Information Retrieval and Organisation

Dell Zhang

Birkbeck, University of London

Motivation

What is Information Retrieval?

- The meaning of the term Information Retrieval (IR) can be quite broad
  - Every time you look up information to get a task done could be considered as IR
  - E.g. getting a credit card out of a wallet to type in the card number
- From an academic point of view the following definition is more useful:
  - Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

What is Information Retrieval?

- Formulated differently: finding exactly the information you need, when you need it

Mundaneum and Google [Video]

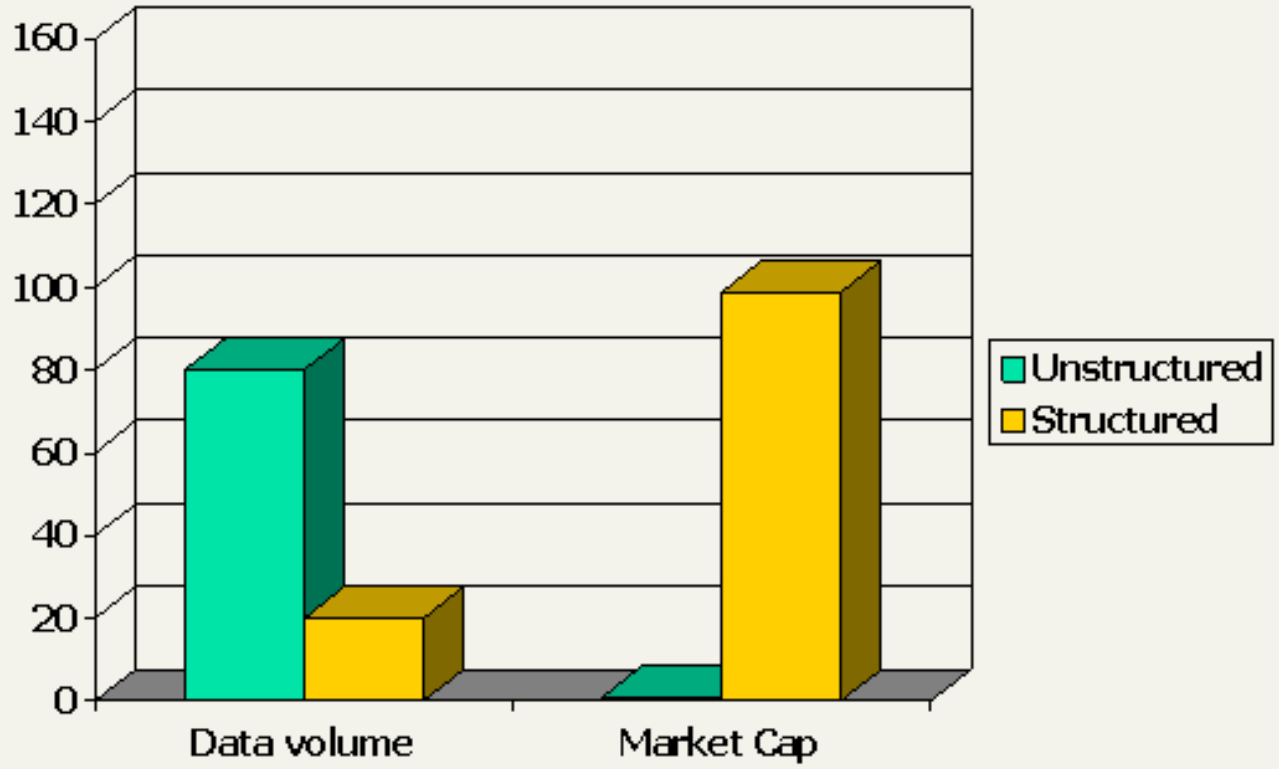
What is Information Retrieval?

- Up until a few decades ago, people preferred to get information from other people
  - e.g. booking travel via a human travel agent
- Used to be an activity that only a few people engaged in:
  - reference librarians, paralegals, and similar professional searchers

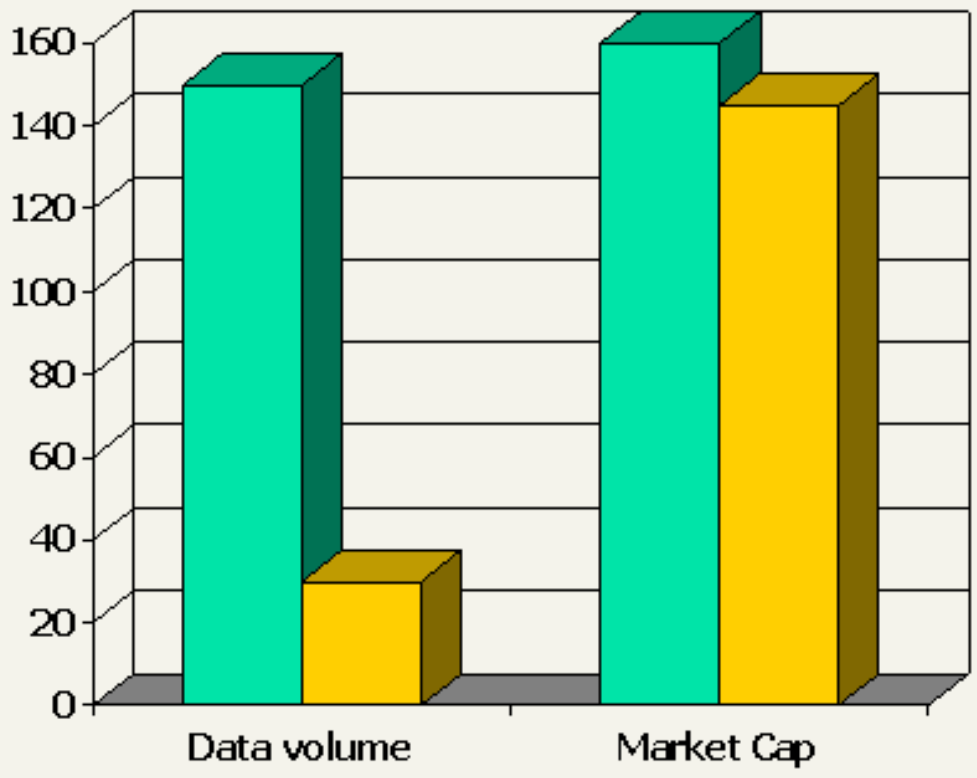
What is Information Retrieval?

- However, the world has changed
- Hundreds of millions of people engage in IR every day, e.g.
  - by using a web search engine
  - by searching their e-mail
- Role of IR has changed from a mostly academic exercise to an important research area
- IR is quickly becoming the dominant form of information access, overtaking database retrieval
- Most information does not reside in database systems

Unstructured (text) vs. structured (database) data in 1996



Unstructured (text) vs. structured (database) data in 2006



Google

YAHOO!

Ask

Data Retrieval vs. Information Retrieval

Data Retrieval	Information Retrieval
syntactical search	semantic search
highly structured data	unstructured documents, content has to be interpreted in the context of user's query
exact queries and results	selection process more vague

Brief history

- IR did not begin with the Web
- IR evolved to give principled approaches to searching various forms of content
- (Not so serious) history of IR follows

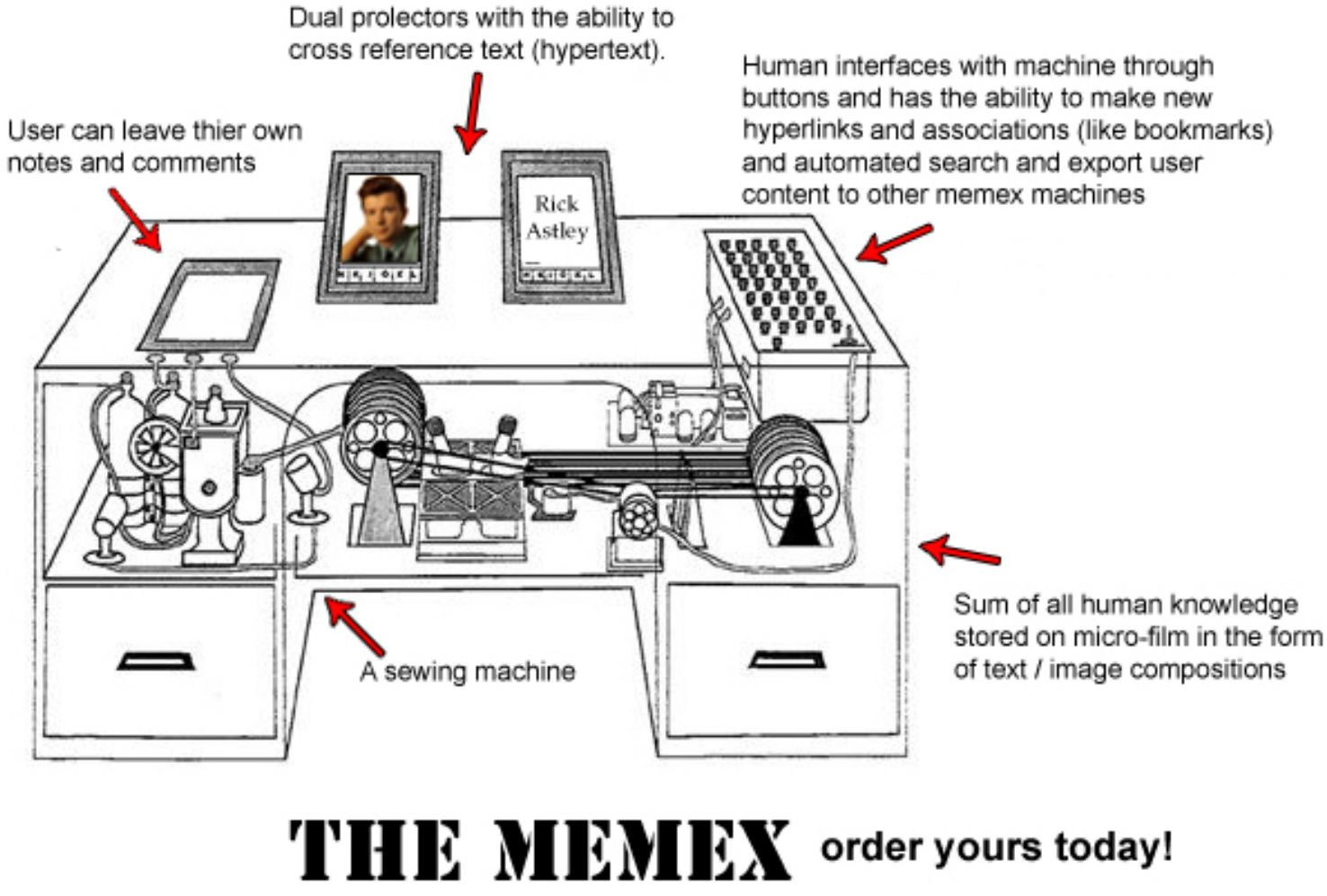
History of IR

- 2000 BC: Sumerian clay tablet lists
- 300 BC?: Library of Alexandria (allegedly contained around 400,000 scrolls)
- 1450 AD: printing press
- 16th century: first indexes in books
- 18th century: book indexes look very similar to today's
- 19th century: construction of concordances

History of IR

- Obviously, we are more interested in high-tech-based IR
- 1934 Paul Otlet: Mundaneum
- 1945: Vannevar Bush "As We May Think", Memex: idea with microfilm reels, screen viewers, and cameras
- 1950: the term IR is coined by Calvin Mooers
- 1958: IR recognized as research area, "International Conference on Scientific Information"
- 1970s: first interactive computer systems (DIALOG, MEDLINE)
- 1990s: the Web

Memex



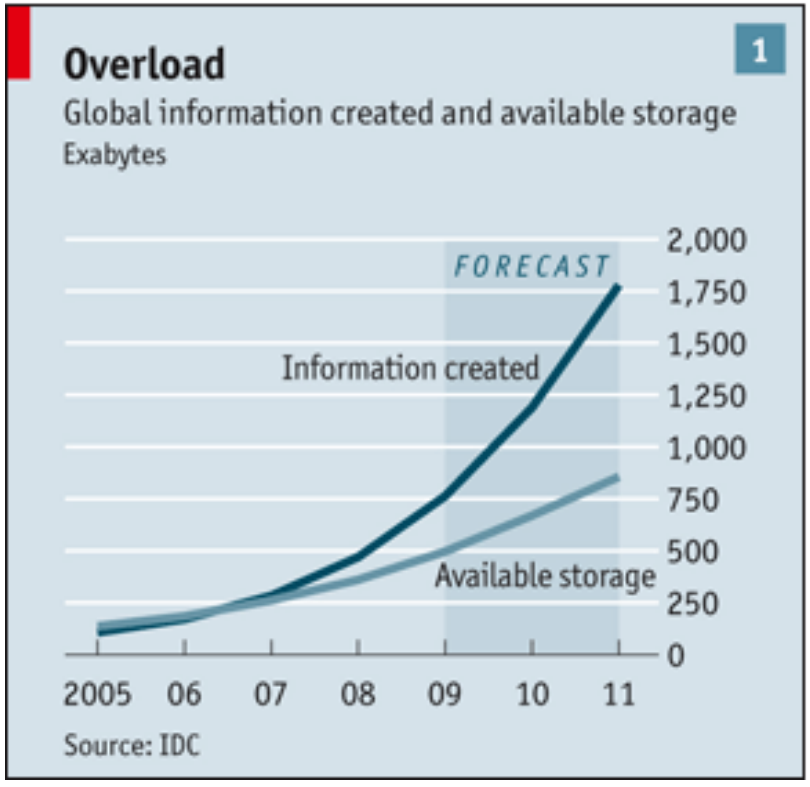
Memex [Video]

Future of IR

- Several challenges lie ahead:
  - Coping with the sheer size of the data volume: "How much information 2003?" at Berkeley to find out how much new information is created each year
    - Result: about 5 exabytes of data in 2002
    - For comparison: Library of Congress contains about 136 terabytes of information: 5 exabytes $\approx$ 37,000 libraries of that size
  - 2008 International Data Corporation (IDC) white paper:
    - Amount of digital data in 2007: 281 exabytes $\approx$ 281 trillion digitized novels

Future of IR

- It seems we're reaching a breaking point:
  - According to another IDC study, the amount of data generated now exceeds the storage capacity



("Data, data everywhere", The Economist, February 25th, 2010)

Future of IR

- Further challenges:
  - Improving semantic search capabilities (e.g. Semantic Web)
  - Categorizing documents (e.g. spam filters)
  - Multi-media retrieval

Overview

- In this module we will talk about:
  - Basics of IR
    - Simple Boolean queries
    - Preprocessing and indexing data
    - More sophisticated retrieval models (ranking)
    - Evaluation of IR systems
  - More advanced topics:
    - Relevance feedback
    - Probabilistic retrieval
    - Classification and clustering

Further Reading

- This module is based on the following textbook:
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. ISBN-10: 0521865719, ISBN-13: 978-0521865715