Lecture 0: Introduction

Fernando Diaz
Microsoft Research NYC

September 13, 2016
Information Retrieval

given a query and a corpus, find relevant documents.
given a **query** and a **corpus**, find **relevant documents**.

- **query**: user’s expression of the information need
Information Retrieval

given a query and a corpus, find relevant documents.

- **query**: user’s expression of the information need
- **corpus**: the repository of retrievable items
Information Retrieval

given a query and a corpus, find relevant documents.

• query: user’s expression of the information need
• corpus: the repository of retrievable items
• relevance: satisfaction of the information need
Examples of Information Retrieval Problems

Web Search

given a keyword and a web crawl, find relevant URLs.
Examples of Information Retrieval Problems

Image Search

given a keyword and image database, find relevant images.
Examples of Information Retrieval Problems

Question Answering

given a question and available text, rules, logic, find an answer.
Examples of Information Retrieval Problems

Job Search

given a resume and job advertisements, find relevant jobs.
Examples of Information Retrieval Problems

Applicant Search

given a advertisement and resumes, find good candidates.
History

• **1950s**: early information work in problem definition, metrics.

• **1960s**: Gerard Salton begins work on SMART; Cranfield evaluation method defined.

• **1970s**: information retrieval research community developed (SIGIR); many fundamental concepts proposed (e.g. cluster-based retrieval, pseudo-relevance feedback).

• **1980s**: development of first commercial information retrieval systems.

• **1990s**: TREC conferences begin, standardizing evaluation; web search engines developed, using many fundamental IR techniques.
Text REtrieval Conference (TREC)

- Started in 1992 as a forum to compare IR systems using standard test collections, ensuring reproducibility.
- Initially focused on ad hoc retrieval (keyword search), the scope has broadened to include multi-lingual retrieval, legal retrieval, and question answering.
- Allowed for accelerated comparison and testing of algorithmic changes across systems.
- Resulted in similar forums in Europe (CLEF), Asia (NTCIR), and India (FIRE).
## IR ≠ DB

<table>
<thead>
<tr>
<th></th>
<th>DB</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>structured</td>
<td>semi-structured</td>
</tr>
<tr>
<td>fields</td>
<td>clear semantics</td>
<td>free text</td>
</tr>
<tr>
<td>queries</td>
<td>structured</td>
<td>free text</td>
</tr>
<tr>
<td>matching</td>
<td>exact</td>
<td>imprecise</td>
</tr>
<tr>
<td>ranking</td>
<td>none</td>
<td>important</td>
</tr>
</tbody>
</table>

Based on a table by James Allan
Fundamental Problems in Information Retrieval Research

- **Effectiveness**: how well does the system satisfy the user’s information need?
  - algorithms
  - interaction
  - evaluation

- **Efficiency**: how efficiently does the system satisfy the user’s information need?
  - indexing architectures
  - fast score computation
  - evaluation
Effectiveness

algorithms

• term importance: which words are important when ranking a document (e.g. frequent vs. discriminative words)?

• stemming: how to collapse words which are morphologically equivalent (e.g. bicycles → bicycle)

• query expansion: how to collapse words which are semantically equivalent (e.g. bicycles → bicycle)

• document structure: do matches in different parts of the document matter (e.g. title vs. body match)?

• personalization: can we exploit user information to improve ranking?
Effectiveness

interaction

- relevance feedback: ask a user which documents are relevant
- disambiguation: how to ask a user which words are important
Effectiveness evaluation

- relevance: how to define a good document
- metrics: how to measure if the ranking is good
- comparison: how to compare two systems
Efficiency
indexing architectures

- parsing: how should a document be split into a set of terms?
- indexing: which words should be kept?
- weighting: what information needs to be stored with terms?
- compression: how to compress the index size?
Efficiency
fast score computation

- inverted indices: fast retrieval and scoring of short queries.
- tiering: can we tier retrieval and ranking to improve performance?
- branch and bound: how to efficiently prevent scoring unnecessary documents?
Modeling

• information retrieval often involves formally modeling the retrieval process in order to optimize performance.

• modeling tasks
  • abstractly represent the documents
  • abstractly represent the queries
  • model the relationship between query and document representations
Modeling
Boolean Retrieval Model

- represent each document as an unweighted bag of words.
- represent the query as an unweighted bag of words.
- retrieve an unordered set of documents containing the query words.
Modeling
Simple Ranked Retrieval Model

• represent each document as an weighted bag of words (based on document frequency).
• represent the query as an unweighted bag of words.
• retrieve a ranking documents containing the query words.
Modeling

- much of the history of information retrieval effectiveness research involves developing new models or extending existing models.
- as modeling becomes more complicated, mathematics and statistics become necessary.
- new models still being developed.
• **Hypothesis**: Incorporating feature $x$ will improve performance.
Information Retrieval Research

• **Hypothesis:** Incorporating feature $x$ will improve performance.
  • how to come up with new features (hypotheses)?
  • how to measure performance?
Information Retrieval Research

• **Hypothesis**: Incorporating feature $x$ will improve performance.
  • how to come up with new features (hypotheses)?
  • how to measure performance?

• **Experiment**: Compare performance using feature $x$ to a strong baseline.
Information Retrieval Research

- **Hypothesis**: Incorporating feature $x$ will improve performance.
  - how to come up with new features (hypotheses)?
  - how to measure performance?

- **Experiment**: Compare performance using feature $x$ to a strong baseline.
  - what is a strong baseline?
  - how to compare?
Search Engines: Information Retrieval in Practice

• **Lecture 1:** Evaluation
  • What are the core offline relevance metrics?
  • What are the core online relevance metrics?

• **Lecture 2:** Ranking
  • What are the core signals for ranking web pages?
  • What are the core algorithms for combining signals?