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

IR

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)

IR

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

IR

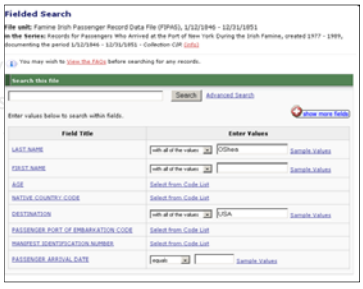
Type of search

- Flat text
  - Query: robot vision
- Quoted phrases
  - Query: "robot vision"

IR

Type of search

- Flat text
  - Query: robot v
- Quoted phrases
  - Query: "robot
- 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

IR

The process

User issues a query

Query is matched to docs in database

Examples:
Book titles in library catalog,
Webpages on the WWW

"Relevant" docs are returned

IR

The process

User issues a query

Query is matched to docs in database

"Relevant" docs are ranked

"Relevant" docs are returned with ranking

IR

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

IR

Document → Vector (simple version)

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

a	and	I	play	saw	sloth	snail	soccer	tortoise	with

IR

Compare document with query

Document:

a	and	I	play	saw	sloth	snail	soccer	tortoise	with

Query:

shell	tortoise

1 match

IR

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

Ranked search result:
Document 3
Document 1

IR

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?

IR

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

IR

Use term frequency to improve search

Document 1:	a	and	I	play	saw	sloth	snail	soccer	tortoise	with
	3	1	1	1	1	1	1	1	1	1

Score: 1

Document 2:	birds	blue	fly	in	sky	the
	1	2	1	1	1	1

Score: 0

Document 3:	a	blue	field	found	jewelry	on	shell	soccer	the	tortoise
	1	2	1	1	1	1	1	1	1	2

Score: 3

Query:	shell	tortoise
	1	1

Ranked search result:
Document 3
Document 1

IR

Two different notations to encode the same vector information

a	and	I	play	saw	sloth	snail	soccer	tortoise	with
3	1	1	1	1	1	1	1	1	1

{ 3:a 1:and 1:I 1:play 1:saw 1:sloth 1:snail 1:soccer 1:tortoise 1:with }

IR

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)

IR

Commonest fifty words

	f		f		f
the	1,130,021	from	96,900	or	54,958
of	547,311	he	94,585	about	53,713
to	516,635	million	93,515	market	52,110
a	464,736	year	90,104	they	51,359
in	390,819	its	86,774	this	50,933
and	387,703	be	85,588	would	50,828
that	204,351	was	83,398	you	49,281
for	199,340	company	83,070	which	48,273
is	152,483	an	76,974	bank	47,940
said	148,302	has	74,405	stock	47,401
it	134,323	are	74,097	trade	47,310
on	121,173	have	73,132	his	47,116
by	118,863	but	71,887	more	46,244
as	109,135	will	71,494	who	42,142
at	101,779	say	66,807	one	41,635
mr	101,679	new	64,456	their	40,910
with	101,210	share	63,925		

IR

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

IR

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?

Report as proportions

IR

Web Search

Artificial Intelligence → Information Retrieval → Web Search

IR

What's special about web search?

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

IR

"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
  - …

IR

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)

IR

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

IR

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

Google can combine this with:
TF
IDF
voodoo
…

IR

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

IR