ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines

Wolf-Tilo Balke
Muhammad Usman

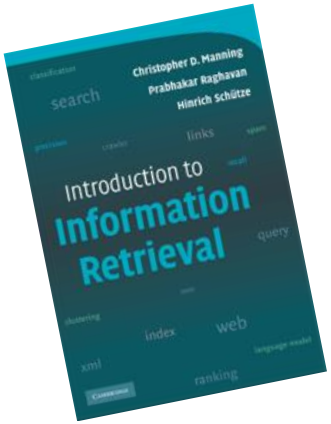
Institut für Informationssysteme
Technische Universität Braunschweig

Organizational Issues

- Course overview
  - 13 lectures (There will be no lectures on 18.05. (Christi Himmelfahrt) and 01.06. (excursion week))
  - Exercises and detours are integrated into the lecture
  - Day and Time: Thursdays, 15:00–17:15
  - Final exam: Oral or Written

- Lecture Slides and Recording will be available at:
  - http://www.ifis.cs.tu-bs.de/teaching/ss-2023/irws

What is Information Retrieval (IR)?



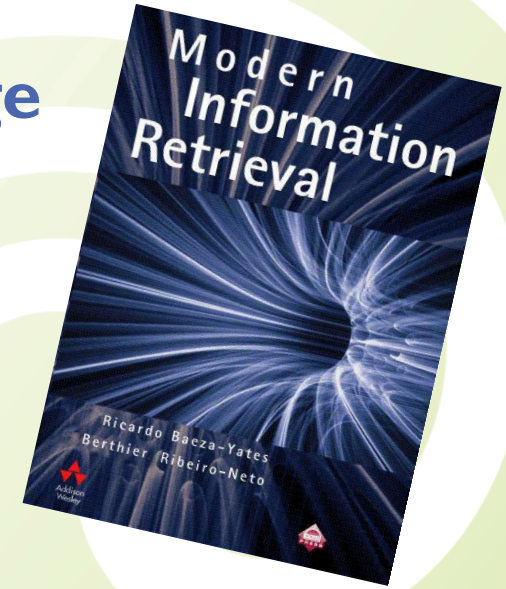
IR is **finding** material of an **unstructured** nature that satisfies an **information need** from within **large** collections (usually text document)

IR is the science of **searching** for **documents,** for **information within documents,** and for **metadata** about documents, as well as that of searching **relational databases** and the **WWW.**

What is Information Retrieval (IR)?

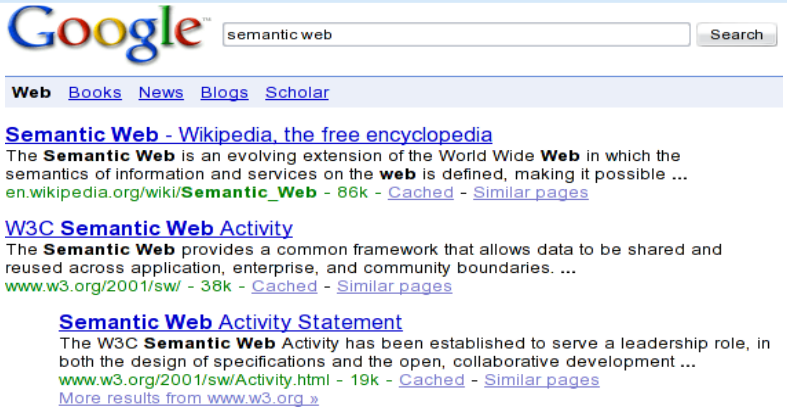
IR: Part of computer science which studies the **retrieval of information (not data)** from a collection of **written documents.** The retrieved documents aim at satisfying a **user information need** usually expressed in **natural language.**

- **Documents, unstructured, text, large**
- **Information need**
- **Store, search, and find**
- **The World Wide Web?**
- **Relational databases?**



Information Retrieval vs. Databases

Information retrieval	Data retrieval
Retrieve all objects **relevant** to some **information need**	Retrieve all objects satisfying some **clearly defined conditions**
Find all documents about the **topic** "semantic web"!	**SELECT** id **FROM** document **WHERE** title **LIKE** '%semantic web%'
Result **list**	**Well-defined** result **set**



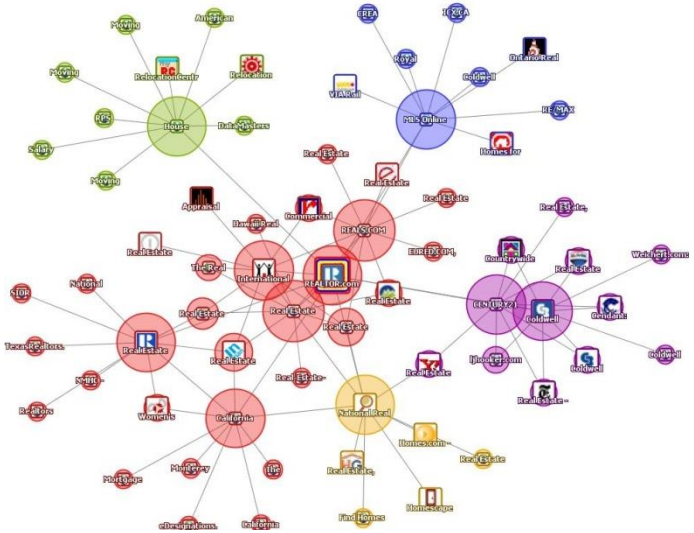

```
[selke@tbdb ~]$ db2 "SELECT id FROM document WHERE title L
IKE '%semantic web%' FETCH FIRST 3 ROWS ONLY"

ID
--------------------
               45489
             9635899
               98556

  3 record(s) selected.
```

Web Search

- Very similar to information retrieval
- Main differences:
  - **Links** between Web pages can be exploited
  - **Collecting,** storing, and **updating** documents is more difficult
  - Usually, the **number of users** is very large
  - **Spam** is a problem



Why Should I Know about All This?

Gartner

- "80% of business is conducted on **unstructured** information"
- "85% of all data stored is held in an **unstructured** format"
- "7 million **Web pages** are being added every day"

Butler Group
a Datamonitor Company

- "**Unstructured** data doubles every three months"

Why Should I Know About All This?

- Managing the **information flood**
- Have you ever tried to **drink from a fire hydrant?**

Why Should I Know about All This?

Google

YAHOO!

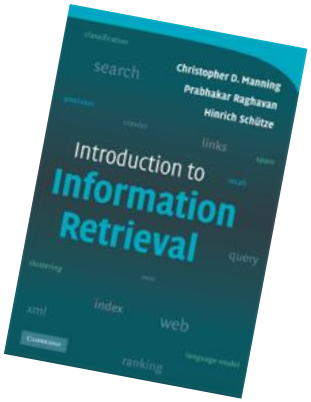
fast
A Microsoft Subsidiary

bing

Ask.com

Literature

Introduction to Information Retrieval



- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze.
- Cambridge University Press, ISBN: 0521865719

Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW

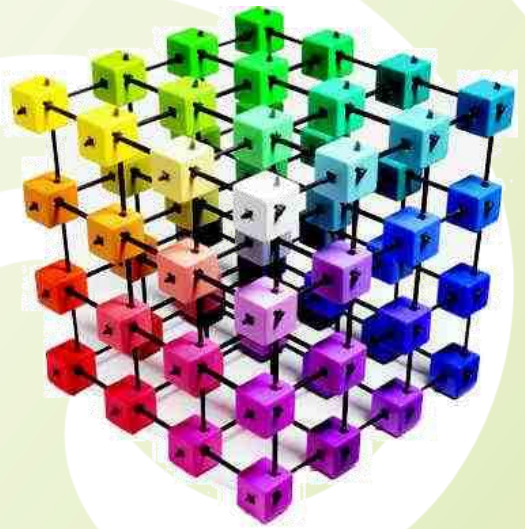
- Richard K. Belew.
- Cambridge University Press, ISBN: 0521630282

Course Overview

1. **Introduction and fundamental notions**
2. Retrieval models: fuzzy, coordination level matching, vector space
3. Probabilistic retrieval models
4. Indexing
5. Latent Semantic Indexing
6. Language models, retrieval evaluation
7. Document clustering
8. Relevance feedback, classification
9. Support vector machines
10. Introduction to Web retrieval
11. Web crawling
12. Link analysis
13. Miscellaneous

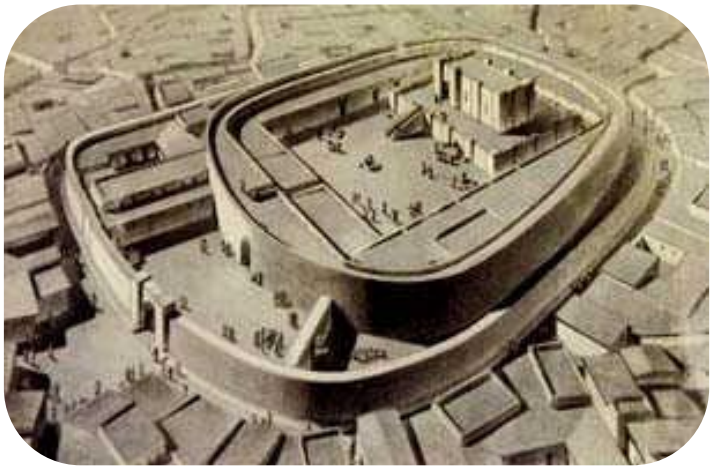
Today's Lecture

1. A Brief History
   - Libraries
   - Information Retrieval
   - Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model



Ancient Libraries

- Sumerian archives
  - Around 3000–2000 BC
  - About **25,000 clay tablets** stored in temple rooms
  - Mostly inventories and records of commercial transactions

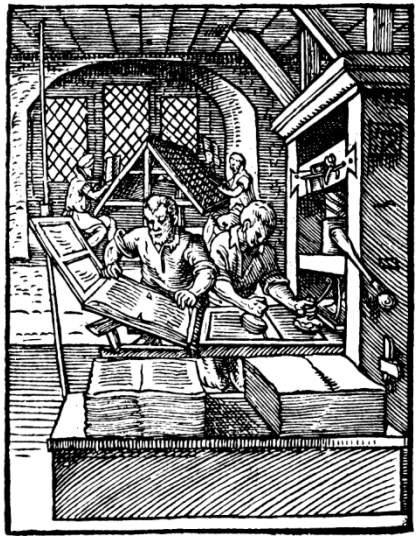


- The Great Library of Alexandria
  - Founded about 300 BC
  - Idea: A **universal library** holding copies of all the world's books
  - At its height, the library held nearly **750,000 scrolls**

Medieval Libraries

- Monastic libraries
  - Educated monks saved many ancient texts from getting lost by **hand-copying**
  - The Vatican Library was formally founded in 1475 but is in fact much older

- Gutenberg's Movable type
  - Around 1450, Johannes Gutenberg introduced **movable type** to Europe for printing
  - The technique spread rapidly, copying books became much easier and less expensive



Modern Libraries

- German National Library
  - **25 million** items
  - Located in Leipzig, Frankfurt (Main), and Berlin
- Library of Congress
  - **150 million** items
  - 20 million new items since 2009
  - The world's **largest library** (according to the Guinness Book)
  - **Classification** system: Library of Congress Classification

Library Catalogs

Items are cataloged by **metadata:**

- **Author/Editor, ISBN,** …
- **Keyword,** e.g. "information retrieval"
- **Subject area,** e.g. "information systems"
- **Specialized classification systems,** e.g. Library of Congress

Titel:	Introduction to information retrieval / Christopher D. Manning; Prabhakar Raghavan; Hinrich Schütze
Verfasser:	Manning, Christopher D. ; Raghavan, Prabhakar ; Schütze, Hinrich
Erschienen:	Cambridge [u.a.] : Cambridge Univ. Press, 2008
Umfang:	XXI, 482 S. : Ill., graph. Darst.
ISBN:	0-521-86571-9, 978-0-521-86571-5*hbk : £ 32.99
Schlagwörter:	*Dokumentverarbeitung / Information Retrieval / Abfrageverarbeitung Text processing (Computer science) Information retrieval Document clustering Semantic Web
Sachgebiete:	54.64 ; Datenbanken 06.74 ; Informationssysteme
Mehr zum Thema:	Klassifikation der Library of Congress: QA76.9.T48 Dewey Dezimal-Klassifikation: 025.04
Link:	http://www.loc.gov/catdir/enhancements/fy0810/2008001257-b.html [Contributor biographical information] http://www.loc.gov/catdir/enhancements/fy0810/2008001257-d.html [Publisher description] http://www.loc.gov/catdir/enhancements/fy0810/2008001257-t.html [Table of contents only] http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html [Companion web site]
Standort:	L all 702
Signatur:	**U 08 B 1668**
Ausleihstatus:	verleihbar Noch nicht verfuegbar

Metadata in the digital world

Detour

- Are small compared to the resource they're describing.
- Traditionally used in libraries (Card Catalogues).
- Used now to describe digital data, due to the increasing conversion of information into digital formats.
- Conforms to some metadata standards as specified per a particular discipline.
- Most search engines use it, when adding pages to their search index.

MEDLINE

Detour

- A life science and biomedical information database containing over 35 million references to journal articles.
- Around 2,000-4,000 references are added each day.
- Accessible online through the PubMed interface.
- Manually indexed by Medical Subject Headings (MeSH) for information retrieval.

- http://www.ncbi.nlm.nih.gov/pubmed/

PubMed

MeSH

Detour

- Controlled vocabulary used for indexing.
- Has a total of 30,000 subject headings (AKA descriptors)
- It can be viewed as a thesaurus, and they are arranged within a hierarchy.
- 10 – 15 subject headings are used to index every entry in MEDLINE.
- Efficiently searching MEDLINE requires familiarity with the MeSH database.

• http://www.nlm.nih.gov/mesh/MBrowser.html

Dublin Core Metadata

Detour

- It's an open, non-profit organization that supports shared innovation in metadata design
- They define a small set of metadata elements for describing information resources
- Dublin Core Metadata Element Set:
  - Used to describe resources
  - Includes 2 levels: Simple (15 elements) and qualified Dublin Core (18 elements)
    e.g. abstract, creator, title, publisher, language, rightsHolder, etc.
    (List: http://dublincore.org/documents/dces/)
  - Endorsed as an ISO standard 15836:2009

Dublin Core® Metadata Initiative
Making it easier to find information.

Encoding DC in HTML

Detour

- In HTML 4.0, META and LINKS tag can be utilized
- META Tag encodes a named metadata element

  E.g.

  <meta name = "prefix.elementName" content = "elementvalue" >

  <meta name = "DC.Title" content = "Information retrieval and web search engines"
  name = "DC.Language" content = "English">

- Link Tag the prefix of the element name to its element set definition

  E.g.

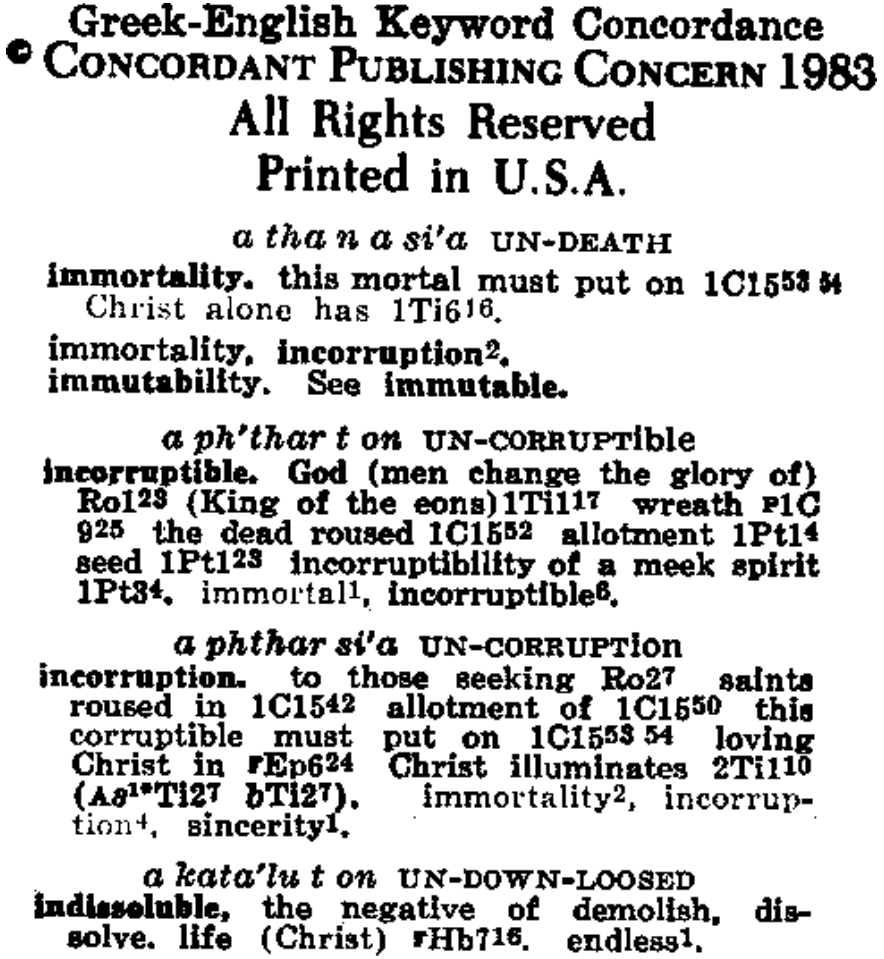
  <link rel = "schema.DC" href = http://purl.org/DC/elements/1.0/>

Full Text Search?

- Catalogue cards are **document proxies**
- Often, they suffice to judge the relevance of a particular item for your information need
- But:
  - A **clever classification scheme** is required:
    - Extensive enough to allow detailed classification
    - Simple enough to be easily understandable
  - **Experts** must catalogue each item individually
- **Problem:** A lot of manual work!
- **Full text search: Every word is a keyword!**

Full Text Search?

- Pre-computer area: **Concordances**
  - Alphabetical list of the **principal words** used in a book
  - Only for works of **special importance,** such as the Bible
  - First Bible concordance by Hugo de Saint Charo, with the help of 500 monks, around 1250



Greek-English Keyword Concordance
© CONCORDANT PUBLISHING CONCERN 1983
All Rights Reserved
Printed in U.S.A.

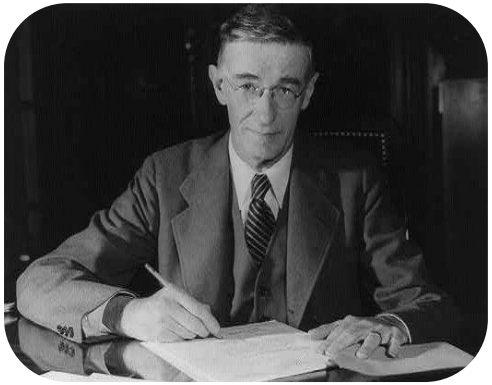
a tha n a si'a UN-DEATH
immortality. this mortal must put on 1C15[53 54] Christ alone has 1Ti6[16].
immortality, incorruption[2], immutability. See **immutable.**

a ph'thar t on UN-CORRUPTible
incorruptible. God (men change the glory of) Ro1[23] (King of the eons) 1Ti1[17] wreath P1C 9[25] the dead roused 1C15[52] allotment 1Pt1[4] seed 1Pt1[23] incorruptibility of a meek spirit 1Pt3[4], immortal[1], **incorruptible[6].**

a phthar si'a UN-CORRUPTion
incorruption. to those seeking Ro2[7] saints roused in 1C15[42] allotment of 1C15[50] this corruptible must put on 1C15[53 54] loving Christ in FEp6[24] Christ illuminates 2Ti1[10] (*As*[1*]Ti2[7] *b*Ti2[7]). immortality[2], incorruption[4], **sincerity[1].**

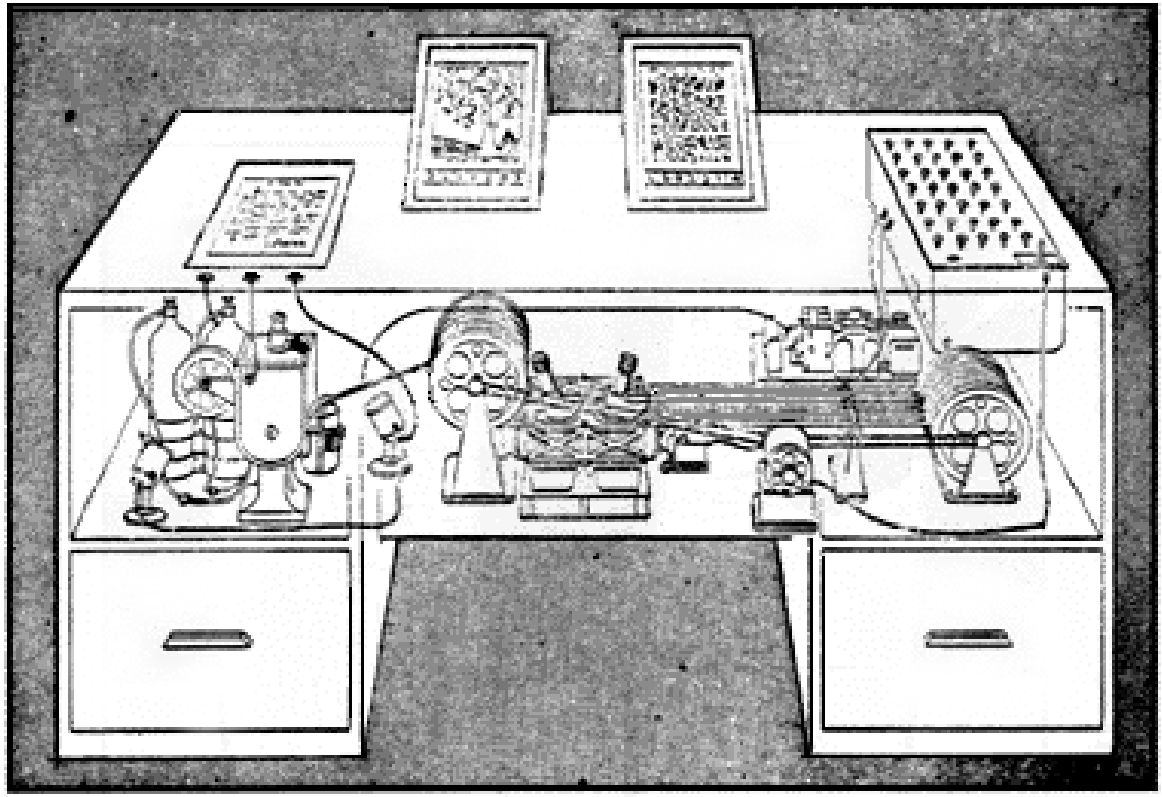
a kata'lu t on UN-DOWN-LOOSED
indissoluble, the negative of demolish, dissolve. life (Christ) FHb7[16]. endless[1].

The Memex



- Vision of a **hypertext-based PDA**
- Proposed by **Vannevar Bush**
  - Director of the **Office of Scientific Research and Development** (USA, 1941–1947)
- Outlined in Bush's famous essay **"As We May Think"** published in The Atlantic Monthly (1945)
- "A device in which an individual stores **all his books, records, and communications,** and which is **mechanized** so that it may be consulted with **exceeding speed and flexibility.**"
- "Selection by **association,** rather than by indexing."

The Memex



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Early Information Retrieval Systems

- 1957: **Hans-Peter Luhn** (IBM) uses **words as indexing units** for documents
  - Measure **similarity** between documents by **word overlap**

- 1960s and 1970s: **Gerard Salton** and his students (Harvard, Cornell) create the **SMART system**
  - Vector space model
  - Relevance feedback

IR Becomes a Research Discipline

- ACM's **SIGIR**
  - Special Interest Group on Information Retrieval
  - Annual **conferences,** beginning in 1978
  - **Gerald Salton award,** first honoree: Gerald Salton (1983)
- **TREC**
  - Annual Text Retrieval Conference, beginning in 1992
  - Sponsored by the **U.S. National Institute of Standards and Technology** as well as the **U.S. Department of Defense**
  - Today: many different **tracks,** e.g., blogs, genomics, spam
  - Provides **data sets** and **test problems**

A Brief History of Web Search

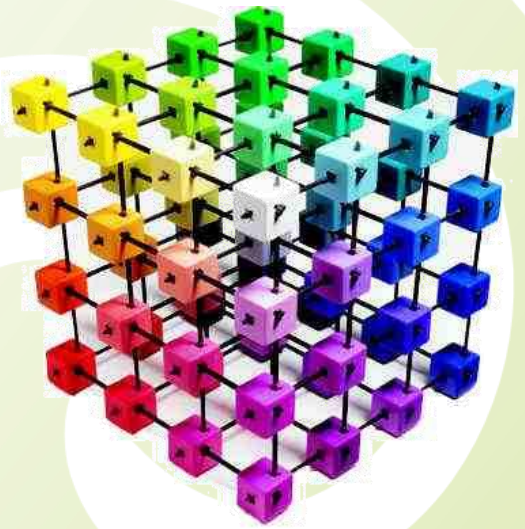
- 1991: **Tim Berners-Lee** "invents" the World Wide Web
- First Web search engines:
  - **Archie:** Query **file names** by regular expressions
  - **Architext/Excite:** Full text search, simple ranking (1993)
- Until 1998, web search meant information retrieval
- 1998: **Google** was founded
  - Exploits **link structure** using the **PageRank** algorithm

Core Problems

- How to store and update **large** document collections?
  - **Small!**
  - **Scalable!**
- How to do **efficient** retrieval?
  - **Fast!**
- How to do **effective** retrieval?
  - **High result quality!**

Today's Lecture

1. A Brief History
   - Libraries
   - Information Retrieval
   - Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model



Document

- A **document** is a **coherent passage of free text**
- "Coherent" means: **is about related topics**
- "Free" means: **natural, written language**
- Examples:
  - Newspaper article
  - Scientific article
  - Dictionary entry
  - Web page
  - Email message

Document Collection

- A **document collection** is a **set of documents**
- Also known as **corpus**
- Usually, all documents within a collection are **similar** with respect to some criterion
- Examples:
  - MEDLINE
  - The articles covered by Google News
  - The Web

Information Need

- An **information need** is the **topic** about which the user desires to know more
- Refers to an individual, hidden **cognitive state**
- Depends on what the users knows **and** doesn't know
- **Ill-defined**
- Examples:
  - What is the capital of Uganda?
  - Is it really true that McDonald's hamburgers contain worm meat?
  - What is "cloud computing"?

Query

- A **query** is what the user conveys to the computer in an attempt to **communicate the information need**
- Stated using a **formal query language**
  - Usually a list of search terms
  - But also: "Panda NEAR Jaguar BUT NOT animal"

Relevance

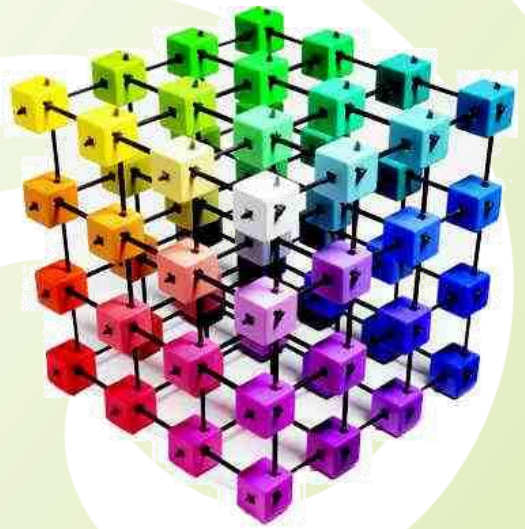
- A document is **relevant**
  with respect to some user's **information need**

  if

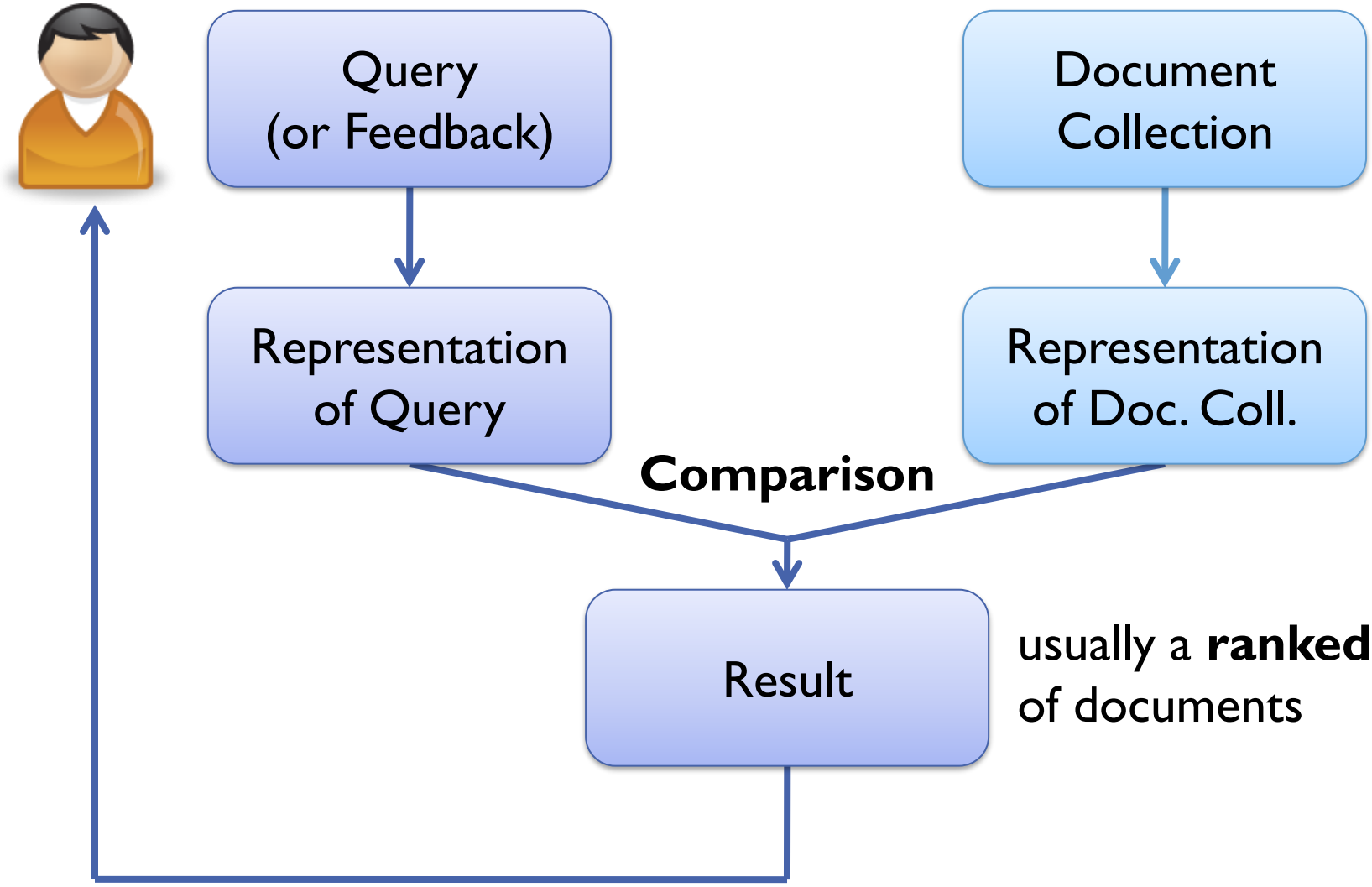
  the user **perceives** it as containing **information of value** with respect to this **information need**
- Usually assumed to be a **binary concept,**
  but could also be graded
- Example:
  - Information need:
    "What is relevance in IR?"
    Relevant document:
    Wikipedia's entry "Relevance (information retrieval)"

Today's Lecture

1. A Brief History
   - Libraries
   - Information Retrieval
   - Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model

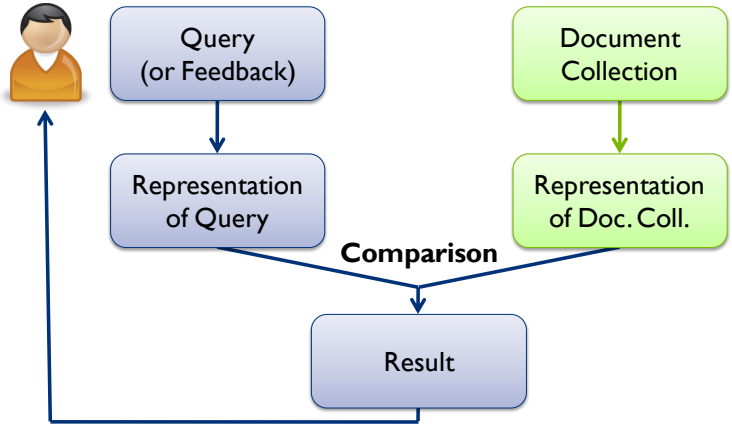


Schematic Diagram of an IR System



IR Models

- Any IR system is based on an **IR model**
- The model defines …
  - a **query language,**
  - an internal **representation of queries,**
  - an internal **representation of documents,**
  - a **ranking function** that associates a real number with each query–document pair.
- Optional: A mechanism for **relevant feedback**



The Bag of Words Representation

- A very popular **representation of documents** is the **bag of words model**
- Each document is represented by a bag (= multiset) of **terms** from a predefined **vocabulary**
- Standard case:
  - Vocabulary
    = set of all the words occurring in the collection's documents
  - Each document is represented by the words it contains

That's one small step for a man, a giant leap for mankind

→ { that's, one, small, step, for (2), a (2), man, giant, leap, mankind }

The Bag of Words Model

- Cons:
  - Word order gets lost
  - Very different documents could have similar representations
  - Document structure (e.g. headings) and metadata is ignored

- Pros:
  - Simple set-theoretic representation of documents
  - Efficient storage and retrieval of individual terms
  - IR models using the bag of words representation work well!

The Bag of Words Model (3)

- Any document can be represented by an **incidence vector:**

vocabulary (aka index terms) →

that's	one	small	step	for	a	man	giant	leap	mankind	taikonaut	Zhai's	is	China

That's one small step for a man, a giant leap for mankind

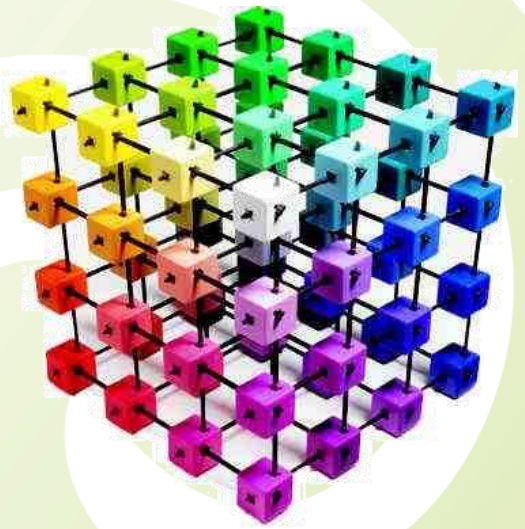
Taikonaut Zhai's small step is a giant leap for China

1	1	1	1	2	2	1	1	1	1	0	0	0	0
0	0	1	1	0	1	0	1	1	0	1	1	1	1

incidence matrix (aka term-document matrix)

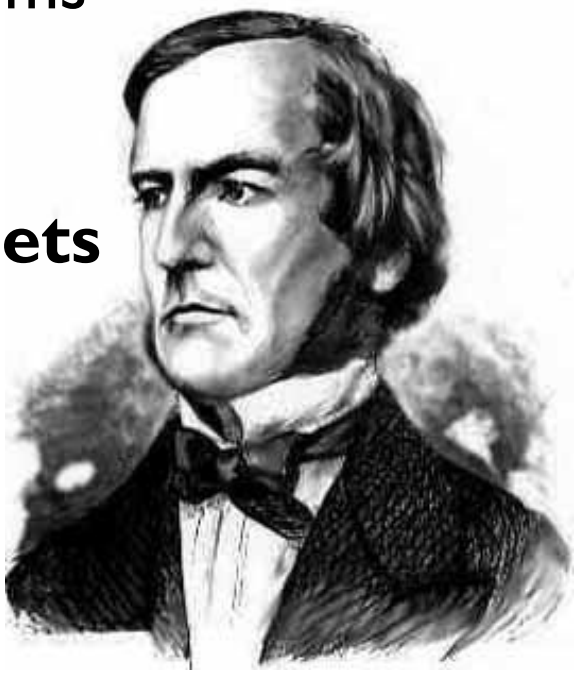
Today's Lecture

1. A Brief History
   - Libraries
   - Information Retrieval
   - Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model



Boolean Retrieval

- The simplest (and arguably oldest) IR model
- Documents = **sets** of words (index terms)
- Query language
  = **Boolean expressions** over index terms
- Binary ranking function, i.e. 0/1-valued
- Retrieval is based on **membership in sets**
  - "Find all documents indexed by the word 'mankind'!"
  - "Find all documents indexed by the word 'man' or 'mankind'!"



Boolean Connectives

- **Boolean connectives:**
  - Conjunction
  - Disjunction
  - Negation

∧	0	1
0	0	0
1	0	1

∨	0	1
0	0	1
1	1	1

¬	
0	1
1	0

Example

- $Document_1$ = {step, mankind}
- $Document_2$ = {step, China}

- $Query_1$ = "step AND mankind"
  - Result set: {$Document_1$}

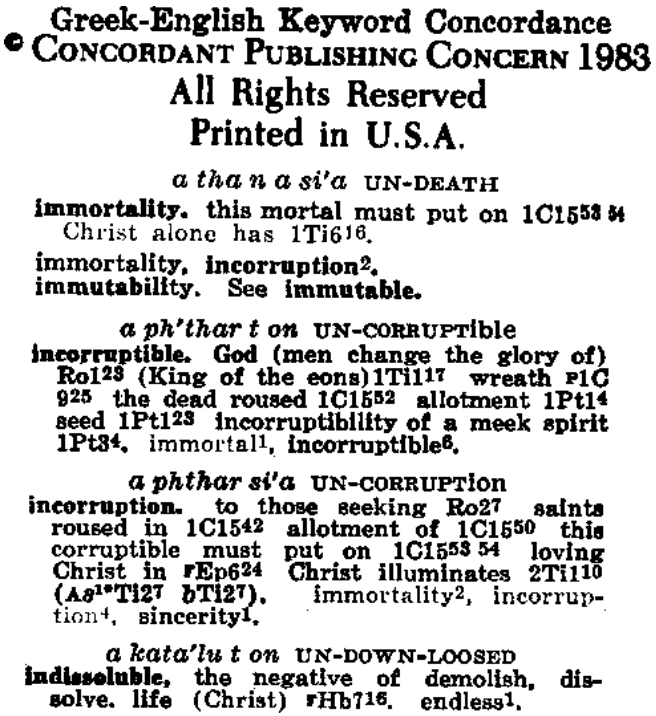
- $Query_2$ = "step OR mankind"
  - Result set: {$Document_1$, $Document_2$}

Boolean Queries in Practice

- **Warning:**
  Exclusive use of negation will result in large result sets!
  - $Query_3$ = "NOT mankind"
- To match natural language better,
  "BUT NOT" can be used instead of "AND NOT"
  - $Query_4$ = "step BUT NOT China"
- Use **"OF"** to search for subsets of a given size:
  - $Query_5$ = "2 of {step, mankind, China}"
  - $Query_5$ ≡ "(step AND mankind)
    OR (step AND China)
    OR (mankind AND China)"

Query Processing

- Usually, documents are indexed by an **inverted index**
  - For each index term, the set of documents containing this term is **pre-computed** and stored on disk
  - This enables **fast query processing**
- Document collection:
  - $Document_1$ = {step, mankind}
  - $Document_2$ = {step, China}
- Inverted index:
  - step: {$Document_1$, $Document_2$}
  - mankind: {$Document_1$}
  - China: {$Document_2$}



Greek-English Keyword Concordance
© CONCORDANT PUBLISHING CONCERN 1983
All Rights Reserved
Printed in U.S.A.

a tha n a si'a UN-DEATH
immortality. this mortal must put on 1C15$^{53\ 54}$ Christ alone has 1Ti6^{16}.
immortality, incorruption^{2}, immutability. See immutable.

a ph'thar t on UN-CORRUPTible
incorruptible. God (men change the glory of) Ro1^{23} (King of the eons) 1Ti1^{17} wreath F1C 9^{25} the dead roused 1C15^{52} allotment 1Pt1^{4} seed 1Pt1^{23} incorruptibility of a meek spirit 1Pt3^{4}. immortal^{1}, incorruptible^{6}.

a phthar si'a UN-CORRUPTion
incorruption. to those seeking Ro2^{7} saints roused in 1C15^{42} allotment of 1C15^{50} this corruptible must put on 1C15$^{53\ 54}$ loving Christ in FEp6^{24} Christ illuminates 2Ti1^{10} (As^{13}Ti^{27} bTi^{27}). immortality^{2}, incorruption^{4}, sincerity^{1}.

a kata'lu t on UN-DOWN-LOOSED
indissoluble, the negative of demolish, dissolve. life (Christ) FHb7^{16}. endless^{1}.

Query Processing

- Thanks to the inverted index, queries of the type
  "Show me all documents containing term *X*"
  can be answered quickly
- Also quick to compute: unions and intersections of sets
- Example:
  - result of "mankind AND step"
    = (result of "mankind") ∩ (result of "step")
  - result of "mankind OR step"
    = (result of "mankind") ∪ (result of "step")
- **Idea:** Convert all queries to **conjunctive normal form** or **disjunctive normal form**

Query Processing

- Conjunctive normal form (CNF)
  - A propositional formula is in **CNF** if it is a **conjunction of clauses**
    - A clause is a disjunction of literal
    - A literal is a variable or its negation
  - **Theorem:** Any propositional formula can be converted into an equivalent formula that is in CNF
- Disjunctive normal form (DNF)
  - A propositional formula is in DNF if it is a **disjunction of conjunctive clauses**
    - A conjunctive clause is a conjunction of literal
  - **Theorem:** Any propositional formula can be converted into an equivalent formula that is in DNF

Query Processing

- $Query_6$ = "step AND ((China AND taikonaut) OR man)"
- **Conjunctive normal form (CNF):**
  $Query_6 \equiv$
  "step AND (China OR man) AND (taikonaut OR man)"
- **Disjunctive normal form (DNF):**
  $Query_6 \equiv$
  "(step AND China AND taikonaut) OR (step AND man)"

Query Processing

- **Conjunctive normal form:**
  "step AND (China OR man) AND (taikonaut OR man)"
  1. Compute unions **(might become very large)**
  2. Compute intersections

- **Disjunctive normal form:**
  "(step AND China AND taikonaut) OR (step AND man)"
  1. Compute intersections **(smaller intermediate results)**
  2. Compute unions

Pros

- Simple query paradigm, **easy to understand**
- If all document representations are mutually distinct, **any possible subset of documents can be retrieved** by a suitable query

  ⇒ cut out the set of relevant documents
- But: This advantage is rather **theoretical,** since the "right" query usually is unknown

Cons

- A binary ranking function returns a **set of results,** i.e. it is **unordered**
- Controlling the **result size** is difficult
- **Similarity queries** are not supported
- Usually, most of the documents found are relevant; but **many relevant documents are not found**

Westlaw

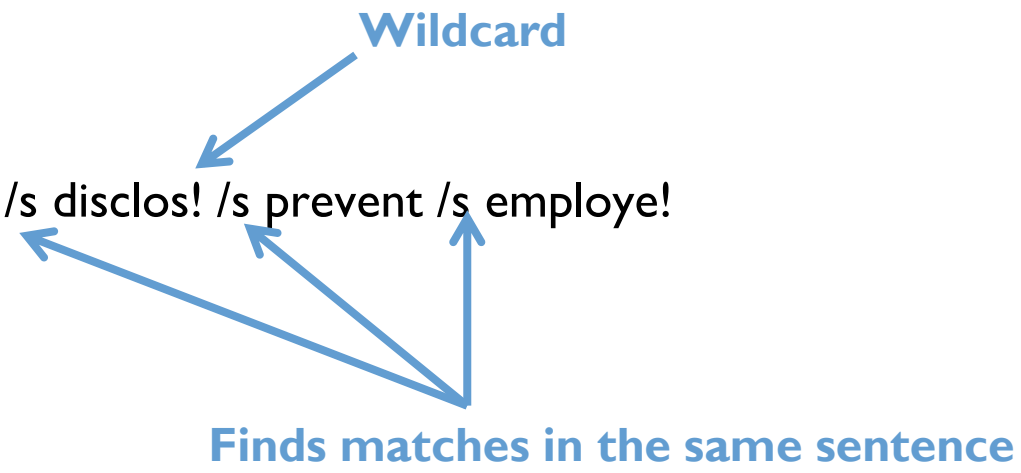
Detour

- Westlaw
  - Online **legal research service** for US law
  - Includes more than **40,000 databases** of case law, state and federal statutes, administrative codes, law journals, newspapers …
  - Offers search by:
    - "Terms and Connectors" – Boolean Search
    - "Natural Language" – Free text querying
  - Boolean search includes the Boolean operators plus some proximity operators
    - space = OR
    - /s, /p, /k = matches in the same sentence, paragraph or within k-words respectively
    - & = AND
    - ! = a trailing wildcard query

Westlaw

Detour

- Example 1:
  - **Information need:**
    Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company
  - **Query:**
    "trade secret" /s disclos! /s prevent /s employe!



Examples taken from
Mannig/Raghavan/Schütze: Introduction to Information Retrieval

Westlaw

Detour

- Example 2:
  - **Information need:**
    Requirements for disabled people to be able to access a workplace
  - **Query:**

Finds matches in the same paragraph

disab! /p access! /s (work-site work-place) (employment /3 place)

Space means disjunction

Finds matches within 3 words

Westlaw

Detour

- Until 2005, Boolean search was the default in Westlaw
- Submitted queries average **about ten words in length**
- Professionals often prefer Boolean search to other methods as they offer **greater control and transparency**
- But: later on, experiments on a Westlaw subcollection found that **free text queries produced better results** for queries prepared by Westlaw's own librarians

Next Lecture

- More retrieval models
  - Fuzzy retrieval model
  - Coordination level matching
  - Vector space model