Web Information Retrieval

Dr. Dr. Sergej Sizov
Summer Term 2013
Why are you here?
What is Web Retrieval all about?

.. discovering useful information from the World-Wide Web and its usage patterns

using

**Text Mining:**

♦ application of Data Mining techniques to unstructured text (e.g. Web sites, blog postings, user comments..)

**Structure Mining:**

♦ taking into account structure and relations in Web data (HTML tags, hyperlinks, friend lists..)

**Usage Mining:**

♦ taking into account user interactions with Web systems (clickstreams, collaborative filtering, ...)
Chapter 1

Motivation & Overview
Theoretic model of IR

\[ \begin{align*}
Q & \xrightarrow{\alpha_Q} Q^r \xrightarrow{\beta_Q} Q^d \\
\mathcal{R} & \xrightarrow{} \mathcal{R} \\
D & \xrightarrow{\alpha_D} D^r \xrightarrow{\beta_D} D^d \\
\rho & \xrightarrow{} \mathcal{R}
\end{align*} \]

- \( \alpha \): Generate representation of documents
- \( \beta \): Further operations on representations
- \( \rho \): Retrieval function
- \( \mathcal{R} \): Actual Relevance
- \( \mathcal{R} \): Computed relevance

Fuhr [8]
How do users seek information

- Simple, linear model
  - Information need remains constant
  - Feedback to clarify information need
- Underlies most current systems
- Not realistic!
- Concept: Anomalous State of Knowledge (ASK)

Belkin [2]
Query:

*A famous car manufacturer born near Koblenz*
famous car manufacturer born near koblenz

About 2,900,000 results (0.22 seconds)

Koblenz - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Koblenz
5 Education; 6 Twin towns – Sister cities; 7 Popular culture; 8 References ... Around 1000 BC, early fortifications were erected on the Festung ... Since 2010 the Koblenz Cable Car is Germanys biggest aerial tramway ... the family house of the Metternichs, where Prince Metternich, the Austrian statesman, was born in 1773.

KOBLENZ RAILWAY STATION Car Rental - Europcar
www.europcar.com/car-rental-KOBLENZ_RAILWAY_STATION.html
Find out where you can collect your car rental vehicle in KOBLENZ ... and are proud to be the first rental car company at KOBLENZ RAILWAY STATION with ... Click and drag to move the map around ... Popular car rental locations in Germany ...

The Koblenz Travel Guide | Live Like a German
www.live-like-a-german.com/destinations/show/koblenz
... house of the Metternichs, where Prince Metternich, the Austrian statesman, was born in 1773. ... Near Koblenz is the Lahneck Castle near Lahnstein, open to visitors from April 1 to October 31. ... am Main international airport, Koblenz is easy and convenient to reach by car. ... Popular Points of Interest in and near Koblenz ...
Query:

*A famous car manufacturer born near Koblenz*

Answer:

*August Horch, born 1868 in Winningen*
famous car manufacturer born in winningen

August Horch - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/August_Horch
August Horch in a Horch automobile, 1908. ... 1 Beginnings; 2 Manufacturing; 3 Post Audi; 4 References. [edit] Beginnings. Horch was born in Winningen.

Audi - Get great deals for Audi on eBay! - Find Popular Products on ...
popular.ebay.com › Popular Items › Miscellaneous AB
Audi is a German automobile manufacturer with headquarters located at ... The credit for establishing Audi goes to August Horch, an engineer born in Winningen.

August Horch, Founder of the Automotive Company Audi - The ...
theworldbiography.blogspot.com/.../august-horchfounder-of-autom...
7 Apr 2012 – Horch was born in Winningen. ... Audi AG is a German car manufacturers that sell cars in various classes ... Positioned as the luxury car class.

Audi's 100th birthday to be marked with showroom bash. - Free ...
www.thefreelibrary.com › ... › Gulf Weekly › July 19, 2009
19 Jul 2009 – He was born in Winningen, Moselle. ... It is notably popular in Britain, Germany, Scandinavia, and Australia. , and of course the 24 hours of Le Mans. ... Audi one of the most successful premium car manufacturers in the world.
Berry picking model

Berrypicking model

Area of Interest

Start

Stop

Bates [1]
Query:

A famous furniture manufacturer born near Koblenz
Who promoted him?
What kind of product made him so famous?
## Types of queries

<table>
<thead>
<tr>
<th>Type</th>
<th>Intention</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Navigational</td>
<td>Go to specific known website</td>
</tr>
<tr>
<td>2</td>
<td>Informational</td>
<td>Learn something by reading web pages</td>
</tr>
<tr>
<td>2.1</td>
<td>Directed</td>
<td>Learn something in particular about my topic</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Closed</td>
<td>Get an single, unambiguous to a question</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Open</td>
<td>Get an answer to an open-ended question</td>
</tr>
<tr>
<td>2.2</td>
<td>Undirected</td>
<td>Learn anything/everything about a topic</td>
</tr>
<tr>
<td>2.3</td>
<td>Advice</td>
<td>Get advice, ideas, suggestions, instructions.</td>
</tr>
<tr>
<td>2.4</td>
<td>Locate</td>
<td>Find real world service or product</td>
</tr>
<tr>
<td>2.5</td>
<td>List</td>
<td>List of web sites helping me achieve some goal</td>
</tr>
<tr>
<td>3</td>
<td>Resource</td>
<td>Obtain a resource (not information) available on web</td>
</tr>
<tr>
<td>3.1</td>
<td>Download</td>
<td>Download a resource for use on computer</td>
</tr>
<tr>
<td>3.2</td>
<td>Entertainment</td>
<td>Be entertained by viewing items</td>
</tr>
<tr>
<td>3.3</td>
<td>Interact</td>
<td>Interact with a program/service available on the web</td>
</tr>
<tr>
<td>3.4</td>
<td>Obtain</td>
<td>Obtain a resource that does not require a computer to use.</td>
</tr>
</tbody>
</table>

Rose, Levinson [20]
How users use the Web?

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

Figure 1 Online Tenure Impacts Search Use And Effectiveness

Search use increases with online tenure
“How have you typically found Web sites you have visited in the past month?”

and most users prefer
a few commercial
large-scale
search engines
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 us 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
noisy inputs
searching over over Eight Billion+ documents
Supporting the user: Result representations
Organization of Search Results

large-scale Web search with authority ranking
http://www.google.com
Organization of Search Results (2)

Cluster search results into topic areas
http://www.yippy.com
Organization of Search Results (4)

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

1. August Horch — Wikipedia, the free encyclopedia
   http://en.wikipedia.org/wiki/August_Horch

2. Август Хорх (August Horch) — Пионер автомобилестроения и основатель...
   Август Хорх (August Horch) — немецкий инженер, один из патриархов мирового автомобилестроения.
   http://vAudi.ru/person/august-horch

3. August HORCH : Автообоз (Автомобильное Обозрение) — Дайджест...
   August HORCH. Август Хорх родился 12 октября 1868 года в Виннингене, на Мозелем, в семье кузнеца.
   http://autooboz.infoimen/8august-horch/

Tag cloud of related terms
http://www.quintura.com
Organization of Search Results (6)

suggest related search directions
http://www.yebol.com
Organization of Search Results (7)

Meta search with previews
http://www.spezify.com
How should it work?
Web IR in a nutshell
Under the hood: crawling and indexing

Crawling

Metternich: One of the famous Austrian politicians

Extraction of relevant words

Linguistic methods: stemming

Statistically weighted features (terms)

Index (e.g. B*-tree)

URLs

Metternich = district of Koblenz?
Koblenz = city in Rhineland-Palatinate?
RP = federal state in Germany?
Observation: Search engines have different data

overlap statistics $\rightarrow$ (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
Observation: they also show different results

http://ranking.thumbshots.com/
Ranking: Content Relevance

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

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

**tf*idf** formula
random walk on the Web graph:
uniformly random choice of links + random jumps

\[ PR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p,q) \]

Authority (page q) = stationary prob. of visiting q
Additionally, consider links between Web nodes:

**Authority Score** \( (d_i) := \text{stationary visit probability} [d_i] \) in the random walk on the Web

.. *reconciliation of relevance and authority by ad hoc weighting*
What is our technology used for?

- **www**: Crawl, extract & clean, index, search, rank, present

- **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**
- **Index-entry and query-result caching for fast search**
Chapter 1: Introduction

Web IR

- Ranking
- Routing
- Classification
- Clustering
- Question answering
- Multimedia analysis
- Web Usage Mining
- Web Content Mining

Tasks in Web IR
Related Fields

Web IR

- HCI
- Network Analysis
- Machine Learning
- Information Science
- Psychology
- Linguistics
- Sociology
- Cultural Anthropology
- Natural Language Processing
- Computational Linguistics
- Data Mining
- Databases
Web IR research problems: an example
Flickr query: August Horch
Flickr query: Koblenz
Recommender scenarios:

- Given a user, recommend photos which may be of interest.
- Given a user, recommend users they may like to contact.
- Given a user, recommend groups they may like to join.

Motivation: multi-modal, sparse Web 2.0 data.
Collaborative content sharing framework: \( Y \subseteq U \times T \times R \)

users \( u \in U \) tags \( t \in T \) resources \( r \in R \)

Folksonomy cloud:

- user-centric \( Y_u^* \subseteq \{u\} \times T \times R \)
- resource-centric \( Y_r^* \subseteq U \times T \times \{r\} \)
- community-specific \( Y_{U^*}^* \subseteq U^* \times T \times R \) (e.g. groups)
- collection-specific \( Y_{R^*}^* \subseteq U \times T \times R^* \) (e.g. favorites)
- arbitrary \( Y_{U^*R^*}^* \subseteq U^* \times T \times R^* \)
The IR background – constructing feature vectors

if(t) = |{(u*, t, r*)}|, (u*, t, r*) ∈ Y*

\[ ii f(t) = \left( \log \frac{|U|}{|U^*|}, \log \frac{|R|}{|R^*|} \right) \]

with \( U^* \subseteq U, R^* \subseteq R : \)
\( u^* \in U^* \Leftrightarrow \exists r \exists t : (u^*, t, r) \in Y^* \)
\( r^* \in R^* \Leftrightarrow \exists t \exists u : (u, t, r^*) \in Y^* \)

weight_{Y^*}(t) = \| if(t) \cdot ii f(t) \|_1

.. defined analogously to tf·idf

- favorites
- groups
- contact lists
- comments
Results: user-focused favorite recommendation

40 Training / 50 Test favorites, 250 contrast (randomly chosen) docs

<table>
<thead>
<tr>
<th>User representation</th>
<th>Training:10 prec@10</th>
<th>Training:10 prec@20</th>
<th>Training:20 prec@10</th>
<th>Training:20 prec@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
</tr>
<tr>
<td>Commented items</td>
<td>0.292</td>
<td>0.280</td>
<td>0.296</td>
<td>0.269</td>
</tr>
<tr>
<td>Favorites</td>
<td>0.757</td>
<td>0.643</td>
<td>0.79</td>
<td>0.752</td>
</tr>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
</tr>
<tr>
<td>Commented items</td>
<td>0.296</td>
<td>0.279</td>
<td>0.269</td>
<td>0.252</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User representation</th>
<th>Training:10 prec@10</th>
<th>Training:10 prec@20</th>
<th>Training:20 prec@10</th>
<th>Training:20 prec@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
</tr>
<tr>
<td>Commented items</td>
<td>0.290</td>
<td>0.273</td>
<td>0.269</td>
<td>0.252</td>
</tr>
<tr>
<td>Favorites</td>
<td>0.840</td>
<td>0.929</td>
<td>0.752</td>
<td>0.752</td>
</tr>
<tr>
<td>User items</td>
<td>0.254</td>
<td>0.233</td>
<td>0.220</td>
<td>0.220</td>
</tr>
</tbody>
</table>
..spatial info?
Same coordinates, different views

German corner?  Old Balduin bridge?

Uni Koblenz?

Nuclear power plant?

What can you see from Ehrenbreitstein?

Steamer „Goethe“?

Fortress?

Restaurant Ferrari?

Shopping malls?
Multi-modal Analysis of Social Media

Deutsches Eck from Ehrenbreitstein Fortress, Koblenz, Germany

by schaengel

121 comments 68 faves

Tagged with koblenz, ehrenbreitstein ...

Taken on November 15, 2008, uploaded November 17, 2008

See more of schaengel's photos, or visit his profile.

When you’re high up on the hill above Koblenz at Ehrenbreitstein Fortress you can get a great panoramic view of the city and the surrounding area.
Course organization
Course organization

Lectures:
- with Summer Academy: June – July 2012 in A-308

Course materials:
- http://www.uni-koblenz.de/~sizov/summeracademy/
- lecture slides & recordings available online

Examination:
- Oral exam at the end of the course (particular slots tba)

Homework:
- No mandatory assignments, classes integrated
- Examples and think-about questions integrated into lectures

Announcements:
- Course Web Site, email, twitter (#webir2013koblenz)

Contact:
- B-112, Wed 14-16 and on appointment (sizov@uni-koblenz.de)
Web information retrieval: course topics

- Motivation and Overview
- Technical basics
- Content processing and analysis
- Link analysis & authority ranking
- Advanced IR models
- SEO
- Web Spam and Advertising
- Web Crawling

basics
making Web Analysis effective and efficient
common scenarios
Related literature
Related literature (2)

- Soumen Chakrabarti: Mining the Web: Analysis of Hypertext and Semi-Structured Data, Morgan Kaufmann, 2002


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

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

evaluation initiatives:
• Text Retrieval Conference (TREC), http://trec.nist.gov
• Cross-Language Evaluation Forum (CLEF), www.clef-campaign.org
• KDD Cup (over many years)
• ECML / PKDD Discovery challenge (over many years)

feel free to contact..
a) lecturer, b) authors of publications, c) members of online communities and mailing lists
Thank you