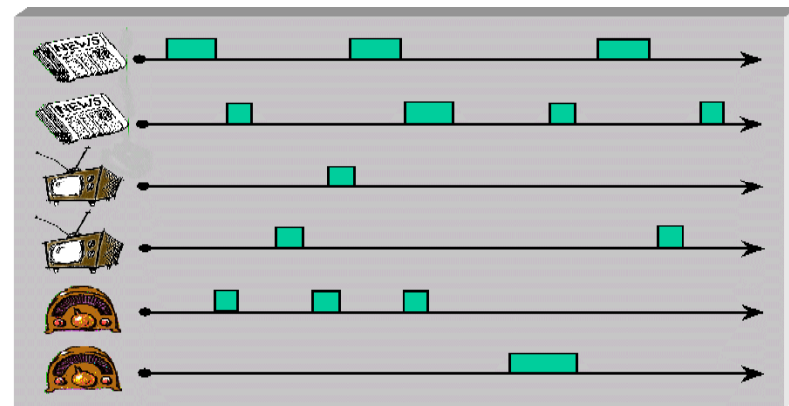


# Information Retrieval and Extraction

Berlin Chen



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

# Textbook and References

- Textbooks

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

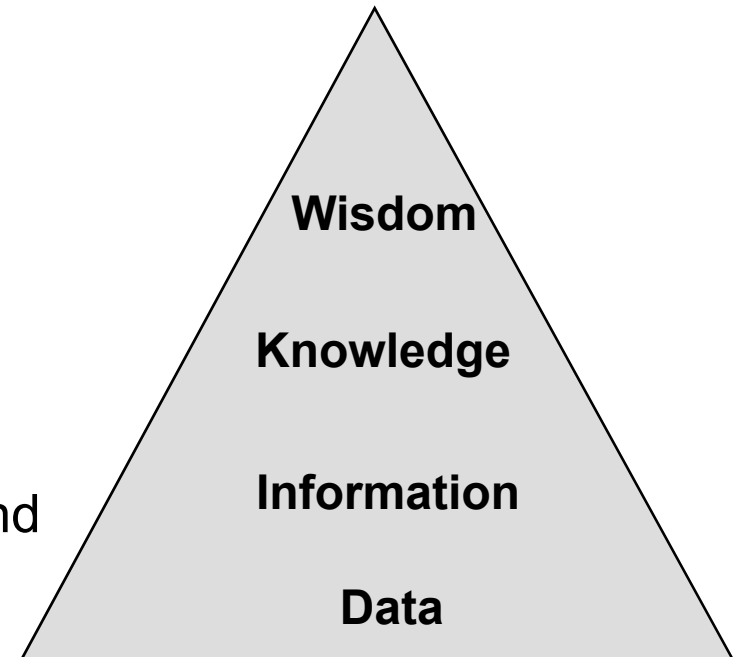
- References

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

# 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



- 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
- 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 (or free 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
  - 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!

# IR system

- Interpret contents of information items (documents)
  - Most of the information in such documents is in the form of text which relatively unstructured
- Generate a ranking (i.e., a ranked list of documents) which reflects relevance
- Notion of *relevance* is most important
  - Relevance judgment (using click-through data ?)
  - The other important issues
    - The vocabulary mismatch problem
    - Evaluation of retrieval performance

# 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

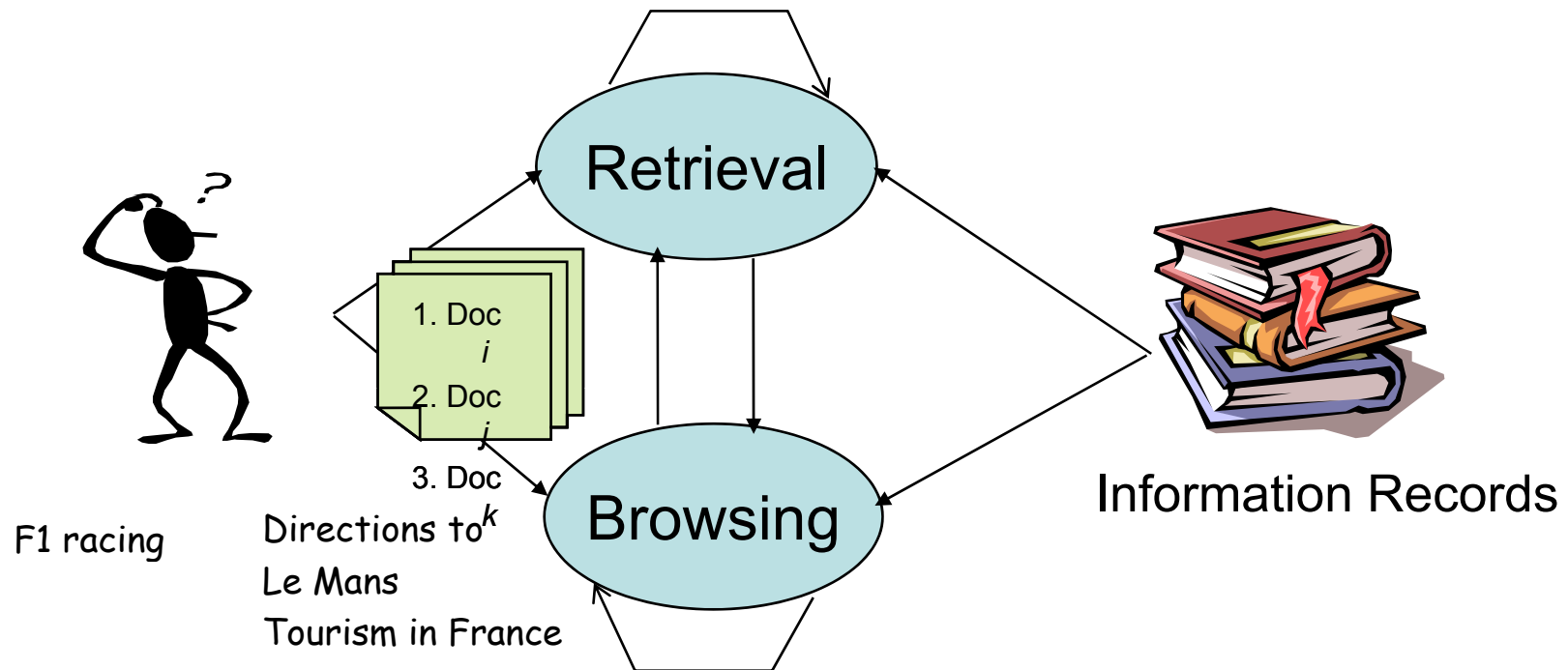


# IR Main Issues

- 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

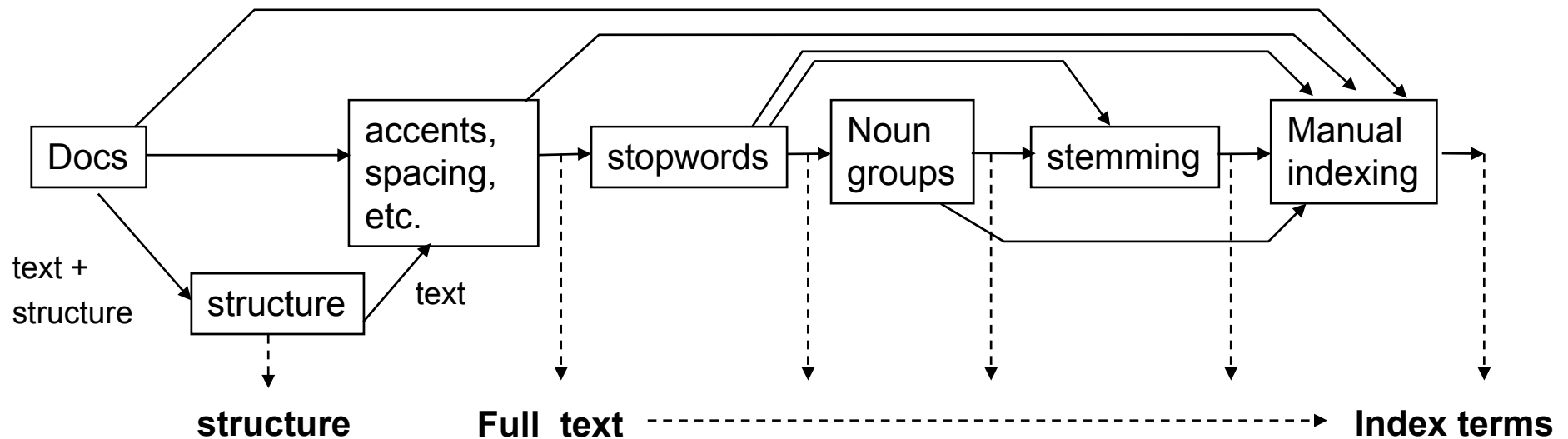


# Logical View of the Documents (1/2)

- 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

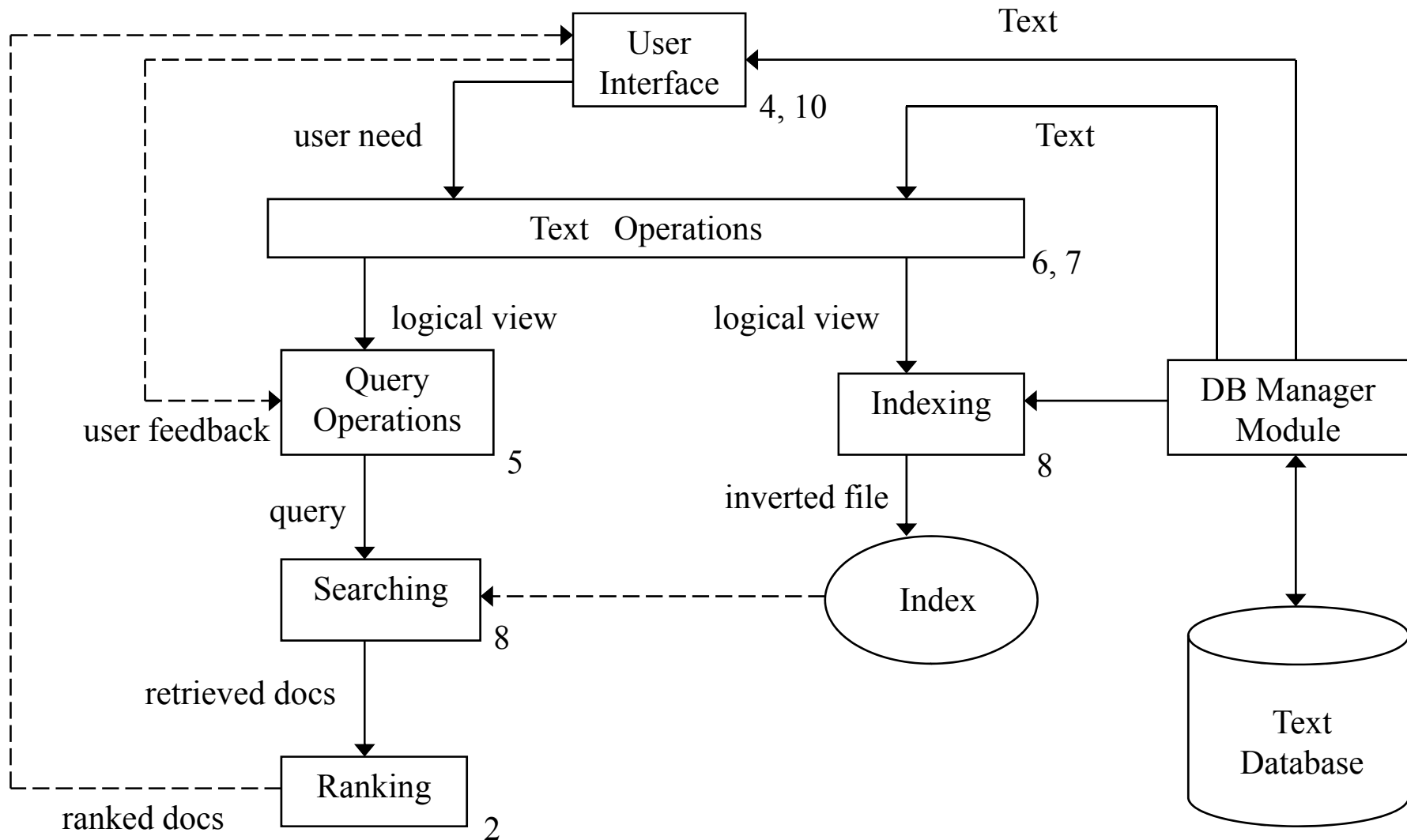
} Library science  
psychology

....

# 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)

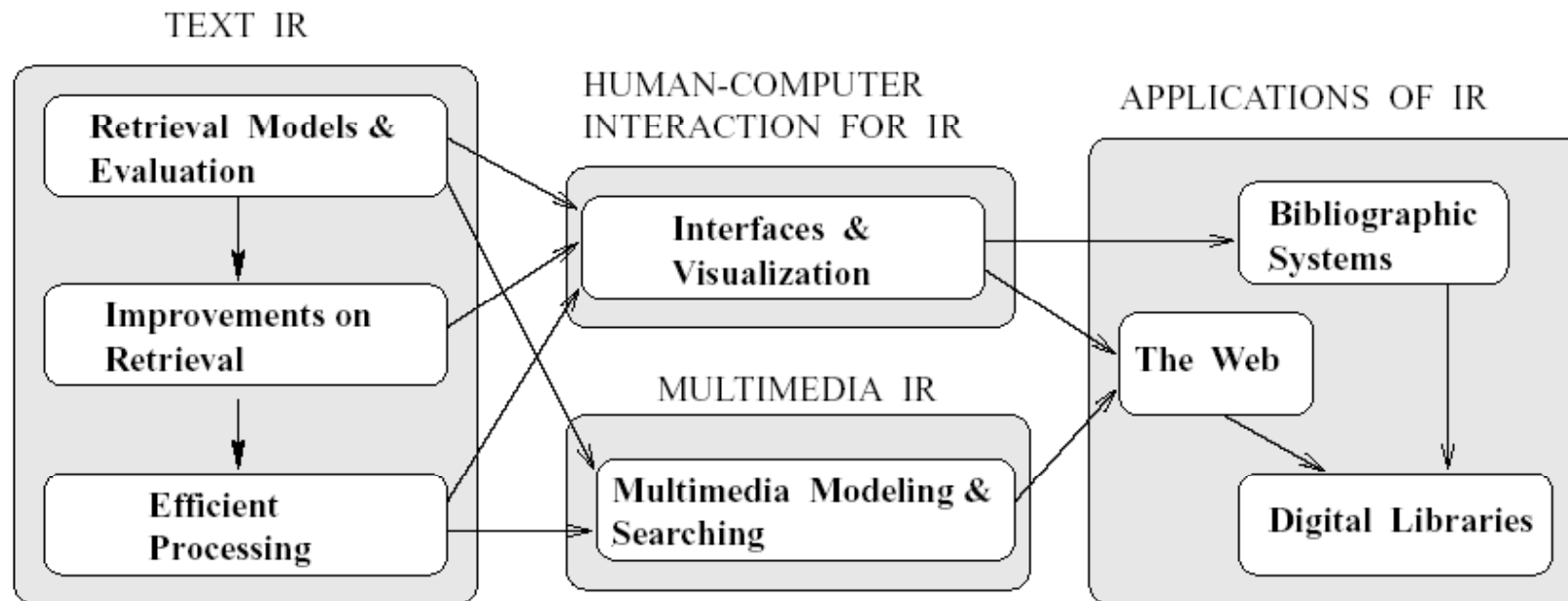
- In current retrieval systems
  - Users almost never declare his information need
    - Only a short queries composed few words (typically fewer than 4 words)
  - Users have no knowledge of the text or query operations

Poor formulated queries lead to poor retrieval !



# Major Topics (1/2)

- Four Main Topics

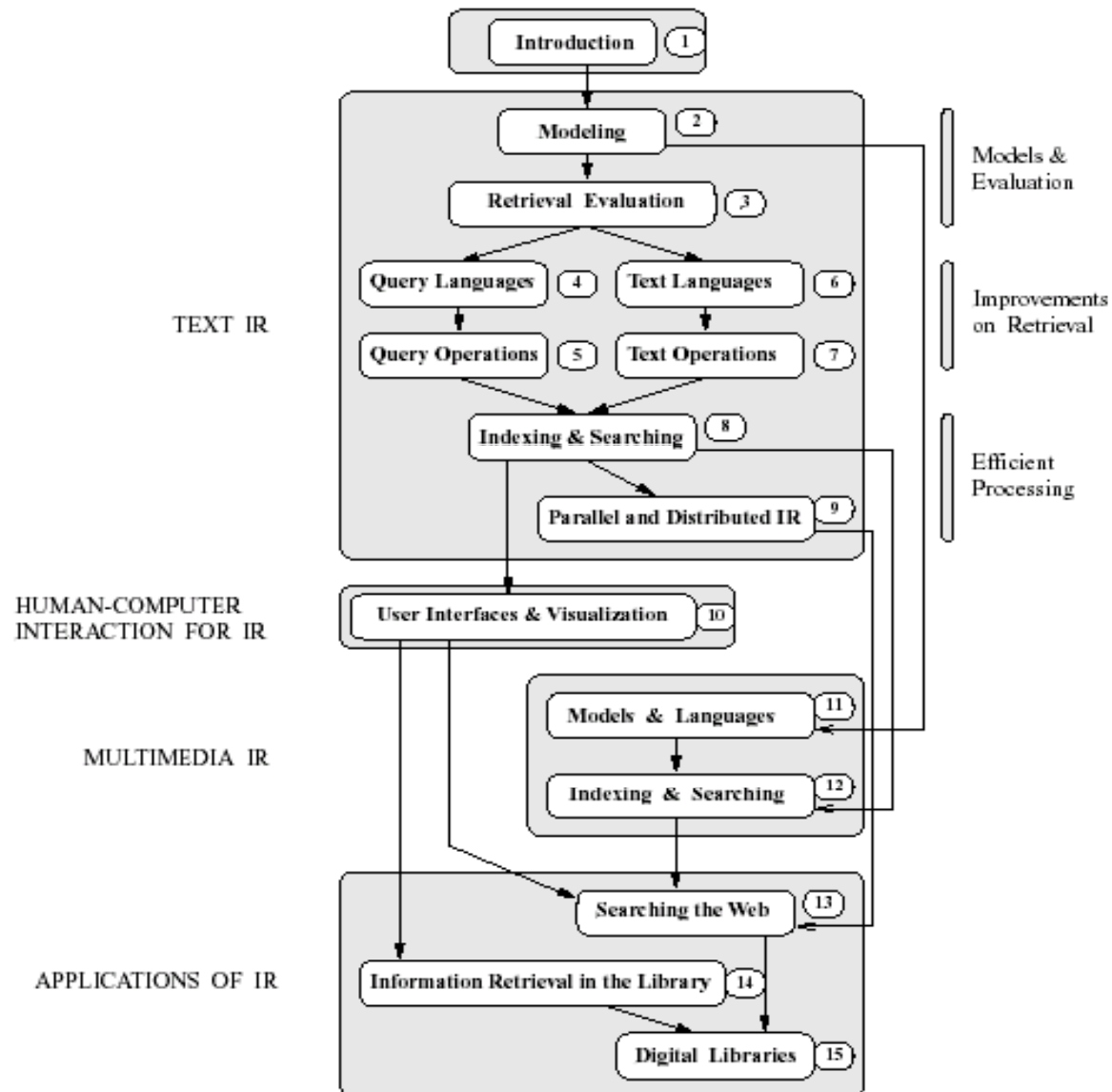


**Figure 1.4** Topics which compose the book and their relationships.

# 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
  - Can multimedia be treated in a unified manner?
- Applications
  - Web, bibliographic systems, digital libraries

# Textbook Topics



# 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	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

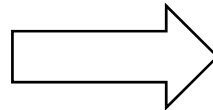
*-Effective ranking*

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

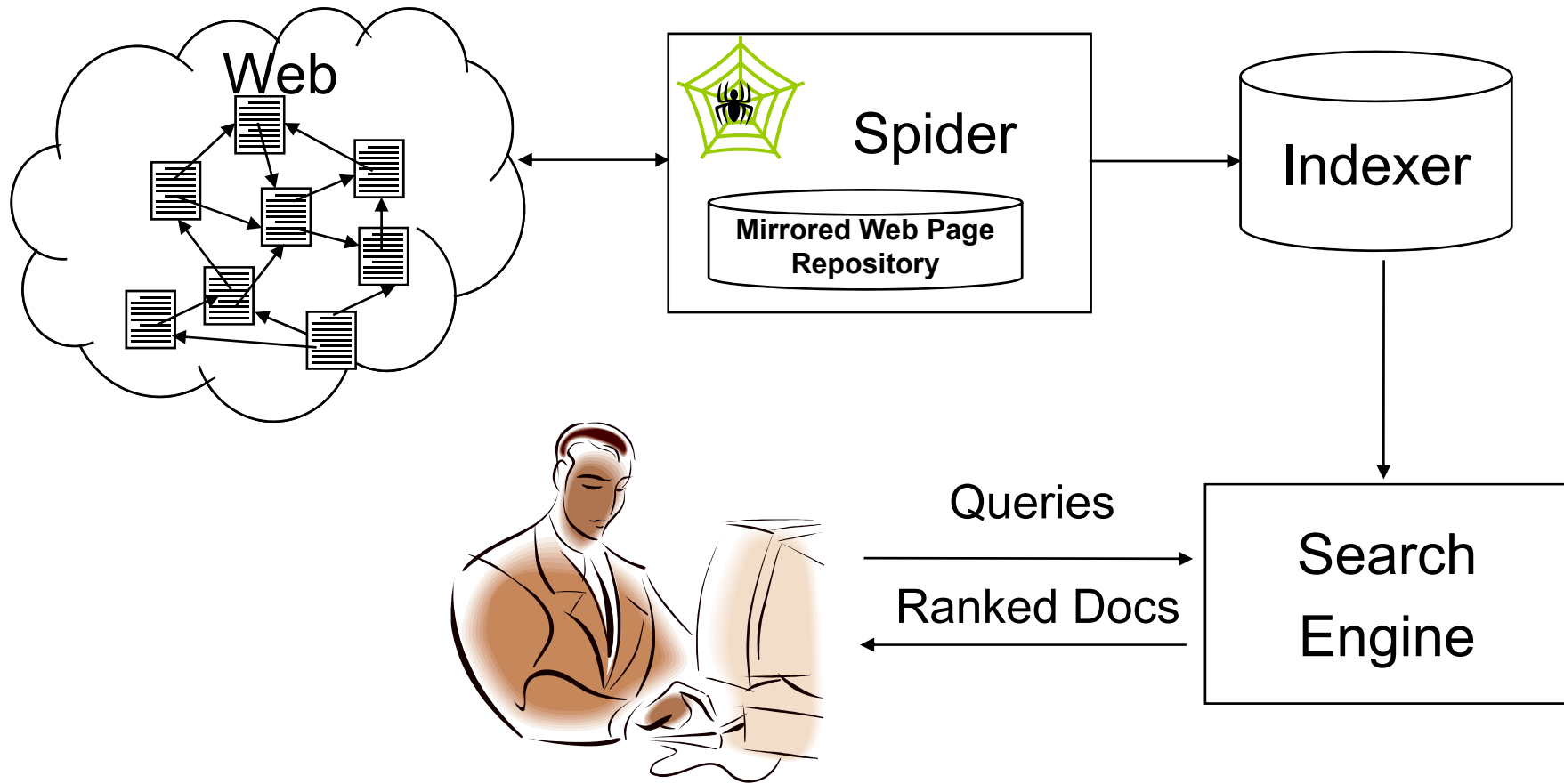
*-Tuning for applications*

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)





# Text Information Retrieval (4/4)

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

The screenshot shows a Baidu search result for the name '陈柏琳'. At the top, there is the Baidu logo and navigation links for '新闻', '网页', '贴吧', 'MP3', and '图片'. The search bar contains '陈柏琳' and the search button is labeled '百度搜索'. Below the search bar, it indicates '找到相关网页156篇, 用时0.158秒'. The main content area lists several search results, including a homepage for Berlin Chen at National Taiwan Normal University, a research paper on spam filtering, and a news article about a film. On the right side, there are several promotional links for eBay, Biz178, and Alibaba, as well as a section for '总有一人知道你问题的答案' and a '发表留言创建陈柏琳贴吧' button. At the bottom right, there is a box with the text '有许多话想对这个人说?' and a link '给陈柏琳传情...'. The footer of the page contains the text '娱乐/中国宁波网'.

设百度为首页 高级搜索 帮助

Baidu 百度 陈柏琳 百度搜索 在结果中找

新闻 网页 贴吧 MP3 图片 找到相关网页156篇, 用时0.158秒

您要找的是不是: [陈柏霖](#)

[陈柏琳 \(Berlin Chen\) 的网页](#)  
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...  
[www.csie.ntnu.edu.tw/~berlin/](http://www.csie.ntnu.edu.tw/~berlin/) 12K 2004-9-21 繁体 - 百度快照  
[www.csie.ntnu.edu.tw 上的更多结果](#)

[Berlin Chen \(陈柏琳\) - Research](#)  
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-... 陈怡婷、黄耀民、叶耀明、陈柏琳, "中文语音文件自动摘要之摘要模型," 投稿至「第十届人工智能与应用...  
[140.122.185.120/berlin\\_research/research\\_...](http://140.122.185.120/berlin_research/research_...) 38K 2005-8-15 繁体 - 百度快照  
[140.122.185.120 上的更多结果](#)

[百度\\_choi吧 \[Charlene Choi相关电影资料\]](#)  
的关机仪式,该片导演刘镇伟偕同主演谢霆锋、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\_...  
[post.baidu.com/?kz=8522392](http://post.baidu.com/?kz=8522392) 125K 2005-8-6 - 百度快照

[娱乐/中国宁波网](#)  
陈柏琳在《...》中饰演... 陈柏琳 经济林均青... 刘镇伟是... 一个非常好的... 拍摄... 陈柏琳

[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](#)

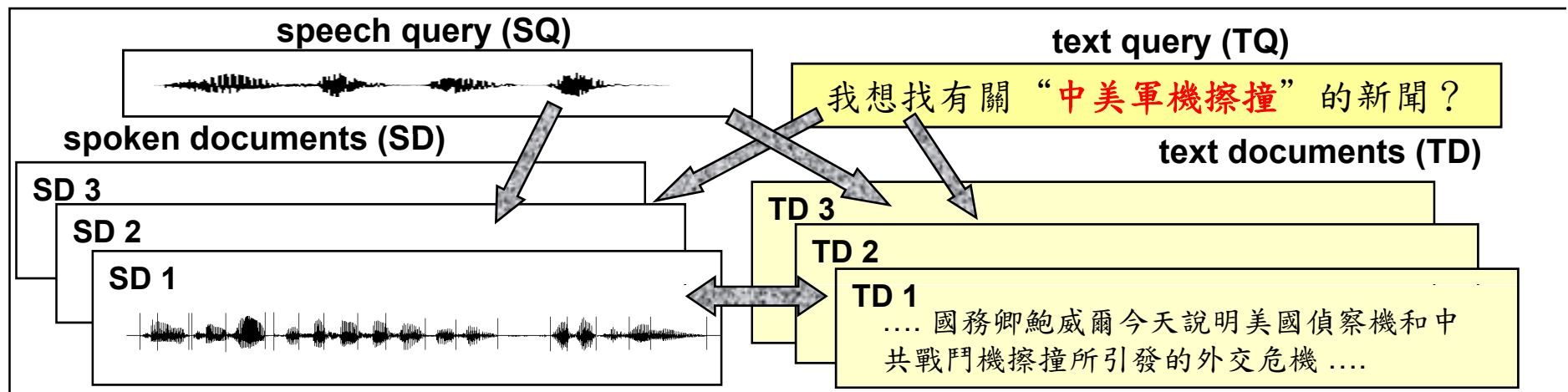
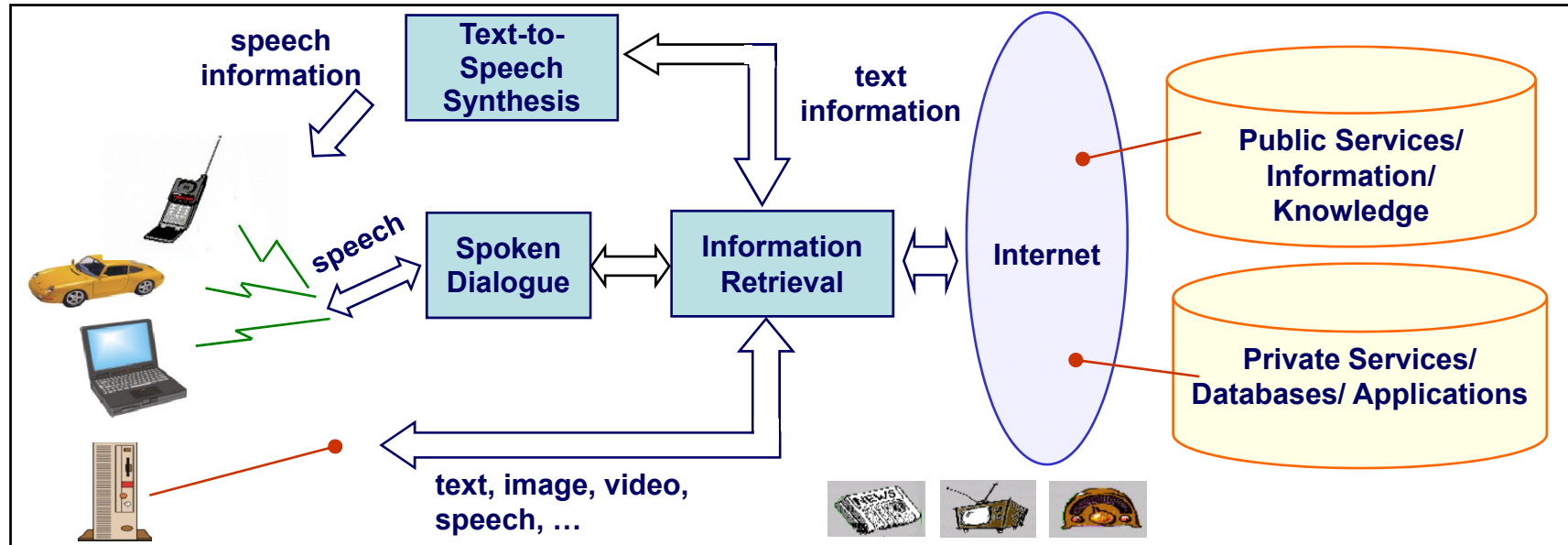
总有一人知道你问题的答案

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..

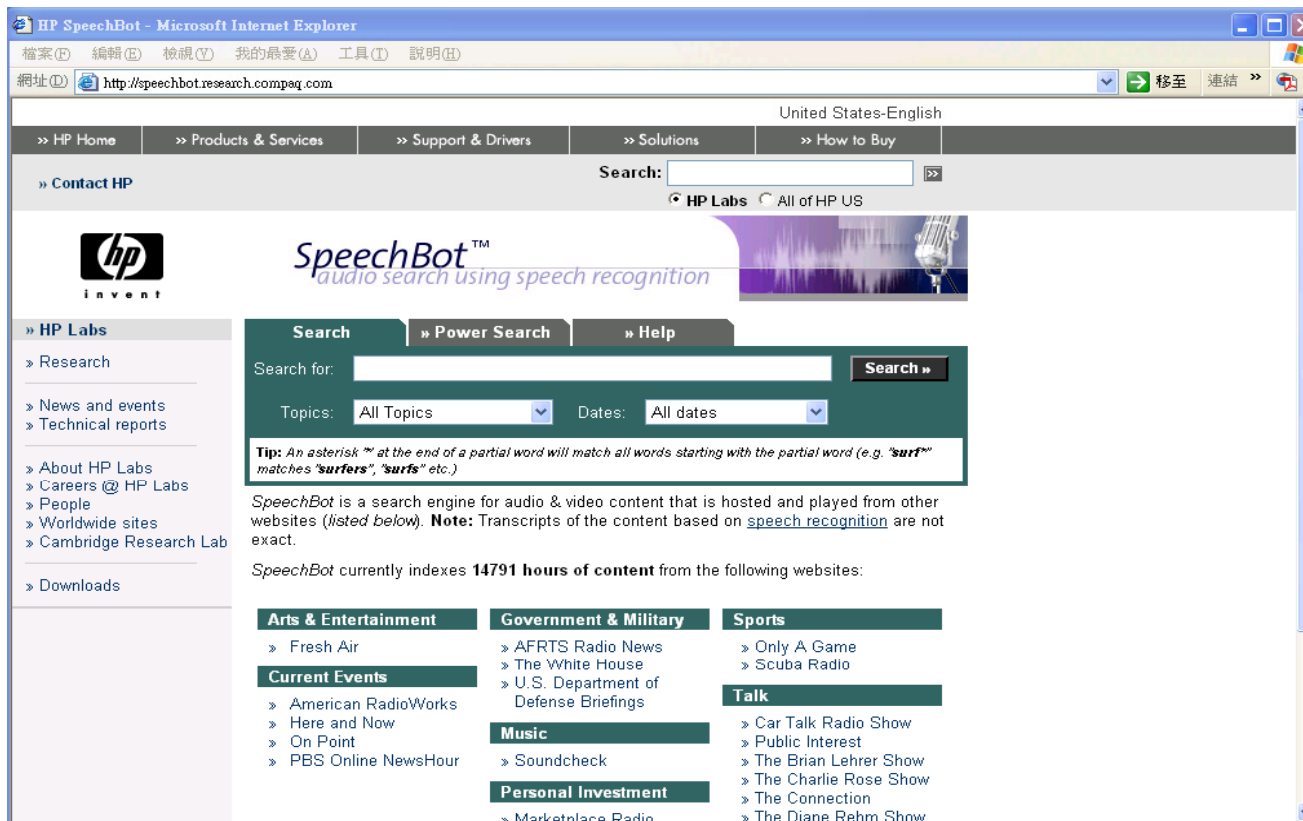
[给陈柏琳传情...](#)

# Speech Information Retrieval (1/4)



# Speech Information Retrieval (2/4)

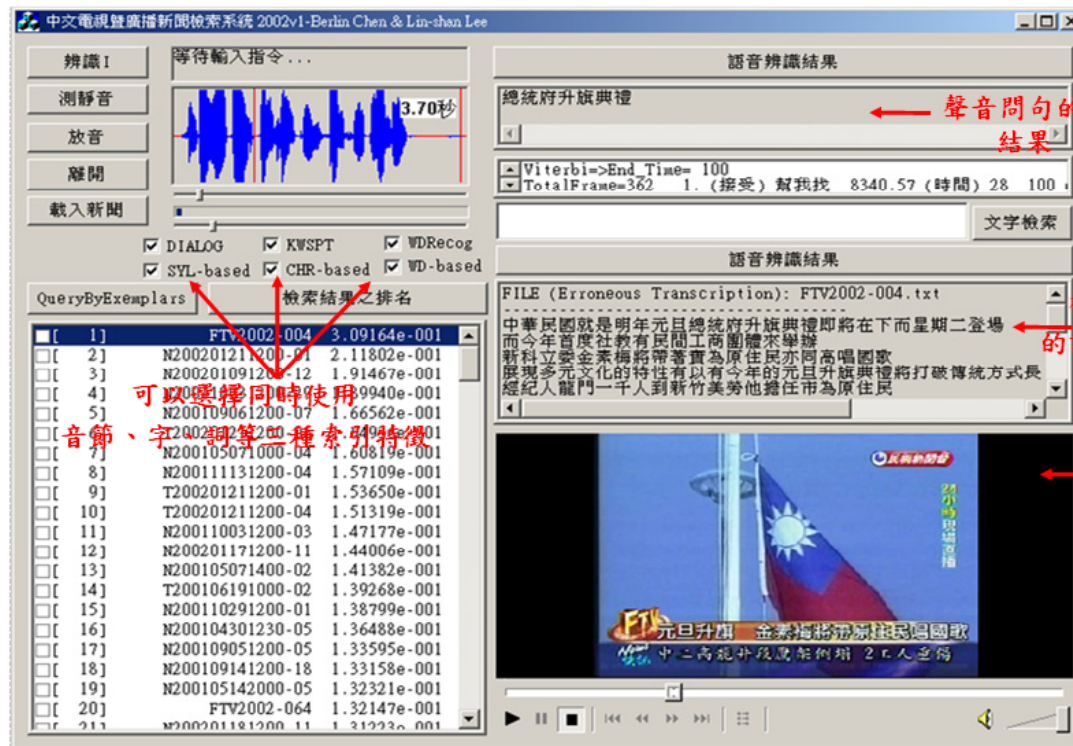
- HP Research Group – Speechbot System  
(Service is No Longer Available)
  - Broadcast news speech recognition, Information retrieval, and topic segmentation (SIGIR2001)
  - Currently indexes **14,791 hours of content** (2004/09/22, <http://speechbot.research.compaq.com/>)



# 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”

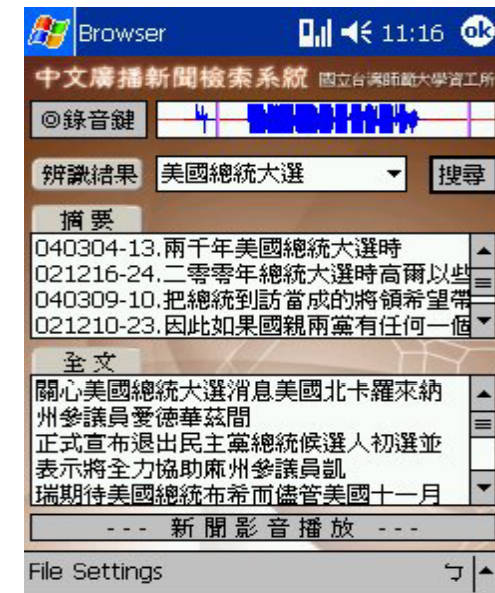


可以選擇同時使用音節、字詞等二種索引特徵

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離形展示系統。

# Speech Information Retrieval (4/4)

- Speech Organization

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

**Topic Map**

- 國外政治 [International Political News]
- 國內政治 [Local Political News]
- 國外財經 [International Business]
- 國內財經 [Local Business]
- 國外影劇 [International Entertainment]
- 國內影劇 [Local Entertainment]
- 國外體育 [International Sports]
- 國內體育 [Local Sports]

**(a)**

**(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)**

go to Level-1

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)**

**(e)**

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-2

- 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

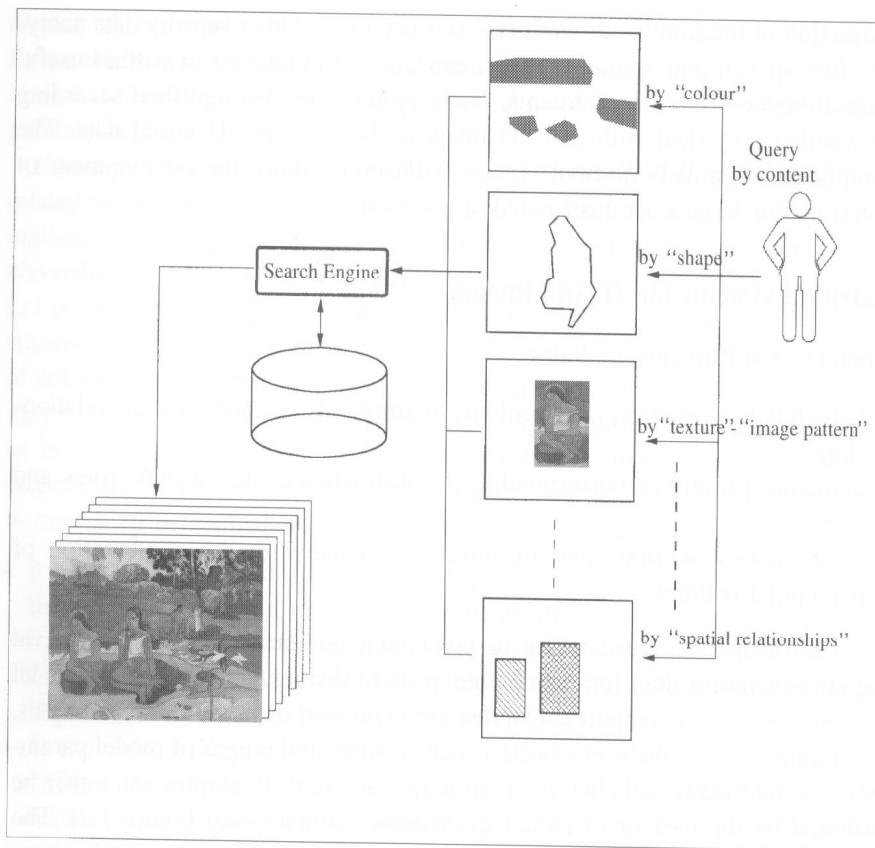


Figure 1.2 Different types of query by example.

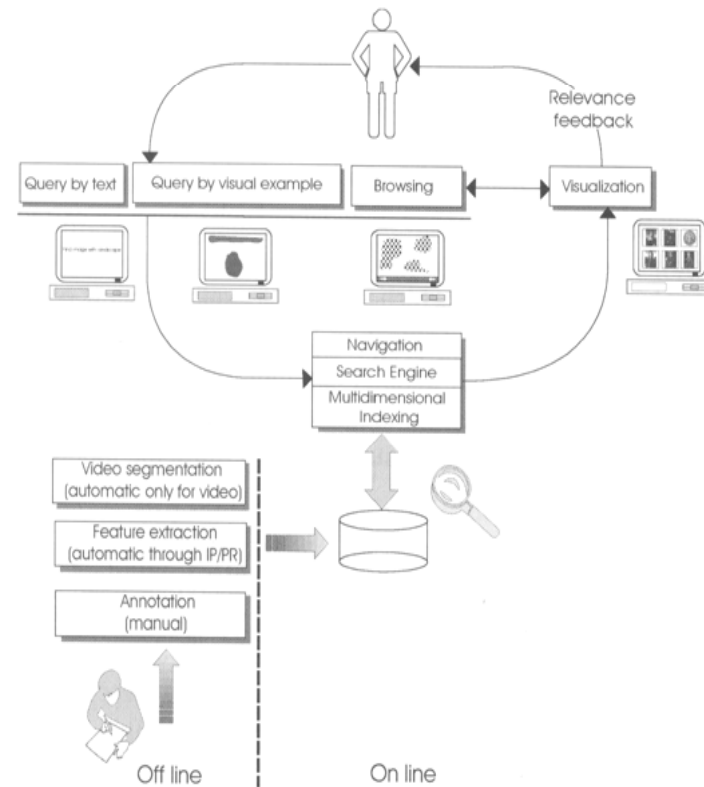
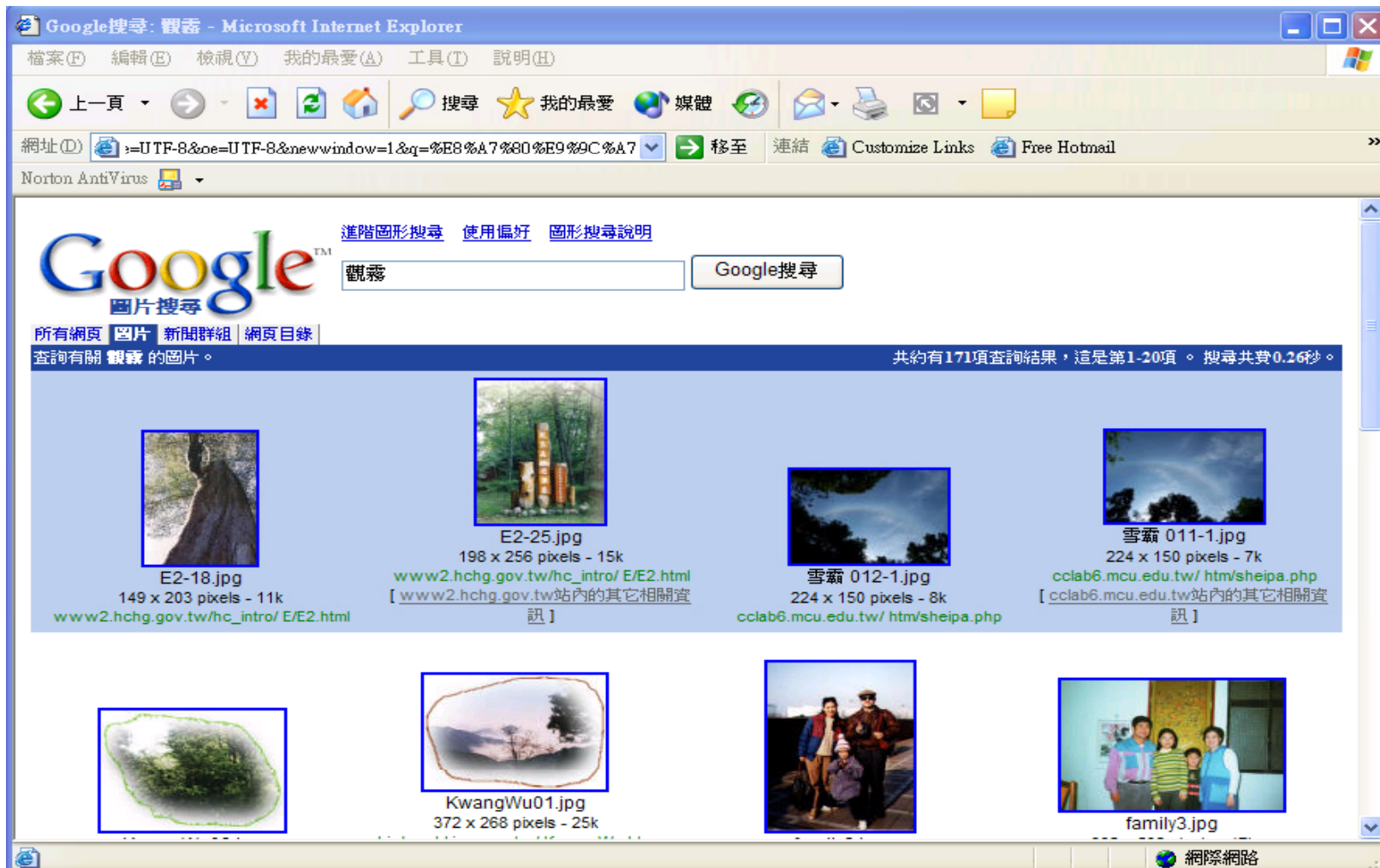


Figure 1.5 Sketch of a new-generation visual information retrieval system for video.

# Visual Information Retrieval (2/4)

- Images with Texts (Metadata)



# Visual Information Retrieval (3/4)

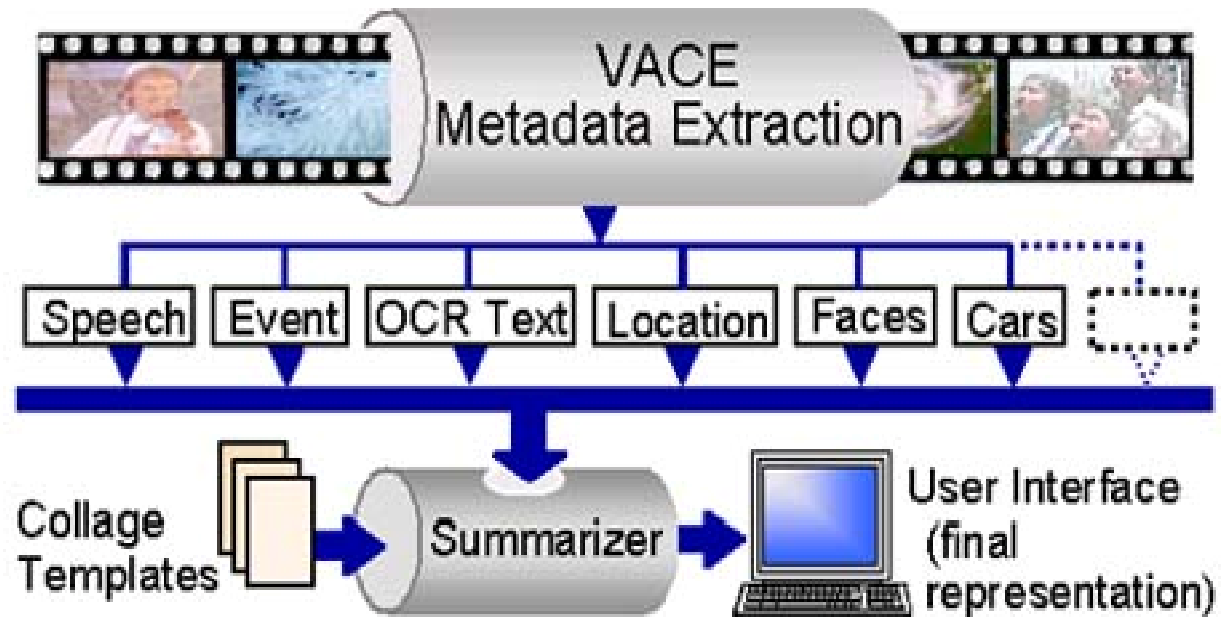
- Content-based Image Retrieval



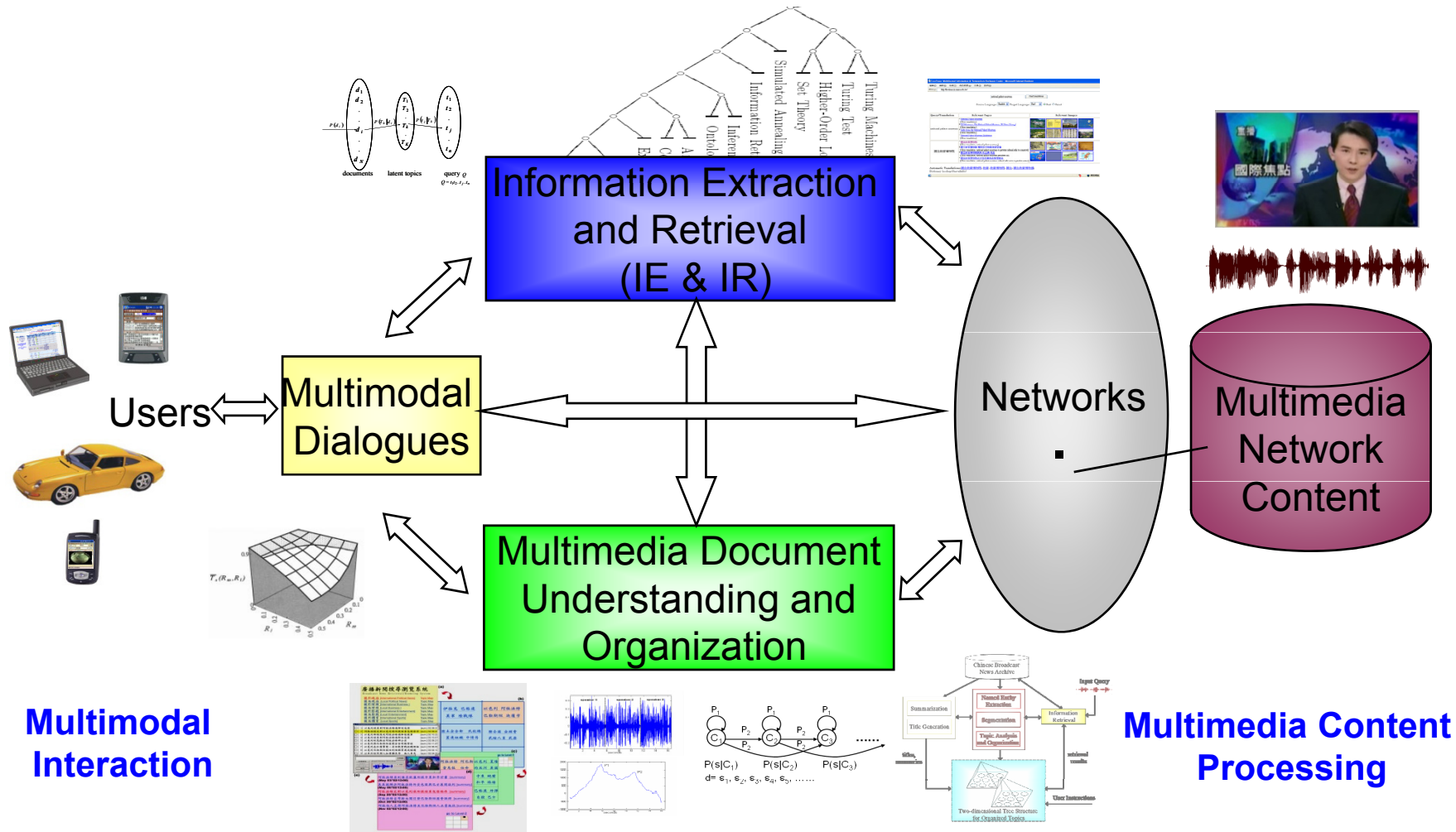


# Visual Information Retrieval (4/4)

## Video Analysis and Content Extraction



# Scenario for Multimedia information access



Multimodal Interaction

Multimedia Content Processing

# Other IR-Related Tasks

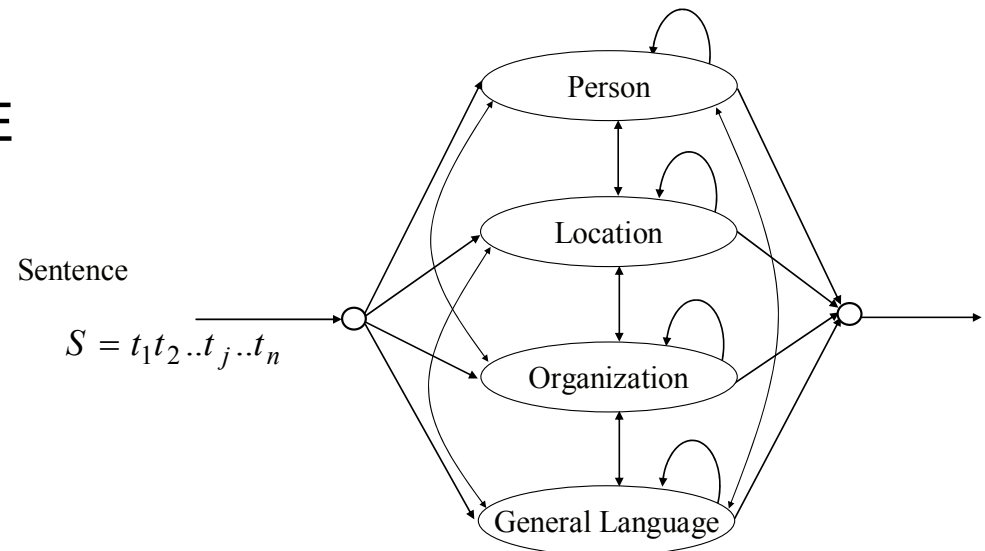
- Information filtering and routing
- **Term/Document categorization**
- **Term/Document clustering**
- **Document summarization**
- **Information extraction**
- Question answering
  - “*What is the height of Mt. Everest?*”
- Crosslingual information retrieval
- .....

# 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

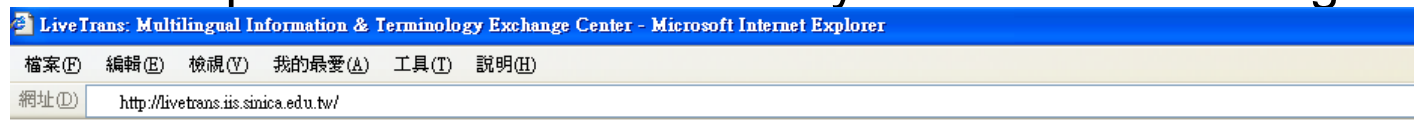
# Information Extraction

- 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



national palace museum    FindTranslations

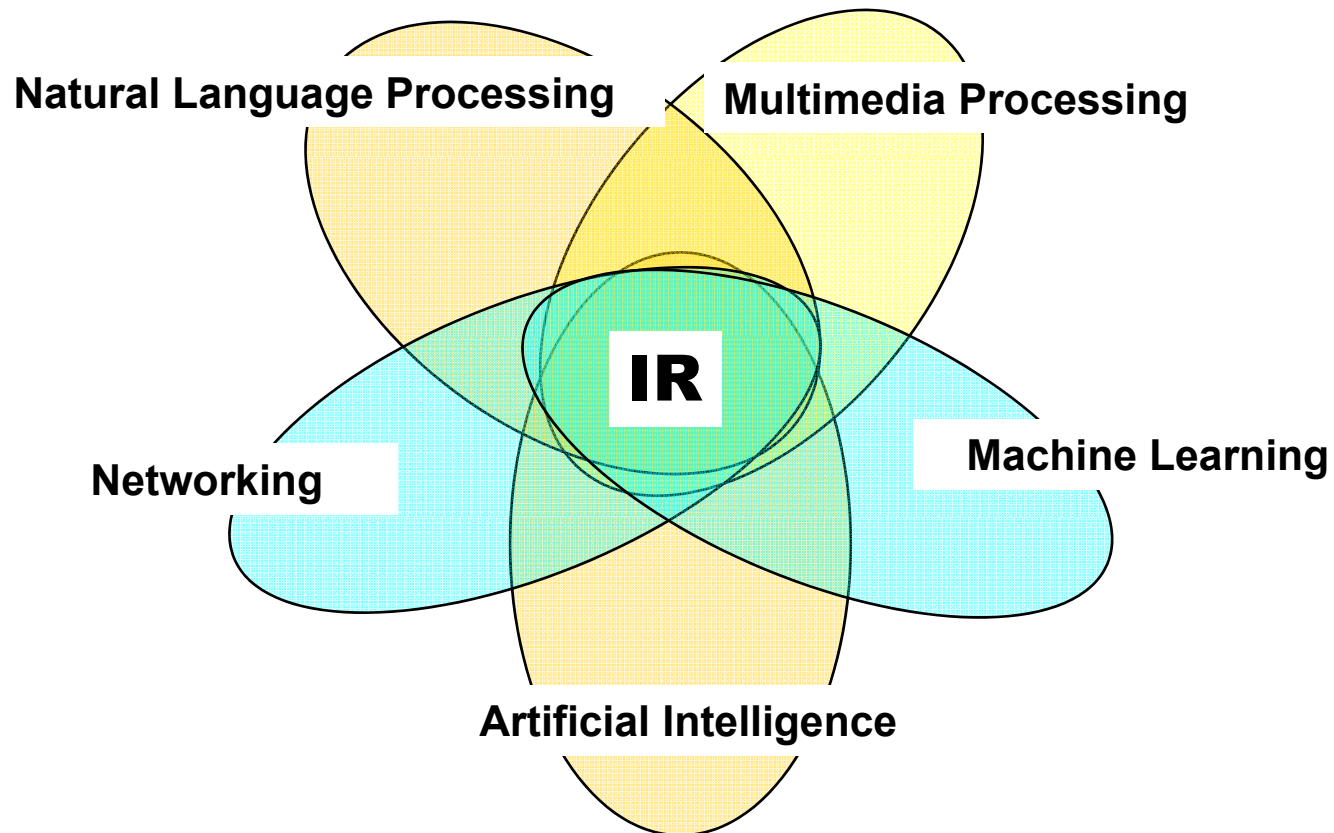
Source Language: English    Target Language: Big5     Fast     Smart

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

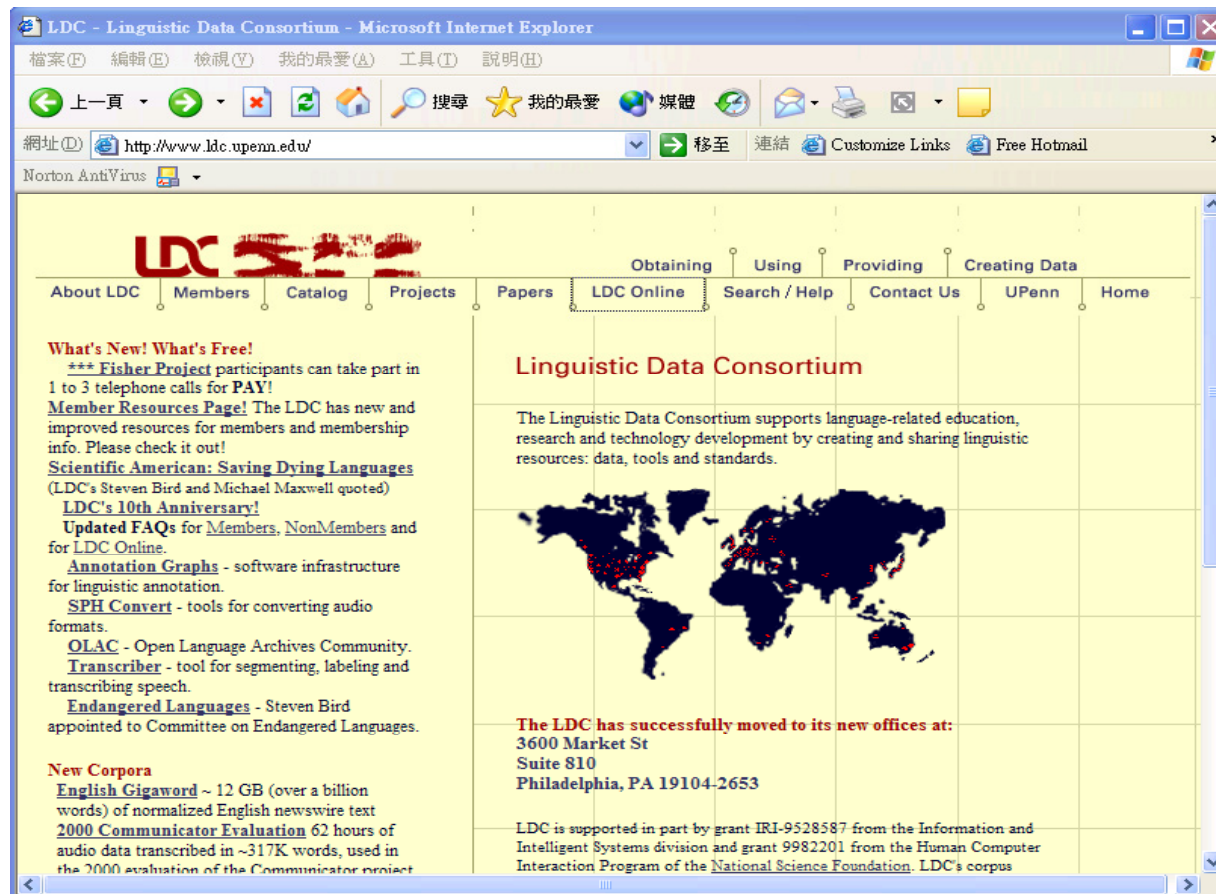
Automatic Translations: [國立故宮博物院](#); [故宮](#); [故宮博物院](#); [國立](#); [國立故宮博物館](#);  
Dictionary Lookup: Unavailable!

# Multidisciplinary Approaches



# Resources

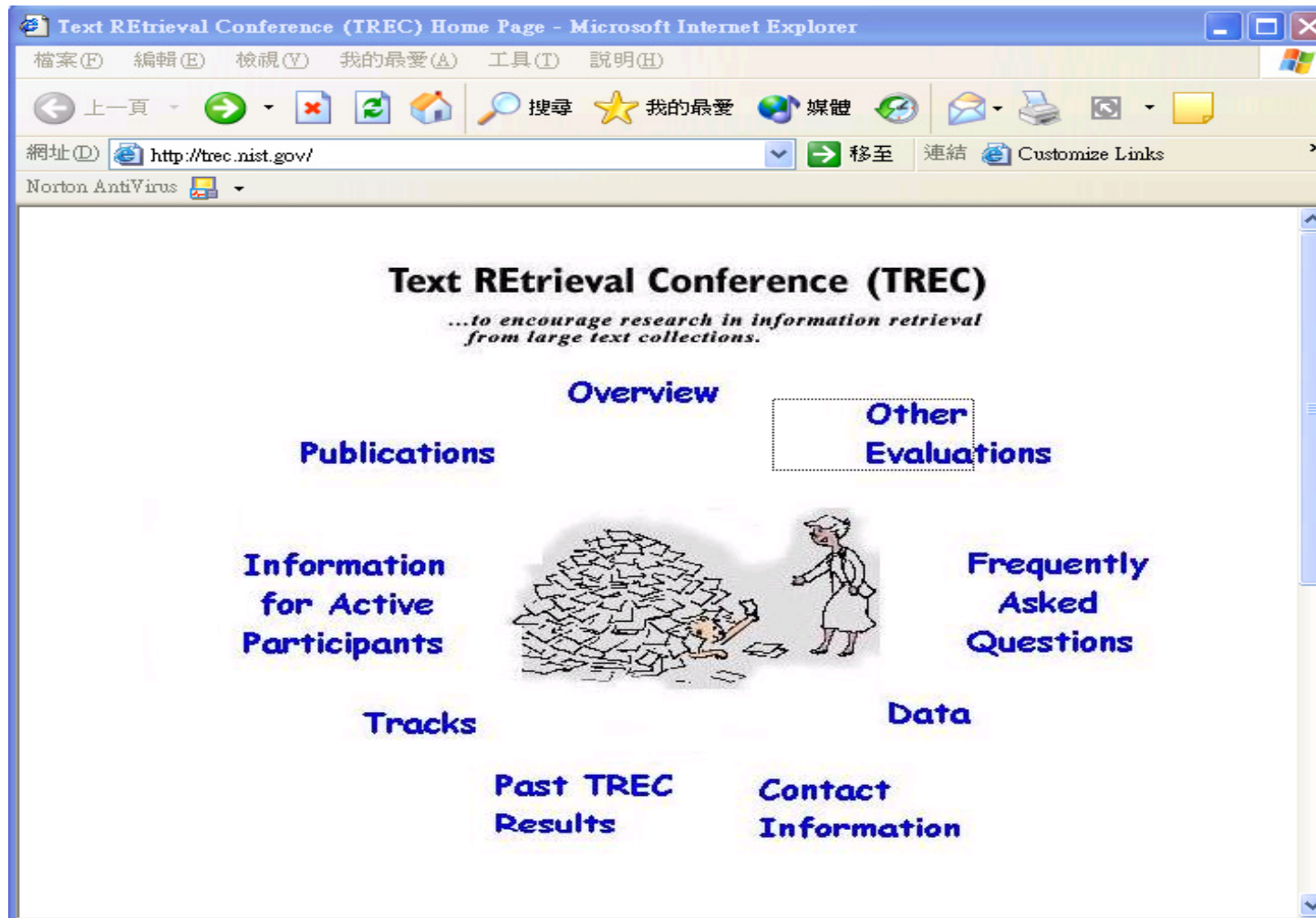
- Corpora (Speech/Language resources)
  - Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them
    - [LDC - Linguistic Data Consortium](http://www ldc.upenn.edu/)





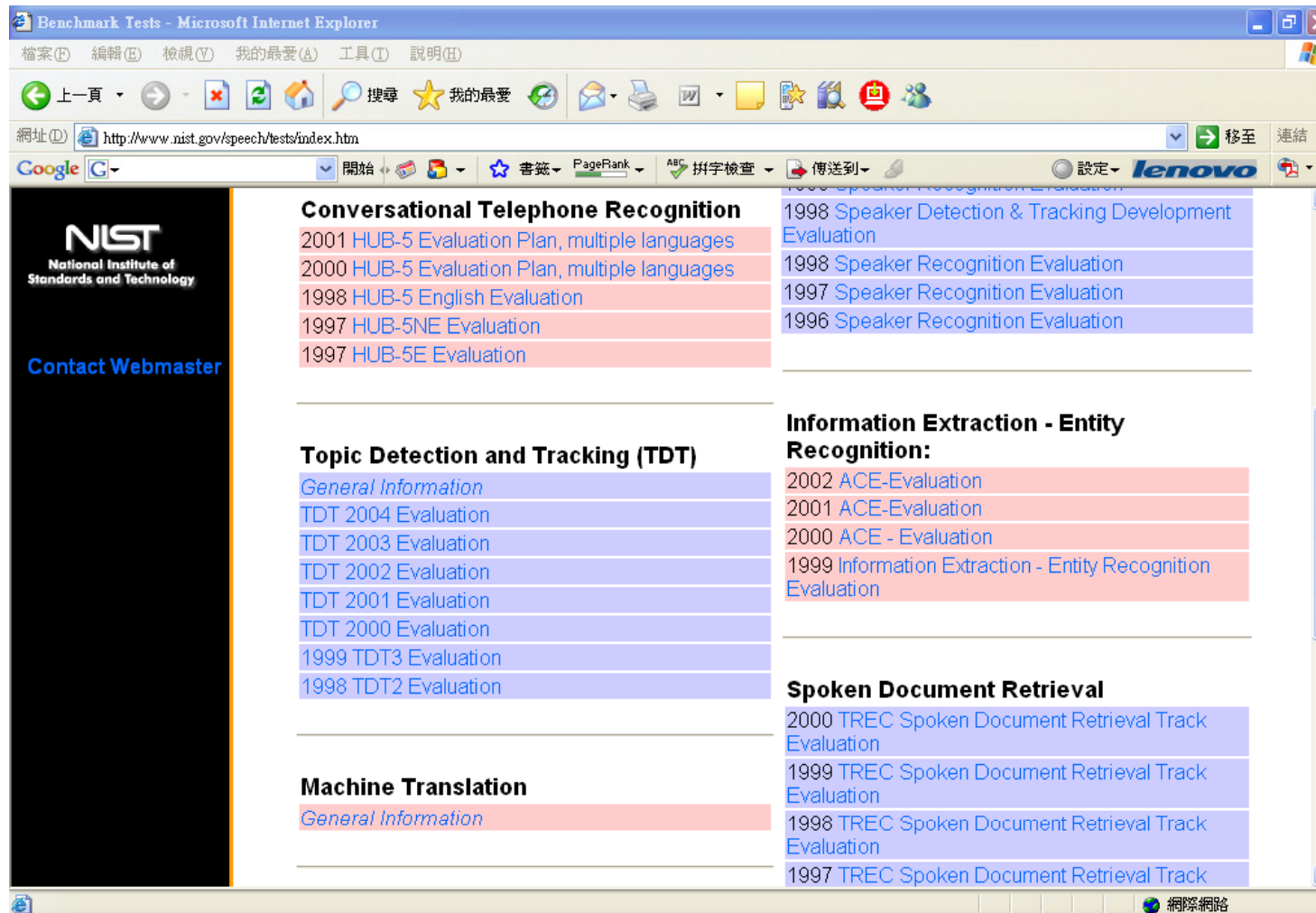
# Contests (1/2)

- [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 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): 20%
- Homework/Projects: 50%
- Presentation: 20%
- Attendance/Other: 10%
  
- TA: 張鈺玫同學
  - E-mail: [cheese0613@gmail.com](mailto:cheese0613@gmail.com)
  - Tel: 29322411ext 208 (資工系208室)