Information Retrieval and Extraction

Berlin Chen

(Picture from the TREC web site)
Textbook and References

• Textbooks

• References
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

![Diagram showing the user task with retrieval and browsing processes]

F1 racing
Directions to Le Mans
Tourism in France

Information Records
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 be 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, …)

```text
Docs -> accents, spacing, etc. -> stopwords -> Noun groups -> stemming -> Manual indexing

structure -> text + structure

structure -> Full text

Index terms
```
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)

- 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
## Some Directions of Information Retrieval

<table>
<thead>
<tr>
<th>Example of Content</th>
<th>Example of Applications</th>
<th>Examples of Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Web search</td>
<td>Ad hoc search</td>
</tr>
<tr>
<td>Images</td>
<td>Vertical search</td>
<td>Filtering</td>
</tr>
<tr>
<td>Video</td>
<td>Enterprise search</td>
<td>Classification</td>
</tr>
<tr>
<td>Scanned documents</td>
<td>Desktop search</td>
<td>Question answering</td>
</tr>
<tr>
<td>Audio (Speech)</td>
<td>Peer-to-peer search</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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
Speech Information Retrieval (1/4)

Speech query (SQ)

我想找有關“中美軍機擦撞”的新聞?

Text query (TQ)

我想找有關“中美軍機擦撞”的新聞？

Text documents (TD)

.... 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 ....

IR – Berlin Chen 26
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)
Speech Information Retrieval (3/4)

• Speech Summarization and Retrieval

Input speech query: "Help me check the president's flag ceremony"
Speech Information Retrieval (4/4)

• Speech Organization

Visual Information Retrieval (1/4)

- Content-based approach
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

- Speech
- Event
- OCR Text
- Location
- Faces
- Cars

Collage Templates
Summarizer
User Interface (final representation)
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 it 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
Multidisciplinary Approaches

Natural Language Processing

Multimedia Processing

Networking

Machine Learning

Artificial Intelligence

IR
Resources

• Corpora (Speech/Language resources)
  – Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them

• LDC - Linguistic Data Consortium
Contests (1/2)

- **Text REtrieval Conference (TREC)**
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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course Overview &amp; Introduction</td>
<td></td>
</tr>
<tr>
<td>Retrieval Models (I) - Classic Retrieval Models</td>
<td>(Boolean, Vector Space and Probabilistic Models)</td>
</tr>
<tr>
<td>Retrieval Performance Evaluation - Measures</td>
<td></td>
</tr>
<tr>
<td>Retrieval Performance Evaluation - Collections</td>
<td></td>
</tr>
<tr>
<td>Retrieval Models (II) - Improved Approaches (Fuzzy Set, Extended Boolean, Generalized Vector Space Models)</td>
<td></td>
</tr>
<tr>
<td>Query Operations (Query Expansion and Term Re-weighting)</td>
<td></td>
</tr>
<tr>
<td>Retrieval Models (III) - Latent Semantic Analysis (LSA)</td>
<td></td>
</tr>
<tr>
<td>Retrieval Models (IV) - Language Models</td>
<td></td>
</tr>
<tr>
<td>Retrieval Models (V) - Learning to Rank</td>
<td></td>
</tr>
<tr>
<td>Clustering for Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>Classification for Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>Efficient Indexing and Searching</td>
<td></td>
</tr>
<tr>
<td>Web Search Basics</td>
<td></td>
</tr>
<tr>
<td>Cross-lingual Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>Spoken Document Recognition, Retrieval and Summarization</td>
<td></td>
</tr>
</tbody>
</table>
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
  – Tel: 29322411ext 208 (資工系208室)