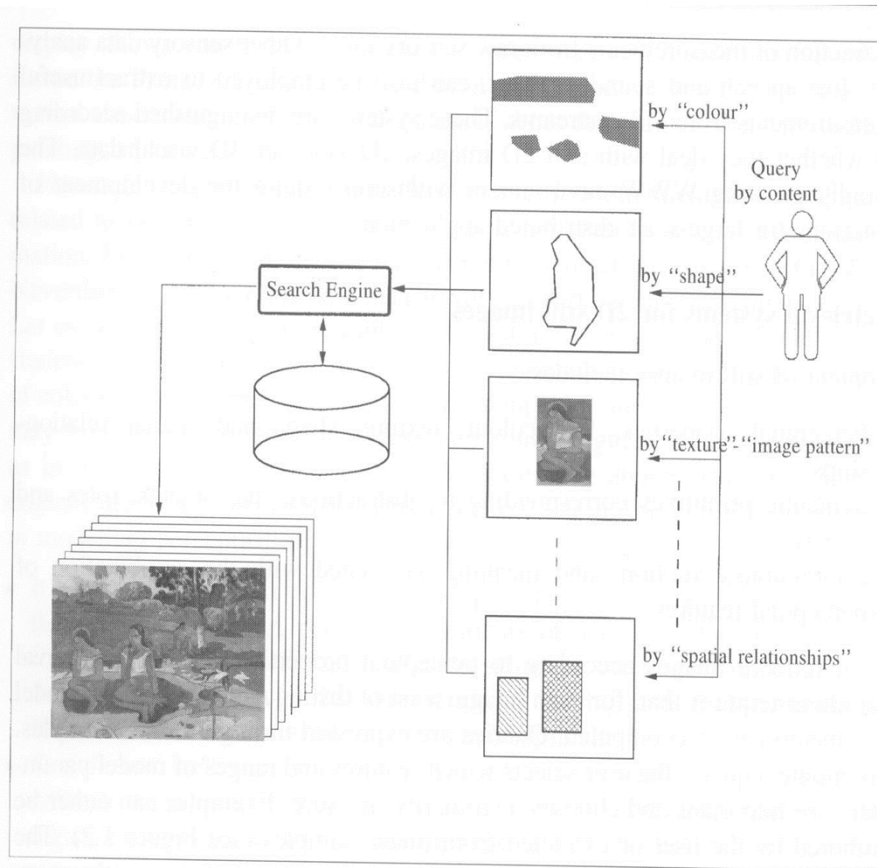


Figure 1.2 Different types of query by example.

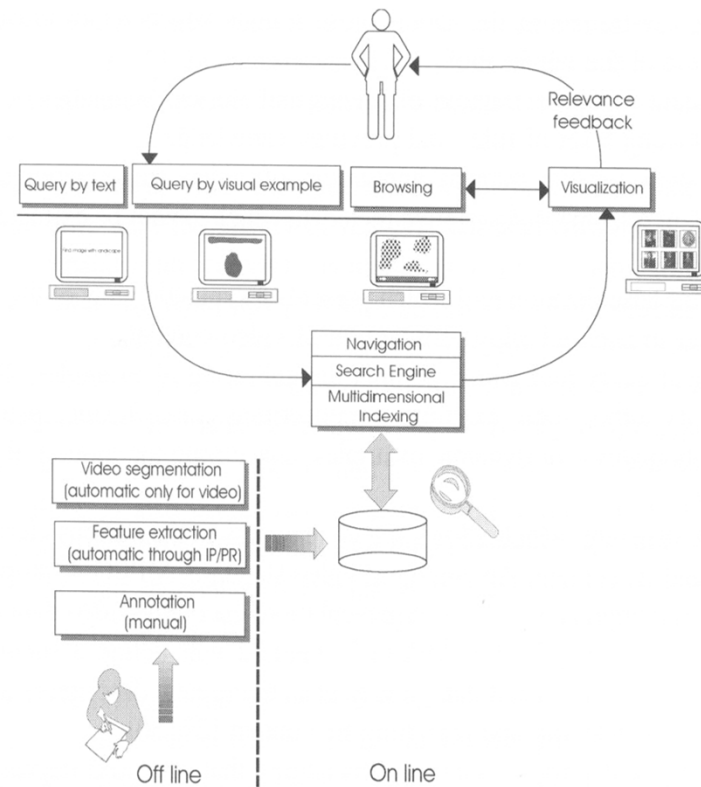
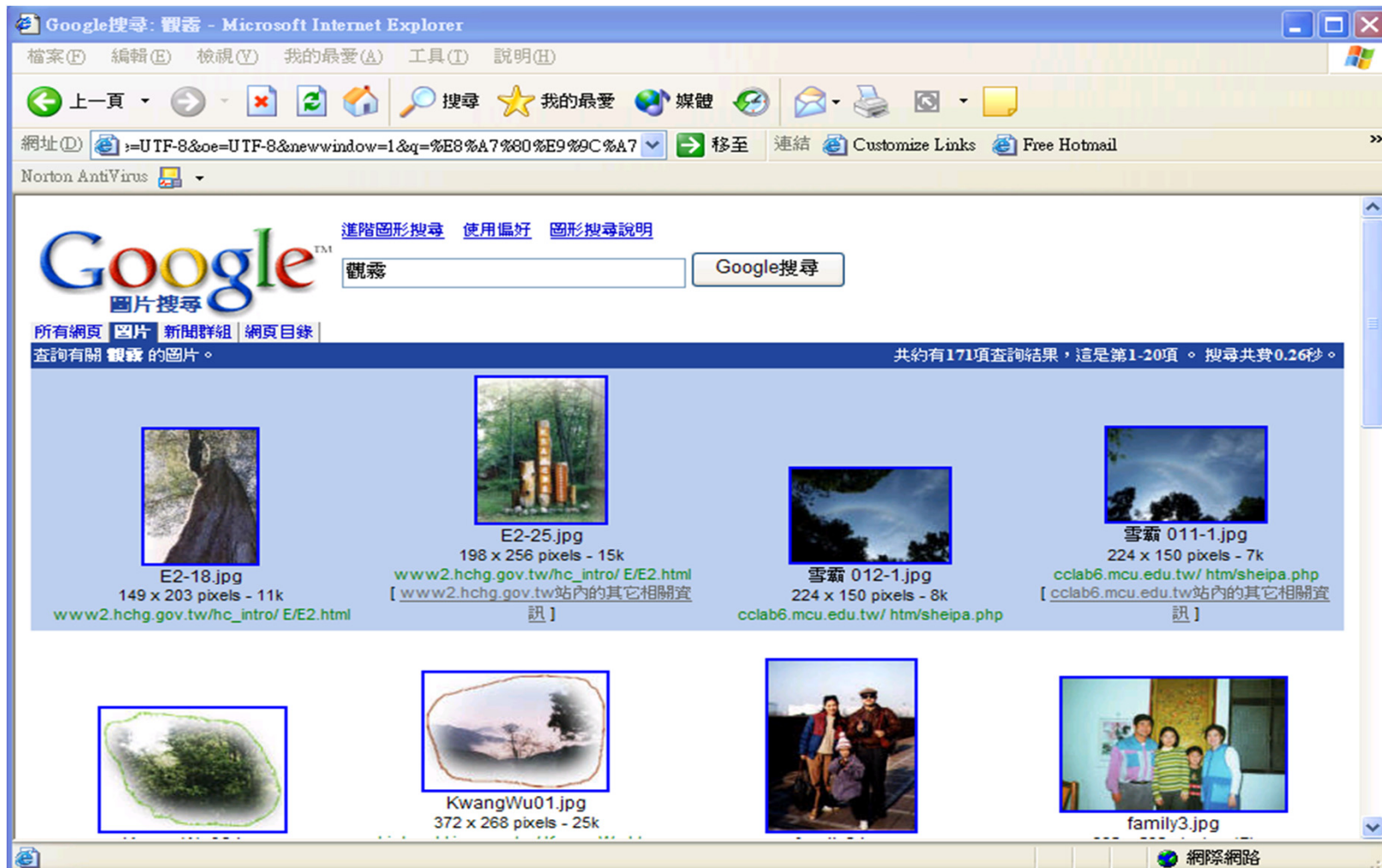


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

Visual Information Retrieval (2/4)

- Images with Texts (Metadata)



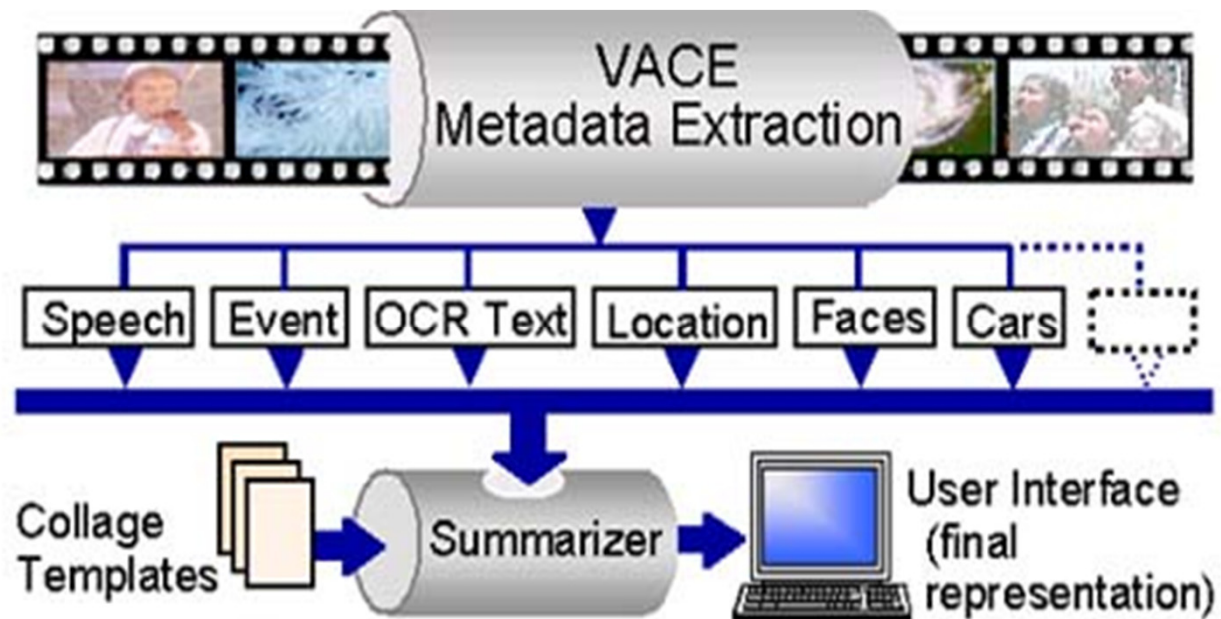
Visual Information Retrieval (3/4)

- Content-based Image Retrieval

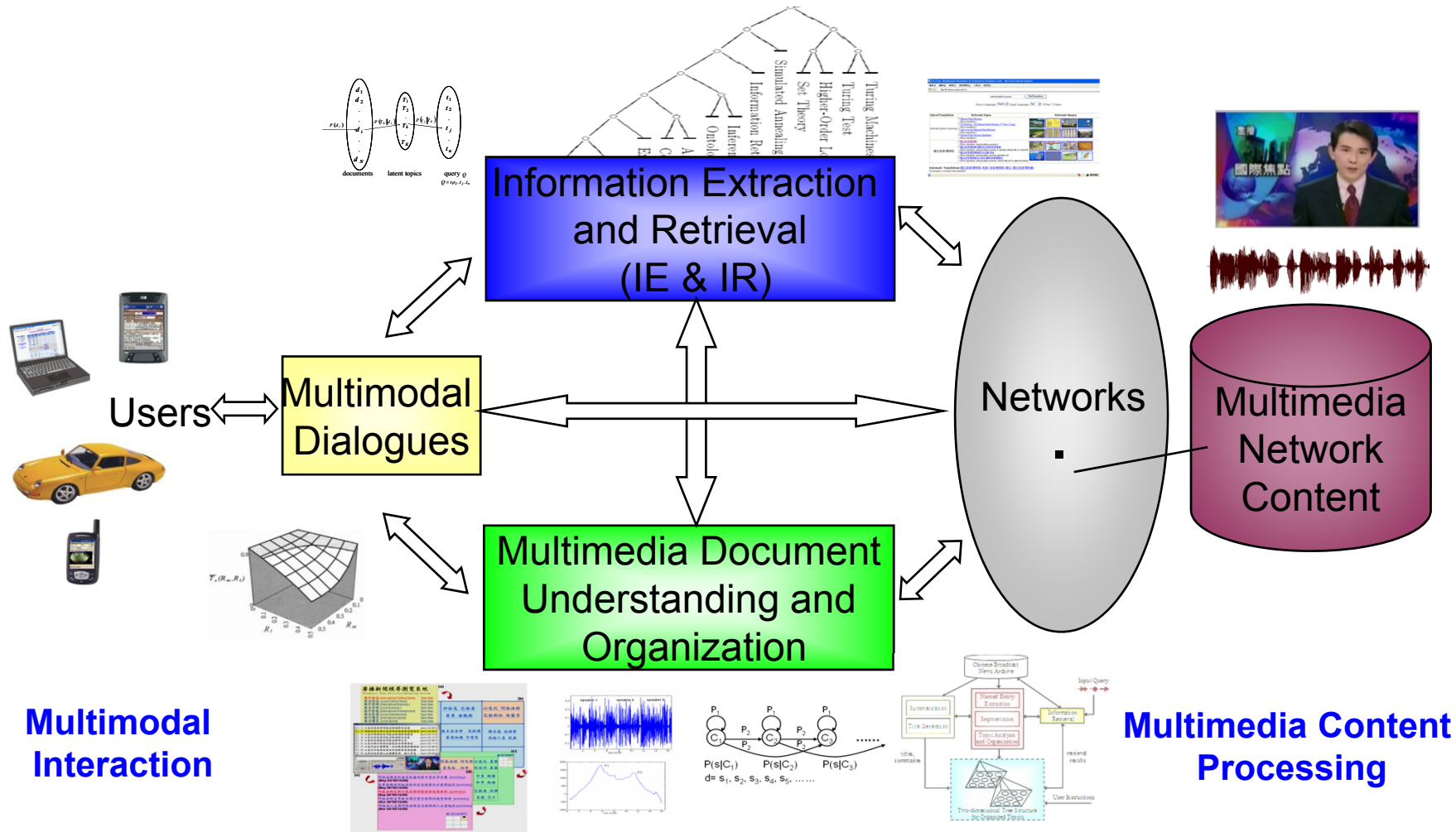


Visual Information Retrieval (4/4)

Video Analysis and Content Extraction



Scenario for Multimedia information access



Other IR-Related Tasks

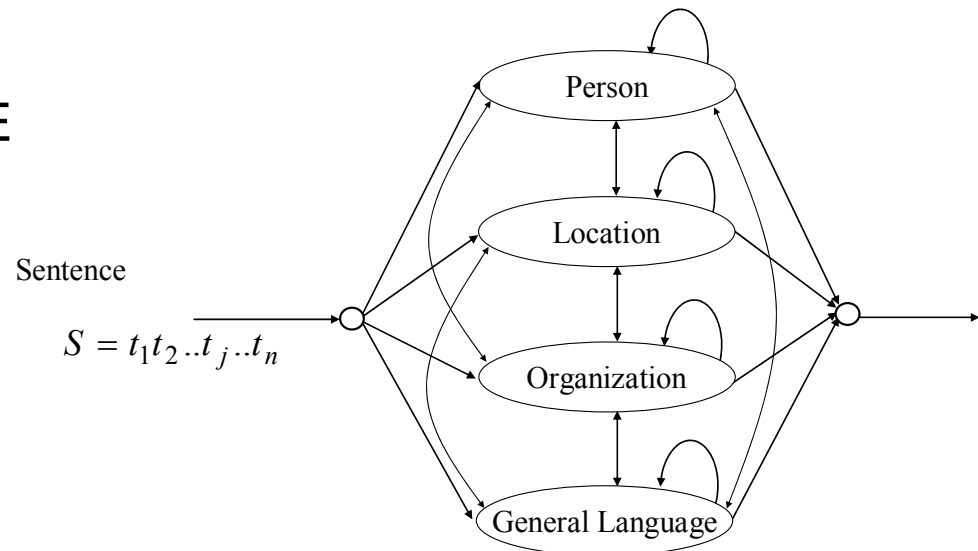
- Information filtering and routing
- **Term/Document categorization**
- **Term/Document clustering**
- **Document summarization**
- **Information extraction**
- Question answering
 - “*What is the height of Mt. Everest?*”
- Crosslingual information retrieval
-

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



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

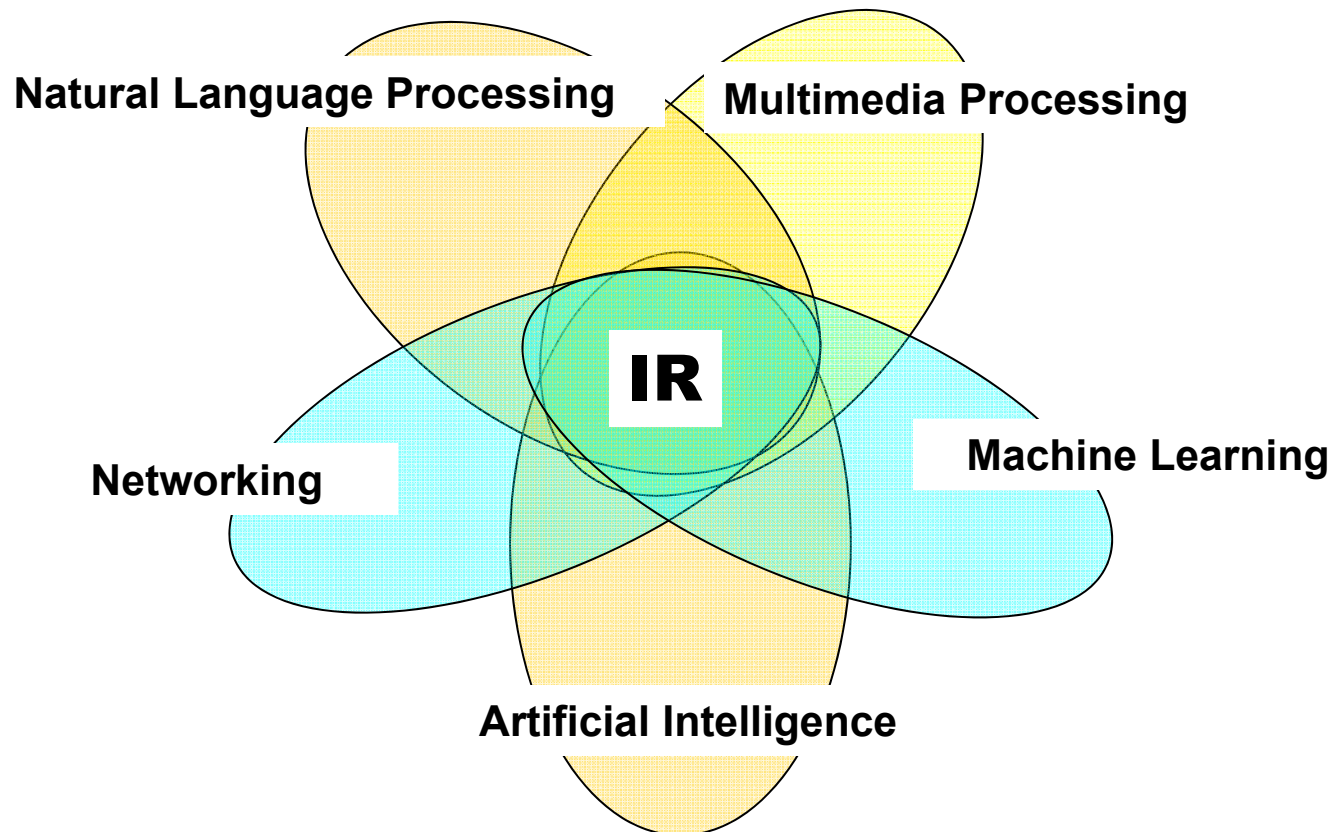
Query/Translation	Relevant Pages	Relevant Images
national palace museum	<ul style="list-style-type: none"> * National Palace Museum [Gloss translation:] * TIT Museums: The National Palace Museum: 70 Years Young! [Gloss translation:] * Jades from the National Palace Museum [Gloss translation:] * National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

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

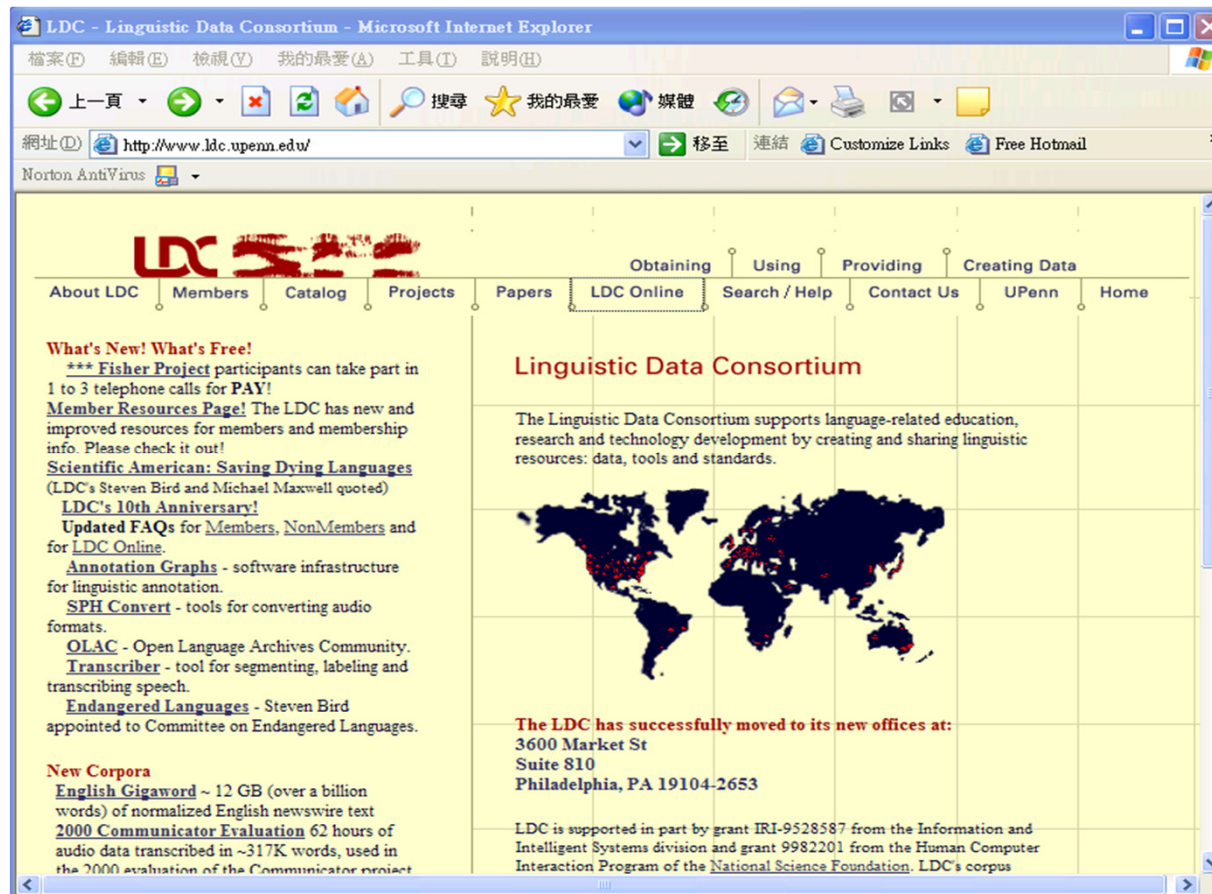


Multidisciplinary Approaches



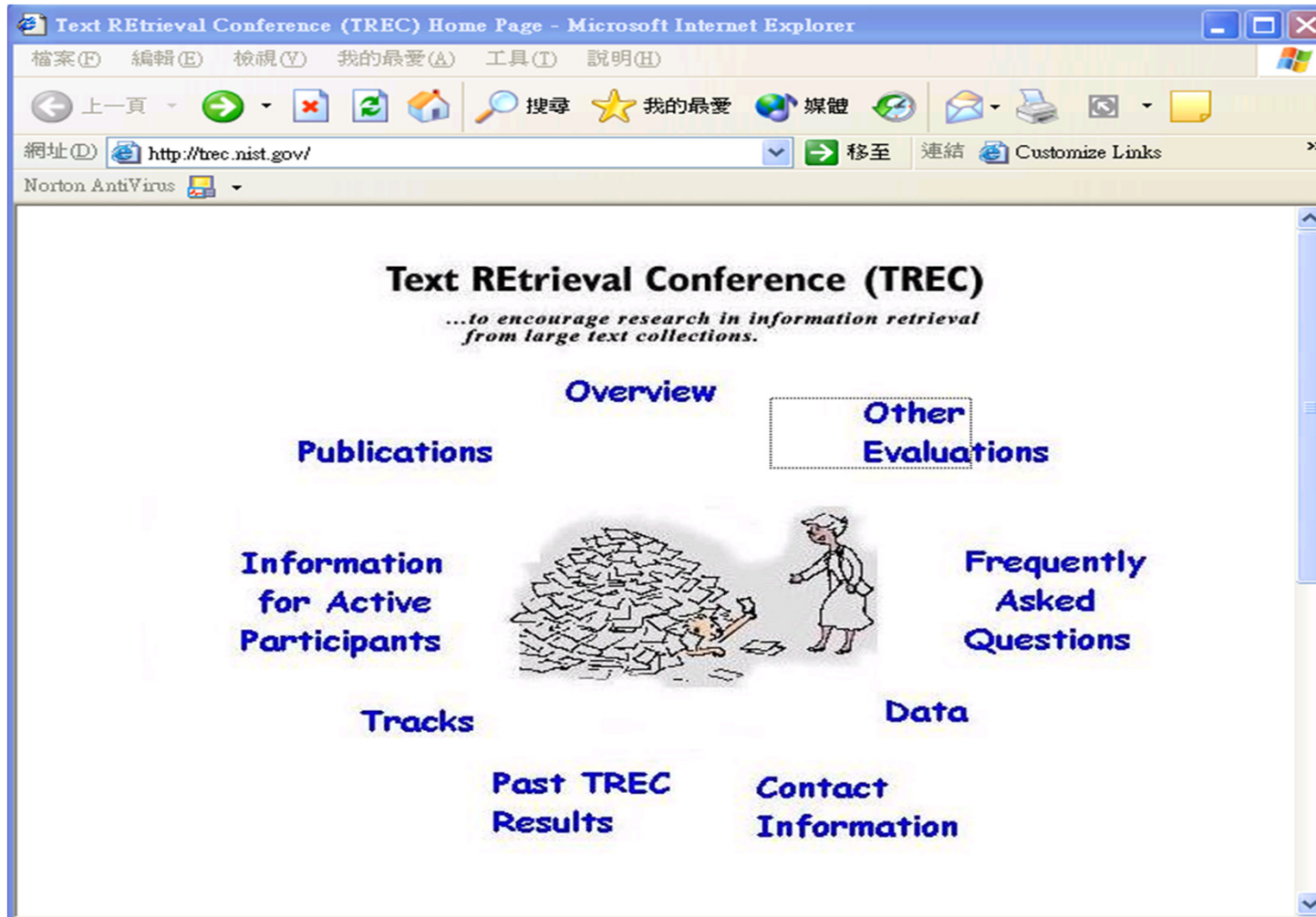
Resources

- 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

NIST
National Institute of Standards and Technology

[Contact Webmaster](#)

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 Evaluation](#)

1998 Speaker Detection & Tracking Development 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

- 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): 45%
- Homework/Projects: 30%
- Presentation: 15%
- Attendance/Other: 10%