Introduction to Information Retrieval

Ulf Leser
Content of this Lecture

• What is Information Retrieval
• Documents
• Queries
• Related topics
Information Retrieval (aka “Search”)

• Naïve: Find all documents containing the following words

• Advanced: “Leading the user to those documents that will best enable him/her to satisfy his/her need for information”[Robertson 1981]
  – A user wants to know something
  – The user needs to tell the machine what he wants to know: query
  – Posing exact queries is difficult: room for interpretation
  – Machine interprets query to compute the (hopefully) best answer
  – Goodness of answer (relevance) depends on original intention of user, not on the query
  – “Leading”: Sensible ranking of all potentially relevant docs
Difference to Database Queries

• Queries: Formal language versus natural language
• Exactly defined result versus loosely described relevance
• Result set versus ranked result list
• DB: Posing the right query is completely left to the user
• IR: Understanding the query is a problem of the software
The Problem

- Help user in **quickly** finding the **requested information** from a **given set of documents**
  - Documents: Corpus, library, collection, …
  - Quickly: Few queries, fast responses, simple interfaces, …
  - Requested: The “best-fitting” documents; the “right” passages; the most “relevant” content

![Diagram](image)
Why is it hard?

- **Homonyms** (context)
  - Fenster (Glas, Computer, Brief, …), Berlin (BRD, USA, …), Boden (Dach, Fussboden, Ende von etwas, …), …

- **Synonyms**
  - Computer, PC, Rechner, Server, …

- **Specific queries** (subfield Question Answering)
  - What was the score of Bayern München versus Stuttgart in the DFB Pokal finals in 1998? Who scored the first goal for Stuttgart?
  - How many hours of sunshine on average has a day in Crete in May?

- Typical web queries have 1-6 terms

- “Information broker” is a profession
Quickly

- **Time to execute a query**
  - Indexing, parallelization, compression, …

- **Time to answer the request** (may involve multiple queries)
  - Understand request, find best matches
  - Success of search engines: Better results (and fast!)
  - **Process-orientation:** User feedback, query history, …

- **Information overload**
  - “We are drowning in data, but starving for knowledge”
  - If the corpus is large, **ranking is a must**
  - Result summarization (grouping on what?)
  - Different **search modes:** What’s new? What’s certain?
IR: An Iterative, Multi-Stage Process

- IR process: “Moving through many actions towards a general goal of satisfactory completion of research related to an information need.”
  - “Berry-picking” [Bates 89]
Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden betrachtet. ... Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen .... auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden können ... Die Unsicherheit resultiert meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten);... Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in Bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte.
Prominent Systems I: Digital Libraries

• E.g. OPAC
  – Combination of structured attributes and IR-style queries

Obviously not perfect
Prominent Systems II: Desktop Search

- Much activity in the last years
- Various search engines and indexing mechanisms
- Important: Search different types of files (txt, doc, mail, ppt, pdf, tex, odp, xls, …)
Prominent Systems III: Web Search Engines
Properties of Information Retrieval (IR)

- IR is about helping a user
- IR is about finding information, not about finding data [BYRN99]
- IR builds systems for end users, not for programmers
  - No SQL
  - IR (web) is used by almost everybody, databases are not
- IR searches unstructured data (e.g. text)
  - But: Keyword search in relational databases
- 90% of all information is presented in unstructured form
  - Claim some analysts
History

• ~300 ad. Library of Alexandria, ~700,000 „documents“
• 1450: Bookprint
• 19th century: Indices / concordance
• Probabilistic models: Maron & Kuhns (1960)
• Boolean queries: Lockheed (~1960)
• Vector Space Model: Salton, Cornell (1965)
  – Faster, simpler to implement, better search results
• 80s-90s: Digital libraries, SGML, metadata standards
• Mid 90s: The web, web search engines, XML
• End 90s: Personalized search engines, recommendations
• 2010: Mobile and context-based search, social networks
Content of this Lecture

• What is Information Retrieval
• Documents
• Queries
• Related topics
Searching only metadata

Searching tokens within documents

Interpreting natural text
Documents

• This lecture: Natural language text
• Might be grammatically correct (books, newspapers) or not (Blogs, Twitter, spoken language)
• May have structure (title, abstract, chapters, ...) or not
• May have associated (explicit or in-text) metadata or not
  – Foreign characters
• May refer to other documents (hyperlinks)
• May have various formats (ASCII, PDF, DOC, XML, ...)
IR Queries

• Users formulate queries
  - Keywords or phrases
  - Logical operations (AND, OR, NOT, ...)
    • Web search: “-ulf +leser”
  - Natural language questions (e.g. MS-Word help)
  - Structured queries (author=… AND title~ …)
  - Voice (Siri)

• Documents as queries: Find documents similar to this one

• Query refinement based on previous results
  - Find documents matching the new query within the result set of the previous search
  - Use relevant answers from previous queries to create next query
# Searching with Metadata (PubMed/Medline)

## Search Field Descriptions and Tags

<table>
<thead>
<tr>
<th>Affiliation [AD]</th>
<th>Issue [IP]</th>
<th>Place of Publication [FL]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Identifier [AID]</td>
<td>Journal Title [TA]</td>
<td>Publication Date [DP]</td>
</tr>
<tr>
<td>All Fields [ALL]</td>
<td>Language [LA]</td>
<td>Publication Type [PT]</td>
</tr>
<tr>
<td>Author [AU]</td>
<td>Last Author [LASTAU]</td>
<td>Secondary Source ID [SI]</td>
</tr>
<tr>
<td>Comment Corrections</td>
<td>Location ID [LID]</td>
<td>Subset [SE]</td>
</tr>
<tr>
<td>Corporate Author [CN]</td>
<td>MeSH Date [MHD]</td>
<td>Substance Name [NM]</td>
</tr>
<tr>
<td>Entrez Date [EDAT]</td>
<td>MeSH Subheadings [SH]</td>
<td>Title [TI]</td>
</tr>
<tr>
<td>Filter [FILTER]</td>
<td>MeSH Terms [MH]</td>
<td>Title/Abstract [TIA]</td>
</tr>
<tr>
<td>First Author Name [AU]</td>
<td>NLM Unique ID [UID]</td>
<td>Transliterated Title [TT]</td>
</tr>
<tr>
<td>Full Author Name [FAU]</td>
<td>Other Term [CT]</td>
<td>UID [PMID]</td>
</tr>
<tr>
<td>Full Investigator Name [FI]</td>
<td>Owner</td>
<td>Volume [VI]</td>
</tr>
<tr>
<td>Grant Number [GR]</td>
<td>Pagination [PG]</td>
<td></td>
</tr>
<tr>
<td>Investigator [IR]</td>
<td>Personal Name as Subject [PS]</td>
<td></td>
</tr>
<tr>
<td>Pharmacological Action MeSH Terms [PA]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

[Image of PubMed interface with circled areas]
Query Refinement

PublMed History - Mozilla Firefox

- Search functionality is available eight hours of inactivity.
- Search numbers may not be continuous, all searches are represented.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use `#search`, e.g., `#3 AND #5` or click query # for more options.

Search

Most Recent Queries

- #12 Search: Myers-g[nau] OR sequence[ta]
- #22 Search: Myers-g[nau] OR sequence[ti] OR Nucleic acids res[ta]
- #23 Search: Myers-g[nau] OR sequence[ti] OR Nucleic acids res[ta]
- #20 Search: Myers-g[nau] OR sequence[ti] OR Nucleic[ta]
- #19 Search: Myers-g[nau] OR sequence[ti] OR Nucleic[ta]
- #18 Search: Myers-g[nau] OR sequence[ti] OR Nucleic[ta]
- #17 Search: Myers-g[nau] OR sequence[ti] OR Nucleic acid[ta]
- #16 Search: Myers-g[nau] OR sequence[ti] OR nature[ta]
- #14 Search: Myers-g[nau] OR sequence[ti] AND nature[ta]
- #13 Search: Myers-g[nau] OR sequence[ti]
- #12 Search: Myers-g[nau] OR string[ti]
- #11 Search: Myers-g[nau] OR genome[ti]
- #10 Search: Myers-g[nau]
Dublin Core Metadata Initiative (W3C), 1995

- **identifier:** ISBN/ISSN, URL/PURL, DOI, …
- **format:** MIME-Typ, media type,
- **type:** Collection, image, text, …
- **language**
- **title**
- **subject:** Keywords
- **coverage:** Scope of doc in space and/or time
- **description:** Free text
- **creator:** Last person manipulating the doc
- **publisher:**
- **contributor:**
- **rights:** Copyright, licenses, …
- **source:** Other doc
- **relation:** To other docs
- **date:** Date or period
Usage in HTML

<head profile="http://dublincore.org/documents/dcq-html/">
<title>Dublin Core</title>
<link rel="schema.DC" href="http://purl.org/dc/..." />
<link rel="schema.DCTERMS" href="http://purl.org/..." />
<meta name="DC.format" scheme="..." content="text/htm" />
<meta name="DC.type" scheme="..." content="Text" />
<meta name="DC.publisher" content="Jimmy Whales" />
<meta name="DC.subject" content="Dublin Core Metadata" />
<meta name="DC.creator" content="Björn G. Kulms" />
<meta name="DCTERMS.license" scheme="DCTERMS.URI" content="http://www.gnu.org/copyleft/fdl.html" />
</head>
Content of this Lecture

• What is Information Retrieval
• Documents
• Queries
• Related topics
Multimedia Retrieval

• Note: Neither searches within images
  - Flickr: tags ("folksonomy")
  - Google: text in neighborhood
„Search by Image“ (10/2014)
Historic Texts

• Sachsenspiegel, ~1250
  - "Swerlenrecht kūnnen wil•d~ volge dis buches lere.alrest sul wi mer ken, daz ..."

• Multiple representations
  - Facsimile
  - Digitalization / diplomacy
    • How well can the facsimile be reproduced from the dig. form?
  - Differences in individual writers (proliferating errors)
  - Different translations
  - Different editions
Other Buzzwords

- **Document management systems (DMS)**
  - Large commercial market, links to OCR, workflow systems, mail routing, etc.
  - Legal issues (compliance, reporting, archival, …)
  - Every DMS includes an IR system

- **Knowledge management**
  - “More sophisticated” DMS with semantic searching
    - Ontologies, thesauri, topic maps, …
  - **Social aspects**: Incentives, communities, standard procedures, enterprise vocabulary, …

- **Digital libraries**
  - Somewhat broader and less technical
  - Includes social aspects, archiving, multimedia, …
Enterprise Content Management

• „The technologies used to capture, manage, store, deliver, and preserve information to support business processes”

• Authorization and authentication
• Business process management and document flow
• Compliance: legal requirements
  – Record management
  – Pharma, Finance, …
• Collaboration and sharing
  – Inter and intra organizations
  – Transactions, locks, …
• Publishing: What, when, where
  – Web, catalogues, mail push, …

Quelle: AIIM International
Technique versus Content

- IR is about techniques for searching a given doc collection
- Creating **doc collections is a business**: Content provider
  - Selection/filtering: classified business news, new patents, …
  - Augmentation: Annotation with metadata, summarization, linking of additional data, …

- Examples
  - **Medline**: >5000 Journals, <20M citations, >500K added per year
  - Institute for Scientific Information (ISI)
    - Impact factors: which journals count how much?
  - Web catalogues ala Yahoo
  - “Pressespiegel”, web monitoring