Web Retrieval

Steffen Staab, Sergej Sizov

http://west.uni-koblenz.de
Course organization

**Lectures:**
- with Summer Academy: 21 June – 14 July 2010

**Course materials:**
- [http://www.uni-koblenz-landau.de/koblenz/fb4/studying/summer-academy/webscience/WebRetrieval](http://www.uni-koblenz-landau.de/koblenz/fb4/studying/summer-academy/webscience/WebRetrieval)

**Examination:**
- Oral exam at the end of the semester

**Homework:**
- No classes or mandatory assignments
- Examples and think-about questions integrated into lectures!

**Announcements:**
- Course Web Site, email

**Contact:**
- B-112, Wed 14-16 and on appointment (sizov@uni-koblenz.de)
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, ...)

Steffen Staab
Chapter 1

Motivation & Overview
What can we use for this?

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

What is this good for?

- Web & Deep Web search
- intranet & enterprise search
- integration of rich data sources
- personalized filtering and recommenders
- user support in Web 2.0
Related areas

- Web Mining is related to Information Retrieval, 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
  - Image understanding, speech recognition, video analysis
  - graph mining (e.g. on Web graph), social network analysis
Retrospective: Web search engines

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, …)

15 years later (2007):

> 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
Observations: Small Diameter of the Web

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} )
The Oracle of Bacon

Johnny Depp has a Bacon number of 2.

Find a different link

Johnny Depp

Public Enemies (2009)

with

Billy Crudup

Sleepers (1996)

with

Kevin Bacon

Kevin Bacon to Johnny Depp

Find link More options >>

Social network of actor collaborations

http://oracleofbacon.org
• power-law distributed degrees: $P[\text{degree}=k] \sim (1/k)^\alpha$
  
  with $\alpha \approx 2.1$ for indegrees and $\alpha \approx 2.7$ for outdegrees
Under the hood: crawling and Indexing

Metternich: One of the famous Austrian politicians

Extraction of relevant words

Linguistic methods: stemming

Metternich = district of Koblenz? 
Koblenz = city in Rhineland-Palatinate? 
RP = federal state in Germany?

Thesaurus (Ontology)
- e.g. synonyms, sub-/super-concepts

Index (e.g. $B^+$-tree)
- politician austria ... 
- URLs
Observation: Search engines have different data

Source: A. Gulli, A. Signorini, WWW 2005

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

Deep Web (Hidden Web) estimated to have 500 Bio. units (> 10 PBytes)
Observation: they also show different results
Ranking: Content Relevance

**Search engine**

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

**Similarity metric:**

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

**Query** $q \in [0,1]^{|F|}$

(Set of weighted features)

**Ranking** by descending relevance

**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
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*
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
What is our technology used for?

- 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

index-entry and query-result caching for fast search
How users use the Web?

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

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 Eight Billion+ documents
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
Organization of Search Results (1)

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

cluster search results into topic areas
http://www.clusty.com
Organization of Search Results (3)

show broader context of results
http://www.exalead.com/search
Organization of Search Results (4)

suggest related search directions
http://www.yebol.com
**Organization of Search Results (5)**

As you type, Google auto-complete queries.

![Google Suggest](http://labs.google.com/suggest/)

<table>
<thead>
<tr>
<th>Web</th>
<th>Images</th>
<th>Video</th>
<th>News</th>
<th>Maps</th>
<th>more »</th>
</tr>
</thead>
<tbody>
<tr>
<td>mottor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mettermich</td>
<td>251,000 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mettermich system</td>
<td>46,600 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metters</td>
<td>32,800 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metter</td>
<td>413,000 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mettermick</td>
<td>138,000 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metter ga</td>
<td>747 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metter advertisor</td>
<td>6,750 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metter high school</td>
<td>33,800 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metter ford</td>
<td>44,000 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metters industries</td>
<td>3,520 results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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:

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

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

Capability to return **all** relevant documents:

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

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

Typical quality

Ideal quality
Evaluation: 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 \( q_1, ..., q_n \) (e.g. TREC benchmark)

**Macro evaluation** (user-oriented) =
\[
\frac{1}{n} \sum_{i=1}^{n} \text{precision}(q_i)
\]

**Micro evaluation** (system-oriented) =
\[
\frac{\sum_{i=1}^{n} \# \text{ relevant & found docs for } q_i}{\sum_{i=1}^{n} \# \text{ found docs for } q_i}
\]

analogous for recall and \( F_1 \)
Problems with IR-like 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 (raw estimate)
great for e-shopping, school kids, scientists, doctors, etc.

high accuracy for simple queries

high scalability
(now >15 Bio. docs, >10000 queries/sec)

continuously enhanced:
Froogle, Google Scholar, alerts, multilingual for >100 languages, query auto-completion, integrated tools (calculator, weather, news, etc. etc.)

Problem solved?
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
Looking for:

Q1: famous 19th century austrian politician born in Koblenz
Q2: famous 19th century austrian politician born in Metternich
Web Retrieval

Dr. Dr. Sergei Sizov
> Institute for Computer Science
> WeST

Outline

Information Retrieval refers to methods and technologies for search,
> analysis, and automatic organization of data collections: text documents,
> multimedia contents, structured and semi-structured knowledge
> representations. It has quickly become one of the most important areas in
> Computer and Information Sciences because of its direct applications in
> e-commerce, e-CRM, corporate knowledge bases and data repositories, Web
> analytics, and Web information systems.
> Recent technological and research trends, such as Linked Open Data and Web
> Science, are closely related to Information Retrieval and offer new
> perspectives for data/knowledge organisation and search. The course will
> introduce mathematical models and algorithms widely used by Web search
> engines, intranets, and modern digital libraries. In doing so, we will
> consider state of the art techniques from linear algebra, statistics, graph
> mining and machine learning. The course will also provide a brief overview
> of other areas in Web mining, such as Web content mining and Web structure
> mining.

open-source tool: GATE/ANNIE
http://www.gate.ac.uk/annie/
Issue: Web spam, advertising, manipulations

..not just for email anymore

Users follow search results
- Money follows users… Spam follows money…

There is value in getting ranked high
- Funnel traffic from SEs to Amazon/eBay/…

Make a few bucks
- Funnel traffic from SEs to a Viagra seller

Make $6 per sale
- Funnel traffic from SEs to a porn site

Make $20-$40 per new member
- Affiliate programs
Let’s do the math..

Assume 500M searches/day on the web
All search engines combined
Assume 5% commercially viable

Much more if you include „adult-only“ queries
- Assume $0.50 made per click (from 5c to $40)
- $12.5M/day or about $4.5 Billion/year
Link Spam: google bombs
Die Jagd auf die Hommingberger Gepardenforelle ist beendet. c’t hatte zum Wettstreit der Suchmaschinenoptimierer aufgerufen. Ziel war es, für den Begriff Hommingberger Gepardenforelle eine Top-Position in den Google.de-, Yahoo.de-, MSN.de- und Seekport.de-Ergebnislisten zu ergattern.


Der Wettbewerb gab einen kleinen Einblick in die Rankingmechanismen der Suchdienste und...
**Personalization:**
- query interpretation depends on personal interests and bias
- need to learn user-specific weights for multi-criteria ranking (relevance, authority, freshness, etc.)
- can exploit user behavior (feedback, bookmarks, query logs, click streams, etc.)

**Personal Information Management (PIM):**
- manage, annotate, organize, and search all your personal data
  - on desktop (mail, files, calendar, etc.)
  - at home (photos, videos, music, parties, invoices, tax filing, etc.)
  - and in smart home with ambient intelligence
Exploiting Query Logs and Clickstreams

from **PageRank**: uniformly random choice of links + random jumps
to **QRank**: + **query-doc transitions** + query-query transitions
  + **doc-doc transitions** on implicit links (w/ thesaurus)
  with probabilities estimated from log statistics

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

\[
QR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \left( \alpha \sum_{p \in \text{explicitIN}(q)} PR(p) \cdot t(p, q) + (1 - \alpha) \sum_{p \in \text{implicitIN}(q)} PR(p) \cdot \text{sim}(p, q) \right)
\]
Scenario: users have a potential interest in certain items
Goal: Provide recommendations for individual users
Examples:
  - recommendations to customers in an on-line store
  - movie recommendations

"If I have 2 million customers on the Web, I should have 2 million stores on the Web" (Jeff Bezos, CEO Amazon.com)

Types of recommendations:
  - display of customer comments
  - personalized recommendations based on buying decisions
  - customers who bought also bought.... (books/authors/artists)
  - email notifications of new items matching pre-specified criteria
  - explicit feedback (rate this item) to get recommendations
  - customers provide recommendation lists for topics
Recommender Systems: Example

Sergej, willkommen bei Amazon.de (Wenn Sie nicht Dr. Dr. Sergej Sizov sind, klicken Sie bitte hier.)

Software Empfehlungen
Lernspaß - 1. Klasse
Aus der Amazon.de-Redaktion
Verhängnisvoll klingt der Titel, bei dem sich wohl alle Eltern ersträuben, es möge den eigenen Kindern zielbewusst so ergeben. Lernen macht Spaß. Diese Software unterstützt Erstklässler in den Fächern Mathematik und Deutsch, steigert ihr Konzentrationsvermögen... Mehr dazu

Mehr gibt es in Kinder & Familie, Schule & Studium, und anderen Software Empfehlungen

DVD-Empfehlungen
The King And 1 [UK IMPORT]
Aus der Amazon.de-Redaktion
Der König und ich ist der dritte Broadway-Hit des berühmten Komponistenduos Rogers & Hammerstein. Der Film zeigt eine schauspielerische Leistung Yol Brynners, die seiner Karriere einen Schub nach oben verleih. Brynner wiederholte seinen Bühnenerfolg in der Hauptrolle und bewies den... Mehr dazu

Mehr gibt es in Originalfassungen, und anderen DVD-Empfehlungen

Buch-Empfehlungen
Guck mal, was hier passiert!
Kurzbeschreibung
Ein Wimmelbilderbuch zum Schauen, Entdecken, Wiedererkennen und natürlich zum Geschichtenverbinden - und -erzählen. (Ab 2 Jahren.)

Mehr gibt es in Kochen & Lifestyle, und anderen Buch-Empfehlungen
Web 1.0 vs. Web 2.0

SERVER

CLIENTS
What makes the difference?
Motivation: multi-modal, sparse Web 2.0 data.

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.
Formalizing the problem..

- Collaborative content sharing framework: $Y \subseteq U \times T \times R$
- $u \in U$, $t \in T$, $r \in R$

- Folksonomy cloud:
  - user-centric: $Y^* \subseteq Y$
  - resource-centric: $Y^*_u \subseteq \{u\} \times T \times R$
  - community-specific (e.g. groups or contacts)
  - collection-specific (e.g. favorites)
  - arbitrary
  - $Y^*_r \subseteq U \times T \times \{r\}$
  - $Y^* \subseteq U^* \times T \times R$
  - $Y^*_U \subseteq U \times T \times R^*$
  - $Y^*_UR^* \subseteq U^* \times T \times R^*$
The IR background – constructing feature vectors

\[ if(t) = \|\{(u^*, t, r^*)\}\|, (u^*, t, r^*) \in Y^* \]

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

with \( U^* \subseteq U, R^* \subseteq R \):

\( u^* \in U^* \iff \exists r \exists t : (u^*, t, r) \in Y^* \)

\( r^* \in R^* \iff \exists t \exists u : (u, t, r^*) \in Y^* \)

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

.. defined analogously to \( tf \cdot idf \)

Clouds of interest:
- 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>Global model</th>
<th>Personal model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training:10</td>
<td>Training:10</td>
</tr>
<tr>
<td></td>
<td>prec@10</td>
<td>prec@20</td>
</tr>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.167</td>
</tr>
<tr>
<td>Commented items</td>
<td>0.292</td>
<td>0.280</td>
</tr>
<tr>
<td>Favorites</td>
<td>0.757</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>Training:20</td>
<td>Training:20</td>
</tr>
<tr>
<td></td>
<td>prec@10</td>
<td>prec@20</td>
</tr>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.167</td>
</tr>
<tr>
<td>Commented items</td>
<td>0.296</td>
<td>0.279</td>
</tr>
<tr>
<td>Favorites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User items</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- German corner?
- Old Balduin bridge?
- Uni Koblenz?
- Nuclear power plant?
- Shopping malls?
- Steamer „Goethe“?
- Fortress?
- Restaurant Ferrari?

What can you see from Ehrenbreitstein?
Multi-modal Analysis of Social Media

Deutsches Eck from Ehrenbreitstein Fortress, Koblenz, Germany

by *schaengel*

121 comments 69 faves

Tagged with *koblenz, ehrenbreitstein...*

Taken on November 15, 2009, uploaded November 17, 2009

See more of *schaengel* photos, or visit his profile.
GeoFolk step by step

- Define the model
- Import training data

- Estimate model parameters (e.g. MCMC method)
  - we use software like JAGS for this..

- Use resulting topic-based feature vectors for tags and resources in IR-like scenarios (similarity based classification, clustering, ranked retrieval)
### Tag recommendation

$$MRR(Q) = \frac{1}{|Q|} \sum_{i} \frac{1}{rank_i}$$

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoFolk</td>
<td>0.212</td>
</tr>
<tr>
<td>LDA</td>
<td>0.119</td>
</tr>
<tr>
<td>Tags</td>
<td>0.073</td>
</tr>
<tr>
<td>Coordinates</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Web retrieval: course topics

- Motivation and Overview
- Technical basics
- Link Analysis and Authority Ranking
- Text processing and analysis
- Advanced IR models
- Classification and Clustering
- Web Spam and Advertising
- Web Crawling
- interactive social media
Recommended Literature (1)

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

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