COM3110/4115/6115: Text Processing

Information Retrieval: Task definition & Document Indexing

Mark Hepple

Department of Computer Science
University of Sheffield

Learning Outcomes

- By the end of the IR sessions, you should be able to:
  - implement a simple IR system (from indexing to retrieval) and
  - evaluate how well it does on a gold-standard dataset
- Assignment addresses a substantial part of the above

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

Google search

jaguar

Jaguar International - Market selector page
www.**jaguar**.com/
Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Jaguar International - Home

www.**jaguar**.com/gl/en/
8 Jul 2009
Our mission at **Jaguar** has been to create and build beautiful fast cars. The XK, XF, and XJ bring the **...**

More videos for **jaguar** »

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars
Jaguar Cars Ltd, known simply as **Jaguar** is a British luxury and sports car manufacturer, headquartered in Whitley, Coventry, England. It is part of the **Jaguar ...**

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**
The **jaguar** is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The **jaguar** is the third-largest feline after the tiger **...**

www.hatfields.co.uk/

Google search (contd)

jaguar south america

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**
The **jaguar's** present range extends from Southern United States and Mexico across much of Central **America** and **south** to Paraguay and northern Argentina.
Jaguar Cars - Jaguar (disambiguation) - Jacksonville Jaguars - Jaguarundi

Images for **jaguar south america** - Report images

South America - Jaguar
library.thinkquest.org/5053/**SouthAmerica/jaguar**.html
Jaguars are magnificent cats that prowl the **South American** jungles. They are fascinating to learn about! To jump to a section, use our Quick Jump below by **...**

Jaguars, **Jaguar** Pictures, **Jaguar** Facts - National Geographic
animals.nationalgeographic.com/animals/mammals/**jaguar**/
Jaguars are the largest of **South America's** big cats. They once roamed from the southern tip of that continent north to the region surrounding the U.S.-Mexico **...**

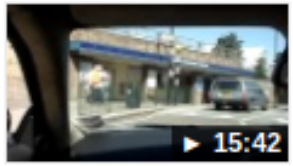
Google search (contd)

black fast jaguar

Jaguars, Jaguar Pictures, **Jaguar** Facts - National Geographic
animals.nationalgeographic.com/animals/mammals/**jaguar**/
Learn all you wanted to know about **jaguars** with pictures, videos, photos, facts, **... Fast** Facts. Type: Mammal; Diet: Carnivore; Average life span in the wild: 12 to **...** Most **jaguars** are tan or orange with distinctive **black** spots, dubbed "rosettes" **...**

Jaguar XKR **black fast** on trackday - YouTube


www.youtube.com/watch?v...
2 Jul 2012 - 16 sec - Uploaded by PrestigeCarCompany
Enjoy this **Jaguar** XKR trackday video.
http://www.prestigecarcompany.co.uk/ Prestige Super Car Sales **...**

Jaguar XKR **Black** Pack **fast** racing - YouTube


www.youtube.com/watch?v...k
16 Aug 2011 - 16 min - Uploaded by MrBobkumar
My XKR being driven hard thru town, hear the sounds of this beast....left window open so noise from air etc....but **...**

Jaguar XKR **black fast** on trackday. Rear Shot - YouTube


www.youtube.com/watch?v=RmvW...
2 Jul 2012 - 17 sec - Uploaded by PrestigeCarCompany
Enjoy this **Jaguar** XKR trackday video.
http://www.prestigecarcompany.co.uk/ Prestige Super Car Sales **...**

More videos for **black fast jaguar** »

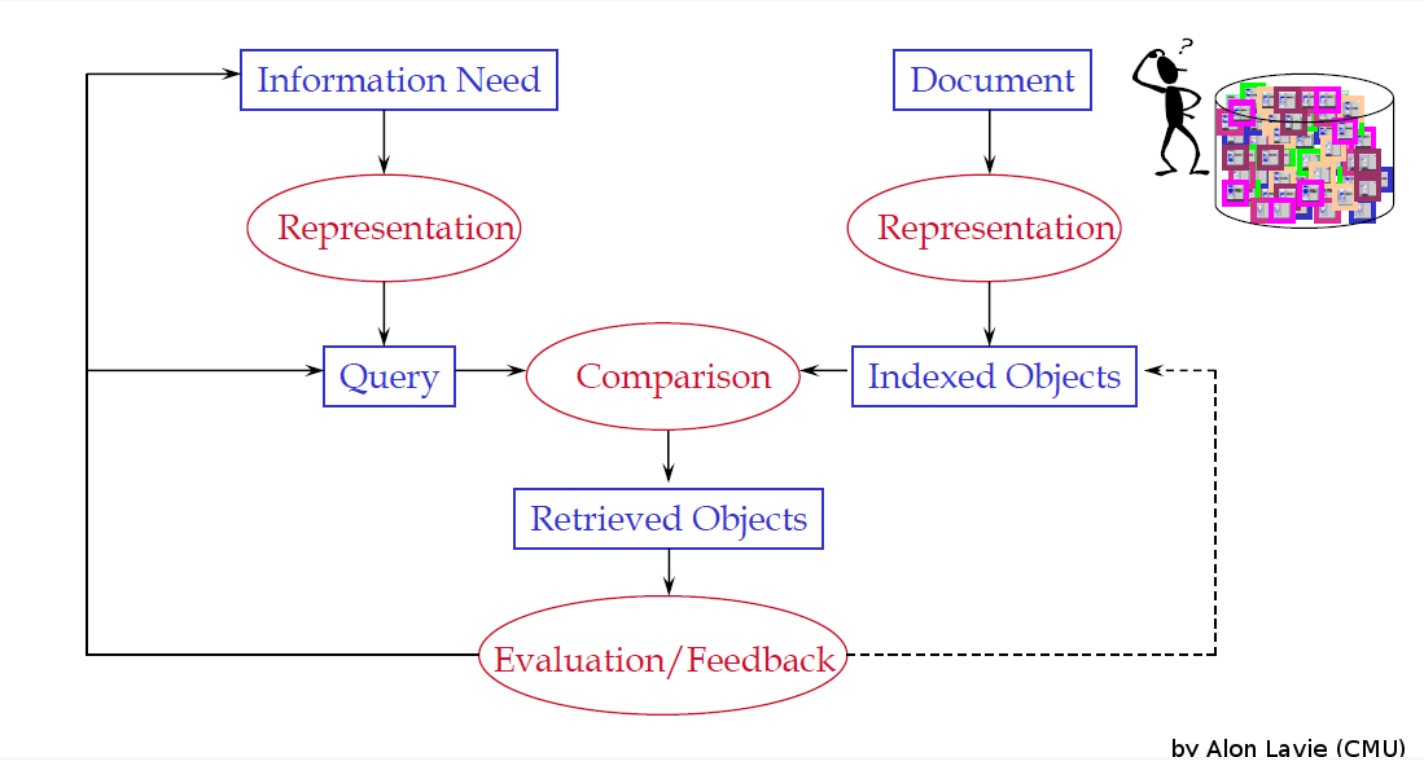
How **fast** can **jaguars** run

Google search (contd)

- What is Google's IR system doing?
  - Finding pages that contain the words in the query.
- How does it rank these pages?
  - By "relevance" to the query.
- How does it do it so fast?
  - By clever indexing (and a lot of hardware!)

Information Retrieval: the task

Text Retrieval: find documents that are "relevant" to a user query.



- Given: a large, static document collection
- Given: an information need (keyword-based query)
- Task: find all and only documents relevant to query

Information Retrieval: the task

Typical IR systems:

- Search a set of abstracts
- Search newspaper articles
- Library search
- Search the Web

Typically: more statistics than 'language', but the object to retrieve (and process) is language

Issues in IR

- How can I formulate a query?
  - query type: normally keywords, could be natural language
- How are the documents represented?
  - indexing
- How does the system find the best-matching document?
  - retrieval model
- How does the system find it *efficiently*?
- How are the results presented to me?
  - unsorted list, ranked list, clusters
- How do we know whether the system is any good?
  - evaluation

Indexing

The task of finding terms that describe documents well

- Manual:
  - indexing by humans (using fixed vocabularies)
  - labour and training intensive
- Automatic:
  - Term manipulation (certain words count as the same term)
  - Term weighting (certain terms are more important than others)
  - Index terms must only derive from text

Manual Indexing

- Large vocabularies (several thousand items)
  - Dewey Decimal System
  - Library of Congress Subject Headings
  - ACM – subfields of CS
  - MeSH – Medical Subject Headings

Example: Manual Indexing — ACM

ACM Computing Classification System (1998)
B Hardware
B.3 Memory structures
B.3.0 General
B.3.1 Semiconductor Memories (NEW) (was B.7.1)
Dynamic memory (DRAM) (NEW)
Read-only memory (ROM) (NEW)
Static memory (SRAM) (NEW)
B.3.2 Design Styles (was D.4.2)
Associative memories
Cache memories
Interleaved memories
Mass storage (e.g., magnetic, optical, RAID)
Primary memory
Sequential-access memory

Example: Manual Indexing — MeSH

MeSH — Medical Subject Headings

- a very large *controlled vocabulary* for describing/indexing medical documents, e.g. journal papers and books
- provides a *hierarchy* of descriptors (a.k.a. *subject headings*)
  - assigned to documents to describe their content
- hierarchy has a number of *top-level* categories, e.g.:
  - Anatomy [A]
  - Organisms [B]
  - Diseases [C]
  - Chemicals and Drugs [D]
  - Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
  - Psychiatry and Psychology [F]
  - Biological Sciences [G]

    ...

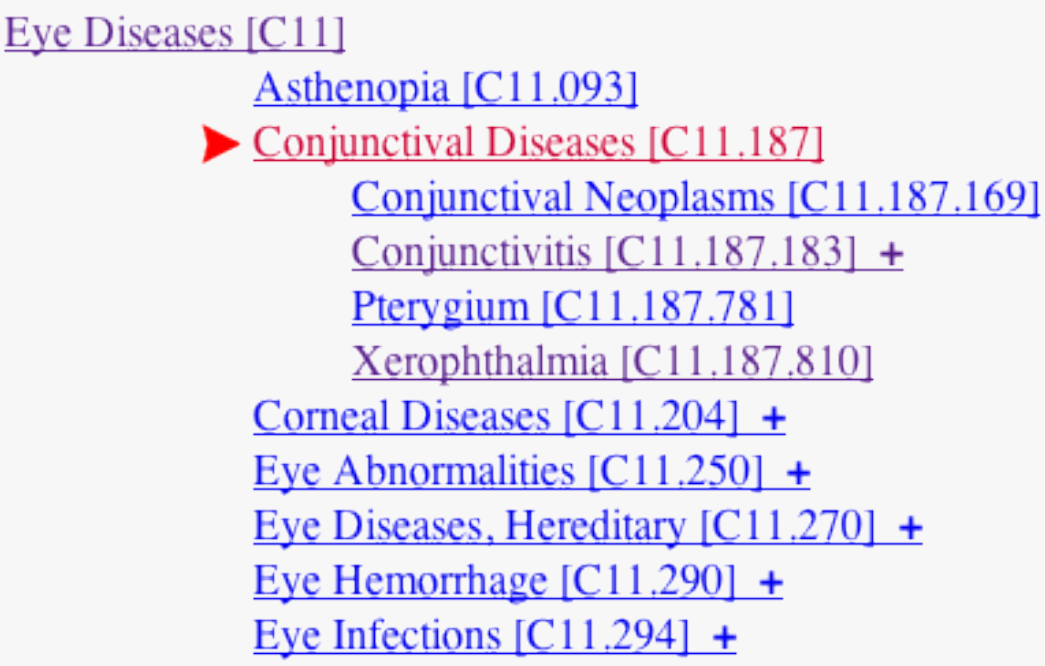
Example: Manual Indexing — MeSH (contd)

- And a number of subcategories (more specific/detailed terms):

  - Diseases [C]
    - **MeSH** *C01* --- bacterial infections and mycoses
    - **MeSH** *C02* --- virus diseases
    - **MeSH** *C03* --- parasitic diseases
    - **MeSH** *C04* --- neoplasms
    - **MeSH** *C05* --- musculoskeletal diseases
    - **MeSH** *C06* --- digestive system diseases
    - **MeSH** *C07* --- stomatognathic diseases
    - **MeSH** *C08* --- respiratory tract diseases
    - **MeSH** *C09* --- otorhinolaryngologic diseases
    - **MeSH** *C10* --- nervous system diseases
    - **MeSH** *C11* --- eye diseases
    - **MeSH** *C12* --- urologic and male genital diseases
    - **MeSH** *C13* --- female genital diseases and pregnancy complications
    - **MeSH** *C14* --- cardiovascular diseases

Example: Manual Indexing — MeSH (contd)

- And a number of subsubcategories (even more specific/detailed terms):



Example: Manual Indexing — MeSH (contd)

- And a number of subsubsubcategories (yet again more specific/ detailed terms):

Eye Diseases [C11]
Conjunctival Diseases [C11.187]
Conjunctival Neoplasms [C11.187.169]
Conjunctivitis [C11.187.183]
Conjunctivitis, Allergic [C11.187.183.200]
Conjunctivitis, Bacterial [C11.187.183.220] +
Conjunctivitis, Viral [C11.187.183.240] +
Keratoconjunctivitis [C11.187.183.394] +
Reiter Syndrome [C11.187.183.749]
Pterygium [C11.187.781]
Xerophthalmia [C11.187.810]

Example: Manual Indexing — MeSH (contd)

- MEDLINE — Medical Literature Analysis and Retrieval System Online
  - international database of literature for medicine and the life sciences
  - includes papers from ≈5600 different sources (mostly journals), in various languages
  - database now holds records for ≈26 million papers
- Each MEDLINE article indexed with 10-15 descriptors from MeSH
  - papers accessed by PubMed search engine interface, using MeSH terms (and other terms, e.g. author name, etc)
  - by default, all descriptors below a given one in the hierarchy are also included in search

Manual Indexing

- Advantages:
  - ◇ High precision searches
  - ◇ Works well for closed collections (books in a library)
- Problems:
  - ◇ Searchers need to know terms to achieve high precision
  - ◇ Labellers need to be trained to achieve consistency
    - Not feasible to expect this from all content creators on the web
  - ◇ Collections are dynamic $\rightarrow$ schemes change constantly

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 ...

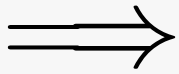
Document	*Text*
1	Pease porridge **hot**, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it **hot**, some like it cold
5	Some like it in the pot
6	Nine days old

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

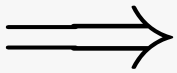


Num	*Token*	*Docs*
1	cold	1, 4
2	days	3, 6
3	hot	1, 4
4	in	2, 5
5	it	4, 5
6	like	4, 5
7	nine	3, 6
8	old	3, 6
9	pease	1, 2
10	porridge	1, 2
11	pot	2, 5
12	some	4, 5
13	the	2, 5

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

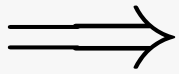


Num	*Token*	*Docs*
1	cold	1:1, 4:1
2	days	3:1, 6:1
3	hot	1:1, 4:1
4	in	2:1, 5:1
5	it	4:2, 5:1
6	like	4:2, 5:1
7	nine	3:1, 6:1
8	old	3:1, 6:1
9	pease	1:2, 2:1
10	porridge	1:2, 2:1
11	pot	2:1, 5:1
12	some	4:2, 5:1
13	the	2:1, 5:1

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.