Search Engines
Chapter 1 – Introduction

12.4.2011
Felix Naumann
Information (r)evolution

- http://www.youtube.com/watch?v=-4CV05HyAbM
Overview

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

Integriertes Informationssystem

Oracle, DB2...

Dateisystem

Web Service

Anwendung

HTML Form

Integriertes Info.-system
### Schematische und Daten-Heterogenität

#### Variante 1

<table>
<thead>
<tr>
<th>Vorname</th>
<th>Nachname</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felix</td>
<td>Naumann</td>
</tr>
<tr>
<td>Jens</td>
<td>Bleiholder</td>
</tr>
</tbody>
</table>

#### Variante 2

<table>
<thead>
<tr>
<th>Vorname</th>
<th>Nachname</th>
<th>Männl.</th>
<th>Weibl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felix</td>
<td>Naumann</td>
<td>Ja</td>
<td>Nein</td>
</tr>
<tr>
<td>Jens</td>
<td>Bleiholder</td>
<td>Ja</td>
<td>Nein</td>
</tr>
<tr>
<td>Melanie</td>
<td>Weis</td>
<td>Nein</td>
<td>Ja</td>
</tr>
<tr>
<td>Jana</td>
<td>Bauckmann</td>
<td>Nein</td>
<td>Ja</td>
</tr>
</tbody>
</table>

#### Variante 3

<table>
<thead>
<tr>
<th>Vorname</th>
<th>Nachname</th>
<th>Geschlecht</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felix</td>
<td>Naumann</td>
<td>Männlich</td>
</tr>
<tr>
<td>Jens</td>
<td>Bleiholder</td>
<td>Männlich</td>
</tr>
<tr>
<td>Melanie</td>
<td>Weis</td>
<td>Weiblich</td>
</tr>
<tr>
<td>Jana</td>
<td>Bauckmann</td>
<td>Weiblich</td>
</tr>
</tbody>
</table>
### Variante 1

<table>
<thead>
<tr>
<th>Vorname</th>
<th>Nachname</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felix</td>
<td>Naumann</td>
</tr>
<tr>
<td>Jens</td>
<td>Bleiholder</td>
</tr>
</tbody>
</table>

### Variante 2

<table>
<thead>
<tr>
<th>FirstNa</th>
<th>Name</th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felix</td>
<td>Naumann</td>
<td>Ja</td>
<td>Nein</td>
</tr>
<tr>
<td>Jens</td>
<td>Bleiho.</td>
<td>Ja</td>
<td>Nein</td>
</tr>
<tr>
<td>Melanie</td>
<td>Weiß</td>
<td>Nein</td>
<td>Ja</td>
</tr>
<tr>
<td>Jana</td>
<td>baukman</td>
<td>Nein</td>
<td>Ja</td>
</tr>
</tbody>
</table>

### Variante 3

<table>
<thead>
<tr>
<th>VN</th>
<th>NN</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.</td>
<td>Naumann</td>
<td>Männlich</td>
</tr>
<tr>
<td>J.</td>
<td>Bleiholder</td>
<td>Männlich</td>
</tr>
<tr>
<td>M.</td>
<td>Weiß</td>
<td>Weiblich</td>
</tr>
<tr>
<td>J.</td>
<td>Bauckmann</td>
<td>Weiblich</td>
</tr>
</tbody>
</table>
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)
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
Search Engines: Information Retrieval in Practice

- Bruce Croft
- Donald Metzler
- Trevor Strohman
- [http://ciir.cs.umass.edu/](http://ciir.cs.umass.edu/)

Addison-Wesley, 2010
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...


Neu kaufen: EUR 83,98 | 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.

Felix Naumann | Search Engines | Summer 2011
Other literature

- *Introduction to Information Retrieval*
  - Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze.

- *Modern Information Retrieval*
  - Addison Wesley (2010)
  - Ricardo Baeza-Yates und Berthier Ribeiro-Neto
1. **Das Google Kompendium: Alles, was Sie über Google wissen müssen** von Jon Smith (Broschiert – 26. 2010)
   - **Neu kaufen**: EUR 19,80
   - **Drauf ghn. EUR 15,90**
   - 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
   - **Bücher**: Alle 7.821 Artikel ansehen

   - **Neu kaufen**: EUR 19,95
   - **Drauf ghn. EUR 12,00**
   - Lieferung bis Dienstag, 12. April: Bestellen Sie innerhalb der nächsten 8 Minuten per Overnight-Express.
   - **★★★★★** (27) Prime
   - **Bücher**: Alle 7.821 Artikel ansehen

   - **Neu kaufen**: EUR 9,90
   - **Drauf ghn. EUR 9,90**
   - Lieferung bis Mittwoch, 13. April: Bestellen Sie innerhalb der nächsten 22 Stunden per Overnight-Express.
   - **★★★★★** (9) Prime
   - **Bücher**: Alle 7.003 Artikel ansehen

   - **Neu kaufen**: EUR 14,80
   - **Drauf ghn. EUR 14,80**
   - Lieferung bis Dienstag, 12. April: Bestellen Sie innerhalb der nächsten 8 Minuten per Overnight-Express.
   - **★★★★★** (4) Prime
   - **Bücher**: Alle 7.821 Artikel ansehen

   - **Neu kaufen**: EUR 9,30
   - **Drauf ghn. EUR 6,79**
   - Lieferung bis Dienstag, 12. April: Bestellen Sie innerhalb der nächsten 8 Minuten per Overnight-Express.
   - **★★★★★** (13) Prime
   - **Bücher**: Alle 22.030 Artikel ansehen

   - **Neu kaufen**: EUR 29,95
   - **Drauf ghn. EUR 29,95**
   - 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 Web 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?

Felix Naumann | Search Engines | Summer 2011
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=LFIFIG22Y9E&hl=de
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
Examples:

- **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.
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.

---

Felix Naumann | Search Engines | Summer 2011
More question answering

Who was chancellor in Germany when Angela Merkel was born?

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 WolframAlpha 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

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]
- Exsite [Defunct as a separate database. Now uses an InfoSpace meta search]
- Exsite 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
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 techniques
  - query expansion
  - query suggestion
  - relevance feedback
- improve ranking

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Bing</th>
<th>Ask</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26.79%</td>
<td>46.76%</td>
<td>49.90%</td>
<td>54.15%</td>
</tr>
<tr>
<td>2</td>
<td>23.39%</td>
<td>18.81%</td>
<td>13.03%</td>
<td>18.11%</td>
</tr>
<tr>
<td>3</td>
<td>18.72%</td>
<td>15.92%</td>
<td>16.09%</td>
<td>12.31%</td>
</tr>
<tr>
<td>4</td>
<td>12.78%</td>
<td>8.40%</td>
<td>6.72%</td>
<td>7.08%</td>
</tr>
<tr>
<td>5</td>
<td>8.23%</td>
<td>5.23%</td>
<td>6.42%</td>
<td>3.73%</td>
</tr>
<tr>
<td>6</td>
<td>4.55%</td>
<td>1.94%</td>
<td>3.77%</td>
<td>2.47%</td>
</tr>
<tr>
<td>7</td>
<td>2.76%</td>
<td>1.40%</td>
<td>0.71%</td>
<td>0.97%</td>
</tr>
<tr>
<td>8</td>
<td>1.36%</td>
<td>0.71%</td>
<td>2.24%</td>
<td>0.68%</td>
</tr>
<tr>
<td>9</td>
<td>1.02%</td>
<td>0.77%</td>
<td>0.81%</td>
<td>0.33%</td>
</tr>
<tr>
<td>10</td>
<td>0.41%</td>
<td>0.06%</td>
<td>0.31%</td>
<td>0.18%</td>
</tr>
</tbody>
</table>

Felix Naumann | Search Engines | Summer 2011
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](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”
- 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. 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 relevance 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
Chapter 3
Crawls and Feeds

- 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

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total documents</td>
<td>84,678</td>
</tr>
<tr>
<td>Total word occurrences</td>
<td>39,749,179</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>198,763</td>
</tr>
<tr>
<td>Words occurring &gt; 1000 times</td>
<td>4,169</td>
</tr>
<tr>
<td>Words occurring once</td>
<td>70,064</td>
</tr>
</tbody>
</table>
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 fish tank (and little surfboards too).

<table>
<thead>
<tr>
<th>Document</th>
<th>Quality Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>tropical fish</td>
<td>14 incoming links</td>
</tr>
<tr>
<td>tropical fish</td>
<td>3 days since last update</td>
</tr>
<tr>
<td>fish</td>
<td>9.7</td>
</tr>
<tr>
<td>tropical</td>
<td>4.2</td>
</tr>
<tr>
<td>seaweed</td>
<td>8.2</td>
</tr>
<tr>
<td>surfboards</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Query

Ranking Function

Document Score: 24.5
Chapter 6
Queries and Interfaces

- Information needs and queries
- Query transformation and refinement
- Showing the results
- Cross-language search
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 \textit{K-Means clustering}
  - \textit{K nearest neighbor clustering}
  - Evaluation
  - How to choose \textit{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
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?

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

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

people, pool, swimmers, water
cars, formula, tracks, wall
clouds, jet, plane, sky
fox, forest, river, water
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