Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

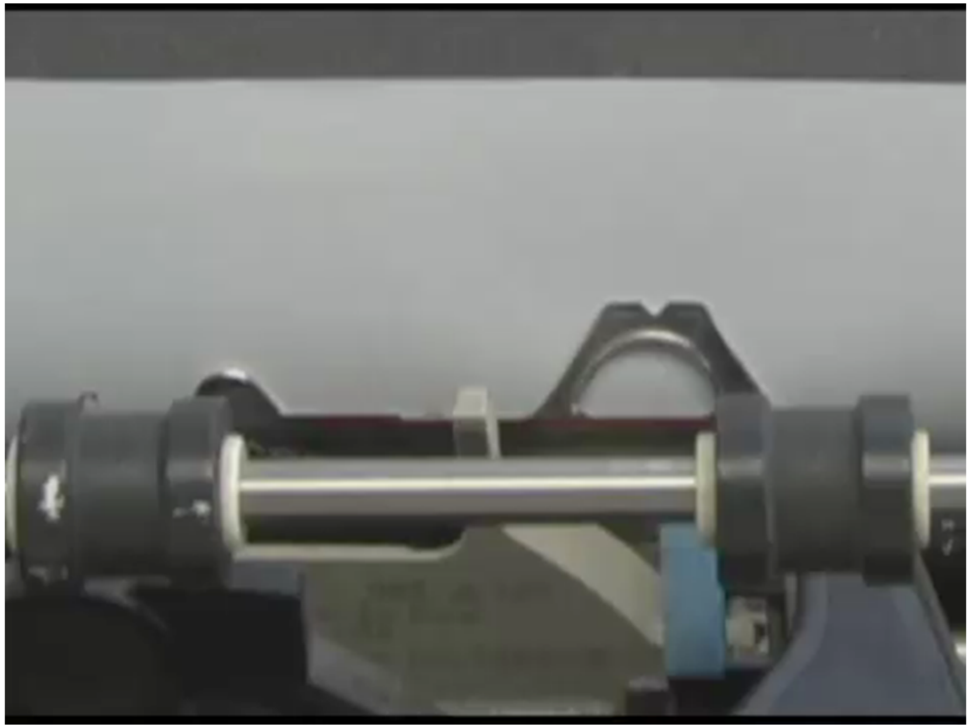
Search Engines
Chapter 1 – Introduction

12.4.2011

Felix Naumann

Anthropology Program at Kansas State University – Michael Wesch

- Information (r)evolution
  - □ http://www.youtube.com/watch?v=-4CV05HyAbM
  - □ http://ksuanth.weebly.com/wesch.html



Overview

- Introduction to team
- Organization
- Information Retrieval & Search Engines
- Overview of semester

Information Systems Team

Prof. Felix **Naumann**

Uwe **Draisbach**

Katrin **Heinrich**

Arvid **Heise**

Dustin **Lange**

Ziawasch **Abedjan**

project **DuDe**

Duplicate Detection

Schufa

Data Fusion

Data Profiling

project **Stratosphere**

Entity Search

Johannes **Lorey**

project
System P

Information Integration

project **HumMer**

ILM

Armin **Roth**

Peer Data Management Systems

Information Quality

Cloud Computing

IBM

Matching

Data Integration for Life Science Data Sources

ETL Management

project **PoSR**

project **M.ETL**

Service-Oriented Systems

project **Aladin**

Christoph **Böhm**

Tobias **Vogel**

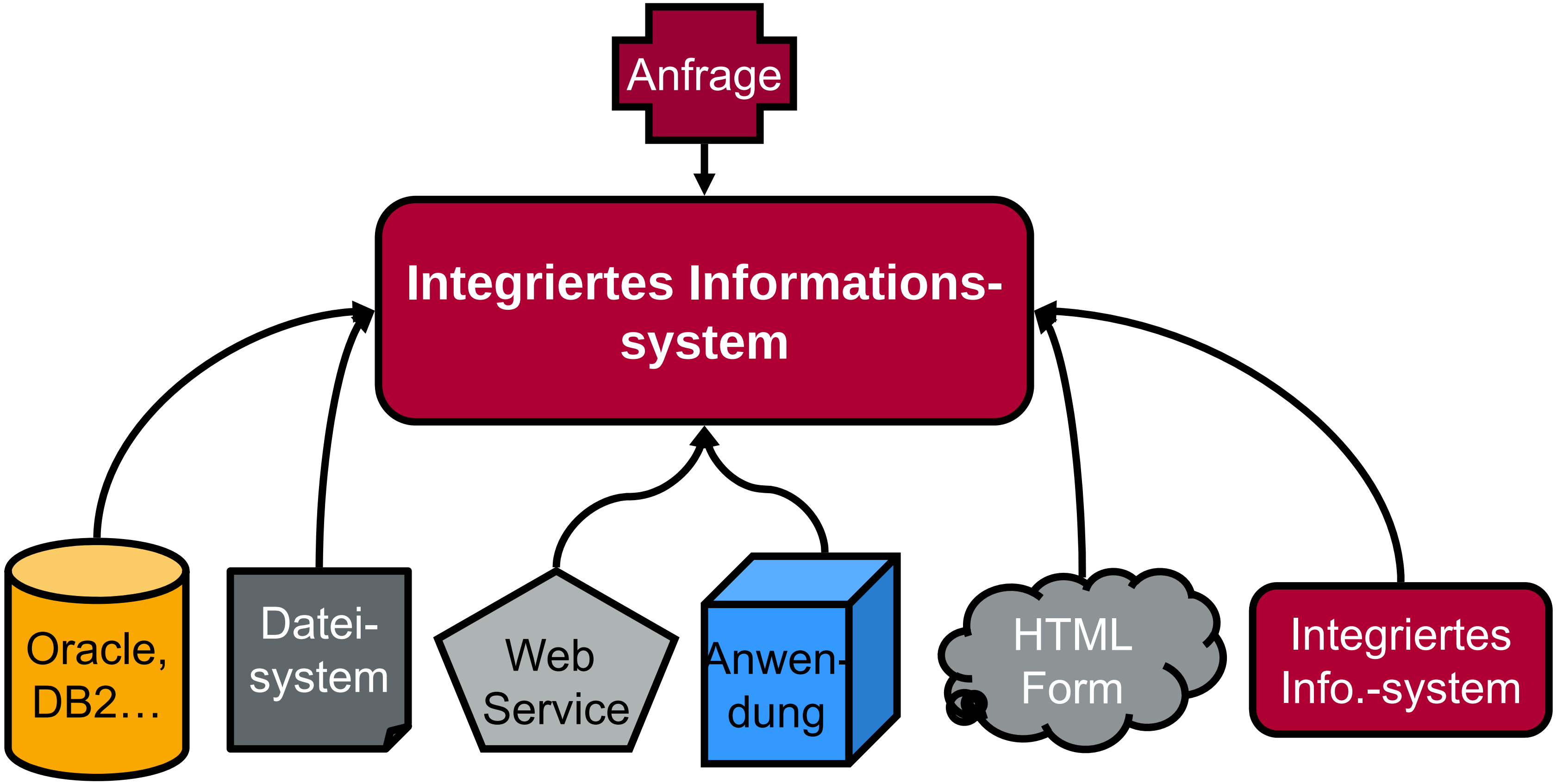
Forschungskolleg

Mohammed **AbuJarour**

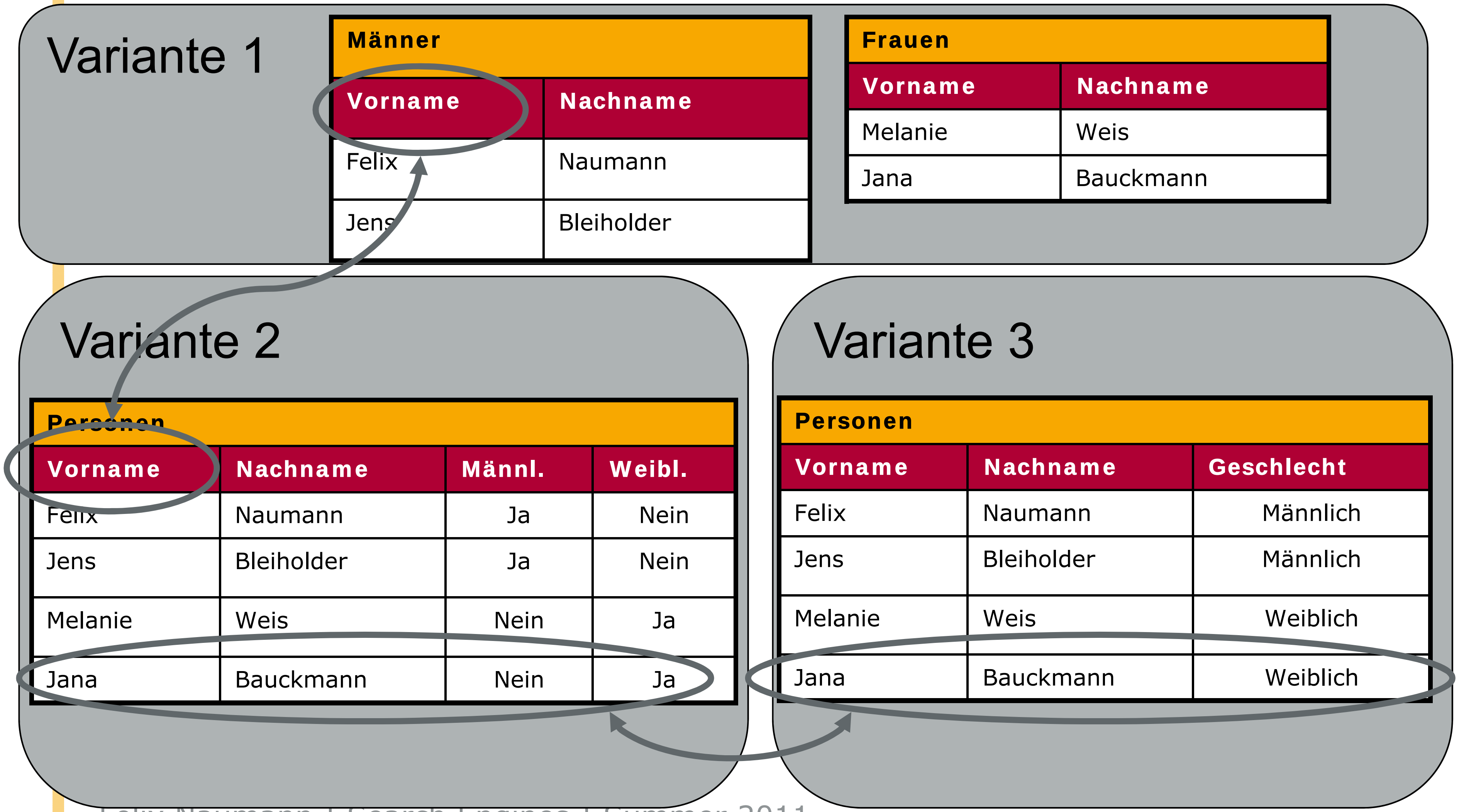
Jana **Bauckmann**

Alexander **Albrecht**

Integrierte Informationssysteme



Schematische und Daten-Heterogenität



Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen	
Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

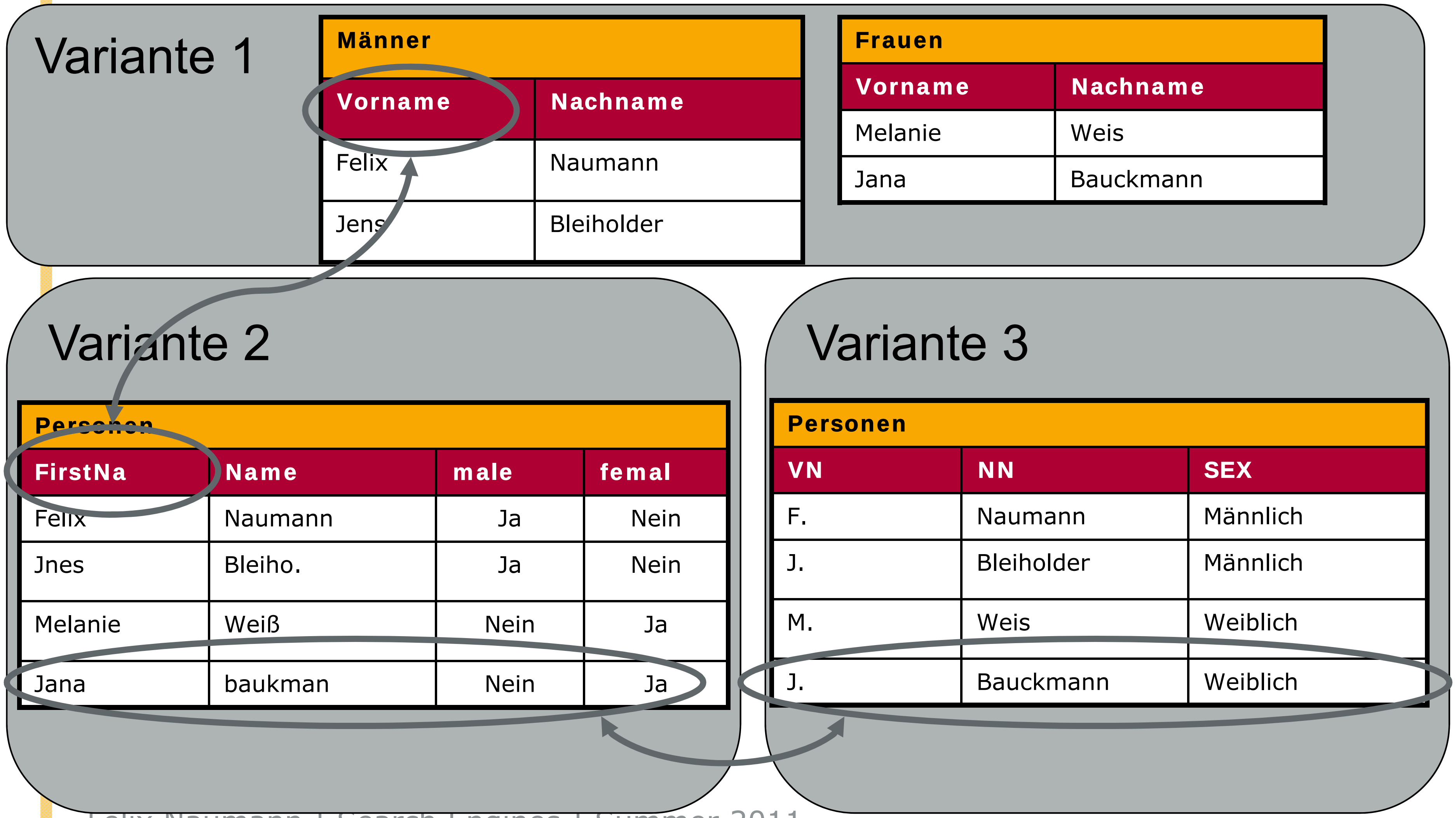
Variante 2

Personen			
Vorname	Nachname	Männl.	Weibl.
Felix	Naumann	Ja	Nein
Jens	Bleiholder	Ja	Nein
Melanie	Weis	Nein	Ja
Jana	Bauckmann	Nein	Ja

Variante 3

Personen		
Vorname	Nachname	Geschlecht
Felix	Naumann	Männlich
Jens	Bleiholder	Männlich
Melanie	Weis	Weiblich
Jana	Bauckmann	Weiblich

Schematische und Daten-Heterogenität



Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen	
Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

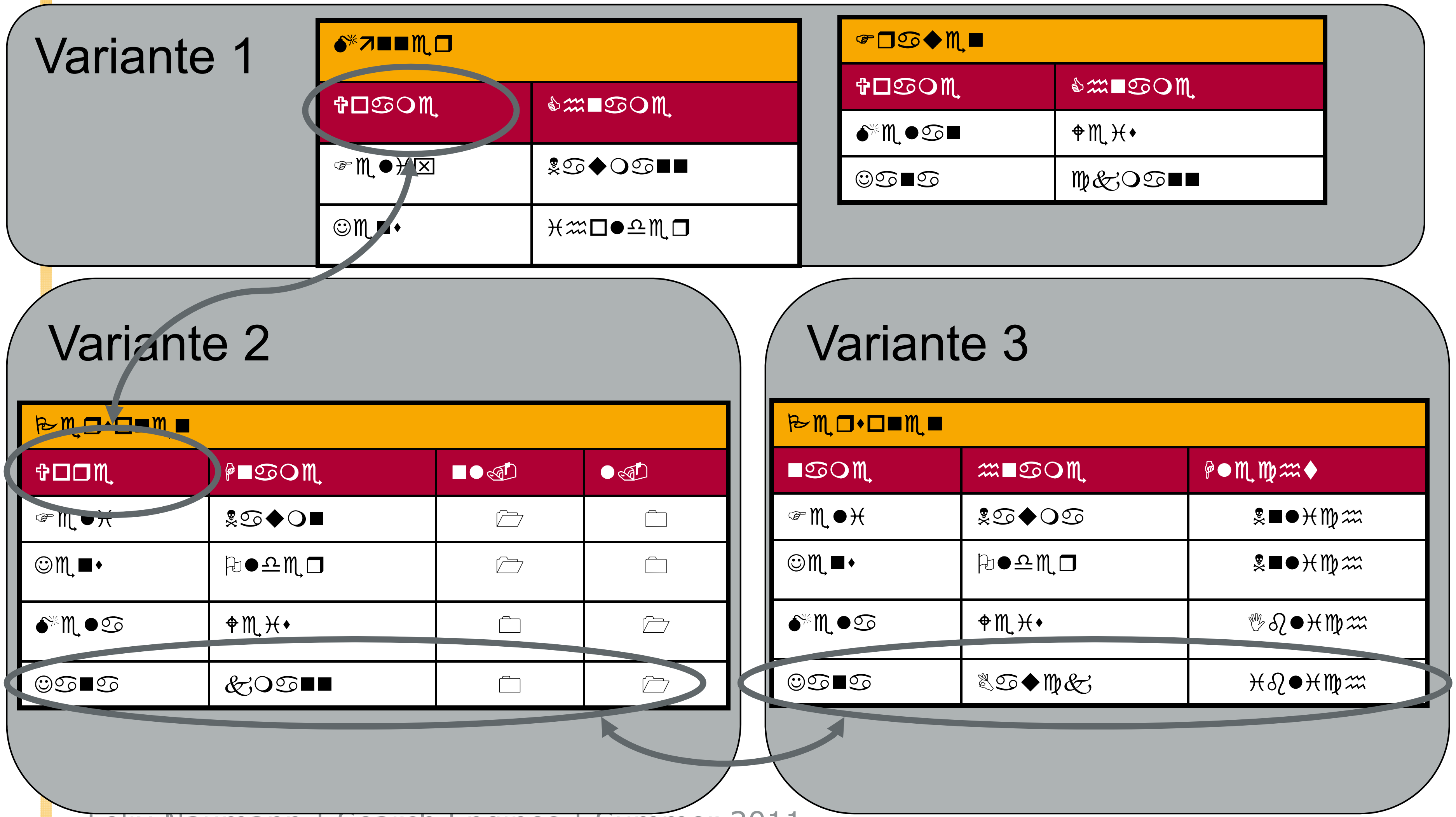
Variante 2

Personen			
FirstNa	Name	male	femal
Felix	Naumann	Ja	Nein
Jnes	Bleiho.	Ja	Nein
Melanie	Weiß	Nein	Ja
Jana	baukman	Nein	Ja

Variante 3

Personen		
VN	NN	SEX
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Schematische und Daten-Heterogenität



Variante 1

💣↗■■♏❒	
✞□♋❍♏	👍♒■♋❍♏
☞♏●♓⌧	☠♋◆❍♋■■
☺♏■⬧	♓♒□●♎♏❒

☞❒♋◆♏■	
✞□♋❍♏	👍♒■♋❍♏
💣♏●♋■	⌖♏♓⬧
☺♋■♋	♍&❍♋■■

Variante 2

🏱♏❒⬧□■♏■			
✞□❒♏	☝■♋❍♏	■●📫	●📫
☞♏●♓	☠♋◆❍■	📂	📁
☺♏■⬧	⚐●♎♏❒	📂	📁
💣♏●♋	⌖♏♓⬧	📁	📂
☺♋■♋	&❍♋■■	📁	📂

Variante 3

🏱♏❒⬧□■♏■		
■♋❍♏	♒■♋❍♏	☝●♏♍♒♦
☞♏●♓	☠♋◆❍♋	☠■●♓♍♒
☺♏■⬧	⚐●♎♏❒	☠■●♓♍♒
💣♏●♋	⌖♏♓⬧	✋♌●♓♍♒
☺♋■♋	👌♋◆♍&	♓♌●♓♍♒

Other courses in this semester

Lectures

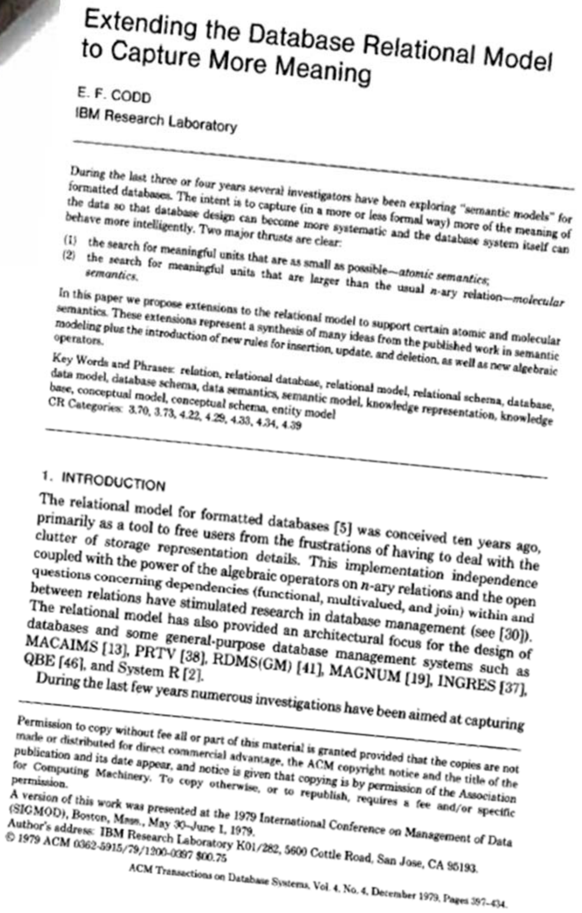
- DBS I
- Search engines

Seminars

- Bachelor: Beauty is our Business
- Bachelor: No SQL
- Master: Collaborative Filtering
- Masterproject: Duplikaterkennung auf GPUs

Bachelorprojects

- LongCat: Data Profiling (IBM)
- Cathbad: Faceted Search (Excentos)



Extending the Database Relational Model to Capture More Meaning

E. F. CODD
IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—*atomic semantics*;
(2) the search for meaningful units that are larger than the usual *n*-ary relation—*molecular semantics*.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model
CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on *n*-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].

During the last few years numerous investigations have been aimed at capturing

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.
Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.
© 1979 ACM 0362-5915/79/1200-0397 $00.75

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 397–434.

Overview

- Introduction to team
- Organization
- Information Retrieval & Search Engines
- Overview of semester

Dates and examination

- Lectures
  - Tuesday 9:15 – 10:45
  - Thursdays 9:15 – 10:45
- Practical work
  - Selected dates – see webpage
- First lecture
  - 12.4.2011
- Last lecture
  - 21.7.2011
- Holidays
  - 2.6. Ascension

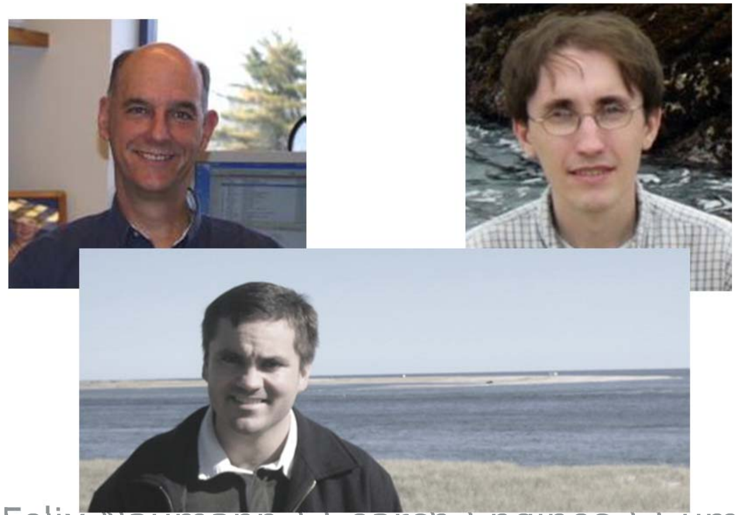
- Exam
  - Oral or written (tbd)
  - First 2 weeks after lectures end
- 7 exercise courses
  - TAs: Dustin Lange
  - Practical work and presentations
  - Teams of two students
- Prerequisites
  - For participation
    - Basic knowledge in databases
  - For exam
    - Attendance of lectures
    - Active participation in exercise courses
    - Successful work on all practical assignments
      - "Success" to be defined

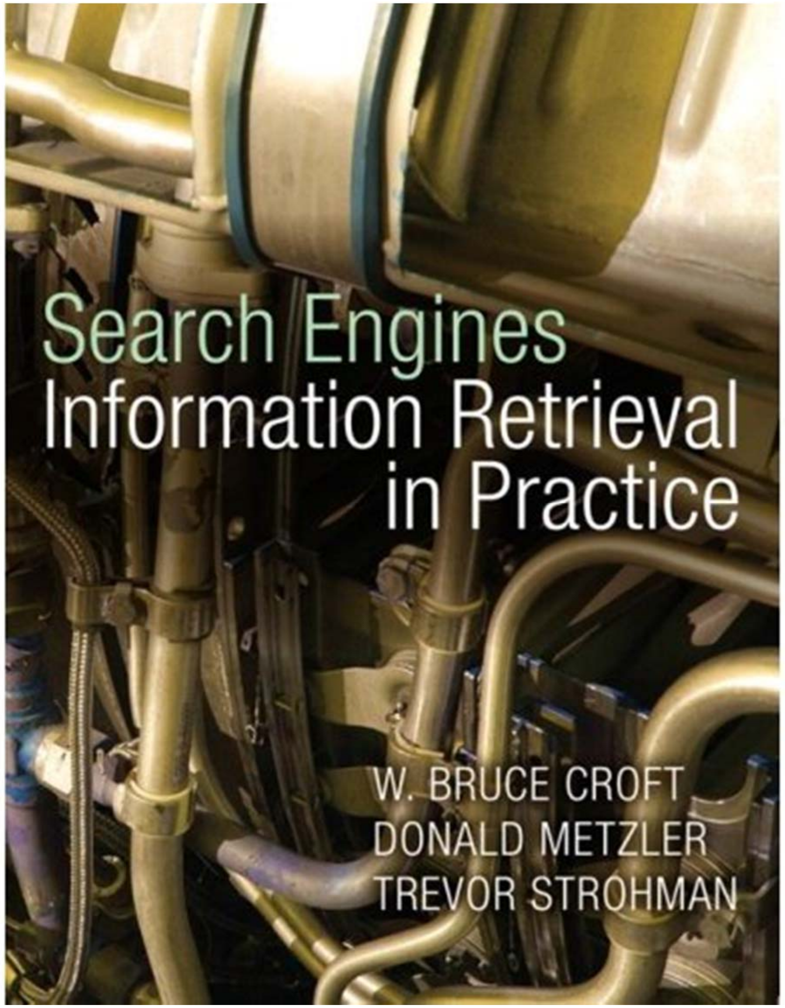
Feedback

- Evaluation at end of semester
- Q&A anytime!
  - During lecture
  - Directly after lecture
  - Consultation: Tuesdays 13-15
  - Email: naumann@hpi.uni-potsdam.de
- Hints on improvements
  - wrt.
    - Slides and their presentation
    - Web information
  - After lecture or during consultation hours
  - Or via email: naumann@hpi.uni-potsdam.de

Textbook

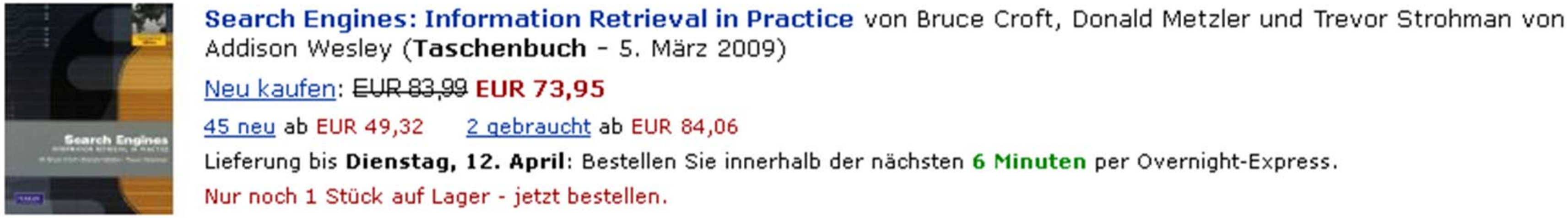
- Search Engines: Information Retrieval in Practice
  - □ **Bruce Croft**
  - □ **Donald Metzler**
  - □ **Trevor Strohman**
  - □ http://ciir.cs.umass.edu/
- Addison-Wesley, 2010





Textbook

- 20 copies in library
- 73,95 € at amazon.de
  - □ Ouch, see http://www.newyorker.com/archive/2005/11/07/051107ta_talk_surowiecki
  - □ „When professors decide which books to assign, the main consideration, they would say, is quality, not price, so any competition occurs on the basis of features rather than of cost. […] When price is no object, professors might as well choose the fanciest textbook around.“
  - □ But: Free delivery…



Search Engines: Information Retrieval in Practice von Bruce Croft, Donald Metzler und Trevor Strohman von Addison Wesley (**Taschenbuch** - 5. März 2009)

Neu kaufen: ~~EUR 83,99~~ **EUR 73,95**

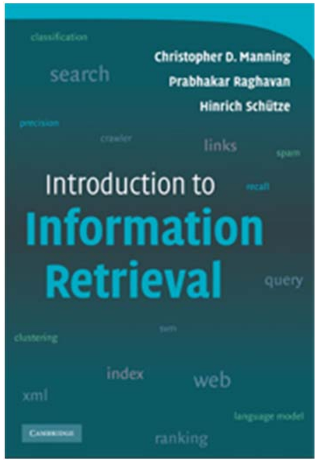
45 neu ab EUR 49,32 2 gebraucht ab EUR 84,06

Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **6 Minuten** per Overnight-Express.

Nur noch 1 Stück auf Lager - jetzt bestellen.

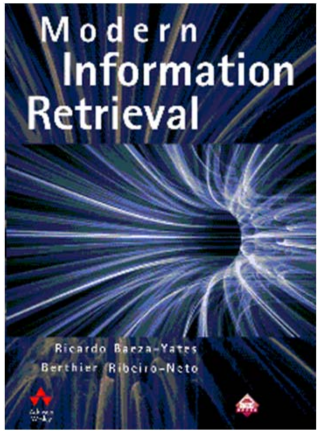
Other literature

- *Introduction to Information Retrieval*
  - □ Cambridge University Press, 2008.
  - □ Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze.
  - □ http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html
- *Modern Information Retrieval*
  - □ Addison Wesley (2010)
  - □ Ricardo Baeza-Yates und Berthier Ribeiro-Neto



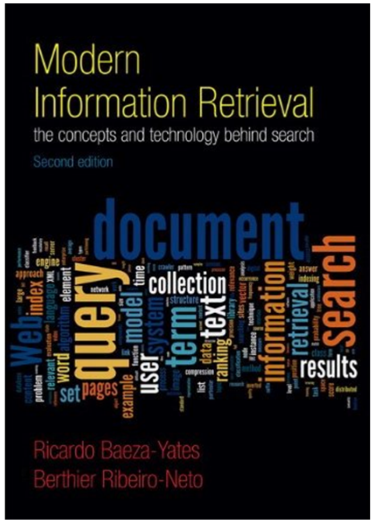
Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze

Introduction to Information Retrieval



Modern Information Retrieval

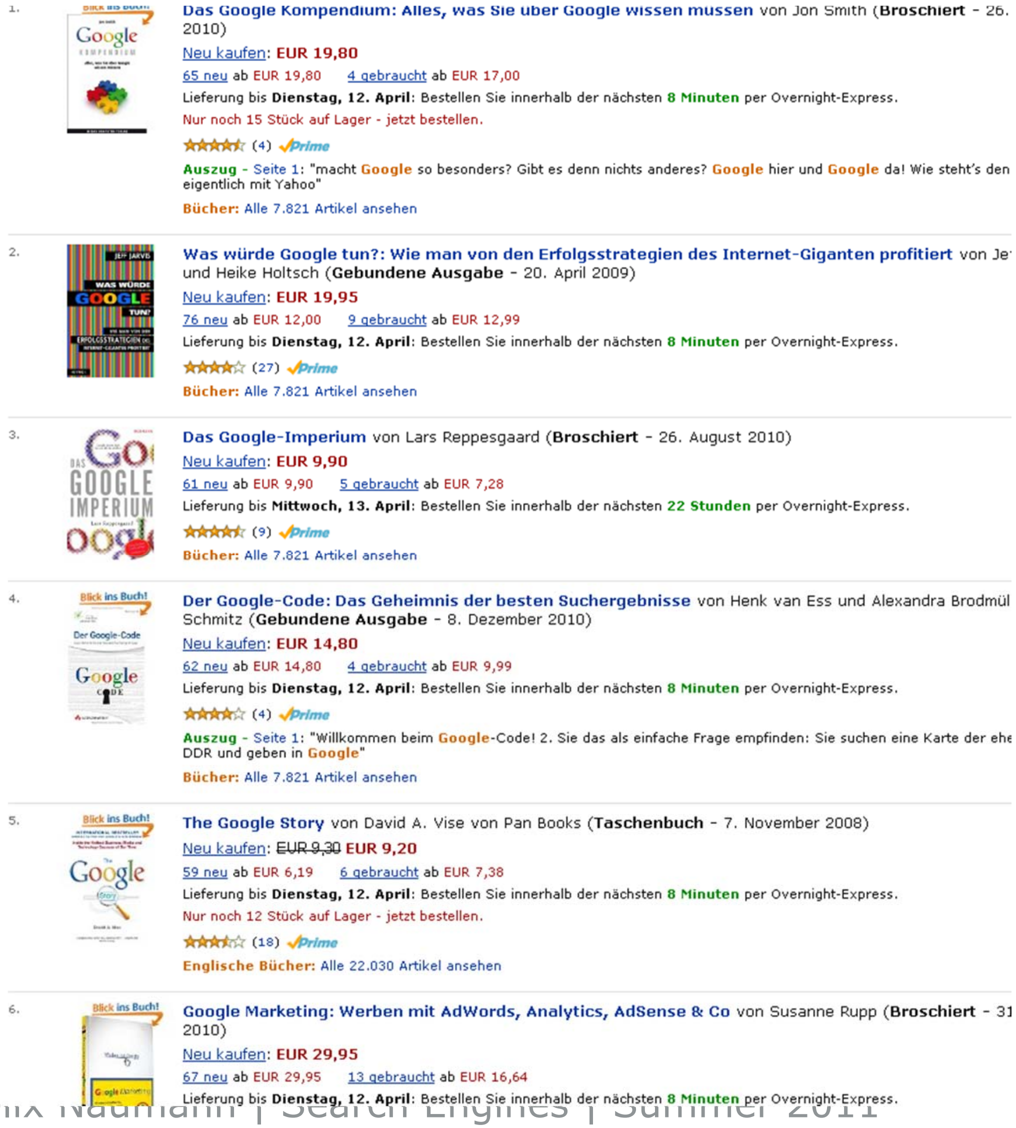
Ricardo Baeza-Yates
Berthier Ribeiro-Neto



Modern Information Retrieval
the concepts and technology behind search
Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Other literature - background



1. Das Google Kompendium: Alles, was Sie über Google wissen müssen von Jon Smith (**Broschiert** - 26. 2010)

Neu kaufen: **EUR 19,80**

65 neu ab EUR 19,80   4 gebraucht ab EUR 17,00

Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **8 Minuten** per Overnight-Express.

Nur noch 15 Stück auf Lager - jetzt bestellen.

(4) Prime

Auszug - Seite 1: "macht **Google** so besonders? Gibt es denn nichts anderes? **Google** hier und **Google** da! Wie steht's den eigentlich mit Yahoo"

Bücher: Alle 7.821 Artikel ansehen

2. Was würde Google tun?: Wie man von den Erfolgsstrategien des Internet-Giganten profitiert von Je und Heike Holtsch (**Gebundene Ausgabe** - 20. April 2009)

Neu kaufen: **EUR 19,95**

76 neu ab EUR 12,00   9 gebraucht ab EUR 12,99

Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **8 Minuten** per Overnight-Express.

(27) Prime

Bücher: Alle 7.821 Artikel ansehen

3. Das Google-Imperium von Lars Reppesgaard (**Broschiert** - 26. August 2010)

Neu kaufen: **EUR 9,90**

61 neu ab EUR 9,90   5 gebraucht ab EUR 7,28

Lieferung bis **Mittwoch, 13. April**: Bestellen Sie innerhalb der nächsten **22 Stunden** per Overnight-Express.

(9) Prime

Bücher: Alle 7.821 Artikel ansehen

4. Der Google-Code: Das Geheimnis der besten Suchergebnisse von Henk van Ess und Alexandra Brodmül Schmitz (**Gebundene Ausgabe** - 8. Dezember 2010)

Neu kaufen: **EUR 14,80**

62 neu ab EUR 14,80   4 gebraucht ab EUR 9,99

Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **8 Minuten** per Overnight-Express.

(4) Prime

Auszug - Seite 1: "Willkommen beim **Google**-Code! 2. Sie das als einfache Frage empfinden: Sie suchen eine Karte der ehe DDR und geben in **Google**"

Bücher: Alle 7.821 Artikel ansehen

5. The Google Story von David A. Vise von Pan Books (**Taschenbuch** - 7. November 2008)

Neu kaufen: ~~EUR 9,30~~ **EUR 9,20**

59 neu ab EUR 6,19   6 gebraucht ab EUR 7,38

Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **8 Minuten** per Overnight-Express.

Nur noch 12 Stück auf Lager - jetzt bestellen.

(18) Prime

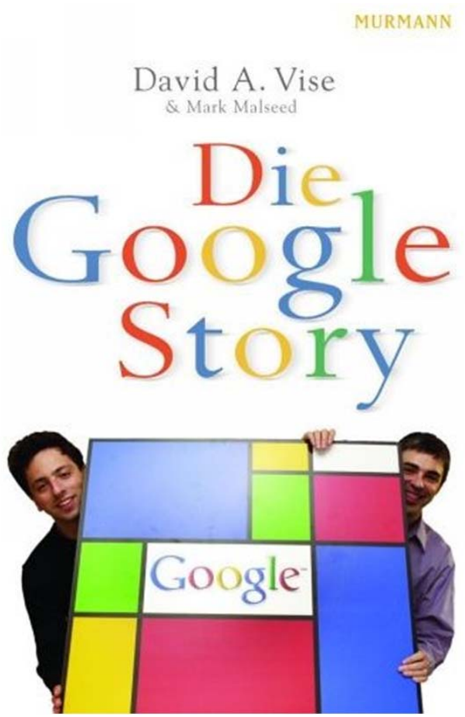
Englische Bücher: Alle 22.030 Artikel ansehen

6. Google Marketing: Werben mit AdWords, Analytics, AdSense & Co von Susanne Rupp (**Broschiert** - 31 2010)

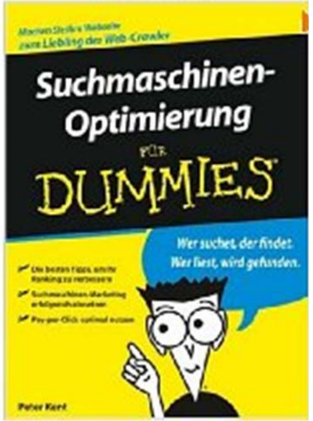
Neu kaufen: **EUR 29,95**

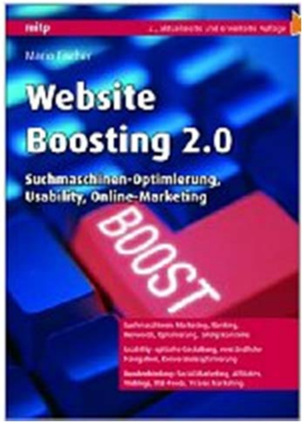
67 neu ab EUR 29,95   13 gebraucht ab EUR 16,64

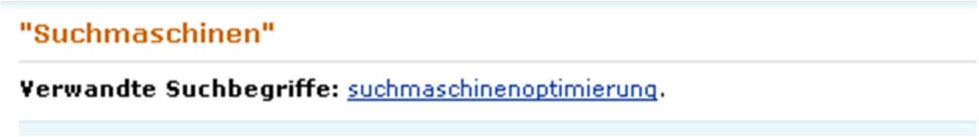
Lieferung bis **Dienstag, 12. April**: Bestellen Sie innerhalb der nächsten **8 Minuten** per Overnight-Express.



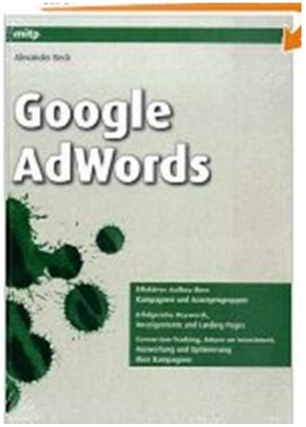
Other literature – Search Engine Optimization






"Suchmaschinen"

Verwandte Suchbegriffe: suchmaschinenoptimierung.



Introduction – Audience

- Which semester?
- HPI or IfI?
- Erasmus / foreign students?
- DB knowledge?
- Other relevant courses?
  - Semantic Web
  - Information Retrieval
- Your motivation?
  - Search engine optimization
  - Behind the scenes
  - Build your own search engine
  - Find a good job
  - Gain knowledge? Start research?

Overview

- Introduction to team
- Organization
- ➡ Information Retrieval & Search Engines
- Overview of semester

Search and Information Retrieval

- Search on the Web1 is a daily activity for many people throughout the world.
  - □ Google: 34,000 searches per second (2 million per minute; 121 million per hour; 3 billion per day; 88 billion per month, figures rounded)
  - □ Yahoo: 3,200 searches per second (194,000 per minute; 12 million per hour; 280 million per day; 8.4 billion per month, figures rounded)
  - □ Bing: 927 searches per second (56,000 per minute; 3 million per hour; 80 million per day; 2.4 billion per month, figures rounded)
- Search and communication are most popular uses of the computer.
- Applications involving search are everywhere.
- The field of computer science that is most involved with R&D for search is information retrieval (IR).

[1] or is it web?

http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009

Brazil

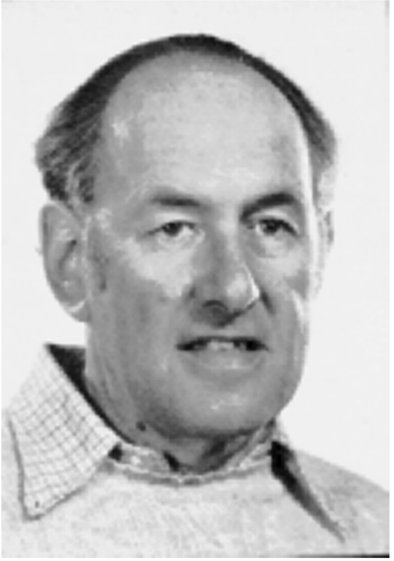
- **Sam Lowry**: My name's Lowry. Sam Lowry. I've been told to report to Mr. Warren.
- **Porter - Information Retrieval**: Thirtieth floor, sir. You're expected.
- **Sam Lowry**: Um... don't you want to search me?
- **Porter - Information Retrieval**: No sir.
- **Sam Lowry**: Do you want to see my ID?
- **Porter - Information Retrieval**: No need, sir.
- **Sam Lowry**: But I could be anybody.
- **Porter - Information Retrieval**: No you couldn't sir. This is Information Retrieval.

- Sources
  - http://en.wikiquote.org/wiki/Brazil_(film)
  - http://www.youtube.com/watch?v=LFlFIG22Y9E&hl=de

Information Retrieval

"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information." (Salton, 1968)

- General definition that can be applied to many types of information and search applications
  - Still appropriate after 40 years.
- Primary focus of IR since the 50s has been on *text* and *documents*



http://www.cs.cornell.edu/Info/Department/Annual95/Faculty/Salton.html

What is a Document?

- Examples:
  - Web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (≈ attributes in DB)
    - Papers: title, author, date
    - Email: subject, sender, destination, date

Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*).
  - Bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics and data types to queries in order to find matches
  - Joins, selection predicates
  - Even duplicate detection is easier.
- Text is more difficult, because unstructured

Documents vs. Database Records

- Example bank database query
  - *Find records with balance > €50,000 in branches located in 14482 Potsdam.*
  - Matches easily found by comparison with field values of records
- Example search engine query
  - *bank scandals in western Germany*
  - This text must be compared to the text of *many, entire* news stories
    - Only "fields" might be *title* and *location*
- Defining the meaning of "balance" is much easier than defining "bank scandal".

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the **core issue** of information retrieval
- Exact matching of words is not enough
  - □ Many different ways to write the same thing in a "natural language" like English
    - ◇ Does a news story containing the text *"bank director in Potsdam steals funds"* match the query "*bank scandals in western Germany*"?
  - □ Some stories are better matches than others
    - ◇ Ranking vs. Boolean
- Defining the **meaning** of a word, a sentence, a paragraph, or a story is more difficult than defining the meaning of a database field.

Dimensions of IR

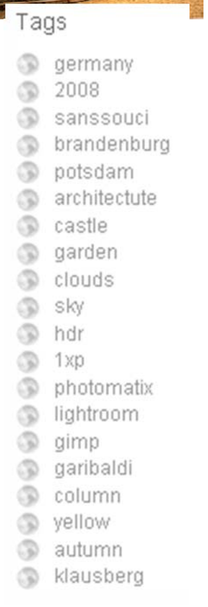
- IR is more than just text, and more than just web search
  - □ although these are central
- People doing IR work with different media, different types of search applications, and different tasks

- Three dimensions of IR
  1. Content
  2. Applications
  3. Tasks

The Content Dimension

- Textual data, but…
- New applications increasingly involve new media
  - Video, photos, music, speech
  - Scanned documents (for legal purposes)
- Like text, content is difficult to describe and compare
  - Text may be used to represent them (e.g., tags)
- IR approaches to search and evaluation are appropriate.

http://www.flickr.com/photos/garibaldi/3122956960/



Tags

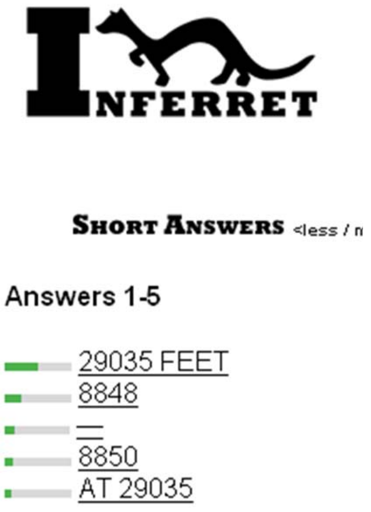
- germany
- 2008
- sanssouci
- brandenburg
- potsdam
- architectute
- castle
- garden
- clouds
- sky
- hdr
- 1xp
- photomatix
- lightroom
- gimp
- garibaldi
- column
- yellow
- autumn
- klausberg

The Application Dimension

- Web search
  - Most common
- Vertical search
  - Restricted domain/topic
  - Books, movies, suppliers
- Enterprise search
  - Corporate intranet
  - Databases, emails, web pages, documentation, code, wikis, tags, directories, presentations, spreadsheets
- Desktop search
  - Personal enterprise search
  - See above plus recent web pages
- P2P search
  - No centralized control
  - File sharing, shared locality
- Literature search
- Forum search
- …

The Task Dimension

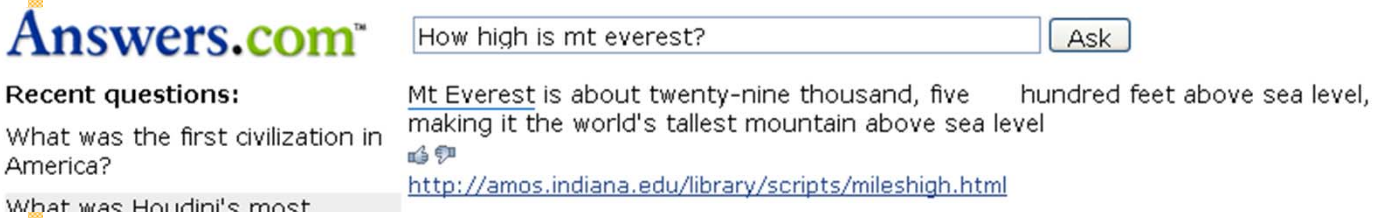
- User queries / ad-hoc search
  - □ Range of query enormous, not pre-specified
- Filtering
  - □ Given a profile (interests), notify about interesting news stories
  - □ Identify relevant user profiles for a new document
- Classification / categorization
  - □ Automatically assign text to one or more classes of a given set.
  - □ Identify relevant labels for documents
- Question answering
  - □ Similar to search
  - □ Automatically answer a question posed in natural language
  - □ Provide concrete answer, not list of documents.



SHORT ANSWERS <less / m

Answers 1-5

- 29035 FEET
- 8848
- —
- 8850
- AT 29035



How high is mt everest?

Recent questions:

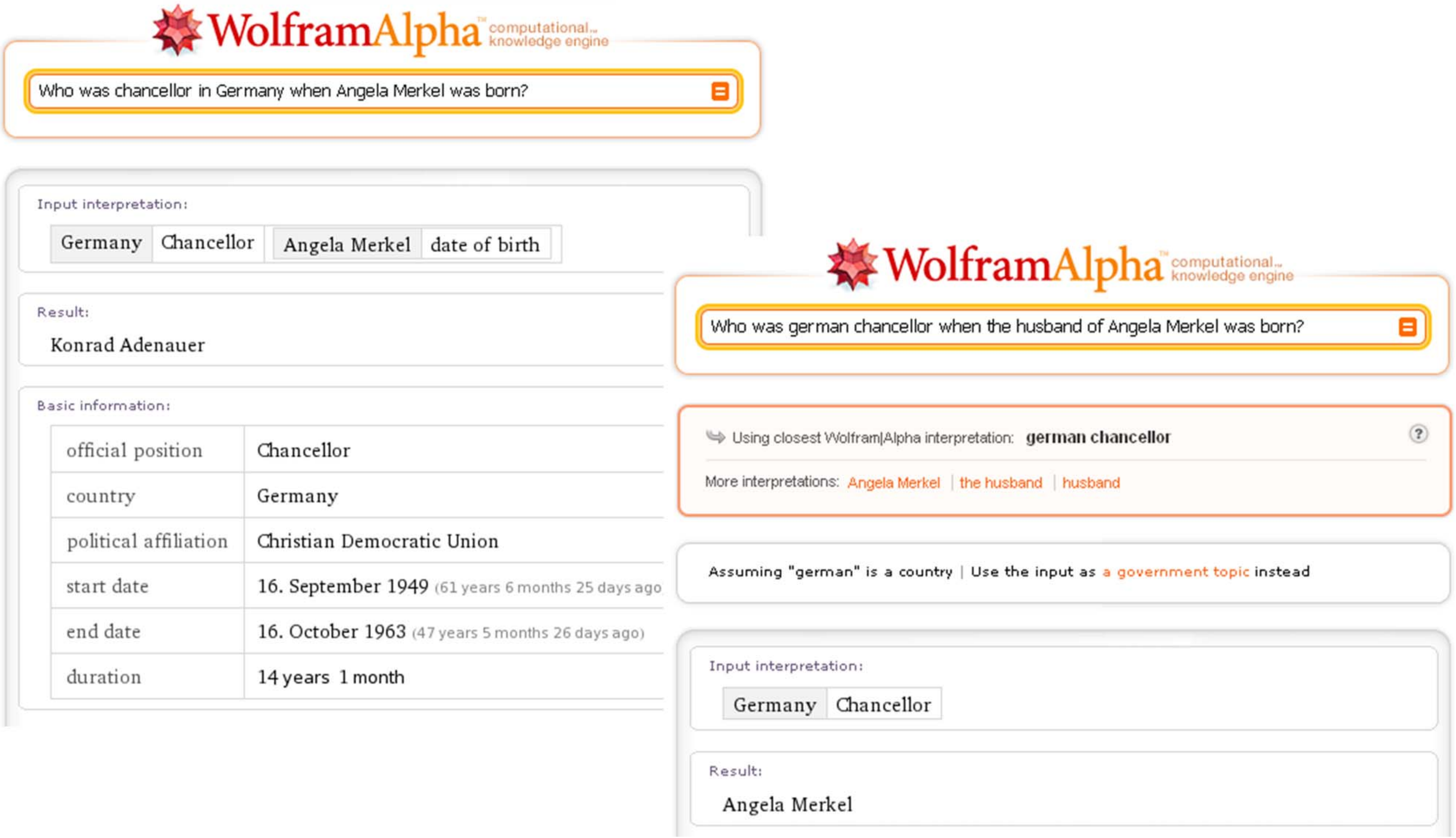
What was the first civilization in America?

What was Houdini's most

Mt Everest is about twenty-nine thousand, five hundred feet above sea level, making it the world's tallest mountain above sea level

http://amos.indiana.edu/library/scripts/mileshigh.html

More question answering



Who was chancellor in Germany when Angela Merkel was born?

Input interpretation:

Germany | Chancellor | Angela Merkel | date of birth

Result:

Konrad Adenauer

Basic information:

official position	Chancellor
country	Germany
political affiliation	Christian Democratic Union
start date	16. September 1949 (61 years 6 months 25 days ago
end date	16. October 1963 (47 years 5 months 26 days ago)
duration	14 years 1 month

Who was german chancellor when the husband of Angela Merkel was born?

Using closest Wolfram|Alpha interpretation: **german chancellor**

More interpretations: Angela Merkel | the husband | husband

Assuming "german" is a country | Use the input as a government topic instead

Input interpretation:

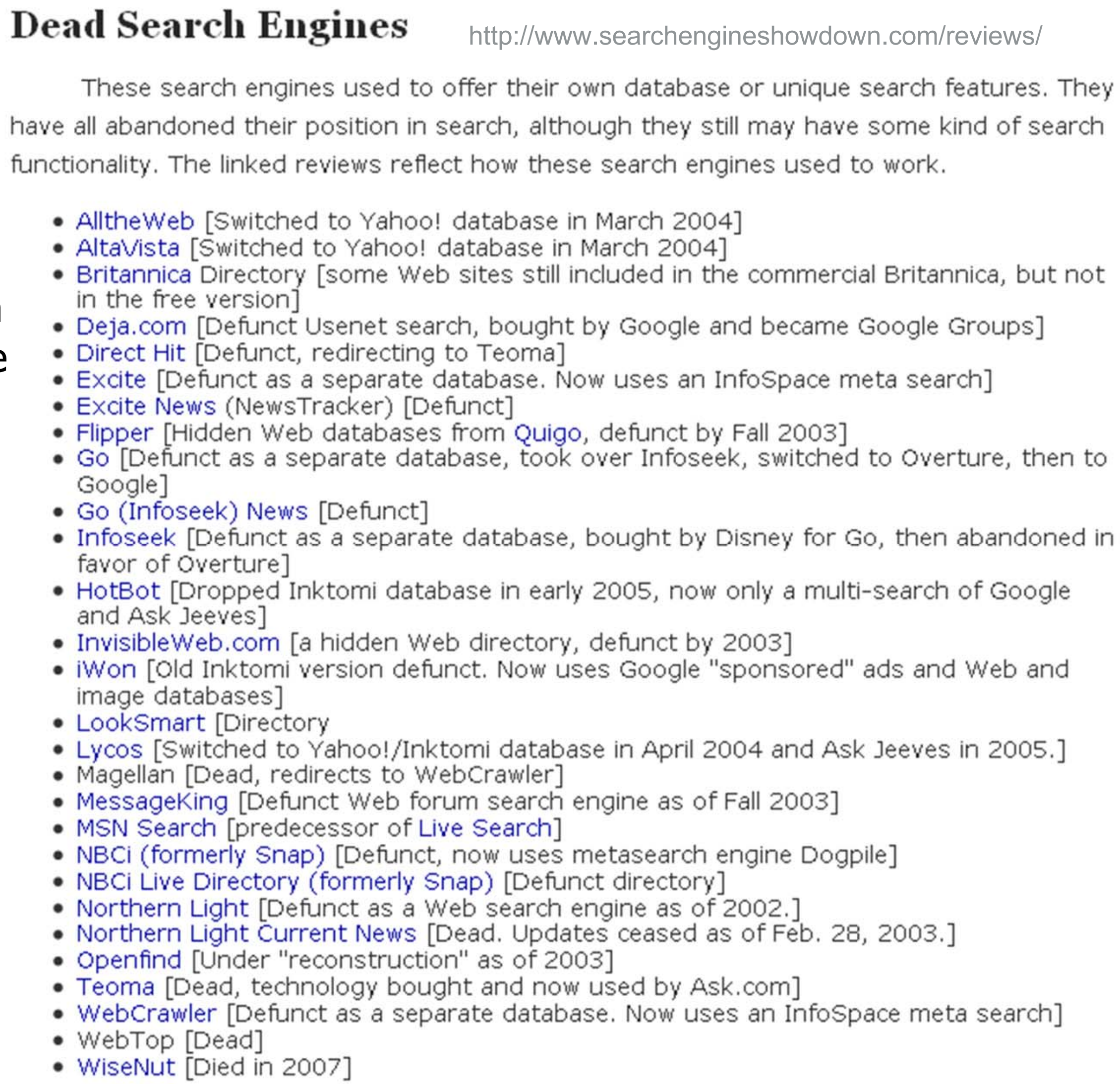
Germany | Chancellor

Result:

Angela Merkel

Big Issues in IR

- Relevance
  - A relevant document contains the information a user was looking for when he/she submitted the query.
- Evaluation
  - How well does the ranking meet the expectation of the user.
- Users and information needs
  - Users of a search engine are the ultimate judges of quality.



Dead Search Engines

http://www.searchengineshowdown.com/reviews/

These search engines used to offer their own database or unique search features. They have all abandoned their position in search, although they still may have some kind of search functionality. The linked reviews reflect how these search engines used to work.

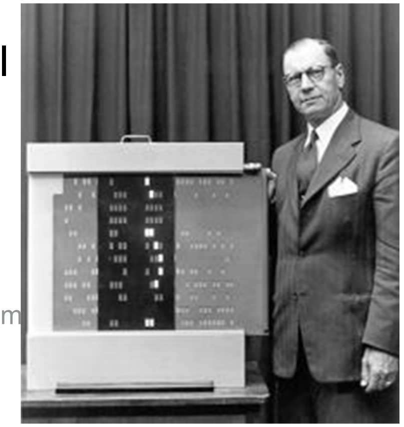
- AlltheWeb [Switched to Yahoo! database in March 2004]
- AltaVista [Switched to Yahoo! database in March 2004]
- Britannica Directory [some Web sites still included in the commercial Britannica, but not in the free version]
- Deja.com [Defunct Usenet search, bought by Google and became Google Groups]
- Direct Hit [Defunct, redirecting to Teoma]
- Excite [Defunct as a separate database. Now uses an InfoSpace meta search]
- Excite News (NewsTracker) [Defunct]
- Flipper [Hidden Web databases from Quigo, defunct by Fall 2003]
- Go [Defunct as a separate database, took over Infoseek, switched to Overture, then to Google]
- Go (Infoseek) News [Defunct]
- Infoseek [Defunct as a separate database, bought by Disney for Go, then abandoned in favor of Overture]
- HotBot [Dropped Inktomi database in early 2005, now only a multi-search of Google and Ask Jeeves]
- InvisibleWeb.com [a hidden Web directory, defunct by 2003]
- iWon [Old Inktomi version defunct. Now uses Google "sponsored" ads and Web and image databases]
- LookSmart [Directory
- Lycos [Switched to Yahoo!/Inktomi database in April 2004 and Ask Jeeves in 2005.]
- Magellan [Dead, redirects to WebCrawler]
- MessageKing [Defunct Web forum search engine as of Fall 2003]
- MSN Search [predecessor of Live Search]
- NBCi (formerly Snap) [Defunct, now uses metasearch engine Dogpile]
- NBCi Live Directory (formerly Snap) [Defunct directory]
- Northern Light [Defunct as a Web search engine as of 2002.]
- Northern Light Current News [Dead. Updates ceased as of Feb. 28, 2003.]
- Openfind [Under "reconstruction" as of 2003]
- Teoma [Dead, technology bought and now used by Ask.com]
- WebCrawler [Defunct as a separate database. Now uses an InfoSpace meta search]
- WebTop [Dead]
- WiseNut [Died in 2007]

Relevance

- Simple (and simplistic) definition:
  A relevant document contains the information that a person was looking for when they submitted a query to the search engine.
- Many factors influence a person’s decision about what is relevant
  - Task at hand, context, novelty, style, serendipity
- *Topical relevance* (same topic)
  - “*Storm in Potsdam last Sunday*” is topically relevant to query “*Wetterereignisse*”…
- Vs. *user relevance* (everything else)
  - … but might not be relevant to user because
    - Read it before
    - Is five years old
    - Is in a foreign language, etc.

Relevance

- *Retrieval models* define a view of relevance
  - Formal representation of the process of matching a query and a document
  - Simple text matching as in DBMS or UNIX **`grep`** is not sufficient: Vocabulary mismatch problem (synonyms and homonyms)
- *Ranking algorithms* used in search engines are based on retrieval models
  - Produce ranked list of documents
  - Real-world search engines consider topical and user relevance
- Most models describe statistical properties of text rather than linguistic
  - i.e. counting simple text features, such as words, instead of parsing and analyzing the sentences
  - Statistical approach to text processing started with Hans Peter Luhn in the 50s
    - Statistical view of text only recently popular in Natural Language Processing (NLP)
  - Linguistic features can be part of a statistical model



http://www.libsci.sc.edu/bob/chemnet/chist10.htm

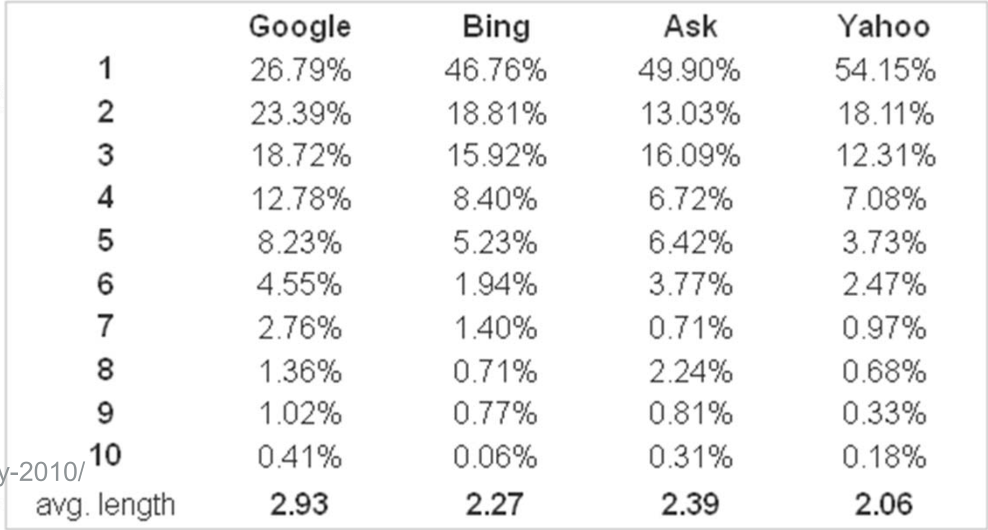
http://www.lunometer.com/

Evaluation

- Experimental procedures and measures for comparing system output with user expectations
  - Originated in Cranfield experiments in the 60s
    - First large scale "benchmark"
- IR evaluation methods now used in many fields
- Typically use *test collection* (corpus) of documents, queries, and relevance judgments
  - Most commonly used are TREC collections (Text REtrieval Conf.)
- *Recall* and *precision* are two examples of effectiveness measures
  - Precision: Proportion of retrieved documents that are relevant
  - Recall: Proportion of relevant documents that are retrieved
    - Assumption: All relevant documents are known. Ouch!
  - F-Measure: Harmonic mean of precision and recall
- Weblog data and clickthrough data to evaluate retrieval models and search engines.

Users and Information Needs

- Search evaluation is user-centered
- Keyword queries are often poor descriptions of actual information needs
  - Query for "cats" could mean places to buy cats or the musical.
  - Search queries (in particular one-word queries) are under-specified.
- Interaction and context are important for understanding user intent
- Query refinement technique
  - *query expansion*
  - *query suggestion*
  - *relevance feedback*
- improve ranking



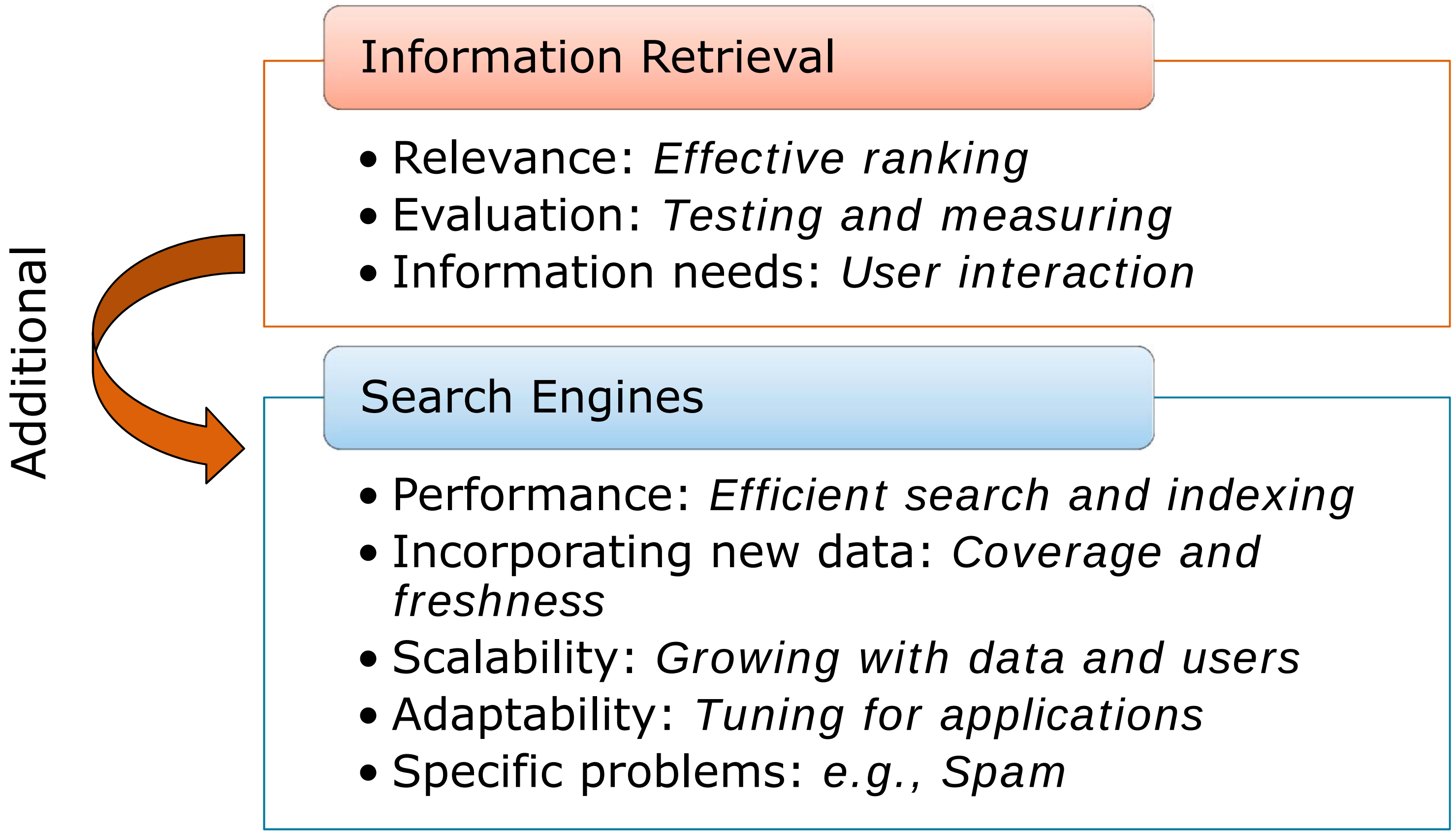
	Google	Bing	Ask	Yahoo
1	26.79%	46.76%	49.90%	54.15%
2	23.39%	18.81%	13.03%	18.11%
3	18.72%	15.92%	16.09%	12.31%
4	12.78%	8.40%	6.72%	7.08%
5	8.23%	5.23%	6.42%	3.73%
6	4.55%	1.94%	3.77%	2.47%
7	2.76%	1.40%	0.71%	0.97%
8	1.36%	0.71%	2.24%	0.68%
9	1.02%	0.77%	0.81%	0.33%
10	0.41%	0.06%	0.31%	0.18%
avg. length	**2.93**	**2.27**	**2.39**	**2.06**

http://blog.alessiosignorini.com/2010/02/average-query-length-february-2010/

IR and Search Engines

- A **search engine** is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others exist
  - Web search: Crawl terabyte of web pages, provide sub-second response times, millions of queries
  - Enterprise search: variety of sources, search, perform data mining / clustering
  - Desktop search: rapidly incorporate new documents, many types of documents, intuitive interface
  - MEDLINE, online medical literature search since 70s
  - *Open source* search engines are important for research and development
    - Lucene, Lemur/Indri, Galago
- Big issues include main IR issues but also some others…

IR and Search Engines



Information Retrieval

- Relevance: *Effective ranking*
- Evaluation: *Testing and measuring*
- Information needs: *User interaction*

Search Engines

- Performance: *Efficient search and indexing*
- Incorporating new data: *Coverage and freshness*
- Scalability: *Growing with data and users*
- Adaptability: *Tuning for applications*
- Specific problems: *e.g., Spam*

Performance

- Measuring and improving the **efficiency** of search
  - Reduce *response time*
  - Increase *query throughput*
  - Increase *indexing speed*
- **Indexes** are data structures designed to improve search efficiency.
  - Designing and implementing them are major issues for search engines.

Dynamic data

- The "collection" for most real applications is constantly changing in terms of updates, additions, deletions.
  - e.g., Web pages
- Acquiring or "crawling" the documents is a major task
  - Typical measures are *coverage* (how much has been indexed)
  - and *recency/freshness* (how recently was it indexed).
- Updating the indexes while processing queries is also a design issue

Scalability

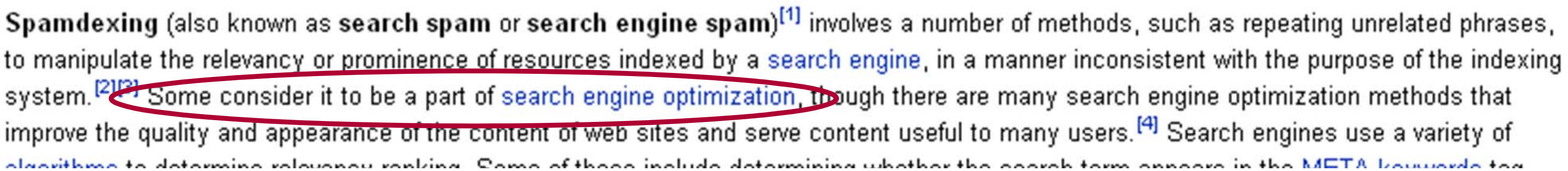
- Making everything work with millions of users every day, and many terabytes of documents
- Distributed processing is essential
- But: Large ≠ scalable
  - Scale gracefully
- Google in 2006
  - > 25 billion pages
  - 400M queries/day
- Google in 2008
  - 1 trillion pages (1,000,000,000,000)
    - http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Adaptability

- Changing and tuning search engine components
  - ranking algorithm
  - indexing strategy
  - interface for different applications
- Adapt to different requirements for different applications / users
  - New APIs
  - New uses for search

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
  - □ e.g., spamdexing or term spam, link spam, "optimization"
  - □ http://en.wikipedia.org/wiki/Spamdexing
- New subfield called *adversarial IR*, since spammers are "adversaries" with different goals



Spamdexing (also known as **search spam** or **search engine spam**)[1] involves a number of methods, such as repeating unrelated phrases, to manipulate the relevancy or prominence of resources indexed by a search engine, in a manner inconsistent with the purpose of the indexing system.[2][3] Some consider it to be a part of search engine optimization, though there are many search engine optimization methods that improve the quality and appearance of the content of web sites and serve content useful to many users.[4] Search engines use a variety of algorithms to determine relevancy ranking. Some of these include determining whether the search term appears in the META keywords tag

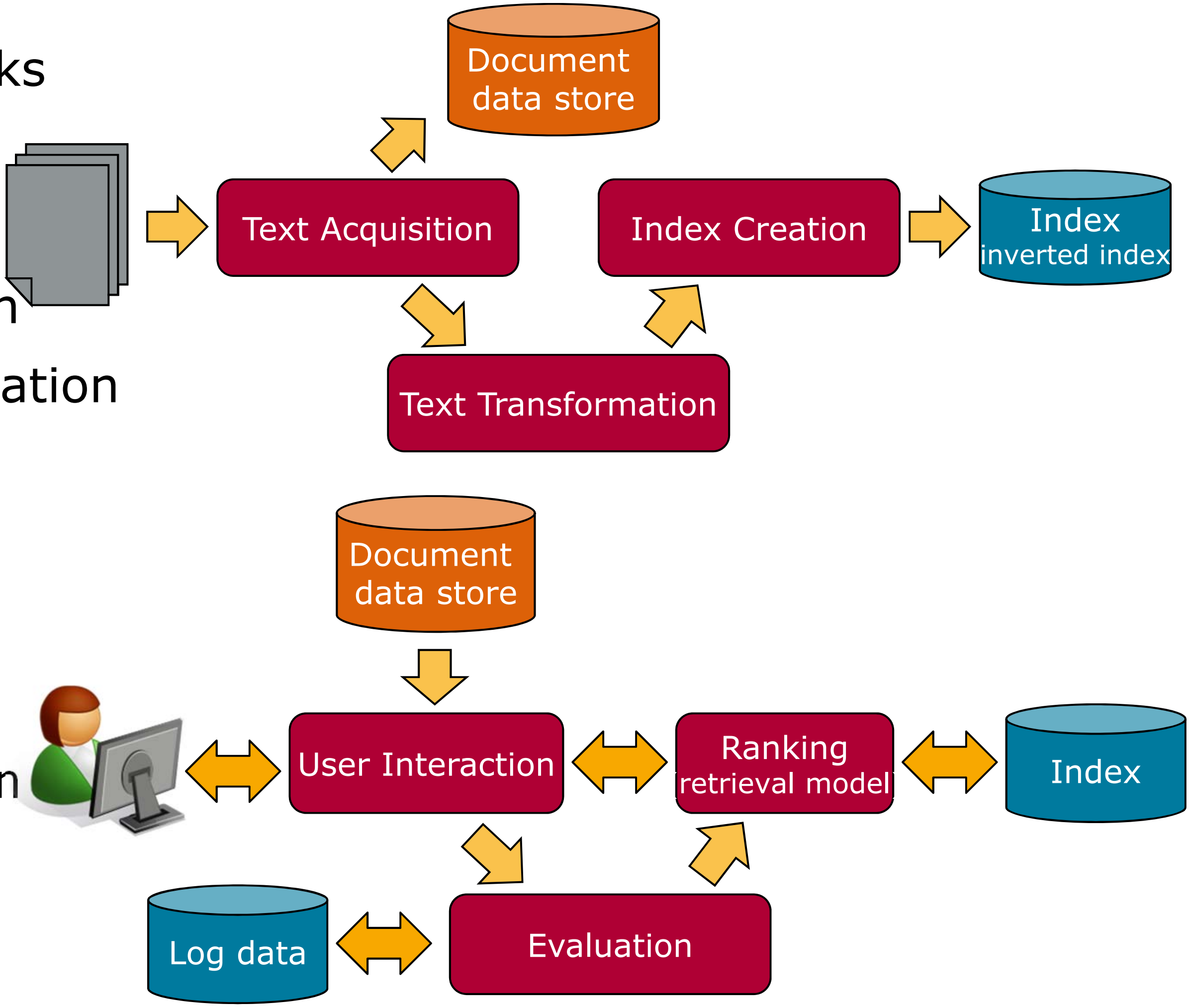
http://en.wikipedia.org/wiki/Spamdexing

Overview

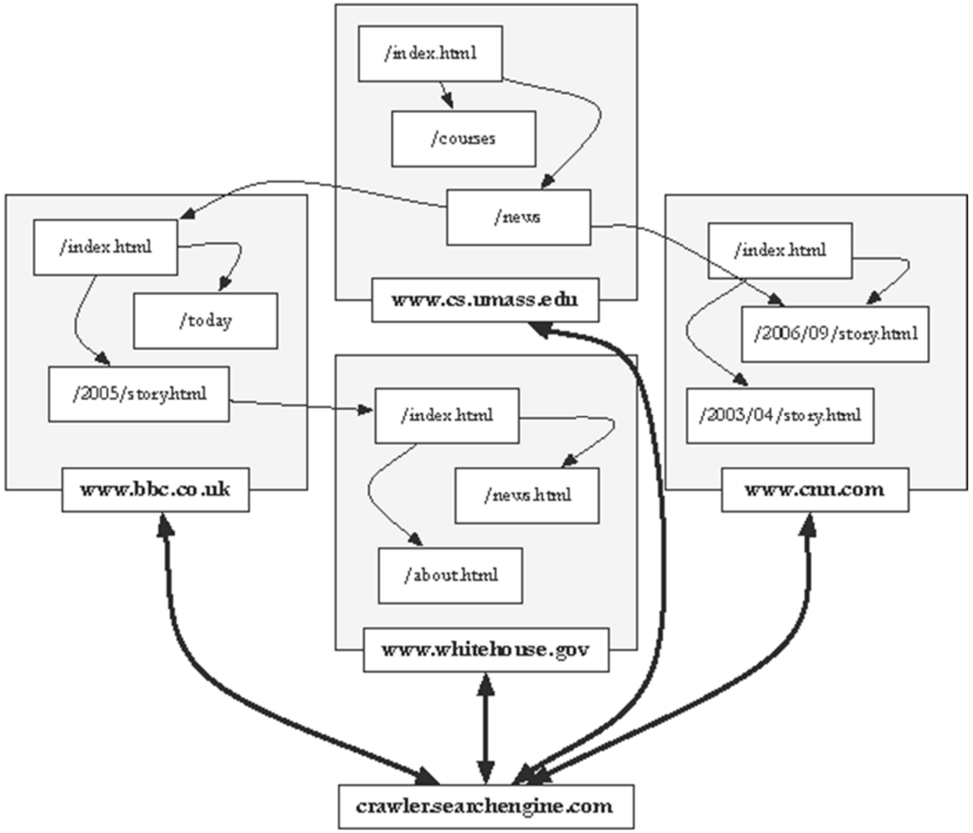
- Introduction to team
- Organization
- Information Retrieval & Search Engines
- Overview of semester

Chapter 2
Architecture of a Search Engine

- Basic building blocks
- Indexing
  - Text acquisition
  - Text transformation
  - Index creation
- Querying
  - User interaction
  - Ranking
  - Evaluation



- Deciding what to search
- Crawling the web
- Directory crawling
- Document feeds
- The conversion problem
- Storing the documents
- Detecting duplicates
- Removing noise



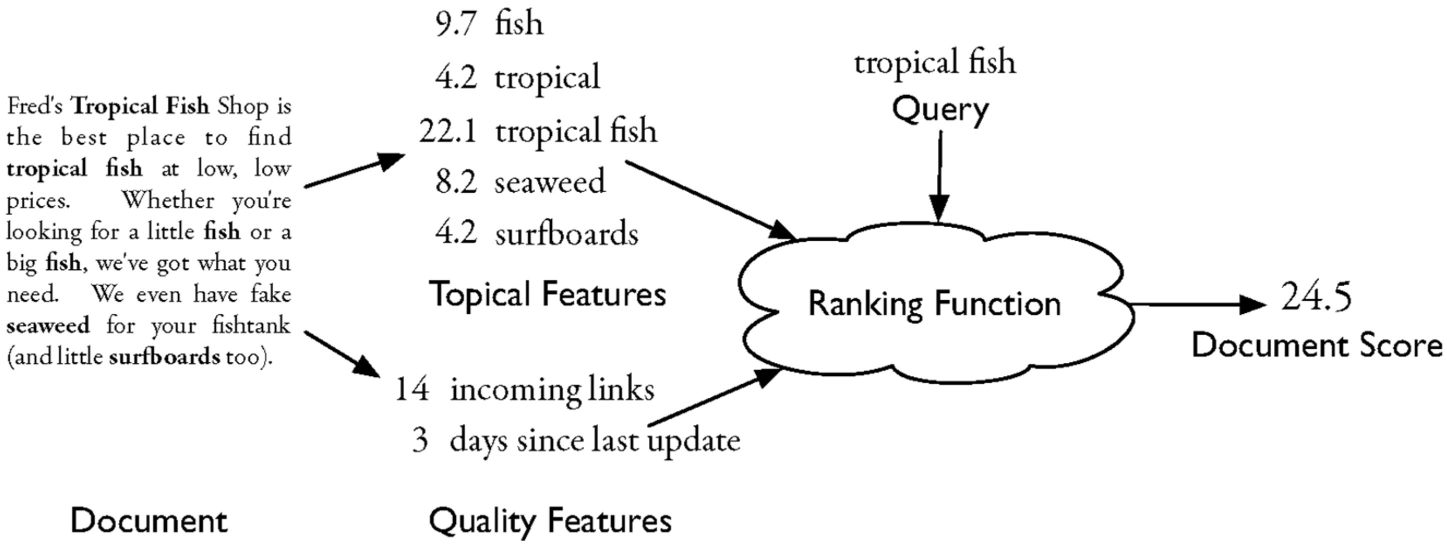
Chapter 4
Processing Text

- From words to terms
- Text statistics
- Document parsing
- Document structure and markup
- Link analysis
- Information extraction
- Internationalization

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Chapter 5
Ranking with Indexes

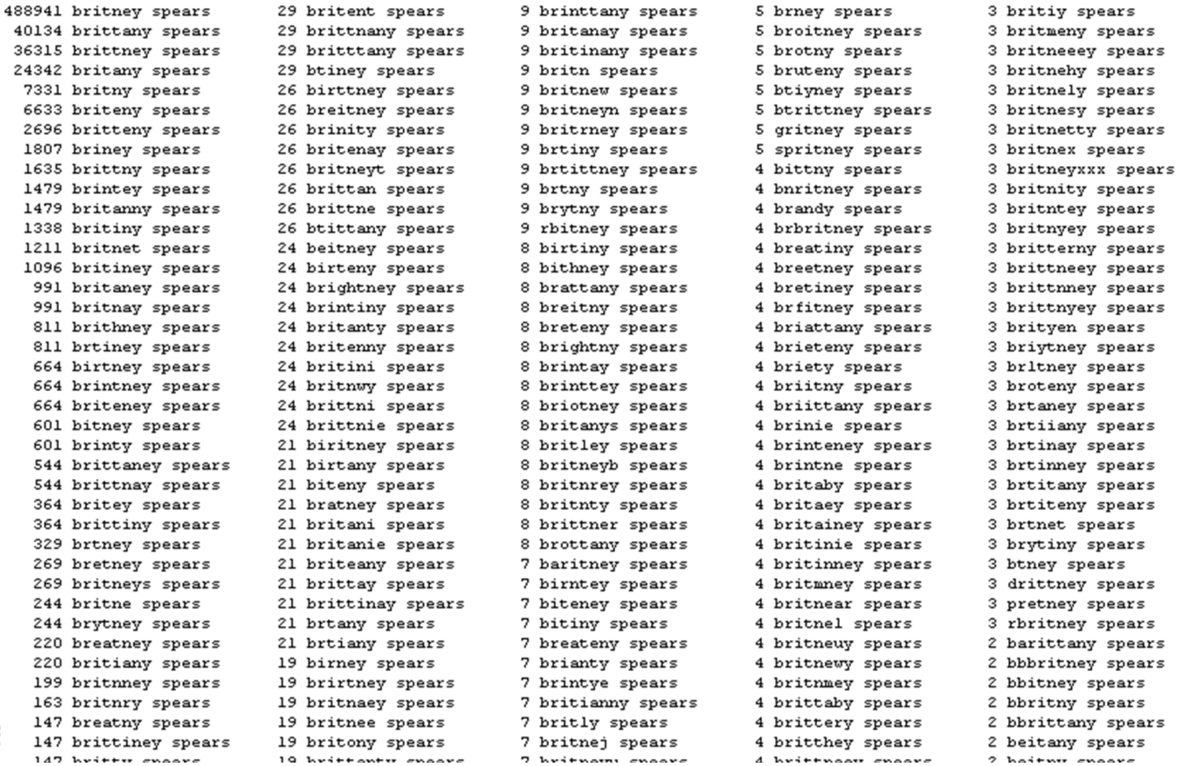
- Abstract model of ranking
- Inverted indexes
- Compression
- Auxiliary structures (index on index)
- Index construction – Map/Reduce
- Query processing



Fred's **Tropical Fish** Shop is the best place to find **tropical fish** at low, low prices. Whether you're looking for a little **fish** or a big **fish**, we've got what you need. We even have fake **seaweed** for your fishtank (and little **surfboards** too).

Chapter 6
Queries and Interfaces

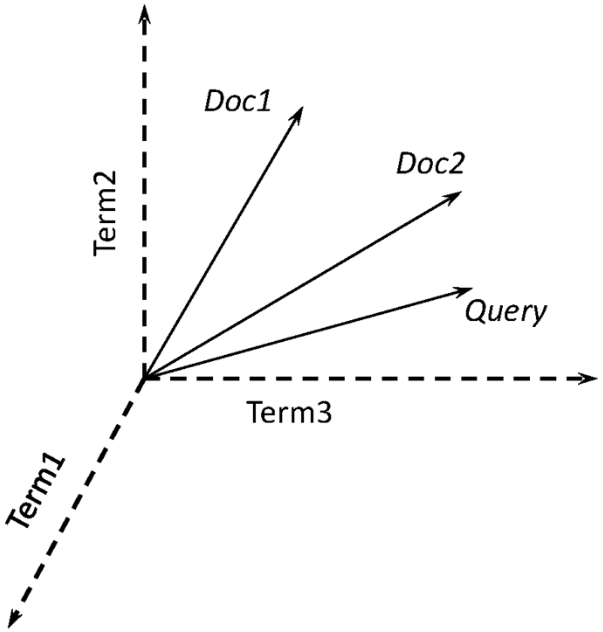
- Information needs and queries
- Query transformation and refinement
- Showing the results
- Cross-language search



```
488941 britney spears
 40134 brittany spears
 36315 brittney spears
 24342 britany spears
  7331 britny spears
  6633 briteny spears
  2696 britteny spears
  1807 briney spears
  1635 brittny spears
  1479 brintey spears
  1479 britanny spears
  1338 britiny spears
  1211 britnet spears
  1096 britiney spears
   991 britaney spears
   991 britnay spears
   811 brithney spears
   811 brtiney spears
   664 birtney spears
   664 brintney spears
   664 briteney spears
   601 bitney spears
   601 brinty spears
   544 brittaney spears
   544 brittnay spears
   364 britey spears
   364 brittiny spears
   329 brtney spears
   269 bretney spears
   269 britneys spears
   244 britne spears
   244 brytney spears
   220 breatney spears
   220 britiany spears
   199 britnney spears
   163 britnry spears
   147 breatny spears
   147 brittiney spears

29 britent spears
29 brittnany spears
29 britttany spears
29 btiney spears
26 birttney spears
26 breitney spears
26 brinity spears
26 britenay spears
26 britneyt spears
26 brittan spears
26 brittne spears
26 btittany spears
24 beitney spears
24 birteny spears
24 brightney spears
24 brintiny spears
24 britanty spears
24 britenny spears
24 britini spears
24 britnwy spears
24 brittni spears
24 brittnie spears
21 biritney spears
21 birtany spears
21 biteny spears
21 bratney spears
21 britani spears
21 britanie spears
21 briteany spears
21 brittay spears
21 brittinay spears
21 brtany spears
21 brtiany spears
19 birney spears
19 brirtney spears
19 britnaey spears
19 britnee spears
19 britony spears

9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears
7 brintye spears
7 britianny spears
7 britly spears
7 britnej spears

5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears
4 britaey spears
4 britainey spears
4 britinie spears
4 britinney spears
4 britmney spears
4 britnear spears
4 britnel spears
4 britneuy spears
4 britnewy spears
4 britnmey spears
4 brittaby spears
4 brittery spears
4 britthey spears

3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 brltney spears
3 broteny spears
3 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears
3 brtiteny spears
3 brtnet spears
3 brytiny spears
3 btney spears
3 drittney spears
3 pretney spears
3 rbritney spears
2 barittany spears
2 bbbritney spears
2 bbitney spears
2 bbritny spears
2 bbrittany spears
2 beitany spears
```

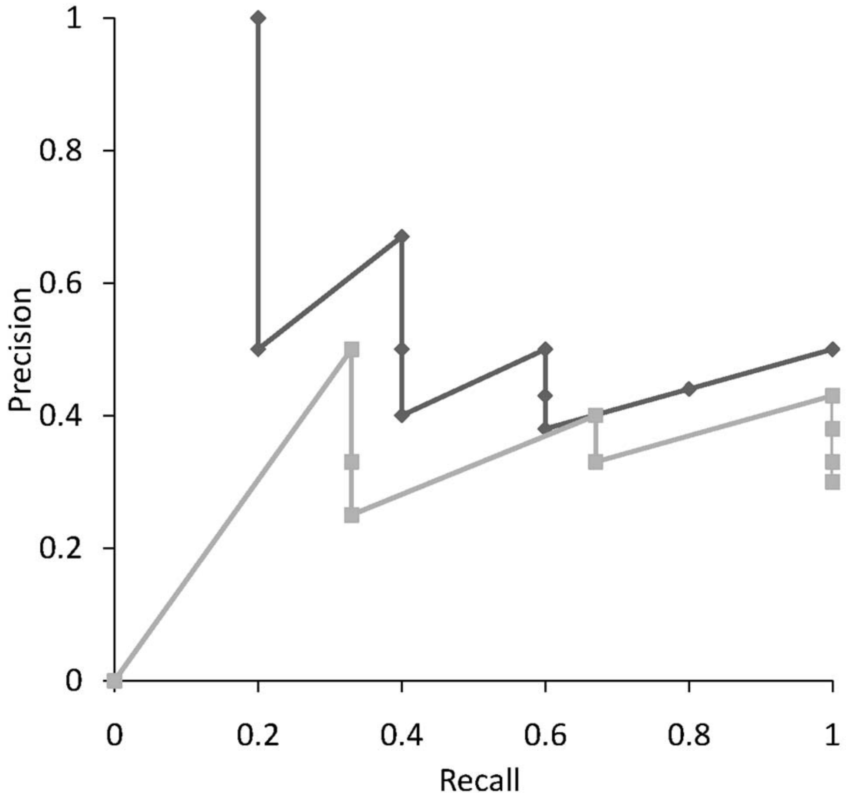
Chapter 7
Retrieval Models

- Boolean retrieval (exact match, no ranking)
- Vector space model (terms as dimensions, spatial proximity)
- Probabilistic models (rank by probability of relevance)
- Ranking based on language models (probability of co-occurring words in particular language, topical relevance)
- Complex queries and combining evidence (inference networks)
- Web search (retrieval models in practice)
- Machine learning and information retrieval (relevance feedback, text categorization)



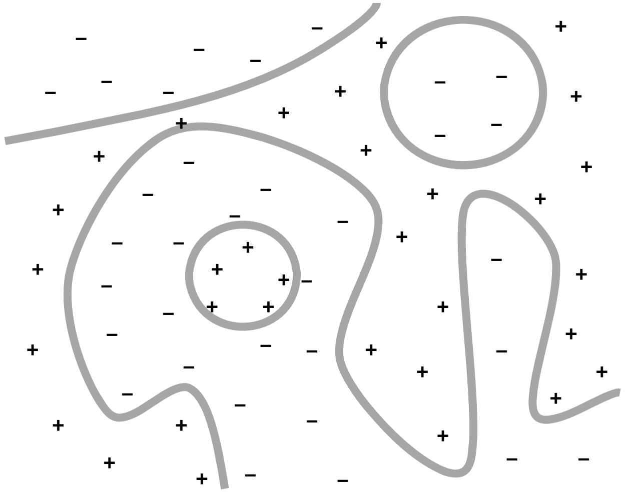
Chapter 8
Evaluating Search Engines

- Evaluation corpora
- Logging
- Effectiveness metrics
- Efficiency metrics
- Training, testing, and statistics



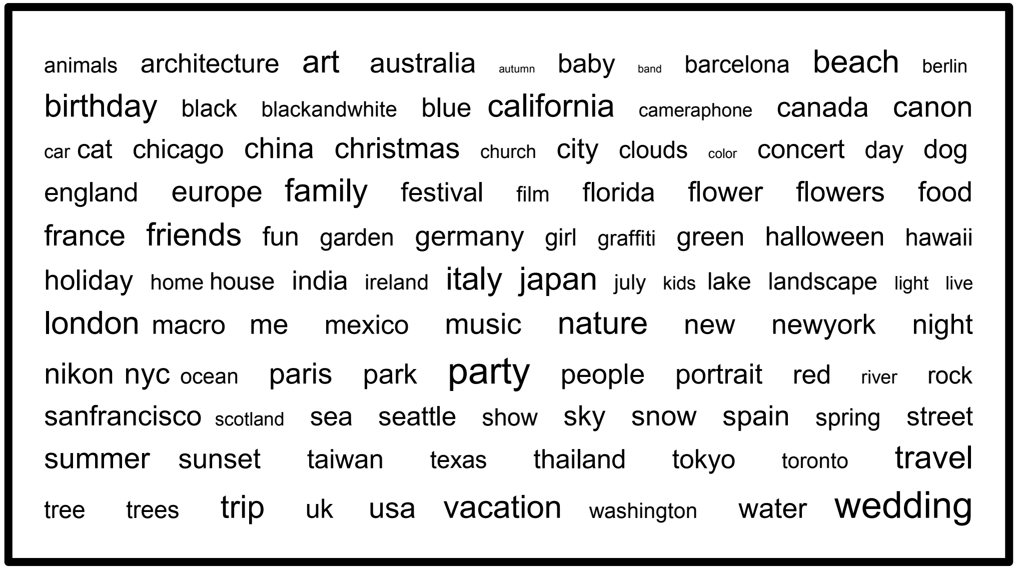
Chapter 9
Classification and Clustering

- Classification and categorization
    - Naïve Bayes
    - Support vector machines
    - Evaluation
    - Classifier and feature selection
    - Spam, sentiment, and online advertising
- Clustering
    - Hierarchical and *K-Means clustering*
    - *K nearest neighbor clustering*
    - Evaluation
    - How to choose K
    - Clustering and search



Chapter 10
Social Search

- User tags and manual indexing
- Searching with communities
- Filtering and recommending
- Personalization
- Peer-to-peer and metasearch



animals architecture art australia autumn baby band barcelona beach berlin birthday black blackandwhite blue california cameraphone canada canon car cat chicago china christmas church city clouds color concert day dog england europe family festival film florida flower flowers food france friends fun garden germany girl graffiti green halloween hawaii holiday home house india ireland italy japan july kids lake landscape light live london macro me mexico music nature new newyork night nikon nyc ocean paris park party people portrait red river rock sanfrancisco scotland sea seattle show sky snow spain spring street summer sunset taiwan texas thailand tokyo toronto travel tree trees trip uk usa vacation washington water wedding

Chapter 11
Beyond Bag of Words

- Feature-based retrieval models
- Term dependence models
- Structure revisited (query structure)
- Longer questions, better answers
- Words, pictures, and music
- One search fits all?

a and as bag could get meaning no-one normal of read representation same sorted text the words

No-one could read a sorted bag of words representation and get the same meaning as normal text.

people, pool, swimmers, water

cars, formula, tracks, wall

clouds, jet, plane, sky

fox, forest, river, water

Anthropology Program at Kansas State University – Michael Wesch

- The machine is Us/ing us
  - □ http://www.youtube.com/watch?v=NLlGopyXT_g

Questions, wishes, …

Now, or …	
Office:	A.1-13
Consultations:	Tuesdays 15-16 Uhr or by arrangement
Email:	naumann@hpi.uni-potsdam.de
Phone:	(0331) 5509 280

The end.