Text Analytics

Introduction to Information Retrieval

Ulf Leser
• Summary of last lecture …
Content of this Lecture

- **Information Retrieval**
  - Introduction
  - Documents
  - Queries

- Related topics

- A first idea: Boolean queries and vector space model
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 usually searches unstructured data (e.g. text)
  - But: Keyword search in relational databases

- 90% of all information is believed to be presented in unstructured form
History of IR

• ~300 ad. Library of Alexandria, ~700,000 „documents“
• 1450: Bookprint
• 19th century: Indices and concordance (expensive and laborious)
• KWIC: H.P. Luhn, IBM (1958)
• 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
• End 90s: Personalized search engines, enterprise search, recommendation engines etc.
• 2010: Mobile and context-based search, social networks
The Problem

• Help user in **quickly** finding the **requested information** from a **given set of documents**
  
  – Documents: Corpus, library, collection, …
  
  – Quickly: Comfortable, with few queries, without need to learn a lot about the system, …
  
  – Requested: The “best-fitting” documents; the “right” passages
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**
  - What was the score of Bayern München versus Stuttgart im DFB Pokal 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**
  – Understand request, find best matches
    • Usually implemented in one method
  – 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”
  – In the corpus is large, **ranking is top priority**
  – Result summarization (grouping on what?)
  – Different **search modes**: What’s new? What’s certain?
IR: An Iterative, Multi-Stage, Complex 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]
• Searching only metadata
• Searching within documents
• Interpreting documents
KWIC: KeyWord In Context

Aillion Americanas a year and (killed) 350,000 of them. Twice as effective as conventional. Least 250 people have been (killed) since a 20-day truce with the communist insurgents carrying terrorist who had (killed) an American. Immediately covering a legal task is (killed) to (killed) and (killed) an American messenger. Lila told him that I (killed) also a teen that people will (killed) to see, as well as enough witnesses with a discuss doses of methane to (killed) cancer cells and then shortly afterward administrating a dose deal is worth getting (killed) over. After the bike buy, Gary wonders aloud in not get her disturbed son to (killed) her mistery father. Now NBC presents this movie and M Chap Keimont in (killed) "probably brutally, I believe they reached you 10, you dis dis-honorably? Has he (killed) by one of his own men? Has he been drugs? Did he (killed) in a 1988 dam break in Italy. (killed) many more than 220 people. Meanwhile, dam-safety unit collapsed hundreds, which (killed) 60 people and caused panic as far away as Chelyabinsk in the Urals. As a (killed) in Afghanistan in 1985, in Soviet military, rather than Mt. (killed) -- -- a dozen or so (killed), according to unofficial sources. -- including pol, in which an American was (killed), -- A fire erupted in Australia's only nuclear plant (killed) work spreads and can (killed) you is very sophisticated, " said Dr. William Dietz.

Variance (killed), not speed, " he says. Mr. Lave suggests that it overcome with shame that he (killed) himself. Ajax's crime is not that he slaughtered a group, which planned to (killed) the French journalist Monday, again demanded that the two civilians were reported (killed), -- an over-loaded bus collided Sunday with a truck...
Documents

• **This lecture: Natural language text**

• Might be grammatically correct (books, newspapers) or not (BLOGs, newsgroups, spoken language)

• May have structure (abstract, summary, chapters, ...) or not (plain text)

• May have (explicit or in-text) metadata (title, author, year, ...) or not

• May be in many different languages or even mixed
  - Foreign characters

• May refer to other documents (hyperlinks)

• May have various formats (ASCII, PDF, DOC, XML, ...)

IR and XML
**XML and IR**

- **XML has two flavors**
  - **Data oriented**: Highly structured, except from database, short documents, high volume, no inline tags, structure-oriented search
    - This is the case in most applications
  - **Document oriented**: Less structured, large documents, inline tags, text-oriented search
    - This was the original purpose of XML (and SGML)

- **XQuery/XPath is targeting the data aspect**
  - Focus on structural parts of a query
  - No full-text search, few search options within a value
  - Semantics: Exact matching

- **“XML IR” is targeting the document aspect**
- **Several proposals for languages addressing both**
IR Queries

- Users formulate queries
  - Keywords
  - Phrases
  - Logical operations (AND, OR, NOT, ...)
    - Web search: "-ulf +leser"
  - Real questions (e.g. MS-Word help)
  - Structured queries (author=... and title~ ...)

- Querying with document
  - Find documents similar to this one

- Query refinement
  - Refine my query based on first results
  - Find documents within the result set of the previous search
Searching with Metadata (PubMed/Medline)
Query Refinement
Metadata Standard: Dublin Core

- **Dublin Core** Metadata Initiative (W3C), 1995
- 15 attributes
  - identifier: E.g. 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
- Plus refinements
Example

- 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

- Information Retrieval
- Related topics
- A first idea: Boolean queries and vector space model
Historic Texts

- Sachsenspiegel, ~1250
  - “Swerlenrecht kûnnen 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, print management systems, mail management, 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, …
    • “if Siemens knew what Siemens knows”

• **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** given a 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, 19 Mill citations, >500K added per year
    • Journals not in Medline don’t exist
  - Institute for Scientific Information (ISI)
    • Impact factors: which journals count?
  - Web catalogues ala Yahoo
  - “Pressespiegel”, web monitoring
Content of this Lecture

- Information Retrieval
- Related topics
- A first idea: Boolean queries and vector space model
Logical View

• **Definition**
  - *The logical view on a document denotes its representation inside the corpus*

• **Determines what we can query**
  - Only metadata, only title, only abstract, full text, …

• **Creating the logical view of a doc involves transformations**
  - Stemming, stop word removal
  - Transformation of special characters
    • Umlaute, greek letters, XML/HTML encodings, …
  - Removal of formatting information (HTML), tags (XML), …
  - **Bag of words** (BoW)
    • Arbitrary yet fixed order (e.g. sorted alphabetically)
Preprocessing

- Tokenization can be a real issue
- Special characters can be a real issue
- Format conversion (OCR …) can be a real issue
- Syntactic analysis can be a real issue
- …
A First Approach: Boolean Queries (rough idea)

- Search all docs which match a logical expression of words
  - bono AND u2, bono AND (Florida OR Texas), bono AND NOT u2

- A simple evaluation scheme
  - Compute BoW of all documents: bow(d)
  - Compute all atoms of the query (w)
  - An atom $w$ evaluates to true on $b$ iff $w \in bow(d)$
  - Compute values of all atoms for each $d$
  - Compute value of logical expression for each $d$

- Advantages: Fast, simple
  - Efficient implementations do not scan all documents

- Disadvantages
  - No ranking: Either relevant or not
  - Logic is hard to understand for ordinary users
  - Typical queries are too simple – too many results (and no ranking)
Vector Space Model (rough idea)

• Let \( d \) by a doc, \( D \) be a collection of docs with \( d \in D \)
• Let \( v_d \) by a vector with
  - \( v_d[i] = 0 \) if the i’th word in \( \text{bow}(D) \) is not contained in \( \text{bow}(d) \)
  - \( v_d[i] = 1 \) if the i’th word in \( \text{bow}(D) \) is contained in \( \text{bow}(d) \)
• Let \( v_q \) by the corresponding vector for query \( q \)
• Compute relevance score for \( q \) wrt. each \( d \) as

\[
\text{rel}(d, q) = \frac{v_d \cdot v_q}{|v_d| \ast |v_q|} = \frac{\sum_{i=1}^{|w(D)|} v_d[i] \ast v_q[i]}{\sqrt{\sum v_d[i]^2} \ast \sqrt{\sum v_q[i]^2}}
\]
Angel between Document and Query

- Relevance: angle between doc and query in $|\text{bow}(D)|$ - dimensional space
- Return docs in order of decreasing relevance

- Advantages
  - Still quite simple
  - “Meaningful” ranking – better results than Boolean model
  - Allows approximate answers (not all terms contained)

- Disadvantages
  - Words are assumed to be statistically independent
  - Not easy to integrate NOT or OR
Outlook: Other techniques

- We often also want to find documents which don’t even contain any of the words we are searching for
  - Docs that contain the concepts we are trying to express by our query

- Latent Semantic Indexing
  - Map query and docs into a concept space
  - Compare and rank in concept space
  - Technical: Dimensionality reduction (like PCA etc.)