Search Engines
Chapter 1 – Introduction

21.4.2009
Felix Naumann
- **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://www.youtube.com/watch?v=LFIFIG22Y9E&hl=de](http://www.youtube.com/watch?v=LFIFIG22Y9E&hl=de)
■ Information (r)evolution
  □ http://www.youtube.com/watch?v=-4CV05HyAbM
  □ http://ksuanth.weebly.com/wesch.html
■ The machine is Us/ing us
  □ http://www.youtube.com/watch?v=NLIGopyXT_g
Overview

- Introduction to team
- Organization
- Information Retrieval & Search Engines
- Overview of semester
Other courses in this semester

Lectures
- DBS I
- Search engines

Seminars
- Bachelor: Beauty is our Business
- Bachelor: Map/Reduce Algorithms on Hadoop
- Master: Linked Data Profiling
- Forschungsseminar

Bachelorproject
- ETL Management
Overview

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

- Lectures
  - Tuesday 9:15 – 10:45
  - Thursdays 13:30 – 15:00
- Practical work
  - Selected dates – see webpage
- First lecture
  - 21.4.2009
- Last lecture
  - 23.7.2009
- Holidays
  - 21.5. Himmelfahrt

- Exam
  - Oral exam, 30 minutes
  - First week after lectures end
- 7 exercise courses
  - TAs: Alexander Albrecht & Mohammed AbuJarour
  - 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: Tuesday 15-16
  - 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
Written by a leader in the field of information retrieval, Search Engines: Information Retrieval in Practice is designed to give undergraduate students the understanding and tools they need to evaluate, compare and modify search engines. The book covers the important issues in IR at a level appropriate for undergraduate computer science or computer engineering majors. Key mathematical models are included. The programming exercises in the book make extensive use of Galago, a Java-based open source search engine.

Authors
Bruce Croft, University of Massachusetts, Amherst
Donald Metzler, Yahoo Research
Trevor Strohman, Google

Estimated pub date is 2/15/09.

A full draft of the textbook is available for review or for use in your information retrieval class.

For more information, please contact Matt Goldstein: matt.goldstein@av.com

Copyright 2008 by Pearson Education • Legal Notice • Privacy Policy • Permissions
■ 20 copies in library
■ 75,99 € 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...
Other literature

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

- **Modern Information Retrieval**
  - Addison Wesley (27. Mai 1999)
  - Ricardo Baeza-Yates und Berthier Ribeiro-Neto
Other literature - background

- The Google Story
"Suchmaschinen"

Verwandte Suchbegriffe: [suchmaschinenoptimierung](#)
Introduction – audience

- Which semester?
- HPI or IfI?
- Erasmus o.ä.?
- 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 on the Web\(^1\) is a daily activity for many people throughout the world.

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*?
“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

1 or is it web?

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
- 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”.

Felix Naumann | Search Engines | SoSe 2009
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?
  - Some stories will be better matches than others
- 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
  - Content
  - Applications
  - Tasks

- 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 Content Dimension

- 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
- ...

Felix Naumann | Search Engines | SoSe 2009
The Task Dimension

- User queries / ad-hoc search
  - Range of query enormous, not prespecified

- 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.
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]
- ExoTea [Defunct]
- ExoTea 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]
- MessageKings [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]
- WiseIlu [Died in 2007]
Relevance

- What is it?
- 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
- *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

- **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
- 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
- *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!
- 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 such as
  - *query expansion*
  - *query suggestion*
  - *relevance feedback*
- improve ranking

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 subsecond response times, millions of queries
- Enterprise search: variety of sources, search, 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*
Measuring and improving the efficiency of search
- reducing *response time*
- increasing *query throughput*
- increasing *indexing speed*

*Indexes* are data structures designed to improve search efficiency
- Designing and implementing them are major issues for search engines
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
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.
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>Total documents</th>
<th>84,678</th>
</tr>
</thead>
<tbody>
<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 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).

Document

Quality Features

Index Construction – Map/Reduce

Query Processing
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
- Vector Space Model
- Probabilistic Models
- Ranking based on Language Models
- Complex Queries and Combining Evidence
- Web Search
- Machine Learning and Information Retrieval
- Application-Based Models
Chapter 8
Evaluating Search Engines

- Motivation
- The Evaluation Corpus
- Logging
- Effectiveness Metrics
- Efficiency Metrics
- Training, Testing, and Statistics
- The Bottom Line
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
- Document Filtering
- Personalization
- Peer-to-Peer and Metasearch

animals architecture art australia australian babby barcelona beach berlin birthday black blackandwhite blue california cameraphone canada canon car cat chicago china christmas church city clouds 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
- Longer Questions, Better Answers
- Words, Pictures, and Music
- One Search Fits All?
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