Information Retrieval
Course presentation
João Magalhães
A bit of history – 20th Century

• 1945 Vannevar Bush’s Memex (Memory Extender)
A bit of history – 20th Century

- 1945 Vannevar Bush’s **Memex** (Memory Extender)
- 1965 Ted Nelson’s **Hypermedia**
- 1975 Keith v. Rijsbergen’s “**Information Retrieval**” book
- 1989 Tim Berners-Lee’s first **HTTP and HTML** implementations
A bit of history – 20th Century

• 1945 Vannevar Bush’s Memex (Memory Extender)
• 1965 Ted Nelson’s Hypermedia
• 1975 Keith v. Rijsbergen’s “Information Retrieval” book
• 1989 Tim Berners-Lee’s first HTTP and HTML implementations
• 1992 Lynx Web Browser
A bit of history – 20th Century

- 1945 Vannevar Bush’s Memex (Memory Extender)
- 1965 Ted Nelson’s Hypermedia
- 1989 Tim Berners-Lee’s first HTTP and HTML implementations
- 1992 Lynx Web Browser
- 1993 NCSA Mosaic Browser, first TREC
- 1994 Netscape, Yahoo!, WebCrawler, ...
- 1995 Altavista search engine, Apache, PHP
- 1996 Wayback machine
- 1997 PageRank
A bit of history – 21st Century

• 2001 Wikipedia
• 2003 MySpace, Hi5, Skype
• 2004 Facebook, Flickr
• 2005 YouTube, Reddit, Mechanical Turk
• 2007 Twitter
• 2010 Instagram, Kaggle
• 2011 SnapChat
• 2012 Vine
• …
What is the best [search space + dissimilarity function] to compute the relevance of documents for a given user information need?
What makes a good search application?

- **Efficiency**: application replies to user queries without noticeable delays.
  - 1 sec is the “limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer”

- **Effectiveness**: application replies to user queries with relevant answers.
  - This depends on the interpretation of the user query and the stored information.
The tasks of a search application

- **Collect** data for storage
  - Crawler

- Analyse collected data and compute the **relevant information**
  - Information analysis

- Store data in an **efficient** manner
  - Indexing

- Process **user** information needs
  - Querying

- Find the documents that best **match** the user information need
  - Ranking
Crawler

• Discovers data for storage and indexing

• Applied when data has to be “discovered”
  • i.e., only a sample of the full data can be collected.

• If the sample is unbiased, it is a faithful representation of the real scenario.

• Implements a strategy to collect relevant data
  • e.g. on the Web the crawler needs to decide which links are more fruitful to follow and if a page should be indexed or not (it can be spam or phishing)
Web crawling

Begin with known “seed” URLs
Fetch and parse them
  Extract URLs they point to
  Place the extracted URLs on a queue
Fetch “robots.txt”
Fetch each URL on the queue and repeat
Information analysis

• This stage deals with the extraction of the information to be made searchable

• Extract meaningful words, pairs of words or n-grams

• Extract images and their main characteristics

• Link visual characteristics and text data

This patient had a sudden loss of her motor functions (she wasn’t able to move her right arms and legs) 2 months before the study. She went thru a slow recovery with lot physical therapy and drugs. She was recovering some of her movements but suddenly all the improvement stop. We performed an MRI that showed the changes expected for a lesion of that time (2 months old) but also showed and increase in the size of the ventricular system (where the Cerebrospinal fluid or CSF flows) that was causing hydrocephalus. Due to this finding, the patient went thru another surgery and had a shunt valve installed, the last word we had from one of her relatives is that she is again on recovery.

The official report included this: T 1 coronal SE (spin echo) sequence that shows an area of infarction in the left parietal lobe. Also enlargement of the ventricular system is observed.
Indexing

• This stage creates an index to quickly locate relevant documents

• An index is an aggregation of several data structures (e.g. several B-trees)

• Index compression is used to reduce the amount of space and the time needed to compute similarities

• The distribution of the index pages across a cluster improves the search engine responsiveness
Querying

• Conversion of the user query into the internal search space
  • Parsing

• Usage history
  • Cookies, profiles, etc.

• User intention
  • What type of task is the user doing?
Ranking

• Once the user query is converted into the internal search space...
  • The ranking function sorts the information according to its relevance to the user query

• Ranking functions should model the human notion of relevance
  • We don’t really know the mathematical form of the human notion of similarity... it is highly subjective and dynamic. 😊

Since all models are wrong, one can only hope for useful approximations.
Putting all together...

Diagram:

- Indexing
- Information analysis
- Crawler
- Indexes
- Ranking
- Query processing
- Application
- Query
- Documents
- User
References

• Slides and articles provided during classes.

• Books:


Course grading

• The course has two mandatory components:
  • Theoretical part (2 tests or 1 exam): 60%
  • Labs (groups of 3 students): 40%

• Tests:
  • First test 25 October
  • Second test 13 December

• Exam: date to be defined
Laboratories: StackOverflow answer search

• Implement a search engine to search for StackOverflow CrossValidated Answers.

• Understand the roles of each component of a search engine in the performance of the search results.

• Labs are done incrementally. Each week new functionalities will be added to the initial implementation.

• There will be 5 submissions throughout the semester.
  • Each submission must provide the code and the answers to the questions.
  • Only the 4 best out of the 5 submissions will count.
## Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>#</th>
<th>Lecture</th>
<th>#</th>
<th>In-class labs</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-Sep-17</td>
<td>1</td>
<td>Introduction</td>
<td>1</td>
<td>Setup environment (Lucene, Luke, dataset)</td>
</tr>
<tr>
<td>27-Sep-17</td>
<td>2</td>
<td>Basic techniques</td>
<td>2</td>
<td>Text pre-processing</td>
</tr>
<tr>
<td>04-Oct-17</td>
<td>3</td>
<td>Evaluation</td>
<td>3</td>
<td>Text pre-processing</td>
</tr>
<tr>
<td>11-Oct-17</td>
<td>4</td>
<td>Query processing</td>
<td>4</td>
<td>Evaluation protocols</td>
</tr>
<tr>
<td>18-Oct-17</td>
<td>5</td>
<td>First test</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>25-Oct-17</td>
<td>6</td>
<td>PRP / Language Models</td>
<td>6</td>
<td>Retrieval models</td>
</tr>
<tr>
<td>01-Nov-17</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>08-Nov-17</td>
<td>7</td>
<td>BIM and BM25</td>
<td>7</td>
<td>Retrieval models</td>
</tr>
<tr>
<td>15-Nov-17</td>
<td>8</td>
<td>Rank fusion</td>
<td>8</td>
<td>Query expansion</td>
</tr>
<tr>
<td>22-Nov-17</td>
<td>9</td>
<td>Linked data</td>
<td>9</td>
<td>Query expansion</td>
</tr>
<tr>
<td>29-Nov-17</td>
<td>10</td>
<td>Index construction</td>
<td>10</td>
<td>Fusion of multiple field ranks</td>
</tr>
<tr>
<td>06-Dec-17</td>
<td>11</td>
<td>Efficient query processing</td>
<td>11</td>
<td>Fusion of multiple field ranks</td>
</tr>
<tr>
<td>13-Dec-17</td>
<td>12</td>
<td>Revisions</td>
<td>12</td>
<td>PageRank</td>
</tr>
<tr>
<td>18-Dec-17</td>
<td>13</td>
<td>Second test</td>
<td>13</td>
<td>PageRank</td>
</tr>
</tbody>
</table>
Summary

• “Information Retrieval” course context

• Course objectives and plan

• Grading

• Labs