



VK Multimedia Information Systems

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

Dienstags, 16.00 Uhr c.t., E.1.42



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. See <http://creativecommons.org/licenses/by-nc-sa/3.0/at/>

Information Retrieval Basics: Agenda



<http://www.uni-klu.ac.at>

- **Information Retrieval History**
- Information Retrieval & Data Retrieval
- Searching & Browsing
- Information Retrieval Models
- Exercise 01



Information Retrieval History



<http://www.uni-klu.ac.at>

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



<http://www.uni-klu.ac.at>

- Searching
 - Indexing, Ranking
- Document Collection
 - Textual, Visual, Auditive
- Particular Needs
 - Query, User based



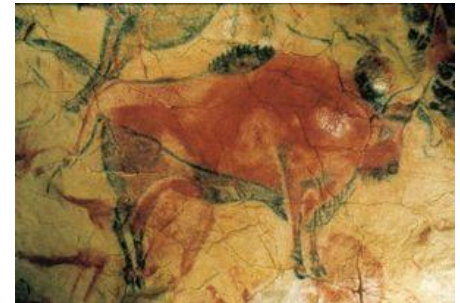
A History of Libraries



<http://www.uni-klu.ac.at>

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, Narmar Palette ~ 3.200 BC



A History of Libraries (ctd.)



<http://www.uni-klu.ac.at>

- 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.)



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- Starts with development of computers.
- Term “Information Retrieval” coined by Mooers in 1952
- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- Information Retrieval History
- **Information Retrieval & Data Retrieval**
- Searching & Browsing
- Information Retrieval Models
- Exercise 01



Organizational: References



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

Information Retrieval

- Information Level
- Search Engine
- Teoma / Google

Data Retrieval

- Data Level
- Data Base
- Oracle / MySQL

Information Retrieval & Data Retrieval



<http://www.uni-klu.ac.at>

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

- Retrieval is nearly always a combination of both.

Information Retrieval Basics: Agenda



<http://www.uni-klu.ac.at>

- Information Retrieval History
- Information Retrieval & Data Retrieval
- **Searching & Browsing**
- Information Retrieval Models
- Exercise 01



Information Retrieval Basics: Searching



<http://www.uni-klu.ac.at>

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

- Two different approaches:
 - Browsing
 - Searching

Searching & Browsing



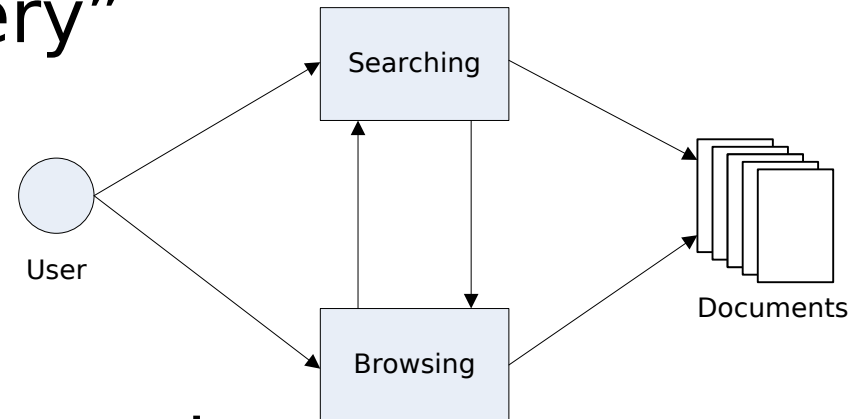
<http://www.uni-klu.ac.at>

Searching

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

Browsing

- Not necessarily explicit need
- Navigation through repositories



- 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



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

- Information Retrieval History
- Information Retrieval & Data Retrieval
- Searching & Browsing
- **Information Retrieval Models**
- Exercise 01



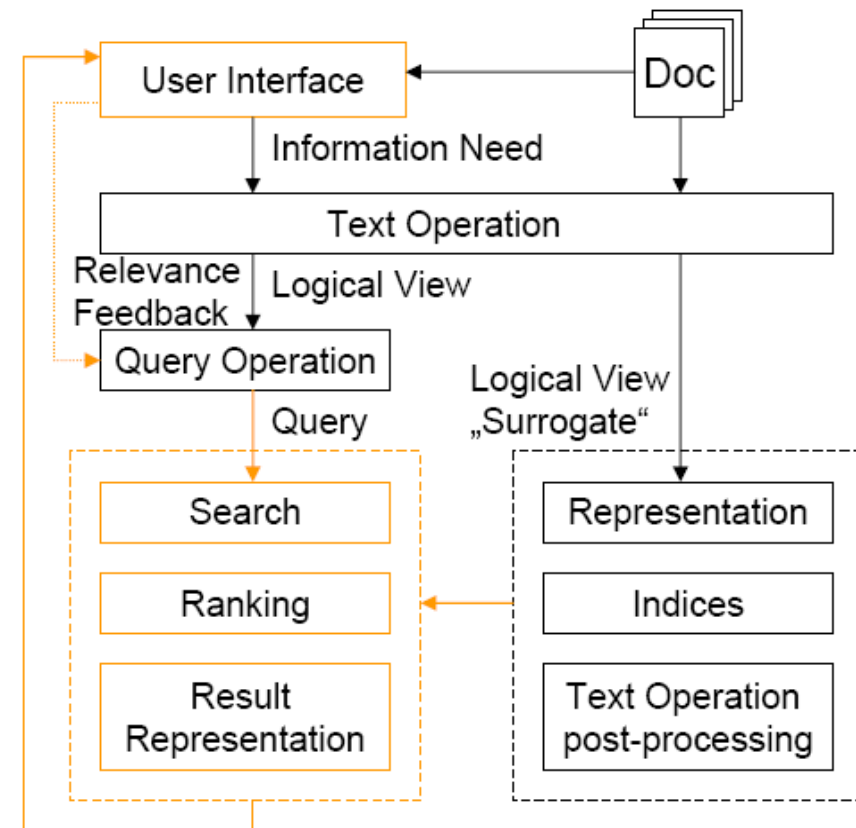
Information Retrieval System Architecture



<http://www.uni-klu.ac.at>

Aspects

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



Information Retrieval Models



<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>

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



<http://www.uni-klu.ac.at>

- **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):

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

Boolean Model



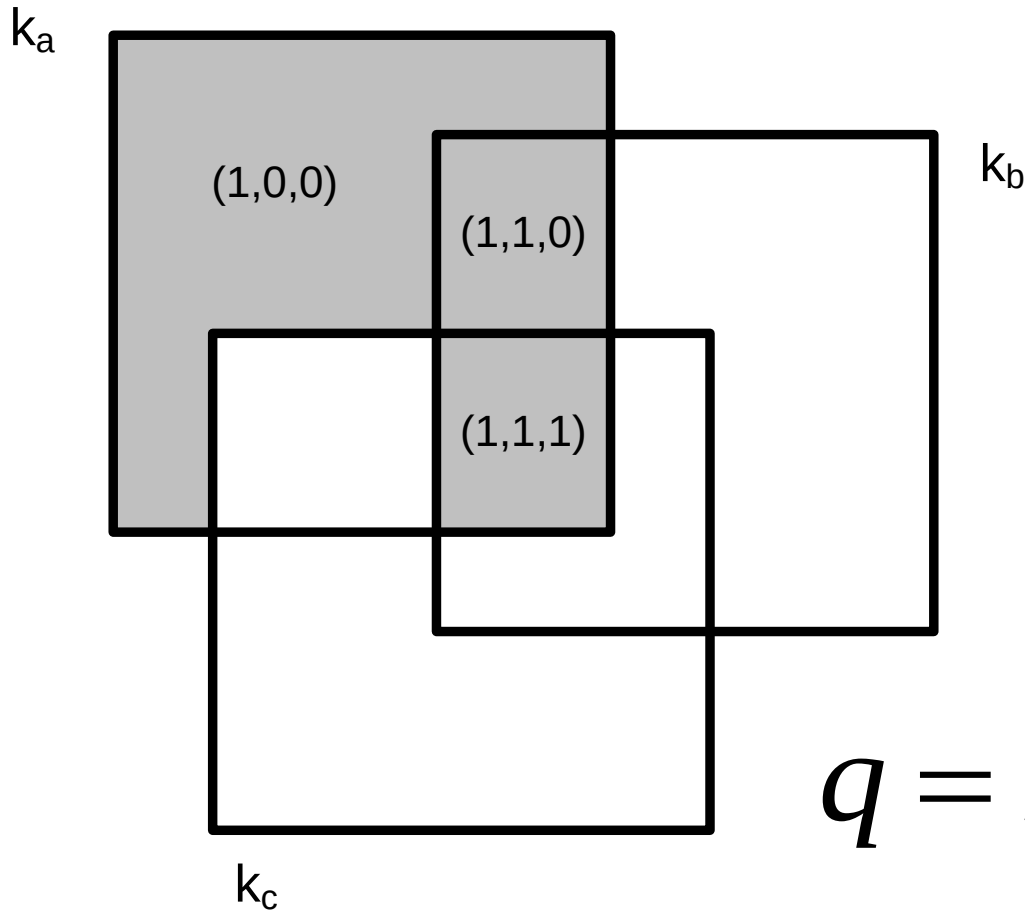
<http://www.uni-klu.ac.at>

- 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



<http://www.uni-klu.ac.at>



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

Boolean Model: DNF



<http://www.uni-klu.ac.at>

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

- 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



<http://www.uni-klu.ac.at>

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

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

Boolean Model



<http://www.uni-klu.ac.at>

- 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.

Vector Model



<http://www.uni-klu.ac.at>

- Integrates the notion of partial match
- Non-binary weights (terms & queries)
- Degree of similarity computed

$$\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

$$\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

Vector model: Similarity



<http://www.uni-klu.ac.at>

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^t w_{i,q}^2}}$$

Vector Model: Example



<http://www.uni-klu.ac.at>

$$\vec{d} = (0.3, 0.4, 0, 0.1, 1)$$

$$\vec{q} = (1, 0, 0, 0.5, 0)$$

$$\text{sim}(\vec{d}, \vec{q}) = \frac{1 \cdot 0.3 + 0.1 \cdot 0.5}{\sqrt{0.3^2 + 0.4^2 + 0.1^2 + 1} \cdot \sqrt{1 + 0.5^2}} \approx \frac{0.35}{2.24} \approx 0.17$$

Another Example:



<http://www.uni-klu.ac.at>

- Document & Query:

- D = “The quick brown fox jumps over the lazy dog”
- Q = “brown lazy fox”

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^t w_{i,q}^2}}$$

- Results:

- $(1,1,1,1,1,1,1,1,1)^t * (1,1,1,0,0,0,0,0,0)^t = 3$
- $\text{sqrt}(9) * \text{sqrt}(3) = 5,196$
- $\text{Similarity} = 3 / 5,196 = 0,577$

Term weighting: TF*IDF



<http://www.uni-klu.ac.at>

Term weighting increases retrieval performance ...

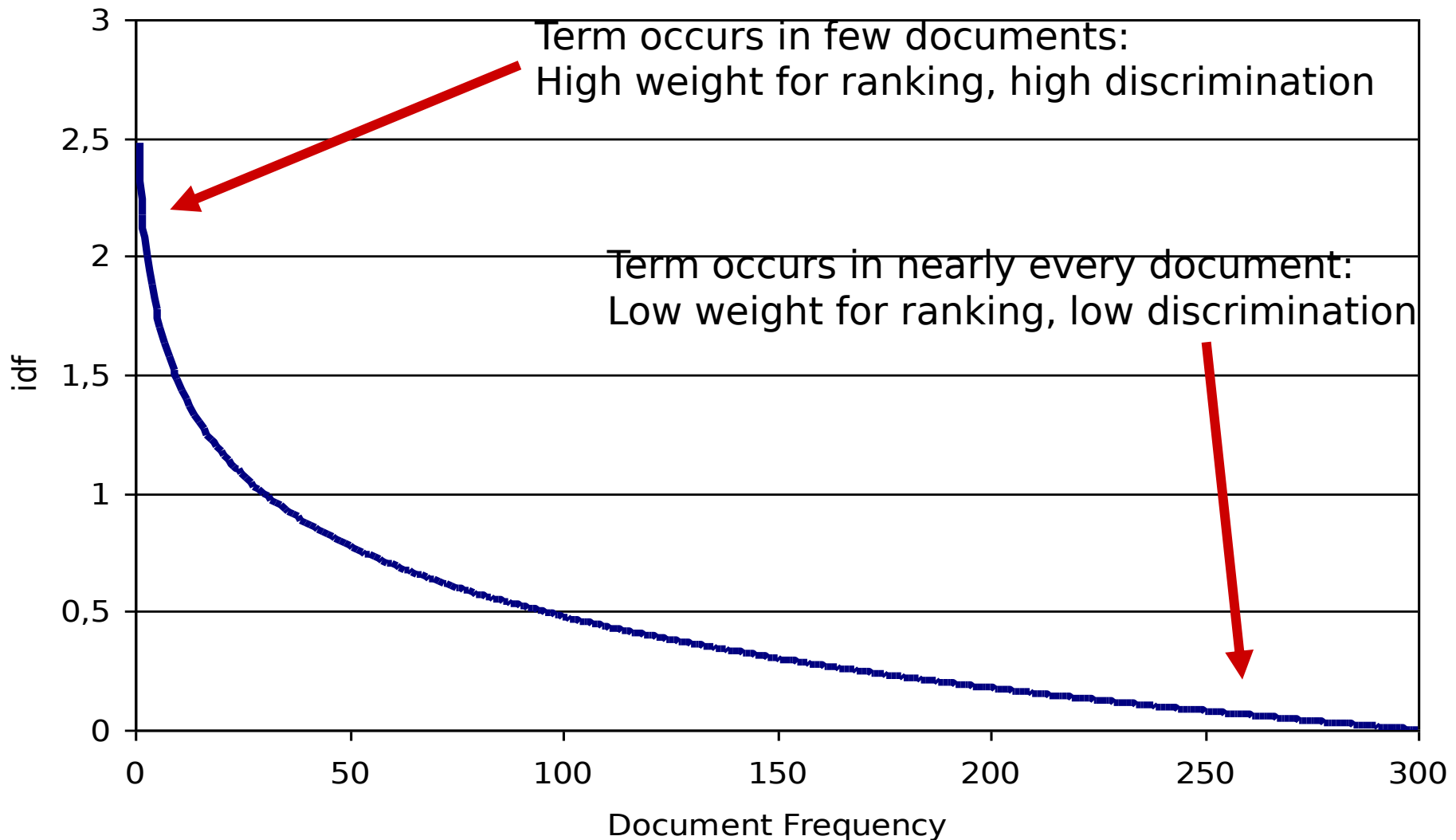
- Term frequency
 - Most intuitive approach
 - How often does a term occur in a document?
- Inverse Document Frequency
 - How much information content has a term for a document collection?
 - Compare to *Information Theory* of Shannon

Example: IDF

300 documents corpus



<http://www.uni-klu.ac.at>



Definitions:

Normalized Term Frequency



<http://www.uni-klu.ac.at>

$$f_{i,j} = \frac{freq_{i,j}}{\max_l(freq_{l,j})} \dots \text{normalized term frequency}$$

$freq_{i,j}$... raw term frequency of term i in document j

- Maximum is computed over all terms in a document
- Terms which are not present in a document have a raw frequency of 0

Definitions:

Inverse Document Frequency



<http://www.uni-klu.ac.at>

$idf_i = \log \frac{N}{n_i}$... inverse document frequency for term i

N ... number of documents in the corpus

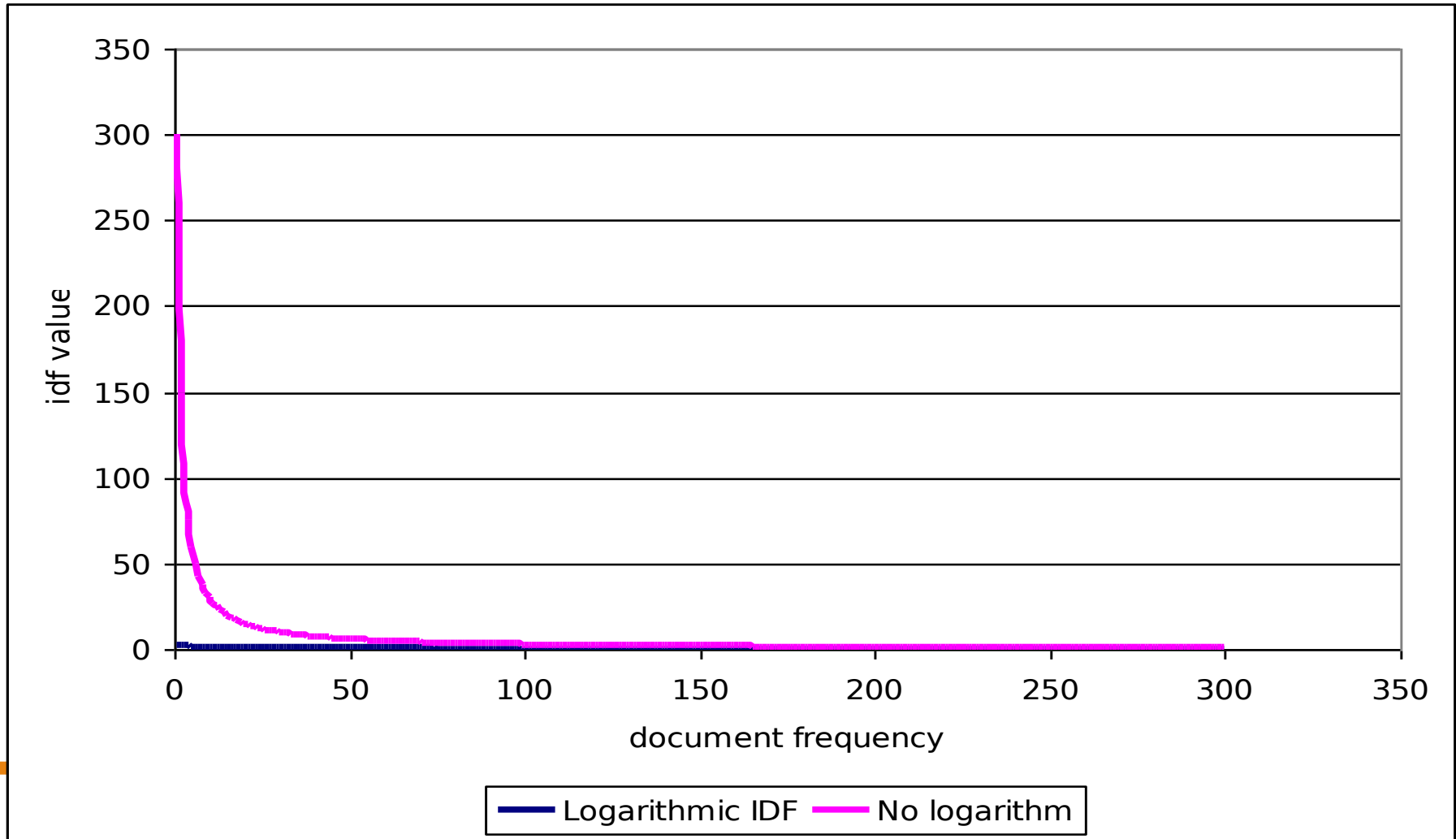
n_i ... number of document in the corpus which contain term i

- Note that idf_i is independent from the document.
- Note that the whole corpus has to be taken into account.

Why $\log(\dots)$ in IDF?



<http://www.uni-klu.ac.at>



- TF*IDF is a very prominent weighting scheme
 - Works fine, much better than TF or Boolean
 - Quite easy to implement

$$w_{i,j} = f_{i,j} \cdot \log \frac{N}{n_i}$$

Weighting of query terms



<http://www.uni-klu.ac.at>

$$w_{i,q} = \left(0.5 + \frac{0.5 \cdot f_{i,q}}{\max_l(f_{l,q})}\right) \cdot \log \frac{N}{n_i}$$

- Also using IDF of the corpus
- But TF is normalized differently
 - $TF > 0.5$
- Note: the query is not part of the corpus!

Vector Model



<http://www.uni-klu.ac.at>

- Advantages
 - Weighting schemes improve **retrieval performance**
 - Partial matching allows retrieving documents that **approximate query** conditions
 - Cosine coefficient allows **ranked list** output
- Disadvantages
 - Term are assumed to be mutually independent

Simple example (i)



<http://www.uni-klu.ac.at>

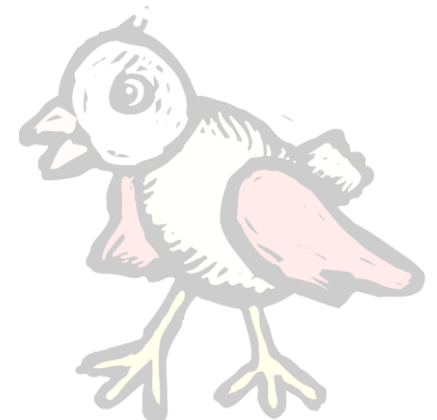
- Scenario
 - Given a **document corpus on birds**: nearly each document (say 99%) contains the word bird
 - someone is searching for a document about sparrow nest construction with a query **“sparrow bird nest construction”**
 - Exactly the document which would satisfy the user needs **does not have the word “bird”** in it.

Simple example (ii)



<http://www.uni-klu.ac.at>

- TF*IDF weighting
 - knows upon the low discriminative power of the term bird
 - The weight of this term is near to zero
 - This term has virtually no influence on the result list.



Information Retrieval Basics: Agenda



<http://www.uni-klu.ac.at>

- Information Retrieval History
- Information Retrieval & Data Retrieval
- Searching & Browsing
- Information Retrieval Models
- **Exercise 02**



Exercise



<http://www.uni-klu.ac.at>

- Given a document collection ...
- Find the results to a query ...
 - Employing the Boolean model
 - Employing the vector model (with $TF*IDF$)
- Some hints:
 - Excel:
 - Sheet on homepage
 - Use functions “Summenprodukt” & “Quadratesumme”

Exercise



<http://www.uni-klu.ac.at>

- Document collection (6 documents)
 - spatz, amsel, vogel, drossel, fink, falke, flug
 - spatz, vogel, flug, nest, amsel, amsel, amsel
 - kuckuck, nest, nest, ei, ei, ei, flug, amsel, amsel, vogel
 - amsel, elster, elster, drossel, vogel, ei
 - falke, katze, nest, nest, flug, vogel
 - spatz, spatz, konstruktion, nest, ei
- Queries:
 - spatz, vogel, nest, konstruktion
 - amsel, ei, nest

Exercise



<http://www.uni-klu.ac.at>

	d1	d2	d3	d4	d6	d6	idf
amsel	1	3	2	1			
drossel	1			1			
ei			3	1		1	
elster				2			
falke	1				1		
fink	1						
flug	1	1	1		1		
katze					1		
konstruktion						1	
kuckuck			1				
nest		1	2		2	1	
spatz	1	1				2	
vogel	1	1	1	1	1		

Thanks ...



<http://www.uni-klu.ac.at>

for your attention!