VK Multimedia Information Systems

Mathias Lux, mlux@itec.uni-klu.ac.at

Information Retrieval Basics: Agenda

- **Information Retrieval History**
- Information Retrieval & Data Retrieval
- Searching & Browsing
- Information Retrieval Models

Information Retrieval History

Currently there are no museums for IR

IR is the process of **searching** through a **document collection** based on a particular **information need.**

IR Key Concepts

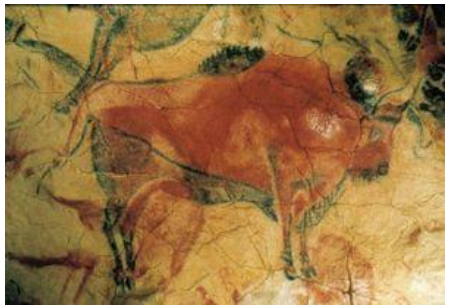
- Searching
  - Indexing, Ranking

- Document Collection
  - Textual, Visual, Auditive

- Particular Needs
  - Query, User based

A History of Libraries

Libraries are perfect examples for document collections.



- Wall paintings in caves
  - e.g. Altamira, ~ 18,500 years old
- Writing in clay, stone, bones
  - e.g. Mesopotamian cuneiforms, ~ 4.000 BC
  - e.g. Chinese tortoise-shell carvings, ~ 6.000 BC
  - e.g. Hieroglyphic inscriptions, Narmer Palette ~ 3.200 BC

A History of Libraries (ctd.)

- Papyrus
  - Specific plant (subtropical)
  - Organized in rolls, e.g. in Alexandria
- Parchment
  - Independence from papyrus
  - Sewed together in books
- Paper
  - Invented in China (bones and bamboo too heavy, silk too expensive)
  - Invention spread -> in 1120 first paper mill in Europe

A History of Libraries (ctd.)

- Gutenberg's printing press (1454)
  - Inexpensive reproduction
  - e.g. "Gutenberg Bible"
- Organization & Storage
  - Dewey Decimal System (DDC, 1872)
  - Card Catalog (early 1900s)
  - Microfilm (1930s)
  - MARC (Machine Readable Cataloging, 1960s)
  - Digital computers (1940s+)

Library & Archives today

- Partially converted to electronic catalogues
  - From a certain time point on (1992 - ...)
  - Often based on proprietary systems
  - Digitization happens slow
  - No full text search available
  - Problems with preservation
    - Storage devices & formats

History of Searching

- Browsing
  - Like “Finding information yourself”
- Catalogs
  - Organized in taxonomies, keywords, etc.
- Content Based Searching
  - `SELECT * FROM books WHERE title='%Search%'`
- Information Retrieval
  - Ranking, models, weighting
  - Link analysis, LSA, ...

History of IR

- Starts with development of computers
- Term “Information Retrieval” coined by Mooers in 1950
  - Mooers, C. (March 1950). "The theory of digital handling of non-numerical information and its implications to machine economics". *Proceedings of the meeting of the Association for Computing Machinery at Rutgers University*.
- Two main periods (Spark Jones u. Willett)
  - 1955 – 1975: Academic research
    - Models and Basics
    - Main Topics: Search & Indexing
  - 1975 – ... : Commercial applications
    - Improvement of basic methods

A Challenge: The World Wide Web

- First actual implementation of **Hypertext**
  - Interconnected documents
  - Linked and referenced
- World Wide Web (1989, T. Berners-Lee)
  - Unidirectional links (target is not aware)
  - Links are not typed
  - Simple document format & communication protocol (HTML & HTTP)
  - Distributed and not controlled

Some IR History Milestones

- Book “Automatic Information Organization and Retrieval”, *Gerard Salton* (1968)
  - Vector Space Model
- Paper “A statistical interpretation of term specificity and its application in retrieval”, Karen Sparck Jones (1972)
  - IDF weighting
  - http://www.soi.city.ac.uk/~ser/idf.html
- Book “Information Retrieval” of *C.J. Rijsbergen* (1975)
  - Probabilistic Model
  - http://www.dcs.gla.ac.uk/Keith/Preface.html

Some IR History Milestones

- Paper “Indexing by Latent Semantic Analysis”, S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, R. Harshman (1990).
  - Latent Semantic Indexing
- Paper “Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval” Robertsen & Walker (1994)
  - BM25 weighting scheme
- Paper “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, *Sergey Brin & Larry Page* (1998)
  - World Wide Web Retrieval

Information Retrieval Basics: Agenda

- Information Retrieval History
- **Information Retrieval & Data Retrieval**
- Searching & Browsing
- Information Retrieval Models

Organizational: References

- in the Library
  - *Modern Information Retrieval*, Ricardo Baeza-Yates & Berthier Ribeiro-Neto, Addison Wesley
  - *Google's Pagerank and Beyond: The Science of Search Engine Rankings*, Amy N. Langville & Carl D. Meyer, University Presses of CA
  - *Distributed Multimedia Database Technologies supported by MPEG-7 and MPEG-21*, Harald Kosch, CRC Press
  - *Readings in Information Retrieval*, Karen Sparck Jones, Peter Willett, Morgan Kaufmann

Organizational: References

- WWW
  - *Skriptum Information Retrieval*, Norbert Fuhr, Lecture Notes on Information Retrieval - Univ. Dortmund, 1996. Updated in 2002
  - *Information Retrieval 2nd Edt.*, C.J. Rijsbergen, Butterworth, London 1979
- Through me:
  - *Lectures on Information Retrieval: Third European Summer-School, Essir 2000 Varenna, Italy, Revised Lectures*, Maristella Agosti, Fabio Crestani & Gabriela Pasi (eds.), Lecture Notes in Computer Science, Springer 2000

Information Retrieval & Data Retrieval

Information Retrieval

- Information Level
- Search Engine
- Teoma / Google

Data Retrieval

- Data Level
- Data Base
- Oracle / MySQL

Information Retrieval & Data Retrieval

Information Retrieval	Data Retrieval
Content Based Search	Search for Patterns and String
Query ambigous	Query formal & unambigous
Results ranked by relevance	Results not ranked
Error tolerant	Not error tolerant
Multiple iterations	Clearly defined result set
Examples	*Examples*
Search for synonyms	Search for patterns
Bag of Words	SQL Statement

- Retrieval is nearly always a combination of both.

Information Retrieval Basics: Agenda

- Information Retrieval History
- Information Retrieval & Data Retrieval
- **Searching & Browsing**
- Information Retrieval Models

Information Retrieval Basics: Searching

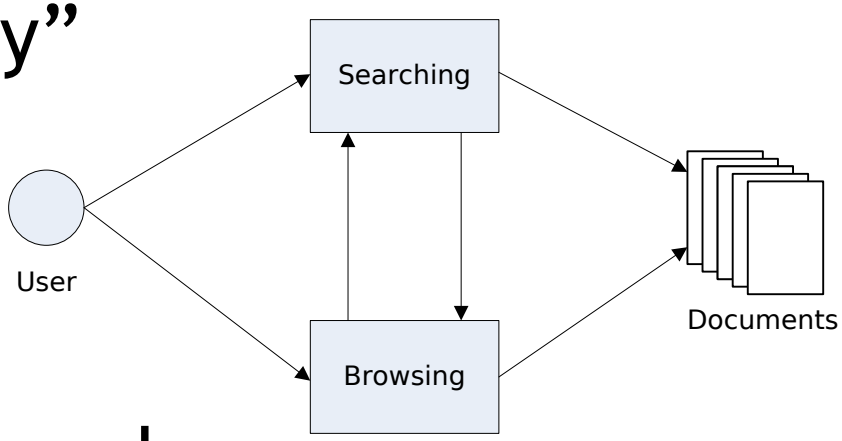
A **user** has an **information need**, which needs to be **satisfied**.

- Two different approaches:
  - Browsing
  - Searching

Searching & Browsing

Searching

- Explicit information need
- Definition through “query”
- Result lists
- e.g. Google



Browsing

- Not necessarily explicit need
- Navigation through repositories

Browsing

- Flat Browsing
  - User navigates through set of documents
  - No implied ordering, explicit ordering possible
  - Examples: One single directory, one single file
- Structure Guided Browsing
  - An explicit structure is available for navigation
  - Mostly hierarchical (file directories)
  - Can be generic digraph (WWW)
  - Examples: File systems, World Wide Web

Searching

- Query defines “Information Need”

- Ad Hoc Searching
  - Search when you need it
  - Query is created to fit the need
- Information Filtering
  - Make sets of documents smaller
  - Query is filter criterion
- Information Push
  - Same as filtering, delivery is different

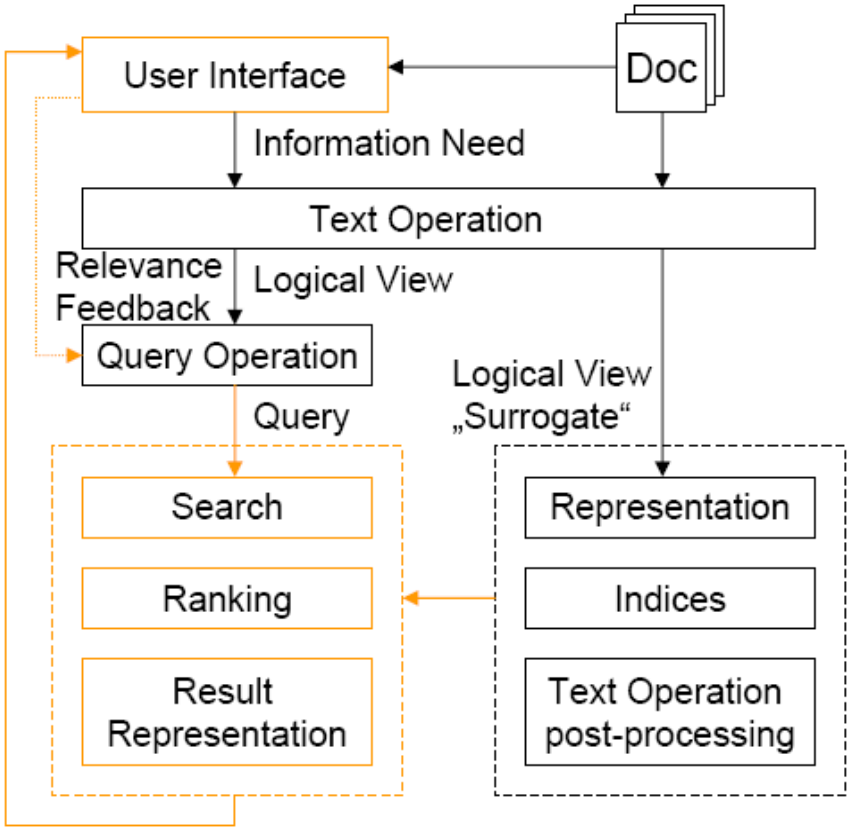
Information Retrieval Basics: Agenda

- Information Retrieval History
- Information Retrieval & Data Retrieval
- Searching & Browsing
- **Information Retrieval Models**

Information Retrieval System Architecture

Aspects

- Query & languages
- IR models
- Documents
- Internal representation
- Pre- and post-processing
- Relevance feedback
- HCI



Information Retrieval Models

- Boolean Model
  - Set theory & Boolean algebra
- Vector Model
  - Non binary weights on dimensions
  - Partial match
- Probabilistic Model
  - Modeling IR in a probabilistic framework

Formal Definition of Models

An information retrieval model is a quadruple [D, Q, F, $R(q_i, d_j)$]

- D is a set of logical views (or representations) for the **documents** in the collection.
- Q is a set of logical views (or representations) for the user needs or **queries**.
- F is a **framework** for modeling document representations, queries and their relationship.
- $R(q_i, d_j)$ is a **ranking function** which associates a real number with a query q_i of Q and a document d_j of D.

Definitions
in Context of Text Retrieval

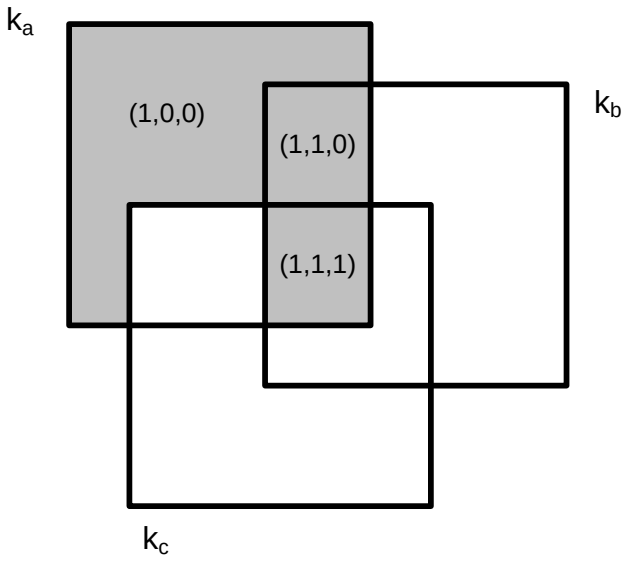
- **index term** – word of a document expressing (part of) document semantics
- **weight** $w_{i,j}$ – quantifies the importance of index term t_i for document d_j
- index term vector for document d_j (having t different terms in all documents):

$$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

Boolean Model

- Based on set theory and Boolean algebra
  - Set of index terms
  - Query is Boolean expression
- Intuitive concept:
  - Wide usage in bibliographic system
  - Easy implementation and simple formalisms
- Drawbacks:
  - Binary decision components (true/false)
  - No relevance scale (relevant or not)

Boolean Model: Example



$$q = k_a \wedge (k_b \vee \neg k_c)$$

Boolean Model: DNF

$$q = k_a \wedge (k_b \vee \neg k_c) \ldots \vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$$

- Express queries in *disjunctive normal form* (disjunction of conjunctive components)
- Each of the components is a binary weighted vector associated with (k_a, k_b, k_c)
- Weights $w_{i,j} \in \{0,1\}$

Boolean Model: Ranking function

$$sim(d_j, q) = \begin{cases} 1 & \text{if } \exists \vec{q}_{cc} \,|\, (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{k_i}, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0 & \text{otherwise} \end{cases}$$

- similarity is one if one of the conjunctive components in the query is exactly the same as the document term vector.

Boolean Model

- Advantages
  - Clean formalisms
  - Simplicity
- Disadvantages
  - Might lead to too few / many results
  - No notion of **partial match**
  - Sequential ordering of terms not taken into account.

Thanks …

for your attention!