Introduction to Information Retrieval

Hongning Wang
CS@UVa
What is information retrieval?

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR syst.
Why information retrieval

• Information overload
  – “It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information.” - wiki
Why information retrieval

- Information overload
Why information retrieval

• Handling unstructured data
  – Structured data: database system is a good choice
  – Unstructured data is more dominant
    • Text in Web documents or emails, image, audio, video...
    • “85 percent of all business information exists as unstructured data” - Merrill Lynch

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Job</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jack</td>
<td>Professor</td>
</tr>
<tr>
<td>3</td>
<td>David</td>
<td>Stuff</td>
</tr>
<tr>
<td>5</td>
<td>Tony</td>
<td>IT support</td>
</tr>
</tbody>
</table>

Table 1: People in CS Department

Why information retrieval

- An essential tool to deal with information overload
History of information retrieval

- Idea popularized in the pioneer article “As We May Think” by Vannevar Bush, 1945
  - “Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.” -> WWW
  - “A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.” -> Search engine
Major research milestones

• Early days (late 1950s to 1960s): foundation of the field
  – Luhn’s work on automatic indexing
  – Cleverdon’s Cranfield evaluation methodology and index experiments
  – Salton’s early work on SMART system and experiments
• 1970s-1980s: a large number of retrieval models
  – Vector space model
  – Probabilistic models
• 1990s: further development of retrieval models and new tasks
  – Language models
  – TREC evaluation
  – Web search
• 2000s-present: more applications, especially Web search and interactions with other fields
  – Learning to rank
  – Scalability (e.g., MapReduce)
  – Real-time search
History of information retrieval

• Catalyst
  – Academia: Text Retrieval Conference (TREC) in 1992
    • “Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.”
    • “... about one-third of the improvement in web search engines from 1999 to 2009 is attributable to TREC. Those enhancements likely saved up to 3 billion hours of time using web search engines.”
    • Till today, it is still a major test-bed for academic research in IR
History of information retrieval

• Catalyst
  – Industry: web search engines
    • WWW unleashed explosion of published information and drove the innovation of IR techniques
    • First web search engine: “Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that periodically mirrored these pages and rewrote them into a standard format.” Sept 2, 1993
    • Lycos (started at CMU) was launched and became a major commercial endeavor in 1994
    • Booming of search engine industry: Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, Yahoo!, Google, and Bing
Major players in this game

• Global search engine market - desktop
  – By http://marketshare.hitslink.com/search-engine-market-share.aspx

![Pie chart showing market share of search engines with Google at 81.12%, Bing at 6.97%, Yahoo at 4.74%, Baidu at 5.82%, Ask at 0.17%, AOL at 0.05%, and Excite at 0.01% in Fall 2017.](image-url)
Major players in this game

• Global search engine market - mobile
  – By http://marketshare.hitslink.com/search-engine-market-share.aspx
How to perform information retrieval

• Information retrieval when we did not have a computer
How to perform information retrieval

Crawler and indexer
- URL Server
- Crawler
- Store Server
- Anchors
- Repository
- URL Resolver
- Indexer

Query parser
- Lexicon

Ranking model
- Barrels
- Sorter

Document Analyzer
- PageRank
- Doc Index
- Searcher
How to perform information retrieval

We will cover:
1) Search engine architecture; 2) Retrieval models;
3) Retrieval evaluation; 4) Relevance feedback;
5) Link analysis; 6) Search applications.
Core concepts in IR

• Query representation
  – Lexical gap: say v.s. said
  – Semantic gap: ranking model v.s. retrieval method

• Document representation
  – Special data structure for efficient access
  – Lexical gap and semantic gap

• Retrieval model
  – Algorithms that find the **most relevant** documents for the given information need
A glance of modern search engine

Yet Another Hierarchical Officious/Obstreperous/Odiferous/Organized Oracle

In old times

Yahoo race of fictional beings from Gulliver's Travels
A glance of modern search engine

Demand of efficiency
Demand of convenience
Demand of understanding
Demand of diversity
Demand of accuracy
Demand of efficiency
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes – Recommendation

Recommended Based on Your Browsing History

- Linear Algebra and Its Applications...
  - David C. Lay
  - Hardcover
  - $141.16

- Linear Algebra: A Modern Introduction
  - David Poole
  - $289.88

- Linear Algebra
  - G. E. Shilov
  - Hardcover
  - $126.55

- Introduction to Linear Algebra...
  - Gilbert Strang
  - Paperback
  - $83.13

- Linear Algebra For Dummies
  - Mary Jane Sterling
  - Paperback
  - $16.23
IR is not just about web search
IR is not just about web search

• Web search is just one important area of information retrieval, but not all
• Information retrieval also includes
  – Question answering

\[
\frac{d}{dx} (\Gamma(x)) = \Gamma(x) \psi^{(0)}(x)
\]

\(\Gamma(x)\) is the gamma function
\(\psi^{(n)}(x)\) is the \(n^{\text{th}}\) derivative of the digamma function
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Text mining

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Online advertising
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Enterprise search: web search + desktop search
IR v.s. DBs

• Information Retrieval:
  – Unstructured data
  – Semantics of objects are subjective
  – Simple keyword queries
  – Relevance-drive retrieval
  – Effectiveness is primary issue, though efficiency is also important

• Database Systems:
  – Structured data
  – Semantics of each object are well defined
  – Structured query languages (e.g., SQL)
  – Exact retrieval
  – Emphasis on efficiency
IR and DBs are getting closer

- **IR => DBs**
  - Approximate search is available in DBs
  - Eg. in mySQL

```sql
mysql> SELECT * FROM articles
       -> WHERE MATCH (title,body) AGAINST ('database');
```

- **DBs => IR**
  - Use information extraction to convert unstructured data to structured data
  - Semi-structured representation: XML data; queries with structured information
IR v.s. NLP

• Information retrieval
  – Computational approaches
  – Statistical (shallow) understanding of language
  – Handle large scale problems

• Natural language processing
  – Cognitive, symbolic and computational approaches
  – Semantic (deep) understanding of language
  – (often times) small scale problems
IR and NLP are getting closer

• IR => NLP
  – Larger data collections
  – Scalable/robust NLP techniques, e.g., translation models

• NLP => IR
  – Deep analysis of text documents and queries
  – Information extraction for structured IR tasks
Text books


What to read?

- Information Retrieval
- Databases
- Library & Info Science
- Machine Learning
- Pattern Recognition
- NLP
- Web Applications, Bioinformatics...
- Statistics
- Optimization
- Algorithms
- Machine Learning
- Pattern Recognition
- ICML, NIPS, UAI
- NLP
- ACL, EMNLP, COLING
- Data Mining
- KDD, ICDM, SDM
- Software engineering
- Computer systems
- Library & Info Science
- Databases
- SIGMOD, VLDB, ICDE
- Systems

- Find more on course website for resource
IR in future

- Mobile search
  - Desktop search + location? Not exactly!!
- Interactive retrieval
  - Machine collaborates with human for information access
- Personal assistant
  - Proactive information retrieval
  - Knowledge navigator
- And many more
  - You name it!
What you should know

• IR originates from library science for handling unstructured data
• IR has many important application areas, e.g., web search, recommendation, and question answering
• IR is a highly interdisciplinary area with DBs, NLP, ML, HCI
Today’s reading

- Introduction to Information Retrieval
  - Chapter 1: Boolean Retrieval