Information Retrieval

Sergej Sizov, Antje Schultz, Klaas Dellschaft
Summer term 2009

sizov@uni-koblenz.de
http://isweb.uni-koblenz.de (teaching)
Course organization and outline

Lectures: Monday 14-16 in E-523
- Lecturer: Sergej Sizov (sizov)
- Office hours: Wed 14-16 and on appointment

Classes: Thursday 12-14 in E-523
- Coordination of paper assignments: Antje Schultz (antjeschultz)
- Coordination of programming excercises: Klaas Dellschaft (klaasd)

Course materials: http://isweb.uni-koblenz.de (teaching)

Examination:
Oral exam at the end of the course
**Recommended Literature (1)**

**Information Retrieval:**

Recommended Literature (2)

Stochastics


Machine Learning


Tools and Programming

**Additional Sources**

**important conferences on IR**
(see DBLP bibliography for full detail, http://www.informatik.uni-trier.de/~ley/db/)
SIGIR, ECIR, CIKM, TREC, WWW, KDD, ICDM, ICML, ECML

**online portals**
DBLP, Google Scholar, SiteSeer search engines
ACM, IEEE portals
Scientific mailing lists (e.g. DBWorld, AK-KDList, SIG-IRList, WebIR, DDLBETAtag, etc.)

**evaluation initiatives:**
- Text Retrieval Conference (TREC), http://trec.nist.gov
- Cross-Language Evaluation Forum (CLEF), www.clef-campaign.org
- Initiative for the Evaluation of XML Retrieval (INEX),
  http://inex.is.informatik.uni-duisburg.de/
  and http://kdd05.lac.uic.edu/kddcup.html
- Language-Independent Named-Entity Recognition
  www.cnts.ua.ac.be/conll2003/ner/

**feel free to contact..**
a) lecturer, b) teaching assistants, c) authors of publications,
d) members of online communities and mailing lists
Major Instruments and Applications

**Major Instruments:**
- document content & structure analysis
- indexing, search, relevance ranking
- classification, grouping, segmentation
- interaction with knowledge bases
- annotation, summarization, visualization
- personalized interaction & collaboration

**Applications:**
- Web & Deep Web search
- intranet & enterprise search
- XML & text integration
- personalized filtering
- P2P search & collaboration
- multimedia search
Information retrieval is related to Web/Data mining, Natural Language Processing (NLP), Graph Theory, Statistical Machine Learning (ML):

- learning predictive models from data
- pattern, rule, trend, outlier detection
- classification, grouping, segmentation
- knowledge discovery in data collections
- information extraction from text and Web
- graph mining (e.g. on Web graph), social network analysis
- Image understanding, speech recognition, video analysis
Web Search Engines

Brief History:

- The Web happened (1992)
- Mosaic/Netscape happened (1993-95)
- Crawler happened (1994): M. Mauldin (founded Lycos)
- Yahoo founded in 1994 as a directory
- Several SEs happened 1994-1996
  (InfoSeek, Lycos, Altavista, Excite, Inktomi, ...)

Today:

- > 11 Billion \((10^9)\) pages
- > 450 Million daily queries
- > 8 Billion US $ annual revenue
Early search engines based on text IR methods

- Field started in the 1950s
- Mostly uses statistical methods to analyze text
  - Repeated words on a page are important
  - Common words are not important (e.g. the, for, ...)

Text IR necessary but not sufficient for Web search

- Doesn’t capture authority
  - An article on BBC as good as a copy www.duck.com
- Doesn’t address Web navigation
  - Query “uni koblenz” seeks www.uni-koblenz.de
  - www.uni-koblenz.de may look less topic-specific than PR releases

Many alternatives have been tried and exist

- Link analysis and authority ranking
- Topics/query suggestion tools (e.g. Vivisimo, Exalead)
- Graphical, 2-D, 3-D user interfaces, ...

..simple and clean preferred by users
Small World Phenomenon (Milgram 1967)
Studies on Internet Connectivity (1999)

suggested small world phenomenon: low-diameter graph
( diameter = max {shortest path (x,y) | nodes x and y} )

Source: Bill Cheswick and Hal Burch,
http://research.lumeta.com/ches/map/index.html

Source: KC Claffy,
http://www.caida.org/outreach/papers/1999/Nae/Nae.html
Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

- power-law distributed degrees: \( P[\text{degree}=k] \sim \frac{1}{k^\alpha} \)
  - with \( \alpha \approx 2.1 \) for indegrees and \( \alpha \approx 2.7 \) for outdegrees
Web Size and Web Coverage

overlap statistics → (surface) Web > **11.5 Bio. pages (> 40 TBytes)**

Deep Web (Hidden Web) estimated to have **500 Bio. units (> 10 PBytes)**

Source: A. Gulli, A. Signorini, WWW 2005
Crawling and Indexing

Crawling

Metternich: One of the famous Austrian politicians

Extraction of relevant words

Linguistic methods: stemming

Metternich

Austrian Politicians

metternich

austria politician

Statistically weighted features (terms)

Thesaurus (Ontology)

e.g. synonyms, sub-/super-concepts

Index (e.g. B+ -tree)

politician austria ...

Metternich = district of Koblenz?
Koblenz = city in Rhineland-Palatinate?
RP = federal state in Germany?
**Content Relevance Ranking**

**Ranking** by descending relevance

**Query** $q \in [0,1]^{|F|}$ (Set of weighted features)

**Documents are feature vectors** $d_i \in [0,1]^{|F|}$

**Similarity metric:**

\[ \text{sim} (d_i, q) := \frac{\sum_{j=1}^{|F|} d_{ij} q_j}{\sqrt{\sum_{j=1}^{|F|} d_{ij}^2} \sqrt{\sum_{j=1}^{|F|} q_j^2}} \]

**e.g., using:**

\[
\begin{align*}
    d_{ij} &:= w_{ij} / \sqrt{\sum_k w_{ik}^2} \\
    w_{ij} &:= \frac{\text{freq}(f_j,d_i)}{\max_k \text{freq}(f_k,d_i)} \log \frac{\# \text{docs}}{\# \text{docs with } f_i}
\end{align*}
\]

**tf*idf formula**
Authority Ranking

**Ranking** by descending relevance & authority

**Query** $q \in [0,1]^{|F|}$
(Set of weighted features)

Additionally, consider links between Web nodes:

**Authority Score** $(d_i) :=$ stationary visit probability $[d_i]$ in the random walk on the Web

.. reconciliation of relevance and authority by ad hoc weighting
random walk on the Web graph:
  uniformly random choice of links + random jumps

\[ \text{Authority (page } q) = \text{stationary prob. of visiting } q \]

\[ PR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p,q) \]
**System Architecture of Web SE**

- **Crawl**: Handle dynamic pages, detect duplicates, detect spam.
- **Extract & Clean**: Strategies for crawl schedule and priority queue for crawl frontier.
- **Index**: Build and analyze Web graph, index all tokens or word stems.
- **Search**: Fast top-k queries, query logging and auto-completion.
- **Rank**: Scoring function over many data and context criteria.
- **Present**: GUI, user guidance, personalization.

Special file system for high-performance storage management caching for fast search.

Server farm with > 10,000 nodes, distributed/replicated data.
Users of the Web

- 70-80% of users use SE to find sites!

- most users prefer a few commercial large-scale search engines
Where are the users?

Google: the users are all over the world

• Search engine serves over 100 different languages
• Should not have a catastrophic failure in any
What are the users asking for?

Google-style Web search:

- Users give a 2-4 word query
- SE gives a relevance ranked list of web pages
- Most users click only on the first few results
- Few users go below the fold
  .. whatever is visible without scrolling down
- Far fewer ask for the next 10 results

over 200 Million queries a day

searching over Eight Billion+ documents
## Themes and Trends

### Zeitgeist This Month

#### Popular Celebrities
- August 2005
1. madonna
2. jessica simpson
3. pamela anderson
4. paris hilton
5. jessica alba

#### Music-Related Queries
- August 2005
1. lyrics
2. my chemical romance
3. beyonce
4. mariah carey
5. green day

#### Popular References
- August 2005
1. dictionary
2. maps
3. weather
4. white pages
5. yellow pages

#### Travel-Related Queries
- August 2005
1. expedia
2. travelocity
3. orbitz
4. southwest airlines
5. american airlines

### Google News Queries

#### Katrina-Related Queries
- August 2005
1. hurricane katrina
2. new orleans
3. hurricane katrina photos
4. slidell
5. french quarter

#### Popular Sports Queries
- August 2005
1. real madrid
2. arsenal
3. cricket
4. nhl
5. nba

#### Popular Newsmakers
- August 2005
1. natalee holloway
2. cindy sheehan
3. peter jennings
4. lance armstrong
5. tiger woods

---

http://www.google.com/press/zeitgeist.html
User intentions

classification of queries [Rose/Levinson: WWW 2004]:

- **navigational**: find specific homepage with unknown URL, e.g. Germanwings
- **informational**: learn about topic (e.g. Information Retrieval)
- **focused**, e.g. Fürst von Metternich, soccer world championship qualification
- **unfocused**, e.g. undergraduate statistics, dark matter, Koblenz
- **seeking advice**, e.g. help losing weight, low-fat food, marathon training tips
- **locating service**, e.g. 6M pixel digital camera, taxi service Koblenz
- **exhaustive**, e.g. Dutch universities, hotel reviews Crete, MP3 players
- **transactional**: find specific resource, e.g. download Lucene source code, Sony Cybershot DSC-W5, Mars surface images, hotel beach south Crete August
- **embedded** in business workflow (e.g. CRM, business intelligence) or personal agent (in cell phone, MP3 player, or ambient intelligence at home) with automatically generated queries
- **natural-language** question answering (QA):
  factoids, e.g. where was Fürst von Metternich born, where is the German Corner, etc
- **list queries**, e.g. in which movies did Johnny Depp play
large-scale Web search with authority ranking
http://www.google.com
Organization of Search Results (2)

cluster search results into topic areas

http://www.vivisimo.com
visualize cluster hierarchies for search results
http://www.grokker.com
Organization of Search Results (4)

show broader context of results
http://www.exalead.com
Organization of Search Results (5)

auto-complete queries
http://labs.google.com/suggest/
Evaluation of Result Quality: Basic Measures

ideal measure is user satisfaction!
heuristically approximated by benchmarking measures
(on test corpora with query suite and relevance assessment by experts)

Capability to return **only** relevant documents:

\[
\text{Precision (Präzision)} = \frac{\text{# relevant docs among top } r}{r}
\]

typically for \( r = 10, 100, 1000 \)

Capability to return **all** relevant documents:

\[
\text{Recall (Ausbeute)} = \frac{\text{# relevant docs among top } r}{\text{# relevant docs}}
\]

typically for \( r = \text{corpus size} \)

**Typical quality**

**Ideal quality**
Evaluation of Result Quality: Aggregated Measures

Combining precision and recall into **F measure** (e.g. with $\alpha=0.5$ harmonic mean $F_1$):

$$F = \frac{1}{\alpha \frac{1}{\text{precision}} + (1-\alpha) \frac{1}{\text{recall}}}$$

**Precision-recall breakeven point** of query $q$:
point on precision-recall curve $p = f(r)$ with $p = r$

for a set of $n$ queries $q_1, \ldots, q_n$ (e.g. TREC benchmark)

**Macro evaluation (user-oriented)** = of precision

$$\frac{1}{n} \sum_{i=1}^{n} \text{precision}(q_i)$$

**Micro evaluation (system-oriented)** = of precision

$$\frac{\sum_{i=1}^{n} \# \text{relevant & found docs for } q_i}{\sum_{i=1}^{n} \# \text{found docs for } q_i}$$

analogous for recall and F1
Problems with common IR-style evaluation on the Web

- Collection is dynamic
  10-20% urls change every month
- Queries are time sensitive
  Topics are hot then they are not
- Spam methods evolve
  Algorithms evaluated against last month’s web may not work today
- Need to keep the collection fresh
- Need to keep the queries fresh
- Search space is extremely large
- Over 100 million unique queries a day
- To measure a 5% improvement at 95% confidence level:
  one would need 2700 judged queries
great for e-shopping, school kids, scientists, doctors, etc.

- high-precision results for simple queries
- high scalability (now >8 Bio. docs, >1000 queries/sec)
- continuously enhanced: Froogle, Google Scholar, alerts, multilingual for >100 languages, query auto-completion, etc.
Modern SE: Limitations

Hard queries (disregarding QA, multilingual, multimedia):

- professors from Koblenz who teach DB and have projects on Semantic Web
- famous 19th century politician that was born in Koblenz
- drama with three women making a prophecy to a British nobleman that he will become king

- best and latest insights on percolation theory
- pros and cons of dark energy hypothesis

- market impact of XML standards in 2002 vs. 2006
- experienced NLP experts who may be recruited for IT staff
Limitations of Modern SE: an Example

Norwich and Koblenz are just twinned cities
Limitations of Modern SE (2)