COM6115: Text Processing

Information Retrieval: Document Indexing — Automatic

Mark Hepple

Department of Computer Science
University of Sheffield

Overview

- Definition of the information retrieval problem
- Approaches to document indexing
  - ◇ manual approaches
  - ◇ automatic approaches
- Automated retrieval models
  - ◇ boolean model
  - ◇ ranked retrieval methods (e.g. vector space model)
- Term manipulation:
  - ◇ stemming, stopwords, term weighting
- Web Search Ranking
- Evaluation

Automatic Indexing

- No predefined set of *index terms*
- Instead: use **natural language** as indexing language
- Words in the document give information about its content
- Implementation of indices: inverted files
- This is what Google's IR system does
  - ◇ at least, it's an important *part* of the story

Automatic Indexing

- A small collection of documents . . .

Document	*Text*
1	Pease porridge **hot**, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it **hot**, some like it cold
5	Some like it in the pot
6	Nine days old

- Say we want to search for word hot. How do we do it?

Inverted files

- A basic inverted file index
  - records for each term, the ids of the documents in which it appears
  - only matters if it *does* or *does not* appear – not how many times

Doc	*Text*
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

⟹

Num	*Token*	*Docs*
1	cold	1, 4
2	days	3, 6
3	hot	1, 4
4	in	2, 5
5	it	4, 5
6	like	4, 5
7	nine	3, 6
8	old	3, 6
9	pease	1, 2
10	porridge	1, 2
11	pot	2, 5
12	some	4, 5
13	the	2, 5

Inverted files (contd)

- A more sophisticated version . . .
  - ◇ also record count of occurrences within each document
  - ◇ help find documents *more relevant* to query

Doc	*Text*
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

⟹

Num	*Token*	*Docs*
1	cold	1:1, 4:1
2	days	3:1, 6:1
3	hot	1:1, 4:1
4	in	2:1, 5:1
5	it	4:2, 5:1
6	like	4:2, 5:1
7	nine	3:1, 6:1
8	old	3:1, 6:1
9	pease	1:2, 2:1
10	porridge	1:2, 2:1
11	pot	2:1, 5:1
12	some	4:2, 5:1
13	the	2:1, 5:1

Inverted files (contd)

- A more sophisticated version . . .
  - ◇ also record *position* of each term occurrence within documents
  - ◇ may be useful for searching for phrases in documents

Doc	*Text*
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

⟹

Num	*Token*	*Docs*
1	cold	1:(6), 4:(8)
2	days	3:(2), 6:(2)
3	hot	1:(3), 4:(4)
4	in	2:(3), 5:(4)
5	it	4:(3, 7), 5:(3)
6	like	4:(2, 6), 5:(2)
7	nine	3:(1), 6:(1)
8	old	3:(3), 6:(3)
9	pease	1:(1, 4), 2:(1)
10	porridge	1:(2, 5), 2:(2)
11	pot	2:(5), 5:(6)
12	some	4:(1, 5), 5:(1)
13	the	2:(4), 5:(5)

Reading

- Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval. New Yorl: ACM Press, 1999.
- C. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- I.H. Witten, A. Moffat and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.