Information Retrieval (Search)

IR

Artificial Intelligence → IR

Information Retrieval

• Search
• Using a computer to find relevant pieces of information
• Text search
• Idea popularized in the article As We May Think by Vannevar Bush in 1945

Where (or for what) do you do text search?

• World Wide Web
  – Using, e.g., Google, Yahoo
• Library catalog
• Personal (desktop) search
  – Email, files
• Within a document
  – Search-and-replace a word
• Specific domain/database
  – Medline (free)
  – Westlaw (for a fee)

Terminology

• Query
  – What you tell the computer to look for
• Document
  – What you are hoping to find
  • A webpage that contains the info you’re after
  • A specific file on your computer
  • A specific email in your mail box

Type of search

• Flat text
  – Query: robot vision
• Quoted phrases
  – Query: “robot vision”

Type of search

• Flat text
  – Query: robot vision
• Quoted phrases
  – Query: “robot vision”
• Fielded search

IR
Type of search

- Flat text
  - Query: robot vision
- Quoted phrases
  - Query: “robot vision”
- Fielded search
  - Boolean operators
    - Query: flu and swine not human

The process

User issues a query
Query is matched to docs in database
“Relevant” docs are returned
Examples:
- Book titles in library catalog
- Webpages on the WWW

Finding and comparing documents

The vector space model is one method that performs a ranked search
- Represent a document as a vector, i.e., a list of individual words
- Represent the query as a vector
- Compare the two vectors mathematically

Document → Vector (simple version)

I saw a sloth play soccer with a tortoise and a snail.

Compare document with query

Document: a and I play saw sloth snail soccer tortoise with

Query: shell tortoise

1 match
Compare document with query

```
I saw a sloth play soccer with a tortoise and a snail.
```

Document 1:
```
a  and  I  play  saw  sloth  snail  soccer  tortoise  with
```

Document 2:
```
blue birds fly in the blue sky
```

Document 3:
```
A blue tortoise found blue tortoise shell jewelry on the soccer field.
```

Query: shell tortoise

Ranked search result:
Document 3
Document 1

Vector space model

- Vectors are very, very long
  - We say it is a “high-dimensional” problem
  - # dimensions = size of vocabulary

- Very computationally intensive
- Any other problems?

Variation: term weighting

Some words are more discriminating than others. E.g., “the” appears in just about every document

- Term frequency (TF)
  - E.g., “Potter” is in the doc, the more likely the doc is about him

- Inverse document frequency (IDF)
  - The more documents there are containing a certain word, the less likely that word is important

Use term frequency to improve search

```
Score: 1
```
```
Score: 0
```
```
Score: 3
```

Preparing documents for vector space model

- Stemming
  - Potter’s = Potters = Potter

- Stop-words
  - Ignore words like “the”, “of”, …

- Use statistical properties of text
  - E.g., Data from Jamie Callan’s Characteristics of Text, 1997 (Sample of 19 million words)
Commonest fifty words

<table>
<thead>
<tr>
<th>f</th>
<th>f</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1,130,021</td>
<td>from</td>
</tr>
<tr>
<td>of</td>
<td>547,311</td>
<td>he</td>
</tr>
<tr>
<td>to</td>
<td>516,635</td>
<td>million</td>
</tr>
<tr>
<td>a</td>
<td>464,736</td>
<td>year</td>
</tr>
<tr>
<td>in</td>
<td>390,819</td>
<td>its</td>
</tr>
<tr>
<td>and</td>
<td>387,703</td>
<td>be</td>
</tr>
<tr>
<td>that</td>
<td>204,351</td>
<td>was</td>
</tr>
<tr>
<td>for</td>
<td>199,240</td>
<td>company</td>
</tr>
<tr>
<td>is</td>
<td>152,483</td>
<td>an</td>
</tr>
<tr>
<td>said</td>
<td>148,302</td>
<td>has</td>
</tr>
<tr>
<td>it</td>
<td>134,323</td>
<td>are</td>
</tr>
<tr>
<td>on</td>
<td>121,173</td>
<td>have</td>
</tr>
<tr>
<td>by</td>
<td>118,863</td>
<td>but</td>
</tr>
<tr>
<td>as</td>
<td>109,135</td>
<td>will</td>
</tr>
<tr>
<td>at</td>
<td>101,779</td>
<td>say</td>
</tr>
<tr>
<td>mr</td>
<td>101,679</td>
<td>new</td>
</tr>
<tr>
<td>with</td>
<td>101,210</td>
<td>share</td>
</tr>
</tbody>
</table>

Criteria for evaluating IR methods

If information retrieval were perfect, every hit would be relevant to the query and every relevant item in the search space would be found.

- **Precision**
  - How many of the returned documents are relevant?
- **Recall**
  - How many of the relevant documents are returned?
  - Cannot be the sole criterion in evaluation
- **Fall-out**
  - How many of the non-relevant documents are returned?
  - Cannot be the sole criterion in evaluation

Example for evaluating IR methods

Suppose I have a collection of 1000 documents, 50 of them are on a specific topic. An ideal search would find these 50 docs and reject all others.

A particular search method finds 25 docs where 20 are relevant but 5 are on other topics, then…

- **Precision**
  - How many of the returned documents are relevant?
- **Recall**
  - How many of the relevant documents are returned?
- **Fall-out**
  - How many of the non-relevant documents are returned?

Web Search

What’s special about web search?

- **Hyperlinks**
- **Size**—scalability issues
- **Dynamic content**
- **Untrained users**
- **Economic model (advertising)**

“Crawling” the web

- Following the links to determine the link structure
- What are some issue and considerations?
  - Broken links, timeouts, … cause failures
  - Update frequency
  - Coverage, duplicate detection
  - Legal issues (owners don’t want their pages indexed)
  - Advertising links
  - Types of content
  - …
Web search through link analysis

- Find relevant webpages by analyzing the link structure, not by the content
- Most famous algorithm is PageRank
- There are other kinds of link analysis
  - E.g., citation analysis—count the number of references to individual research papers (CiteSeer)

PageRank

- Important part of Google’s success (although most search engines use something like PageRank nowadays)
- Rank pages not just by how relevant they are, but also by how important they are
- Estimate importance by considering a link as a vote
  - The more pages link to you, the more important you are

The PageRank idea

- Many pages link to my page
- ➔ there are many ways to get to my page
- ➔ the probability of getting to my page is high
- ➔ I am important

Start from a random page
Repeat:
Click on a random link ➔ go to that page

Do a large number of such simulations. Where do you end up after a large number of clicks? For each page, how many visitors end up there? ➔ Give the ranks by importance of all the pages

Web search is big business! Advertising

- The advertiser
  - Buy words (e.g., “digital camera”)
  - Then if my search has those words, I’ll see their ad
- The webmaster
  - I want to put ads on my site (revenue)
  - I give space on my site to a search engine company and they fill it with relevant ads
- The user
  - Sees sponsored results
  - Sees customized results