4D2b - Information Retrieval

Owen.Conlan@scss.tcd.ie
Outline

• Introduction to the principles of Information Retrieval and Web Search
• Web Crawling
• The Future of IR and The Web
IR – Past and Present

• The Fundamental Paradox of Information Retrieval
  – “The need to describe that which you do not know in order to find it” - Roland Hjerrpe

• circa 1955-1975
  – Foundational research
  – Fundamental IR concepts advanced in research

• 1975 to present
  – Slow adoption of IR research into operational systems
  – Accelerated in mid-1990’s due to WWW search engines
  – Accelerated in mid 2000’s due to OS level indexing applications for individual workstations
Information Lifecycle

- **Active**
  - Authoring
  - Modifying
  - Using
  - Creating

- **Inactive**
  - Retention
  - Mining
  - Accessing
  - Filtering

- **Semi-Active**
  - Organising
  - Indexing
  - Storing
  - Retrieval

- **Searching**
  - Networking
  - Distribution

- **Utilisation**
  - Disposition
  - Discard
Key Issues....

• How to organise and describe collections of Information Objects for ease of discovery and utilisation
  – Organising / Indexing / Storing

• How to find Information Objects which are relevant to a user's explicit needs
  – Retrieval / Accessing / Filtering
Information Lifecycle

- Creation
  - Authoring
  - Modifying
- Using
  - Creating
- Retention
  - Mining
- Accessing
  - Filtering
- Storing
  - Retrieval
- Organising
  - Indexing
- Searching
  - Distribution
  - Networking
- Utilisation
  - Disposition
  - Discard
  - Active
  - Semi-Active
  - Inactive
Information is a Relationship

“..information is not *stuff* contained in books as marbles might be contained within a bag – even though we sometimes speak of it that way. It is, rather a relationship.”

“..the impact of a given message on an individual is relative to what he already knows...the same message could convey different amounts of information to different receivers”
The Role of an IR System

“An IR system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (non-existence) and whereabouts of documents relating to his request”

- C.J. van Rijsbergen 1975
Types of IR Systems

- Human IR Systems
  - Librarian
- Digital Libraries or Content Repositories
  - NDLR
  - JORUM
  - MERLOT
- Web IR Systems
  - Search Engines
“The Library of Alexandria was the first time humanity attempted to bring the sum total of human knowledge together in one place at one time..... Our latest attempt? Google.”

– Brewster Kahle

• Entrepreneur and Founder, The Internet Archive
Evaluating retrieval

- **Precision** = \( \frac{\text{number of relevant items retrieved}}{\text{number of items retrieved}} \)

- **Recall** = \( \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}} \)

There will almost always be a trade-off between precision and recall.

On the web you can’t know this.
Searching the Web

• Challenges
  – Distributed data
  – High percentage of volatile data
    • 60%+ changes every month
  – Large volume
  – Unstructured and redundant data
    • 30% near duplicates
  – Quality of data
    • 1 in 200 typos
  – Heterogeneous data
    • Media types
    • Languages
Crawler-Indexer Architecture

- Crawler
- Interface
- Query Engine
- Index
- Indexer

Online

Offline
IR Systems at their most basic...

Information Retrieval System

Information Need or Query

Information Objects

Highly Relevant Documents (Hopefully!!)
Information Objects or Documents

- Books, Journals
- Articles, Reports
- Web Pages
- Office Documents
- Pictures Graphics
- Audio or Video Files
- Blogs, RSS Feeds
Doc Representation v Full Text

• Document Representation
  – Uses a Surrogate to represent each Doc
  – Surrogate may include....
    • Abstract, Keywords, Index Terms
  – This Process may be manual or automatic

• Full-Text Retrieval
  – Uses the full text of each document
  – Searches can be performed on given terms or strings that occur anywhere in the document
Data versus Information

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Salary</th>
<th>Date of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chandler Bing</td>
<td>M</td>
<td>$45,000</td>
<td>11/03/1972</td>
</tr>
<tr>
<td>Joey Tribbiani</td>
<td>M</td>
<td>$20,000</td>
<td>18/07/1974</td>
</tr>
<tr>
<td>Rachel Green</td>
<td>F</td>
<td>$50,000</td>
<td>23/10/1973</td>
</tr>
</tbody>
</table>

“James Joyce was born in Dublin in 1882. His works include Ulysses and Finnegans Wake. He died in 1941 in Switzerland.”
Data versus Information

SQL

• SELECT Name
  FROM Character
  WHERE Salary
  BETWEEN 40000
  AND 60000

• Artificial Language
• Known Data Types
• Exact Criteria
Data versus Information

SQL

• Definitive Results
• Returns the Complete Set of Data that meets search criteria
• No estimation of Relevancy
Structure of IR systems

- Obtain Information Objects
- Represent / Index the Information Objects
- User provides an Information Request
- Attempt to match this request against the indexed Information Objects
- Produce a list of recommended Information Objects based on this match
Structure of an IR System

- Information Objects
- Query
- Object Indexing
- Query Indexing
- Refined Query
- Matching / Retrieval
- Recommended Objects
- Information Need
- User Review

Information Object Description Files
Why is IR hard?

• Documents, Images, Videos, Audio Clips, etc etc are all very complex

• Computers cannot extract meaning or context from these Information Objects
  – Semantics
    • What words and images mean
  – Natural Language
    • How humans structure sentences
Why is IR hard?

• Consider the image search
  – “A Cow called Margaret”
Why is IR hard?

• ...or the search terms
  – “Homer’s Odyssey”
Why is IR hard?

• Individual preferences and Opinion....

Funny?  
Talented?  
Honest?
Why is IR hard?

• Information needs must be expressed as a query which can be input to an IR system
  – but users often aren’t sure what they want!

• Issues....
  – Verbalising Information needs
  – Understanding Query Syntax
  – Understanding Search Engines
Why is IR hard?

• Queries can often be...
  – Underspecified
    • “premier league” “oc”
  – Ambiguous
    • “jordan” “jaguar”
  – Context-sensitive
  – Represent different Information needs
    • Question answering
    • Background search......
Why is IR hard?

• Information is often dynamic
  – Blogs, News, Traffic Reports, Timetables....

• Queries can also be dynamic
  – Information needs can evolve as users search

• Scalability is also an issue
  – 357 terabytes of print info produced annually
  – **224,749,695** web sites as of March 2009! – stats from netcraft.com
  – Plus TV, Radio, Satellite Images.......
IR as a Business

• Information Retrieval is massive business. Some well known companies whose business is based upon or relies on IR.....
IR as a Business

• Google earned $21.8 Billion in 2008 – up 31% on 2007! Its market value is over $95 Billion.

• That is more than the traditional media companies Time Warner, Viacom, CBS and New York Times combined!

• Most people claim that IR and search is less than 10% solved, it is an industry with enormous potential growth.
Postulates of Impotence
(according to Swanson, 1988)

• An information need cannot be expressed independent of context
• It is impossible to instruct a machine to translate a request into adequate search terms
• A document’s relevance depends on other seen documents
• It is never possible to verify whether all relevant documents have been found
• Machines cannot recognise meaning
  – can’t beat human indexing etc
Summary

• Information Retrieval concerns the structure, analysis, organization, storage, searching, and retrieval of information.

• Information Retrieval is different to Data Access

• Information Retrieval concerns People

• Information Retrieval is hard!

• Information Retrieval is Big Business
Recommended Reading

• C. J. van Rijsbergen. Information Retrieval. 2nd Edition (1979). Full text available free online....
  – http://www.dcs.gla.ac.uk/Keith/Preface.html


• Chakrabarti, S; Mining the Web: Discovering Knowledge from Hypertext Data (2002).