Web Mining

Introduction and Course Outline

Prof. Dr. Simone Ponzetto
Prof. Dr. Christian Bizer
Dr. Goran Glavaš
Dr. Stefano Faralli
FFS 2017
Hallo

Prof. Dr. Christian Bizer
Professor for Information Systems V
Research Interests:
- Data and Web Mining
- Web Data Integration
- Linked Data Technologies
Room: B6, 26 - B1.15
Consultation: Wednesday, 13:30-14:30
eMail: chris@informatik.uni-mannheim.de

Will teach the theory blocks on Web Usage Mining and Web Structure Mining
Hallo

Prof. Dr. Simone Ponzetto
Professor for Information Systems III

Research Interests:
- Knowledge Acquisition
- Natural Language Processing
- NLP Applications in DH and PolText

Room: B6, 26 - B1.14
Consultation: Tuesday, 13:30-14:30

NOTE: Notify me in advance via email if you plan to come to my office hours!

eMail: simone@informatik.uni-mannheim.de

Will teach the theory block on Web Content Mining
Hallo

- Dr. Stefano Faralli
- Postdoctoral Researcher
- Research Interests:
  - Natural Language Processing
  - Word Sense Disambiguation
  - Ontology Learning
- Room: B6, 26, C 1.11
- eMail: stefano@informatik.uni-mannheim.de

- Will teach the exercise on Recommender Systems and Social Network Analysis and coaches the student projects
Hallo

- Dr. Goran Glavaš
- Postdoctoral Researcher
- Research Interests:
  - Statistical Natural Language Processing
  - Information Retrieval
  - Machine Learning
- Room: B6, 26, C 1.02
- eMail: goran@informatik.uni-mannheim.de

- Lectures and exercises on Web Content Mining (sentiment analysis and information extraction); coaching student projects
Introduction and Course Organization

1. Course Organization

2. The World Wide Web
   1. The Classic Document Web
   2. The Web of Data
   3. Web 2.0 Applications

3. What is Web Mining?
   1. Web Usage Mining
   2. Web Structure Mining
   3. Web Content Mining

4. The Web Mining Process
1. Course Organization

■ Lecture
  ■ covers different types of Web Mining methods
  ■ presents examples of Web Mining applications
  ■ discusses how to evaluate learned models

■ Exercise
  ■ students experiment with given tools and given data sets

■ Project Work
  ■ teams of four students realize a Web Mining project
  ■ teams may choose their own data sets and tasks
    (in addition, we will propose some suitable data sets and tasks)
  ■ write summary about the project, present the project results

■ Final exam
  ■ 50 % written exam
  ■ 50 % project work
Course Organization

Course Webpage
- contains general information and lecture slides

Exercise Material
- ILIAS eLearning System, https://ilias.uni-mannheim.de/

Time and Location
- Tuesday, 15:30 to 17:00, Room: B 6, A104
- Thursday, 15:30 to 17:00, Room: B 6, A104
Waiting List

- There are currently 40 people on the waiting list
  - if you decide not to attend, please leave ILIAS group this week
  - so others can have your place

- Policy: Two strikes out
  - you have to attend this lecture or the lecture next Tuesday (22.02.2017)
  - If you are not attending, you will be deleted from the participants list
  - Please sign now next to your name on the participants list

- If you are on the waiting list
  - you may get assigned a place next Wednesday
  - The waiting list is cleared after this semester
    (i.e., no priority for next semester!)
<table>
<thead>
<tr>
<th>Week</th>
<th>Topic Tuesday</th>
<th>Topic Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.02.2017</td>
<td>Lecture: Introduction to Web Mining</td>
<td>Lecture: Web Usage Mining</td>
</tr>
<tr>
<td>07.03.2017</td>
<td>Lecture: Social Network Analysis</td>
<td>Exercise: Introduction to Pajek</td>
</tr>
<tr>
<td>14.03.2017</td>
<td>Exercise: Social Network Analysis</td>
<td>Exercise: Social Network Analysis</td>
</tr>
<tr>
<td>21.03.2017</td>
<td>Lecture: Web Content Mining: Sentiment analysis</td>
<td>Exercise: Sentiment Analysis</td>
</tr>
<tr>
<td>04.04.2017</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outlines</td>
</tr>
<tr>
<td></td>
<td>- Easter break -</td>
<td></td>
</tr>
<tr>
<td>25.04.2017</td>
<td>Feedback about Project Outlines</td>
<td>Project work</td>
</tr>
<tr>
<td>02.05.2017</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>09.05.2017</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>16.05.2017</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>23.05.2017</td>
<td>Project work</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>30.05.2017</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>14.02.2017</td>
<td>Lecture: Introduction to Web Mining</td>
<td>Lecture: Web Usage Mining</td>
</tr>
</tbody>
</table>
Lecture Videos

Video recordings of all lectures from FSS 2016

http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/
Literature


Software

- Open-source machine learning library which can be deployed on a single machine as well as on Hadoop and Spark clusters
- Implements various recommendation algorithms
- Download: http://mahout.apache.org/
Software

- Tool for the analysis and visualization of large networks
- Download: http://mrvar.fdv.uni-lj.si/pajek/
Software

- Stanford CoreNLP: A Suite of Core NLP Tools
- Download: http://stanfordnlp.github.io/CoreNLP/

Named Entity Recognition:

1. President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Coreference:

1. President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Basic Dependencies:
Software

- Powerful data mining suite
- Download: https://rapidminer.com/

![RapidMiner Software](https://rapidminer.com/)

**Features:**
- Powerful data mining suite
- Download at [https://rapidminer.com/](https://rapidminer.com/)
Questions?
2. The World Wide Web

The Web is a global information space build on a set of technical standards for the identification, retrieval and representation of content.

- **Uniform Resource Identifiers (URIs):** Globally unique identification of Web resources.

- **Hypertext Transfer Protokoll (HTTP):** Protocol for interacting with Web resources.

- **Content Formats:** HTML, XML, RDF, ...

- **The Web was invented in 1989 at CERN by Tim Berners-Lee**

- **Architectural Principles of the Web**

Topology of the Web Today

The Classic Document Web

The Web of Data

Web 2.0 Applications

HTML

RDFa

Twitter

Facebook

Google+
2.1 The Classic Document Web

Global information space consisting of interlinked Web documents (text, images, multimedia).

The Size of the Web

- **Overall Size**: 1 trillion URLs on the Web at once
  - announced by Google in 2008
  - [http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html](http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html)

- **Indexed Web**: approx. 50 billion pages
  - estimate based on search engine hit counts for popular words
  - Example: The word „the“ appears in 67% of all English pages and has 25.2 billion hits on Google
  - [http://www.worldwidewebsize.com/](http://www.worldwidewebsize.com/)
The link distribution follows (kind of) a power law.
- A small number of pages is target of many links.
- A large number of pages is target of only a few or no links.

Classic Paper:
- AltaVista crawl with over 200 million pages and 1.5 billion links
- Conclusion: Log-log scale plot shows power-law.
In-Degree Distribution

Broder et al. (2000)
Power law with exponent 2.1
(200 million pages and 1.5 billion links from Altavista crawl 2000)

WDC Hyperlink Graph (2012)
Best power law exponent 2.24
(3 billion pages and 128 billion links from Common Crawl 2012)

Too small number of high in-degree pages for power law
Four major components (Border at al., WWW2000)

- **Central Strongly Connected Component (SCC)**
  - Pages that can reach one another along directed links
  - About 30% of the Web (normal pages)

- **IN Group**
  - Can reach SCC but cannot be reached from it
  - About 20% (maybe new pages or boring ones)

- **OUT Group**
  - Can be reached from SCC but cannot reach it
  - About 20% (maybe company pages that don’t link)

- **Tendrils**
  - Cannot reach SCC and cannot be reached by it
  - About 20%

- **Unconnected**
  - About 10%

Probabilty of path between nodes is 24%
A strongly connected component (SCC) in a directed graph is a subset of the nodes such that:

1. every node in the subset has a path to every other node
2. the subset is not part of some larger set with the property that every node can reach every other.
Largest Strongly Connected Component

Largest SCC

- Broder, 2000: 27.7%
- WDC, 2012: 51.3 %

→ Factor 1.8 larger

→ Also, factor 4.9 more links/page

→ The Web has become denser.

- SCC calculated using a machine with 1 TB RAM.
Hands-on: Accessing the Document Web for Mining

1. Crawl the Web Yourself
   - See Bing Liu: Web Data Mining, Chapter 8.
   - Small focused crawls
     - Single-machine Java library: Crawler4j
   - Large-scale crawls
     - Apache Nutch (runs on hadoop clusters)

2. Use Existing Crawls
   1. Common Crawl: Non-profit organization which provides large web crawls on Amazon S3 for free.
      - 2012 CC Corpus: 3.0 billion HTML pages (48 Terabyte compressed)
      - 2017 CC Corpus: 3.1 billion HTML pages (250 Terabyte uncompressed)
   2. ClueWeb 2012 Crawl: 733 million English HTML pages
   3. Internet Archive 2011: 2.7 billion URIs (80 Terabyte)
More and more Websites
- semantically markup the content of their HTML pages
- publish structured data in addition to HTML pages
Microformats

- Microformat effort dates back to 2003
- Small set of fixed formats
  - hcard: people, companies, organizations, and places
  - XFN: relationships between people
  - hCalendar: calendaring and events
  - hListing: small-ads; classifieds
  - hReview: reviews of products, businesses, events

- Shortcoming of Microformats
  - can not represent any kind of data.

- indexed by Google and Yahoo since 2009
RDFa

- serialization format for embedding RDF data into HTML pages
- proposed in 2004, W3C Recommendation in 2008
- can be used together with any vocabulary

```html
<html xmlns="http://www.w3.org/1999/xhtml"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">
  ...
  <div about="http://example.com/Peter" typeof="foaf:Person">
    <span property="foaf:name">Peter Smith</span> knows
  </div>
  ...
</html>
```
Open Graph Protocol

- allows site owners to determine how entities are described in Facebook
- relies on RDFa for encoding data in HTML pages
- available since April 2010
Microdata

- alternative technique for embedding structured data
- proposed in 2009 by WHATWG as part of HTML5 work
- tries to be simpler than RDFa (5 new attributes instead of 8)

```html
1  <div itemscope itemtype="http://schema.org/Person" itemid="http://example.com/Peter">
2    <span itemprop="name">Peter Smith</span>
3    <a href="http://example.com/Paula" itemprop="knows">Paula Jones</a>
4  </div>
```
Schema.org

- ask site owners to embed data to enrich search results.
- 200+ Types: Event, Organization, Person, Place, Product, Review
- Encoding: Microdata, RDFa, JSON-LD
Usage of Schema.org Data @ Google

**Gramercy Tavern - Flatiron - New York, NY | Yelp**

*www.yelp.com > Restaurants > American (New)*

★★★★★ Rating: 4.5 - 1,288 reviews - Price range: $$$$

Jeff C and I were in New York for vacation, and I wanted to treat him to a nice dinner for..... Gramercy Tavern is certainly a legendary NY dining establishment.

**Gramercy Tavern Restaurant - New York, NY | OpenTable**

*www.opentable.com > ... > Gramercy restaurants*

★★★★★ Rating: 4.7 - 508 reviews - Price range: $50 and over

Book now at Gramercy Tavern in New York, explore menu, see photos and read 508 reviews: “The menu was so limited but it was worth trying, food was deli...”

---

**The Black Keys**

**Band**

The Black Keys is an American rock duo formed in Akron, Ohio in 2001. The group consists of Dan Auerbach and Patrick Carney. [Wikipedia](#)

**Origin:** Akron, Ohio, United States

**Members:** Dan Auerbach, Patrick Carney

**Record labels:** Fat Possum Records, Nonesuch Records, V2 Records, Alive Naturalsound Records

**Awards:** Grammy Award for Best Rock Album, more

**Upcoming events**

<table>
<thead>
<tr>
<th>Date</th>
<th>The Black Keys</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun 20</td>
<td>The Black Keys</td>
<td>Neuhausen ob Eck</td>
</tr>
<tr>
<td>May 16</td>
<td>The Black Keys</td>
<td>Gulf Shores, AL</td>
</tr>
<tr>
<td>Jun 22</td>
<td>The Black Keys</td>
<td>Scheeßel</td>
</tr>
</tbody>
</table>
Microformat, Microdata, RDFa Deployment

- **WebDataCommons.org Project**
  - extracts all Microformat, Microdata, RDFa data from the Common Crawl
  - provides the extracted data for download

- **Four extractions runs**
  - 2012 CC Corpus: 3.0 billion HTML pages → 7.3 billion RDF triples
  - 2013 CC Corpus: 2.2 billion HTML pages → 17.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 billion RDF triples
  - 2016 CC Corpus: 3.2 billion HTML pages → 44.0 billion RDF triples

- **uses 100 machines on Amazon EC2**
  - approx. 3000 machine/hours
    (spot instances of type c3.xlarge) → 550 Euro

- [http://www.webdatacommons.org/structureddata/](http://www.webdatacommons.org/structureddata/)
Overall Adoption 2016

1.2 billion HTML pages out of the 3.2 billion pages provide semantic annotations (38%).

5.6 million pay-level-domains (PLDs) out of the 32 million pay-level-domains covered by the crawl provide annotations (16.5%).
## Top Classes

### Topics:
- CMS and blog metadata
- products and offers
- ratings and reviews
- business listings
- address data
- ...and a massive long tail

<table>
<thead>
<tr>
<th>Class</th>
<th>2014 Instances #</th>
<th>2013 PLDs #</th>
<th>2014 PLDs %</th>
<th>2013 PLDs %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. schema:WebPage</td>
<td>51,757,000</td>
<td>148,893</td>
<td>18,16%</td>
<td>69,712</td>
</tr>
<tr>
<td>2. schema:Article</td>
<td>54,972,000</td>
<td>88,7</td>
<td>10,82%</td>
<td>65,930</td>
</tr>
<tr>
<td>3. schema:Blog</td>
<td>3,787,000</td>
<td>110,663</td>
<td>13,50%</td>
<td>64,709</td>
</tr>
<tr>
<td>4. schema:Product</td>
<td>288,083,000</td>
<td>89,608</td>
<td>10,93%</td>
<td>56,388</td>
</tr>
<tr>
<td>5. schema:PostalAddress</td>
<td>48,804,000</td>
<td>101,086</td>
<td>12,33%</td>
<td>52,446</td>
</tr>
<tr>
<td>6. dv:Breadcrumb</td>
<td>269,088,000</td>
<td>76,894</td>
<td>9,38%</td>
<td>44,187</td>
</tr>
<tr>
<td>7. schema:AggregateRating</td>
<td>59,070,000</td>
<td>50,510</td>
<td>6,16%</td>
<td>36,823</td>
</tr>
<tr>
<td>8. schema:Offer</td>
<td>236,953,000</td>
<td>62,849</td>
<td>7,66%</td>
<td>35,635</td>
</tr>
<tr>
<td>9. schema:LocalBusiness</td>
<td>20,194,000</td>
<td>62,191</td>
<td>7,58%</td>
<td>35,264</td>
</tr>
<tr>
<td>10. schema:BlogPosting</td>
<td>11,458,000</td>
<td>65,397</td>
<td>7,98%</td>
<td>32,056</td>
</tr>
<tr>
<td>11. schema:Organization</td>
<td>101,769,000</td>
<td>52,733</td>
<td>6,43%</td>
<td>24,255</td>
</tr>
<tr>
<td>12. schema:Person</td>
<td>115,376,000</td>
<td>47,936</td>
<td>5,85%</td>
<td>21,107</td>
</tr>
<tr>
<td>13. schema:ImageObject</td>
<td>35,356,000</td>
<td>25,573</td>
<td>3,12%</td>
<td>16,084</td>
</tr>
<tr>
<td>14. dv:Product</td>
<td>124,111,000</td>
<td>16,003</td>
<td>1,95%</td>
<td>13,844</td>
</tr>
<tr>
<td>15. schema:Review</td>
<td>42,561,000</td>
<td>20,124</td>
<td>2,45%</td>
<td>13,137</td>
</tr>
<tr>
<td>16. dv:Review-aggregate</td>
<td>3,964,000</td>
<td>14,094</td>
<td>1,72%</td>
<td>13,075</td>
</tr>
<tr>
<td>17. dv:Organization</td>
<td>3,155,000</td>
<td>10,649</td>
<td>1,30%</td>
<td>9,582</td>
</tr>
<tr>
<td>18. dv:Offer</td>
<td>7,170,000</td>
<td>11,64</td>
<td>1,42%</td>
<td>9,298</td>
</tr>
<tr>
<td>19. dv:Address</td>
<td>2,138,000</td>
<td>9,674</td>
<td>1,18%</td>
<td>8,866</td>
</tr>
<tr>
<td>20. dv:Rating</td>
<td>1,732,000</td>
<td>9,367</td>
<td>1,14%</td>
<td>8,360</td>
</tr>
</tbody>
</table>

schema: = Schema.org  
dv: = Google Rich Snippet Vocabulary (deprecated)
Adoption by E-Commerce Websites

Distribution by Top-Level Domain

<table>
<thead>
<tr>
<th>TLD</th>
<th>#PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>38344</td>
</tr>
<tr>
<td>co.uk</td>
<td>3605</td>
</tr>
<tr>
<td>net</td>
<td>1813</td>
</tr>
<tr>
<td>de</td>
<td>1333</td>
</tr>
<tr>
<td>pl</td>
<td>1273</td>
</tr>
<tr>
<td>com.br</td>
<td>1194</td>
</tr>
<tr>
<td>ru</td>
<td>1165</td>
</tr>
<tr>
<td>com.au</td>
<td>1062</td>
</tr>
<tr>
<td>nl</td>
<td>1002</td>
</tr>
</tbody>
</table>

Adoption by Top-15: 60%

Alexa Top-15 Shopping Sites

<table>
<thead>
<tr>
<th>Website</th>
<th>schema:Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon.com</td>
<td>✗</td>
</tr>
<tr>
<td>Ebay.com</td>
<td>✓</td>
</tr>
<tr>
<td>NetFlix.com</td>
<td>✗</td>
</tr>
<tr>
<td>Amazon.co.uk</td>
<td>✗</td>
</tr>
<tr>
<td>Walmart.com</td>
<td>✓</td>
</tr>
<tr>
<td>etsy.com</td>
<td>✗</td>
</tr>
<tr>
<td>Ikea.com</td>
<td>✓</td>
</tr>
<tr>
<td>Bestbuy.com</td>
<td>✓</td>
</tr>
<tr>
<td>Homedepot.com</td>
<td>✓</td>
</tr>
<tr>
<td>Target.com</td>
<td>✓</td>
</tr>
<tr>
<td>Groupon.com</td>
<td>✗</td>
</tr>
<tr>
<td>Newegg.com</td>
<td>✓</td>
</tr>
<tr>
<td>Lowes.com</td>
<td>✗</td>
</tr>
<tr>
<td>Macys.com</td>
<td>✓</td>
</tr>
<tr>
<td>Nordstrom.com</td>
<td>✓</td>
</tr>
</tbody>
</table>
Adoption by Travel Websites

<table>
<thead>
<tr>
<th>Top 15 Travel Websites</th>
<th>schema:Hotel</th>
<th>Any Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Booking.com (uses DataVoc)</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>TripAdvisor</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Expedia</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Agoda</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Hotels.com</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Kayak</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Priceline</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Travelocity</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Orbitz</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>HolidayCheck</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>InterContinental Hotels Group</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Marriott International</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Global Hyatt Corp.</td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>

Adoption: 73 %
Alternative Approach: Linked Data

Extend the Web with a single global data graph

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.
Entities are identified with HTTP URIs

HTTP URIs take the role of global primary keys.

pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

dbpedia:Berlin = http://dbpedia.org/resource/Berlin
URIs can be looked up on the Web

By following RDF links applications can

- navigate the global data graph
- discover new data sources

Richard Cyganiak

By following RDF links applications can

- navigate the global data graph
- discover new data sources
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>event</td>
<td>...</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>G1</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/cygri.rdf">http://richard.cyganiak.de/cygri.rdf</a></td>
<td>G2</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/foaf.rdf">http://richard.cyganiak.de/foaf.rdf</a></td>
<td>G3</td>
</tr>
<tr>
<td>nearest airport</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>phone</td>
<td>tel:+49-175-5630408</td>
<td>G1</td>
</tr>
<tr>
<td>sameAs</td>
<td>Richard Cyganiak</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G1</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs">http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs</a></td>
<td>G3</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq">http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq</a></td>
<td>G3</td>
</tr>
<tr>
<td>depiction</td>
<td>[Image]</td>
<td>G4</td>
</tr>
</tbody>
</table>
### Berlin

**URI:** http://dbpedia.org/resource/city/Berlin

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>population</strong></td>
<td>3398888</td>
<td>G2</td>
</tr>
<tr>
<td><strong>type</strong></td>
<td><a href="http://dbpedia.org/City">http://dbpedia.org/City</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>comment</strong></td>
<td>Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union.</td>
<td>G2</td>
</tr>
<tr>
<td><strong>comment</strong></td>
<td>Berlin ist die deutsche Bundeshauptstadt und als Stadtstaat ein eigenständiges Land der Bundesrepublik Deutschland. Berlin ist die bevölkerungsreichste und flächengrößte Stadt Deutschlands und nach Einwohnern die zweitgrößte Stadt der EU.</td>
<td>G2</td>
</tr>
<tr>
<td><strong>label</strong></td>
<td>Berlin</td>
<td>G2</td>
</tr>
<tr>
<td><strong>sameAs</strong></td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/Berlin">http://dbpedia.org/resource/category/Berlin</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/Capitals_in_Europe">http://dbpedia.org/resource/category/Capitals_in_Europe</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/Cities_in_Germany">http://dbpedia.org/resource/category/Cities_in_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/German_state_capitals">http://dbpedia.org/resource/category/German_state_capitals</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games">http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>subject</strong></td>
<td><a href="http://dbpedia.org/resource/category/States_of_Germany">http://dbpedia.org/resource/category/States_of_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td><strong>sourceURL</strong></td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td><strong>depiction</strong></td>
<td><img src="http://dbpedia.org/resource/city/Berlin" alt="Image" /></td>
<td>G2</td>
</tr>
<tr>
<td><strong>is birthplace of</strong></td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Linked Data Search Engine: SigMa

Chris Bizer

given name: Chris [3,5,9,10,16]
family name: Bizer [3,5,9,10,16]
is creator of: DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food [6,18]
The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]
D2R Server - Publishing Relational Databases on the Semantic Web. [16]
Named Graphs, Provenance and Trust [16]
LOD Datasets on the Web (April 2014)

http://lod-cloud.net/
## Categorization by Topical Domain

<table>
<thead>
<tr>
<th>Category</th>
<th>Datasets 2014</th>
<th>Percentage</th>
<th>Datasets 2011</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>24 (-2)</td>
<td>2%</td>
<td>25</td>
<td>-4%</td>
</tr>
<tr>
<td>Government</td>
<td>199 (-16)</td>
<td>18%</td>
<td>49</td>
<td>306%</td>
</tr>
<tr>
<td>Publications</td>
<td>138 (-42)</td>
<td>13%</td>
<td>87</td>
<td>59%</td>
</tr>
<tr>
<td>Geographic</td>
<td>27 (-6)</td>
<td>2%</td>
<td>31</td>
<td>-13%</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>85 (-2)</td>
<td>8%</td>
<td>41</td>
<td>107%</td>
</tr>
<tr>
<td>Cross-domain</td>
<td>47 (-6)</td>
<td>4%</td>
<td>41</td>
<td>15%</td>
</tr>
<tr>
<td>User-generated Content</td>
<td>51 (-3)</td>
<td>5%</td>
<td>20</td>
<td>155%</td>
</tr>
<tr>
<td>Social Networking</td>
<td>520 (0)</td>
<td>48%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1091 (-77)</strong></td>
<td><strong>294</strong></td>
<td><strong>271%</strong></td>
<td></td>
</tr>
</tbody>
</table>

**More statistics**

[http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/](http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/)
Hands-on: How to access the Data for Mining?

- **Download the Billion Triples Challenge Dataset**
  - 4 billion triples (57 GB gzipped)
  - crawled from the public Web of Linked Data in February/June 2014

- **Download the Web Data Commons Dump**
  - 44 billion triples (900 GB gzipped)
  - RDFa, Microdata, Microformat data crawled October 2016
  - [http://webdatacommons.org/structureddata/](http://webdatacommons.org/structureddata/)
2.3. Web 2.0 Applications

- A multitude of Web-based applications has sprung up which enable users to share information.

- These applications
  - collect large amounts of data using proprietary schemata.
  - form separate data spaces that are only partly accessible from the Web via:
    1. HTML interfaces
    2. Web APIs
Example: Facebook

- **Users (September 2012)**
  - 1 billion monthly active users
  - including 600 million mobile users
  - **140.3 billion** friend connections
  - **1.13 trillion** likes since launch in February 2009
  - 219 billion photos uploaded
  - 17 billion location-tagged posts, including check-ins

- **Data Volume**
  - over 100 Petabyte
  - including profile data, communication, usage logs, photos, ...

**Source**
Web APIs

- Provide limited access to the collected data
  - restricted to specific queries
  - restricted number of queries

- Web API Catalog
  - lists over 14,000 Web APIs (2016)
  - lists over 7,300 Mashups
Mashups are based on a fixed set of data sources

Web APIs expose proprietary interfaces

No single global data space

Not index-able by generic crawlers

No automatic discovery of additional data sources
Web APIs slice the Web into Data Silos
3. What is Web Mining?

- Definition

Non-trivial extraction of implicit, previously unknown and potentially useful information from
  - Web content,
  - Web structure and
  - Web usage data.

- Recurring Challenges

1. huge amount of available data ➔ requires sampling or multiple machines
2. semi-structured nature of data ➔ often mixture of data and text mining
3. heterogeneity of data ➔ data integration might be a challenge
4. distributed nature of data ➔ often requires large-scale crawling
Web Mining is a Multi-Disciplinary Field

- Draws ideas and techniques from
  - Machine Learning
  - Database Systems
  - Natural Language Processing
  - Social Network Analysis

- Sub-Fields
  1. Web Usage Mining
  2. Web Structure Mining
  3. Web Content Mining
3.1 Web Usage Mining

Definition

Discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more web sites.

Sources of Data

1. automatically generated data stored in server access logs
2. e-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-throughs, purchases)
3. user profiles (e.g. Facebook) and/or user ratings (likes)
4. page attributes, page content, site structure
5. additional domain knowledge and demographic data
Leading Usage Data Collections

Enable the
- analysis of the current interests and behavior of the world’s population.
- identification of suspected terrorists.
Simple Summarization Statistics

- is the most common form of analysis.
- gives a quick overview of how a site is being used.
Web Usage Mining: Going deeper

- Prediction of the next event
- Discovery of associated events or application objects
- Recommendation of products and content
- Discovery of visitor groups with common properties and interests
- Discovery of visitor groups with common behaviour
- Characterization of visitors into predefined classes
- Card fraud detection

- Sequence mining
  - Markov chains
  - Association rules
  - Recommendation Algorithms
  - Clustering
  - Session Clustering
  - Classification

- Prediction of the next event
- Discovery of associated events or application objects
- Recommendation of products and content
Example Application: Website Improvement

- Discovery of common navigation paths
- Discovery of pages on which users leave the site or discontinue shopping process
- Provide personalized websites tailored to the user's needs

Technique: Sequential Pattern Mining
Example Application: Product Recommendations

Technique: Collaborative Filtering
Example Application: Personalized Search

Google

Search

50 personal results and 419,000 other results (0.61 seconds)

Manilkara zapota - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Manilkara_zapota

Sapodilla is known as chikoo ("चिकू" or chiku, "ชิโค้") in India and Pakistan and sapota in some parts of India (Tamil Nadu, Kerala, Karnataka, Andhra ... Description - Other names - See also - References

You've visited this page 3 times. Last visit 12/4/11

Images for chikoo - Report images

Mountain View, CA
Change location

Chikoo - a simple file organizer for the Mac
codingturtle.com/chikoo/
Example Application: Search Log Mining

Analysis of search queries entered by a user.

- **Examples:**
  1. Query completion using association analysis
  2. Query topic detection using classification
3.2 Web Structure Mining

- Definition

Discovery of patterns in

- the hyperlink structure of webpages
- the structure of communities that interact on the Web

- Exploits the graph structure, but can of course also be combined with content or usage mining techniques.

- Typical Sources of Data
  1. Web crawls including HTML pages and hyperlinks
  2. Crawls of the blogosphere
  3. Social networks including explicit relations between actors (your Facebook friend network)
  4. Other types of community data (discussion forums, email conversations, …)
Identification of Prominent Nodes

Question: Who are the “most important” actors in a social network?

Centrality

- A central actor is one involved in many edges.
- The direction of lines is not considered.

Prestige

- A prestigious actor is one who is the target of many arcs.
- The direction of arcs is considered.
Example Application: Ranking Search Results

- PageRank
  - exploits the hyperlinks of the Web to rank pages according to their level of “prestige”
  - a page is prestigious if many other prestigious pages link to it
  - initial algorithm used by Google
  - today, one important factor amongst many others in the Google ranking algorithm.
Community Detection

A community is a set of actors between which interactions are (relatively) frequent.

- Finding a community in a social network is to identify a set of nodes such that they interact with each other more frequently than with those nodes outside the group.

- Methods: Components, K-Cores, Islands, ...

- Applications: Recommendation based on communities, visualization of huge networks, network compression
3.3 Web Content Mining

- **Definition**

  Automatic extraction of useful information (facts, patterns) from Web content (text, images, multimedia).

- **Content Mining Tasks**
  - Content Clustering
  - Content Classification
  - Sentiment Analysis
  - Information Extraction
Content Clustering

- Unsupervised Learning: Given a set of documents and a similarity measure among documents find clusters such that:
  - documents in one cluster are more similar to one another
  - documents in separate clusters are less similar to one another

- Example Application
  - Google News: Find similar, but not too similar news stories

- Techniques
  - Algorithms: K-Means, K-Medoids, DBScan
  - Similarity measures: Cosine, Jaccard
Supervised Learning: Given a collection of labeled documents/images (training set) find a model for the class as a function of the values of the features.

Goal: Previously unseen documents/images should be assigned a class as accurately as possible.

Applications
- Classification of news into categories
- SPAM detection
- Filtering ‘boring’ documents based on personal profiles
- Product categorization exploiting product images

Classification methods commonly used for
- Text: Naive Bayes, Support Vector Machines, Deep Neural Nets
- Images: Support Vector Machines, Deep Neural Nets
Sentiment Analysis

The basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level.

- **Polarity Values**
  - Positive, neutral, negative
  - Likert scale (1 to 10)

- **Application Examples**
  - **Document-Level**
    - tweet analysis about politicians
  - **Feature/Aspect-Level**
    - analysis of product reviews
Information extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured machine-readable documents.

- **Subtasks**
  - **Named Entity Recognition and Disambiguation**
    - “M. Smith likes fishing“
    - Which M. Smith?
  - **Coreference Resolution**
    - “M. Smith likes fishing. But he doesn't like biking.”
    - Does he refer to M. Smith?
  - **Relation Extraction**
    - PERSON works for ORGANIZATION
    - PERSON located in LOCATION
Example: Named Entity Recognition

- Allows you to automatically interlink blog posts with background knowledge (like Wikipedia pages)

```
Batman became a very popular character soon after his introduction and gained his own comic book title, Batman, in 1940. A cultural icon, Batman has been licensed and adapted into a variety of media, from radio to television and film, and appears on a variety of merchandise sold all over the world, such as toys and video games.

The late 1960s Batman television series used a camp aesthetic which continued to be associated with the character for years after the show ended. Various creators worked to return the character to his dark roots, culminating in the 1986 miniseries The Dark Knight Returns, by Frank Miller, while the successes of Tim Burton's 1989 film Batman and Christopher Nolan's 2005 reboot Batman Begins also helped to reignite popular interest in the character.
```
Example: Relation and Fact Extraction

- **Google Squared** (former Google Labs prototype)

- **Paper about information extraction from all HTML tables on the Web**

- **WebDataCommons Web Tables Corpus containing 233 million tables**
  - [http://webdatacommons.org/webtables/](http://webdatacommons.org/webtables/)
Example: Information Extraction from Wikipedia
Wikipedia Information Extraction Projects

- DBpedia
- YAGO
- Wikidata
- Universität Mannheim
- Universität Leipzig
- max planck institut informatik
- Google
- Wikimedia Foundation
The DBpedia Knowledge Base - Version 3.9

- describes 4.00 million things, out of which 3.22 million are classified in a consistent ontology using 529 classes and 2217 different properties
  - 832,000 persons
  - 639,000 places
  - 209,000 organizations
  - 116,000 music albums
- Altogether 2.46 billion pieces of information (RDF triples)
  - 24,000,000 links to external web pages
  - 27,200,000 external links into other RDF datasets
- DBpedia Internationalization
  - provide data from 119 Wikipedia language editions for download
  - 24 popular languages we provide cleaned infobox data
**Highcliff**

Highcliff is a 252.4-metre (828-foot) tall skyscraper located on a south slope of Happy Valley on the Hong Kong Island in Hong Kong. The 75 storey (70 floors of which are livable space) building's construction began in 2000 and was completed in 2003 under a design by DLN Architects & Engineers. It was the Silver Winner of the 2003 Emporis Skyscraper Award, coming in second to 30 St Mary Axe in London.

**The Harbourside**

The Harbourside is a 255 m (836.6 ft) tall residential skyscraper located at 1 Austin Road West, in Union Square complex on Kowloon peninsula. The building is erected on the West Kowloon Reclamation west of Kwun Chung. Construction of the 74 storey building began in 2000 and was completed in 2003 under the design by P & T Architects & Engineers. The building is, in fact, three towers joined at the base, middle
Applications of Google’s Knowledge Graph

1. Answer fact queries: “birthdate michael douglas”

2. Compare things: „compare eiffel tower vs empire state building”

<table>
<thead>
<tr>
<th></th>
<th>Eiffel Tower</th>
<th>Empire State Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>324 m</td>
<td>381 m</td>
</tr>
<tr>
<td>Floors</td>
<td>3</td>
<td>102</td>
</tr>
<tr>
<td>Architect</td>
<td>Stephen Sauvestre</td>
<td>William F. Lamb</td>
</tr>
<tr>
<td>Construction started</td>
<td>January 20, 1887</td>
<td>1929</td>
</tr>
<tr>
<td>Architecture firms</td>
<td>Eiffel &amp; Cie, Barber, Benard and Turenne</td>
<td>Shreve, Lamb &amp; Harmon</td>
</tr>
<tr>
<td>Function</td>
<td>Observation tower, Transmitter station</td>
<td>Tower block</td>
</tr>
<tr>
<td>Opened</td>
<td>March 31, 1889</td>
<td>May 1, 1931</td>
</tr>
<tr>
<td>Engineering firm</td>
<td>Eiffel &amp; Cie</td>
<td>Starrett Brothers and Eken</td>
</tr>
</tbody>
</table>
Applications of Google’s Knowledge Graph

3. Enrich search results with infoboxes and lists
   • Infoboxes might also contain Microdata/RDFa data, e.g. concerts of a band

4. Rank search results using new Hummingbird ranking algorithm.
Equal to the standard data mining process with the difference that data is gathered from the Web.
4.1 Gathering and Exploration

- **Gathering of Web Content**
  - Crawl documents or data
  - Retrieve data via Web API
  - Download pre-gathered data sets

- **Exploration**
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records
4.2 Preprocessing and Transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)

- Methods
  - Attribute transformation / text to term vector / information extraction
  - Aggregation, sampling
  - Dimensionality reduction / feature subset selection

- Good data preparation is key to producing valid and reliable models.

- Data gathering and preparation estimated to take 70-80% of the time and effort of a Web Mining project!
4.3 Data Mining

- **Input:** Preprocessed Data
- **Output:** Model / Patterns

1. Apply data mining method.
2. Evaluate resulting model / patterns.
3. Iterate
   - Experiment with different parameter settings.
   - Experiment with different alternative methods.
   - Improve preprocessing and feature generation.
   - Combine different methods.
Questions to you

- What do you study?
- In which semester are you?

- What experience do you already have with Data Mining?
  - Practical projects?
  - Lectures?
  - Tools?

- What experience do you already have with Web Mining?
  - Practical projects?
  - Lectures?
  - Tools?