The Course

Lecturers

Klaus Berberich
kberberi@mpi-inf.mpg.de

Pauli Miettinen
pmiettin@mpi-inf.mpg.de

Teaching Assistants

Amy Siu
sui@mpi-inf.mpg.de

Arunav Mishra
amishra@mpi-inf.mpg.de

Erdal Kuzey
ekuzey@mpi-inf.mpg.de

Kai Hui
khui@mpi-inf.mpg.de

Kaustubh Beedkar
kbeedkar@mpi-inf.mpg.de

Sourav Dutta
sdutta@mpi-inf.mpg.de

D5: Databases & Information Systems Group
Max Planck Institute for Informatics
Organization

• Lectures:
  – Tuesday 16-18 and Thursday 14-16
    in Building E1.3, HS-002

• Office hours:
  – Tuesday 14-16

• Assignments/tutoring groups
  – Monday 12-14 / 14-16 / 16-18, R021, E1.4 (MPI-INF building)
  – Friday 12-14 / 14-16, R021, E1.4 (MPI-INF building)

Assignments given out in Thursday lecture, to be solved until next Thursday
  – First assignment sheet given out on Thursday, Oct 17
  – First meetings of tutoring groups on Friday, Oct 25
Requirements for Obtaining 9 Credit Points

- **Pass 2 out of 3 written tests**
  Tentative dates: **Tue, Nov 12; Thu, Dec 12; Tue, Jan 28**
  (45-60 min each)

- **Pass the final written exam**
  Tentative date: **Tue, Feb 13** (120-180 min)

- **Must present solutions to 3 assignments**, more possible
  (You must return your assignment sheet and have a correct solution in order to present in the exercise groups.)
  - 1 bonus point possible in tutoring groups
  - **Up to 3 bonus points** possible in tests
  - Each bonus point earns one mark in letter grade
    (0.3 in numerical grade)
Register for Tutoring Groups


- Register for one of the tutoring groups until Oct 22
- Check back frequently for updates & announcements
Agenda

I.  Introduction
II. Probability theory, statistics, linear algebra
II. Ranking principles
III. Link analysis
IV. Indexing & searching
V.  Information extraction
VI. Frequent itemsets & association rules
VII. Unsupervised clustering
VIII. (Semi-)supervised classification
IX.  Advanced topics in data mining
X.  Wrap-up & summary

Information Retrieval

Data Mining
Literature (I)

• **Information Retrieval**


Literature (II)

• Data Mining

  – Mohammed J. Zaki, Wagner Meira Jr.
    *Data Mining and Analysis: Fundamental Concepts and Algorithms*
    Manuscript (will be made available during the semester)

  – Pang-Ning Tan, Michael Steinbach, Vipin Kumar.
    *Introduction to Data Mining*
    Addison-Wesley, 2006.
Literature (III)

• Background & Further Reading

  – Jiawei Han, Micheline Kamber, Jian Pei. 
    Website: [http://www.cs.sfu.ca/~han/dmbook](http://www.cs.sfu.ca/~han/dmbook)

  – Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack.
    *Information Retrieval: Implementing and Evaluating Search Engines*,
    MIT Press, 2010

  – David B. Skillicorn.
    *Understanding complex datasets: data mining with matrix decomposition*,
    Chapman & Hall/CRC, 2007

  – Christopher M. Bishop.

  – Larry Wasserman.
    *All of Statistics*, Springer, 2004
    Website: [http://www.stat.cmu.edu/~larry/all-of-statistics/](http://www.stat.cmu.edu/~larry/all-of-statistics/)
Quiz Time!

• Please answer the **20 quiz questions** during the rest of the lecture.

• The quiz is completely **anonymous**, but keep your id on the top-right corner. There will be a **prize for the 3 best answer** sheets.
Chapter I:
Introduction – Information Retrieval and Data Mining in a Nutshell

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2013/14
Chapter I: Information Retrieval and Data Mining in a Nutshell

• 1.1 Information Retrieval in a Nutshell
  – Search & beyond

• 1.2 Data Mining in a Nutshell
  – Real-world DM applications

„We are drowning in information, and starved for knowledge.“
-- John Naisbitt
I.1 Information Retrieval in a Nutshell

- Web, intranet, digital libraries, desktop search
- Unstructured/semi-structured data

Server farms with 10,000s (2002) – 100,000s (2010) computers, distributed/replicated data in high-performance file system (GFS, HDFS,…), massive parallelism for query processing (MapReduce, Hadoop,…)

strategies for crawl schedule and priority queue for crawl frontier
handle dynamic pages, detect duplicates, detect spam
build and analyze web graph, index all tokens or word stems
scoring function over many data and context criteria

GUI, user guidance, personalization
Politicians are worried that the Web is now dominated by search engine companies ...

Extraction of salient words

Linguistic methods: stemming, lemmas

Statistically weighted features (terms)

Search Engines

Politicians

worry

web

...
Vector Space Model for Relevance Ranking

**Ranking** by descending relevance

Search engine

Documents are feature vectors

**Query** \( q \in [0, 1]^{|F|} \) (feature vector)

**Similarity metric:**

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

with \( d_i \in [0, 1]^{\lfloor F \rfloor} \)

**e.g., using:**

\[
d_{ij} := w_{ij} / \sqrt{\sum_k w_{ik}^2}
\]

\[
w_{ij} := \log \left( 1 + \frac{\text{freq}(f_j, d_i)}{\max_k \text{freq}(f_k, d_i)} \right) \log \frac{\#\text{docs}}{\#\text{docs with } f_i}
\]

TF-IDF formula
Link Analysis for Authority Ranking

**Ranking** by descending relevance & authority

Search engine

Query \( q \in [0,1]^{|F|} \)

(=feature vector)

+ Consider in-degree and out-degree of web pages:

**Authority** \((d_i) := \)

Stationary visiting probability \([d_i]\)

in random walk on the Web (ergodic Markov Chain)

+ Reconciliation of relevance and authority by ad hoc weighting
Google’s PageRank [Page and Brin 1998]

• **Ideas:** (i) Hyperlinks are endorsements  
  (ii) Page is important if many important pages link to it  

• **Random walk** on web graph $G(V, E)$ with random surfer that randomly follows outgoing link or jumps to another random page

$$P(v) = (1 - \varepsilon) \sum_{(u, v) \in E} \frac{P(u)}{\text{out}(u)} + \frac{\varepsilon}{|V|}$$

• **PageRank** $P(v)$ corresponds to the stationary visiting probability of state $v$ in an ergodic Markov chain
Inverted Index

Vector space model suggests **term-document matrix**, but data is sparse and queries are even very sparse → better use **inverted index** with terms as keys for B+ tree

$q$: professor research xml

B+ tree on terms

<table>
<thead>
<tr>
<th>professor</th>
<th>...</th>
<th>research</th>
<th>...</th>
<th>xml</th>
</tr>
</thead>
<tbody>
<tr>
<td>17: 0.3</td>
<td></td>
<td>12: 0.5</td>
<td></td>
<td>11: 0.6</td>
</tr>
<tr>
<td>44: 0.4</td>
<td></td>
<td>14: 0.4</td>
<td></td>
<td>17: 0.1</td>
</tr>
<tr>
<td>52: 0.1</td>
<td></td>
<td>28: 0.1</td>
<td></td>
<td>28: 0.7</td>
</tr>
<tr>
<td>53: 0.8</td>
<td></td>
<td>44: 0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55: 0.6</td>
<td></td>
<td>51: 0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>52: 0.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Google:
> 10 Mio. terms
> 20 Bio. docs
> 10 TB index

index lists with postings (DocId, Score) sorted by DocId

terms can be full words, word stems, word pairs, substrings, N-grams, etc. (whatever “dictionary terms” we prefer for the application)

• index-list entries in **DocId order** for fast Boolean operations
• many techniques for excellent **compression** of index lists
• additional **position index** needed for phrases, proximity, etc.

(or other pre-computed data structures)
Evaluation of Search Result Quality

Ideal measure is “satisfaction of user’s information need” 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
Beyond Web Search…

• Find answers to “knowledge queries” and natural language questions (e.g., by scientists or journalists)
  – Who was German chancellor when Angela Merkel was born?
  – How are Max Planck, Angela Merkel, and the Dalai Lama related?
  – Which politicians are also entrepreneurs?
  – What was the population of Munich in 1972?
  – …

• Knowledge about entities (e.g., persons and locations), classes, attributes, relationships between them is required
  – focus on structured data sources (e.g., relational, XML, RDF)
  – perform information extraction on semi-structured & textual data
Google Knowledge Graph

About 127,000,000 results (0.44 seconds)

Albert Einstein - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Albert_Einstein

Albert Einstein (ˈaɪlˌbɜːrt ˈaɪnstaɪn; German: [ˈalbɛrt ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist who ... Hans Albert Einstein - Eduard Einstein - Mileva Marč - Elsa Einstein

Albert Einstein – Wikipedia
dw.wikipedia.org/wiki/Albert_Einstein ▪ Translate this page
Albert Einstein (14. März 1879 in Ulm; † 18. April 1955 in Princeton, New Jersey) war ein theoretischer Physiker. Seine Forschungen zur Struktur von Materie, ...
Relativitätstheorie - Ulm - Zionismus - Thomas Harvey

Albert Einstein - Biographical - Nobelprize.org
www.nobelprize.org/nobel_prizes/physics/laureates/.../einstein-bio.html ▪
Albert Einstein - Biographical. Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where ...

News for albert einstein
Providence digital-content company chosen for online publishing aspects of
Albert Einstein collection
The Providence Journal - 17 hours ago
PROVIDENCE – As Princeton University Press works to publish "The Collected Papers of Albert Einstein," it has selected a Providence ...
Maybe he's a relative? Caterpillar hears incredible resemblance to Albert Ein ...
Dally Mail - 3 days ago
Congress proves Albert Einstein's definition of insanity
San Jose Mercury News - 1 day ago

Einstein Archives Online
www.alberteinstein.info/ ▪
The homepage of the repository of the personal papers of the great scientist, humanist and Jew, Albert Einstein.

http://www.google.com
Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics. While best known for his mass-energy equivalence formula \( E = mc^2 \), he received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect". The latter was pivotal in establishing quantum theory. Near the beginning of his career, Einstein thought that Newtonian mechanics was no longer enough to reconcile the laws of classical mechanics with the laws of the development of his special theory of relativity. He realized, however, that the principle of relativity could also be extended to gravitational fields, and with his subsequent theory of gravitation in 1916, he published a paper on the general theory of relativity. He continued to develop quantum theory, which led to his explanations of photoelectric effect and the motion of molecules. He also investigated the thermal properties of light which laid the foundation of the photon theory of light. In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. Wikipedia [1]

http://www.freebase.com
http://www.yago-knowledge.org
### About: Dave Grohl

#### An Entity of Type: musical artist, from Named Graph: http://live.dbpedia.org, within Data Space: live.dbpedia.org

David Eric “Dave” Grohl (born January 14, 1969) is an American rock musician, multi-instrumentalist, singer-songwriter and film director, who is the lead vocalist, guitarist, primary or main songwriter and founder of the band Foo Fighters. Prior to Foo Fighters, Grohl was the drummer for the grunge band Nirvana. He is also the drummer and co-founder of the rock supergroup Them Crooked Vultures.

#### Property | Value
--- | ---
**dbpedia-owl:abstract** | The author of the property is not specified.
**dbpedia-owl:activeYearsStartYear** | 1984-01-01 00:00:00 (xsd:date)
**dbpedia-owl:activeYearsEndYear** | 1984-01-01 00:30:00 (xsd:date)
**dbpedia-owl:alias** | David Grohl, Dale Nixon, Latef (pseudonym for his solo album Pocketwatch), and Dr. G (as Tenacious D's drummer).
**dbpedia-owl:associatedBand** | Paul McCartney, Stevie Nicks, The Prodigy, Trent Reznor, Tom Petty and the Heartbreakers, Rick Springfield, Killing Joke, Probot, Mondo Generator, Tenacious D, Foo Fighters, Queens of the Stone Age, Them Crooked Vultures, Daim, Bramage, Nirvana (band), Slash (musician), Screem (band).
**dbpedia-owl:associatedMusicalArtist** | Paul McCartney, Stevie Nicks, The Prodigy, Trent Reznor, Tom Petty and the Heartbreakers, Rick Springfield, Killing Joke, Probot, Mondo Generator, Tenacious D, Foo Fighters, Queens of the Stone Age, Them Crooked Vultures, Daim, Bramage, Nirvana (band), Slash (musician), Screem (band).
**dbpedia-owl:background** | solo singer.
**dbpedia-owl:birthDate** | 1969-01-14 (xsd:date)
**dbpedia-owl:birthPlace** | United States, Ohio, Warren, Ohio, Norderico, Sweden.
**dbpedia-owl:genre** | hardcore punk, alternative rock, hard rock, heavy metal music, post-grunge, grunge.

[http://dbpedia.org](http://dbpedia.org)
The Linked Data Project

as of 2011:
• 295 sources
• 32 billion triples
• 504 million links

http://linkeddata.org
Jeopardy!

A big US city with two airports, one named after a World War II hero, and one named after a World War II battle field?

Chicago Midway International Airport (IATA: MDW, ICAO: KMDW, FAA LID: MDW), also known simply as Midway Airport or Midway, is an airport in Chicago, Illinois, United States, located on the city's southwest side, eight miles (13 km) from Chicago's Loop. The airport's current IATA code MDW has been used since 1949 when Chicago Municipal Airport was renamed Chicago Midway Airport,[8] although the airline schedule books continued to call it CHI until airline flights began at O'Hare. It is bordered by 55th Street, Cicero Avenue (terminal entrance), 63rd Street, and Central Avenue. The airport's northern half is within the Garfield Ridge community area, and the southern half is within the Clearing community area. The airport is managed by the Chicago Airport System, which also oversees operations at O'Hare International Airport and Gary/Chicago International Airport.[4] The airport is named after the Battle of Midway during World War II.

Midway is dominated by low-cost carrier Southwest Airlines. AirTran Airways and Delta Air Lines are the

A big US city with two airports, one named after a World War II hero, and one named after a World War II battle field?
Deep-QA in NL

William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel

This town is known as "Sin City" & its downtown is "Glitter Gulch"

As of 2010, this is the only former Yugoslav republic in the EU

99 cents got me a 4-pack of Ytterlig coasters from this Swedish chain

IRDM Research Literature

Important **conferences** on IR and DM
(see DBLP bibliography for full detail, http://www.dblp.org)
SIGIR, WSDM, ECIR, CIKM, WWW, KDD, ICDM, ICML, ECML

Important **journals** on IR and DM
(see DBLP bibliography for full detail, http://www.dblp.org)
TOIS, TOW, InfRetr, JASIST, InternetMath, TKDD, TODS, VLDBJ

Performance **evaluation/benchmarking** initiatives:
• Text Retrieval Conference (TREC), http://trec.nist.gov
• Cross-Language Evaluation Forum (CLEF), http://www.clef-campaign.org
• Initiative for the Evaluation of XML Retrieval (INEX),
  http://www.inex.otago.ac.nz/
• KDD Cup, http://www.kdnuggets.com/datasets/kddcup.html
  & http://www.sigkdd.org/kddcup/index.php