Information Retrieval (Search)

IR

Artificial Intelligence \(\rightarrow\) 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-n-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”
• Fielded search

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
  - 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

"Relevant" docs are ranked
"Relevant" docs are returned with ranking

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  
1 match

"Blue birds fly in the blue sky"  
Document 2:  birds blue fly in sky the  
0 match

"A blue tortoise found blue tortoise shell jewelry on the soccer field"  
Document 3:  a blue field found jewelry on shell soccer the tortoise  
2 matches

Query:  shell tortoise

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., The more times “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

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

Document 2:  birds blue fly in sky the  
Score: 0

Document 3:  a blue field found jewelry on shell soccer the tortoise  
Score: 3

Query:  shell tortoise

Prepared 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)
### Finding documents

- **Brute-force approach?**
  - Look through every single document every time you have a query

- **Efficient way?**
  - Make an index

### Criteria for evaluating IR methods

- **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?

- Can combine these criteria

### 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
- \( \Rightarrow \) there are many ways to get to my page
- \( \Rightarrow \) the probability of getting to my page is high
- \( \Rightarrow \) I am important

Start from a random page
Repeat:
  - Click on a random link \( \Rightarrow \) 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? \( \Rightarrow \) Give the ranks by importance of all the pages

Google can combine this with:
- TF
- IDF
- voodoo
- …

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