Web Mining

Introduction and Course Outline

Prof. Dr. Christian Bizer
Cäcilia Zirn
Robert Meusel
Oliver Lehmberg

FFS 2014
Hallo

- Prof. Dr. Christian Bizer
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Linked Data Technologies
  - Data and Web Mining
  - E-Business und E-Government
- Room: B6, 26 - B1.15
- Consultation: Friday 10:30-11:30
- eMail: chris@informatik.uni-mannheim.de
Hallo

- M.A. Cäcilia Zirn
- Graduate Research Associate
- Research Interests:
  - Natural Language Processing
  - Social Media Monitoring
  - Sentiment Analysis
- Room: B6, 26, C 1.06
- eMail: caecilia@informatik.uni-mannheim.de

- Will teach the block on Web Content Mining
Hallo

- Dipl.-Wi.-Inf. Robert Meusel
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Social Network Analysis
  - Linked Data Technologies
- Room: B6, 26, C 1.04
- eMail: robert@informatik.uni-mannheim.de

Will teach the exercise on Recommender Systems
Hallo

- Dipl.-Wi.-Inf. Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Social Network Analysis
  - Web Data Integration
- Room: B6, 26, C 1.07
- eMail: oli@informatik.uni-mannheim.de

- Will teach the exercise on Social Network Analysis
Introduction and Course Outline

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

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

3. The Web Mining Process

4. Course Outline and Organization
1. The World Wide Web

The Web is a global information space built 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 Protocol (HTTP):** Protocol for interacting with Web resources.
- **Content Formats:** HTML, RDF, RSS, ...
- 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
1.1 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**: 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.worldwidewebssize.com/](http://www.worldwidewebssize.com/)
The link structure follows a power law.

- Small number of pages is target of many links.
- Large number of pages is target of only a few links.

Classic large-scale analysis:

- AltaVista crawl with over 200 million pages and 1.5 billion links
- Log-log scale shows 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%
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.
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 that provides large web crawls on Amazon S3 for free.
      - 2012 CC Corpus: 3.0 billion HTML pages (48 Terabyte compressed)
      - 2013 CC Corpus: 2.0 billion HTML pages (102 Terabyte uncompressed)
   2. ClueWeb 2012 Crawl: 733 million English HTML pages
   3. Internet Archive 2011: 2.7 billion URIs (80 Terabyte)
1.2. The Web of Data

More and more Websites
- semantically markup the content of their HTML pages
- publish structured data in addition to HTML pages

- Microformats
- RDFa
- Linked Data
- Microdata
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
- can assign URIs as global primary keys to entities

```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>

... 
```
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)
- W3C currently tries to reconcile the two alternative proposals

```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 or alternatively RDFa
Usage of Schema.org Data @ Google

Data snippets within search results

Data tables within search results
Microformat, Microdata, RDFa Deployment

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

- **Two extractions runs**
  - 2009/2010 CC Corpus: 2.5 billion HTML pages $\rightarrow$ 5.1 billion RDF triples
  - 2012 CC Corpus: 3.0 billion HTML pages $\rightarrow$ 7.3 billion RDF triples

- **used 100 machines on Amazon EC2**
  - approx. 3000 machine/hours (spot instances of type c1.xlarge) $\rightarrow$ 550 EUR

- **Jointed project of**

![Universität Mannheim](image1.jpg) ![KIT](image2.jpg)
Websites containing Structured Data

2.29 million websites (PLDs) out of 40 million provide Microformat, Microdata or RDFa data (5.65%)

- Grouped by Alexa Website Popularity Rank
  (the site rank based on amount of page views)

<table>
<thead>
<tr>
<th>First $x$ in AL</th>
<th>PLDs in CC #</th>
<th>% AL</th>
<th>% containing structured data overall</th>
<th>RDFa</th>
<th>$\mu D$</th>
<th>$\mu F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99</td>
<td>99.00</td>
<td>74.75</td>
<td>34.34</td>
<td>55.56</td>
<td>68.69</td>
</tr>
<tr>
<td>1k</td>
<td>963</td>
<td>96.30</td>
<td>62.62</td>
<td>40.08</td>
<td>31.67</td>
<td>46.11</td>
</tr>
<tr>
<td>10k</td>
<td>9,294</td>
<td>92.94</td>
<td>47.34</td>
<td>30.47</td>
<td>15.55</td>
<td>29.75</td>
</tr>
<tr>
<td>100k</td>
<td>85,058</td>
<td>85.01</td>
<td>31.94</td>
<td>16.46</td>
<td>7.20</td>
<td>20.07</td>
</tr>
<tr>
<td>1m</td>
<td>734,882</td>
<td>73.49</td>
<td>20.56</td>
<td>7.55</td>
<td>3.04</td>
<td>14.18</td>
</tr>
</tbody>
</table>
**RDFa Topics (2012)**

- **Top Classes:**
- **Topics**
  - CMS and Blog metadata
  - Product data
  - Ratings
  - Navigational metadata
  - Company listings

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total #</th>
<th>PLDs in Alexa #</th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1     og:“article”</td>
<td>183,046</td>
<td>17,002</td>
<td>35.24</td>
<td>30.29</td>
</tr>
<tr>
<td>2     og:“blog”</td>
<td>58,971</td>
<td>5,820</td>
<td>11.35</td>
<td>10.37</td>
</tr>
<tr>
<td>3     og:“website”</td>
<td>56,573</td>
<td>9,533</td>
<td>10.89</td>
<td>16.98</td>
</tr>
<tr>
<td>5     foaf:Image</td>
<td>44,644</td>
<td>2,794</td>
<td>8.60</td>
<td>4.98</td>
</tr>
<tr>
<td>6     sioc:Item</td>
<td>33,141</td>
<td>2,188</td>
<td>6.38</td>
<td>3.90</td>
</tr>
<tr>
<td>7     sioc:UserAccount</td>
<td>19,331</td>
<td>1,327</td>
<td>3.72</td>
<td>2.36</td>
</tr>
<tr>
<td>8     og:“product”</td>
<td>19,107</td>
<td>3,389</td>
<td>3.68</td>
<td>6.04</td>
</tr>
<tr>
<td>9     skos:Concept</td>
<td>13,477</td>
<td>1,135</td>
<td>2.59</td>
<td>2.02</td>
</tr>
<tr>
<td>10    dv:Breadcrumb</td>
<td>9,054</td>
<td>2,123</td>
<td>1.74</td>
<td>3.78</td>
</tr>
<tr>
<td>11    sioc:Post</td>
<td>6,994</td>
<td>691</td>
<td>1.35</td>
<td>1.23</td>
</tr>
<tr>
<td>12    og:“company”</td>
<td>6,758</td>
<td>1,067</td>
<td>1.30</td>
<td>1.90</td>
</tr>
<tr>
<td>13    dv:Review-aggregate</td>
<td>6,236</td>
<td>1,410</td>
<td>1.20</td>
<td>2.51</td>
</tr>
<tr>
<td>14    dv:Rating</td>
<td>4,139</td>
<td>845</td>
<td>0.80</td>
<td>1.51</td>
</tr>
<tr>
<td>15    sioc:BlogPost</td>
<td>3,936</td>
<td>308</td>
<td>0.76</td>
<td>0.55</td>
</tr>
<tr>
<td>16    sioc:Comment</td>
<td>3,339</td>
<td>456</td>
<td>0.64</td>
<td>0.81</td>
</tr>
<tr>
<td>17    og:“activity”</td>
<td>3,303</td>
<td>606</td>
<td>0.64</td>
<td>1.08</td>
</tr>
<tr>
<td>18    vcard:Address</td>
<td>3,167</td>
<td>401</td>
<td>0.61</td>
<td>0.71</td>
</tr>
<tr>
<td>19    gr:BusinessEntity</td>
<td>3,155</td>
<td>392</td>
<td>0.61</td>
<td>0.70</td>
</tr>
<tr>
<td>20    dv:Organization</td>
<td>2,502</td>
<td>367</td>
<td>0.48</td>
<td>0.65</td>
</tr>
</tbody>
</table>

og = Facebook Open Graph Protocol
## Microdata Topics (2012)

### Top Classes:

- CMS and Blog metadata
- Navigational metadata
- Company listings
- Products and offers
- Ratings
- Places
- Events

### Topics

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total #</th>
<th>PLDs in Alexa #</th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema:BlogPosting</td>
<td>25,235</td>
<td>1,502</td>
<td>17.98</td>
<td>6.63</td>
</tr>
<tr>
<td>datavoc:Breadcrumb</td>
<td>21,729</td>
<td>5,244</td>
<td>15.49</td>
<td>23.13</td>
</tr>
<tr>
<td>schema:PostalAddress</td>
<td>19,592</td>
<td>1,404</td>
<td>13.96</td>
<td>6.19</td>
</tr>
<tr>
<td>schema:Product</td>
<td>16,612</td>
<td>3,038</td>
<td>11.84</td>
<td>13.40</td>
</tr>
<tr>
<td>schema:LocalBusiness</td>
<td>16,383</td>
<td>845</td>
<td>11.68</td>
<td>3.73</td>
</tr>
<tr>
<td>schema:Article</td>
<td>15,718</td>
<td>3,025</td>
<td>11.20</td>
<td>13.35</td>
</tr>
<tr>
<td>datavoc:Review-aggregate</td>
<td>8,517</td>
<td>2,376</td>
<td>6.07</td>
<td>10.48</td>
</tr>
<tr>
<td>schema:Offer</td>
<td>8,456</td>
<td>1,474</td>
<td>6.03</td>
<td>6.50</td>
</tr>
<tr>
<td>datavoc:Rating</td>
<td>7,711</td>
<td>1,726</td>
<td>5.50</td>
<td>7.61</td>
</tr>
<tr>
<td>schema:AggregateRating</td>
<td>7,029</td>
<td>1,791</td>
<td>5.01</td>
<td>7.90</td>
</tr>
<tr>
<td>schema:Organization</td>
<td>7,011</td>
<td>1,270</td>
<td>5.00</td>
<td>5.60</td>
</tr>
<tr>
<td>datavoc:Product</td>
<td>6,770</td>
<td>1,156</td>
<td>4.82</td>
<td>5.10</td>
</tr>
<tr>
<td>schema:WebPage</td>
<td>6,678</td>
<td>2,112</td>
<td>4.76</td>
<td>9.32</td>
</tr>
<tr>
<td>datavoc:Organization</td>
<td>5,853</td>
<td>654</td>
<td>4.17</td>
<td>2.89</td>
</tr>
<tr>
<td>datavoc:Address</td>
<td>5,559</td>
<td>654</td>
<td>3.96</td>
<td>2.89</td>
</tr>
<tr>
<td>schema:Person</td>
<td>5,237</td>
<td>890</td>
<td>3.73</td>
<td>3.93</td>
</tr>
<tr>
<td>schema:GeoCoordinates</td>
<td>4,677</td>
<td>312</td>
<td>3.33</td>
<td>1.38</td>
</tr>
<tr>
<td>schema:Place</td>
<td>4,131</td>
<td>488</td>
<td>2.94</td>
<td>2.15</td>
</tr>
<tr>
<td>schema:Event</td>
<td>4,102</td>
<td>659</td>
<td>2.92</td>
<td>2.91</td>
</tr>
<tr>
<td>datavoc:Person</td>
<td>2,877</td>
<td>523</td>
<td>2.05</td>
<td>2.31</td>
</tr>
<tr>
<td>datavoc:Review</td>
<td>2,816</td>
<td>783</td>
<td>2.01</td>
<td>3.45</td>
</tr>
</tbody>
</table>

datavoc = Google’s Rich Snippet Vocabulary
schema = Schema.org
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
URI can be looked up on the Web

By following RDF links applications can

- navigate the global data graph
- discover new data sources
### The Disco – Hyperdata Browser

**Richard Cyganiak**

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>event</code></td>
<td>...</td>
<td>G2</td>
</tr>
<tr>
<td><code>type</code></td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>G1 G2 G3 G4</td>
</tr>
<tr>
<td><code>seeAlso</code></td>
<td><a href="http://richard.cyganiak.de/cygri.rdf">http://richard.cyganiak.de/cygri.rdf</a></td>
<td>G2</td>
</tr>
<tr>
<td><code>seeAlso</code></td>
<td><a href="http://richard.cyganiak.de/foaf.rdf">http://richard.cyganiak.de/foaf.rdf</a></td>
<td>G3</td>
</tr>
<tr>
<td><code>nearest_airport</code></td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td><code>phone</code></td>
<td>tel:+49-175-5630408</td>
<td>G1</td>
</tr>
<tr>
<td><code>sameAs</code></td>
<td>Richard Cyganiak</td>
<td>G1</td>
</tr>
<tr>
<td><code>based_near</code></td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td><code>based_near</code></td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td><code>based_near</code></td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G1</td>
</tr>
<tr>
<td><code>currentProject</code></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><code>currentProject</code></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><code>depiction</code></td>
<td></td>
<td>G4</td>
</tr>
<tr>
<td>Property</td>
<td>Value</td>
<td>Sources</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>population</td>
<td>3398888</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://dbpedia.org/City">http://dbpedia.org/City</a></td>
<td>G2</td>
</tr>
<tr>
<td>comment</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>comment</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>label</td>
<td>Berlin</td>
<td>G2</td>
</tr>
<tr>
<td>sameAs</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Berlin">http://dbpedia.org/resource/category/Berlin</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</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>subject</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>subject</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>subject</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>subject</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>sourceURL</td>
<td>Berlin</td>
<td>G2</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="http://en.wikipedia.org/wiki/Berlin" alt="Image of Berlin" /></td>
<td>G2</td>
</tr>
<tr>
<td>birthplace</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Linked Data Search Engine: SigMa

Chris Bizer

picture:

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]
Grassroots community effort to
- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources
- maintain a data set catalog on the CKAN DataHub
LOD Datasets on the Web (May 2007)

- Over 500 million RDF triples
- Around 120,000 RDF links between data sources
LOD Datasets on the Web (September 2011)

- 31,6 billion RDF triples
- 503 million RDF links
### Distribution by Topical Domain (September 2011)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Data Sets</th>
<th>Triples</th>
<th>Percent</th>
<th>RDF Links</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>25</td>
<td>1,841,852,061</td>
<td>5.82 %</td>
<td>50,440,705</td>
<td>10.01 %</td>
</tr>
<tr>
<td>Geographic</td>
<td>31</td>
<td>6,145,532,484</td>
<td>19.43 %</td>
<td>35,812,328</td>
<td>7.11 %</td>
</tr>
<tr>
<td>Government</td>
<td>49</td>
<td>13,315,009,400</td>
<td>42.09 %</td>
<td>19,343,519</td>
<td>3.84 %</td>
</tr>
<tr>
<td>Library</td>
<td>87</td>
<td>2,950,720,693</td>
<td>9.33 %</td>
<td>139,925,218</td>
<td>27.76 %</td>
</tr>
<tr>
<td>Cross-domain</td>
<td>41</td>
<td>4,184,635,715</td>
<td>13.23 %</td>
<td>63,183,065</td>
<td>12.54 %</td>
</tr>
<tr>
<td>Life sciences</td>
<td>41</td>
<td>3,036,336,004</td>
<td>9.60 %</td>
<td>191,844,090</td>
<td>38.06 %</td>
</tr>
<tr>
<td>User content</td>
<td>20</td>
<td>134,127,413</td>
<td>0.42 %</td>
<td>3,449,143</td>
<td>0.68 %</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td><strong>295</strong></td>
<td><strong>31,634,213,770</strong></td>
<td><strong>503,998,829</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### More statistics
- [http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/](http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/)

### Newer statistics
- LODstats (University of Leipzig, 2014): 928 data sets
- LDspider Crawl (University of Mannheim, 2013): 850 data sets
Hands-on: How to access the Data for Mining?

- **Download the Billion Triples Challenge Dataset**
  - 1.4 billion triples (17 GB gzipped)
  - crawled from the public Web of Linked Data in May/June 2012

- **Download the Web Data Commons Dump**
  - 7 billion triples (101 GB gzipped)
  - RDFa, Microdata, Microformat data crawled February-June 2012

- **Download the Sindice Dump**
  - 12 billion triples (164GB gzipped, ~1,16TB uncompressed)
  - Linked Data, RDFa, Microdata, Microformat crawled 2009-2011
<table>
<thead>
<tr>
<th>LOD Cloud</th>
<th>Microdata, Microformats, RDFa</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1000 sources</td>
<td>millions of sources</td>
</tr>
<tr>
<td>covers wider range of specific topics</td>
<td>focused on search engines and Facebook</td>
</tr>
<tr>
<td>contains more complex data structures</td>
<td>very simple and shallow data structures</td>
</tr>
<tr>
<td>partial ontology agreement</td>
<td>strong ontology agreement</td>
</tr>
<tr>
<td>data integration eased by RDF links</td>
<td>data integration requires NLP techniques</td>
</tr>
</tbody>
</table>
1.3. Web 2.0 Applications

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

- These applications form separate data spaces that are only partly accessible via the Web.
  - HTML interfaces
  - 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, ...

Sources
- https://s3.amazonaws.com/OneBillionFB/Facebook+1+Billion+Stats.docx
Web APIs

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

- Web API Catalog
  - lists over 10,900 Web APIs
  - 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
2. 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

- huge amount of available data \(\rightarrow\) requires sampling or multiple machines
- un-/semi-structured nature of data \(\rightarrow\) often mixture of data and text mining
- heterogeneity of data \(\rightarrow\) data integration might be a challenge
- distributed nature of data \(\rightarrow\) often requires large-scale crawling
Web Mining is a Multi-Disciplinary Field

- Draws ideas and techniques from

- Sub-Fields
  1. Web Content Mining
  2. Web Structure Mining
  3. Web Usage Mining
2.1 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.

Typical Sources of Data

- automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
- e-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-throughs, purchases)
- user profiles and/or user ratings
- page attributes, page content, site structure
- additional domain knowledge and demographic data
The Web Usage Mining Process

Data Preparation Phase

Site Content & Structure

Domain Knowledge

Web & Application Server Logs

Data Cleaning
Pageview Identification
Sessionization
Data Integration
Data Transformation

User Transaction Database

Pattern Discovery Phase

Aggregate User models

Pattern Analysis

Pattern Filtering Aggregation Characterization

Patterns

Usage Mining

Transaction Clustering
Pageview Clustering
Correlation Analysis
Association Rule Mining
Sequential Pattern Mining
Simple Statistical Analysis

- 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: Card fraud detection
Discovery of associated events or application objects: Recommendation of products and content
Recommendation of products and content: Card fraud detection
Discovery of visitor groups with common properties and interests: Characterization of visitors into predefined classes
Discovery of visitor groups with common behaviour: Card fraud detection
Characterization of visitors into predefined classes: Card fraud detection
Example Application: Website Improvement

- Discovery of common navigation paths
- Discovery of pages on which users leave the site or discontinue shopping process

Technique: Sequential Pattern Mining
Website Personalization and Product Recommendation

Web Personalization: “personalizing the browsing experience of a user by dynamically tailoring the look, feel, and content of a Web site to the user’s needs and interests.”

Why Personalize?

- broaden and deepen customer relationships
- provide continuous relationship marketing to build customer loyalty
- help automate the process of proactively market products to customers
  - lights-out marketing
  - cross-sell/up-sell products
- provide the ability to measure customer behavior and track how well customers are responding to marketing efforts
Product Recommendations

Amazon Recommended for You

These recommendations are based on items you own and more.

   - by Kevin P. Murphy (August 24, 2012)
   - Average Customer Review: ★★★★★ (9)
   - In Stock
   - Price: $90.00
   - 11 used & new from $89.97

2. **Data Mining with R: Learning with Case Studies (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)**
   - by Luis Torgo (November 9, 2010)
   - Average Customer Review: ★★★★★ (6)
   - Usually ships in 1 to 3 months
   - List Price: $83.95
   - Price: $67.02
   - 53 used & new from $61.98
Personalized News

WH gun plan: Out-organize the NRA
President Barack Obama is trying an end run around the NRA - rallying groups as varied as churches, medical organizations, retailers and the Rotary Club to build support for new gun regulations.

Opinion: The NRA’s game plan Chicago Tribune - by charles madigan
In Depth: Obama gun plan may feature background checks on all buyers
Los Angeles Times

See realtime coverage »

Boeing’s 787 Dreamliner suffers more mishaps
LONDON - Another two incidents struck Boeing’s 787 Dreamliner plane on Friday when an All Nippon Airways aircraft suffered a crack to its windscreen during a flight in Japan and an oil leak was found coming from the engine of a separate plane after it ...

In Depth: More Problems for Boeing’s 787 Surface in Japan New York Times
Wikipedia: Boeing 787 Dreamliner
See realtime coverage »

California school shooter targeted bullies, sheriff says
TAFT, Calif. - The 16-year-old boy had allegedly wounded the teenager he claimed had bullied him, fired two more rounds at students fleeing their first-period science class, then faced teacher Ryan Heber.

Highly Cited: Youth fires shotgun at 2 high school students, hits one; suspect in custody CNN International
In Depth: Sheriff: High school gunman felt he’d been bullied CBS News

Weather for Mannheim, Germany

37° 30’

The Weather Channel - Weather Underground - AccuWeather

Baden-Württemberg, Germany » - Change location

Stuttgart: Slight increase in lease prices
Property Magazine International - Jan 10, 2013

Eishockey: Adler Mannheim bauen Tabellenführung aus
ZEIT ONLINE - Jan 6, 2013

One dead as helicopter crashes on autobahn
Personalized Search

Google search for chikoo:
- Suggested search results include Manilkara zapota - Wikipedia, the free encyclopedia.
- Images for chikoo - Report images.
Standard Recommendation Approaches

Rule-based filtering
- provide content to users based on predefined rules (e.g., “if user has clicked on A and the user’s zip code is 90210, then add a link to C”)

Collaborative filtering
- give recommendations to a user based on responses/ratings of other “similar” users

Content-based filtering
- track which pages the user visits and recommend other pages with similar content

Hybrid Methods
- usually a combination of content-based and collaborative
Search Log Mining

Analysis of search queries entered by a user.

Examples:

1. Query completion using association analysis
2. Query topic detection using classification
Summary

- Web usage mining has emerged as the essential tool for realizing more personalized, user-friendly and business-optimal Web sites.
- The key is to use the user-clickstream data for many mining purposes.
- Traditionally, Web usage mining is used by e-commerce sites to improve their sites and to increase profits.
- It is now also used by search engines to improve search quality and to evaluate search results, etc., and by many other applications.
2.2 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
- Web crawls including HTML pages and hyperlinks
- crawls of the blogosphere
- social networks including explicit relations between actors (your Facebook friend network)
- other types of community data (discussion forums, email conversations, ...)
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 levels of “prestige”
  - a page is prestigious if many other prestigious pages link to it
  - initial algorithm used by Google
  - Today, one factor amongst 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, network compression, visualization of huge networks
2.3 Web Content Mining

- **Definition**

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

- **Content Mining Tasks**
  - Content Clustering
  - Content Classification
  - Information Extraction
  - Sentiment Analysis
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 (training set) find a model for the class as a function of the values of the features.

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

Applications
- Classification of news into categories
- SPAM detection
- Filtering ‘boring’ documents based on personal profiles

Classification methods commonly used for text
- Naive Bayes
- Support Vector Machines
Mixture of Document Clustering and Classification
Information Extraction

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?
  - **Relationship Extraction**
    - PERSON works for ORGANIZATION
    - PERSON located in LOCATION
The degree of document structure determines the difficulty of the information extraction task.
Example: Named Entity Recognition

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

![Zemanta social blogging interface with an example of named entity recognition. Text about Batman and his introduction, cultural icon, and adaptations into various media formats are shown. The interface highlights links and tags for Batman, his creators, and related articles.]
Example: Relationship Extraction

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

- **Paper about information extraction from all HTML tables on the Web**
Example: Information Extraction from Wikipedia
Information Extraction Projects

- Wikipedia
- DBpedia
- YAGO
- Freebase
- Wikidata
- Universität Mannheim
- Max Planck Institute for Informatics
- Universität Leipzig
- Google
- Wikimedia Foundation
DBpedia – Querying Wikipedia like a Database

- Joined project to
  - create a huge, multi-lingual knowledge base
  - by extracting structured information from Wikipedia
  - make the knowledge base available on the Web as Linked Data under an open license

- Structured Information in Wikipedia
  - infoboxes
  - geo-coordinates
  - categorization of articles
  - inter-language links
  - links to images and external webpages
  - titles and abstracts
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”
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 of search results using new Hummingbird ranking algorithm
Which factors correlate with unemployment in France?

Input Table

<table>
<thead>
<tr>
<th>Row No.</th>
<th>region</th>
<th>unempl...</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Bretagne</td>
<td>0.097</td>
</tr>
<tr>
<td>19</td>
<td>Pays de la L</td>
<td>0.098</td>
</tr>
<tr>
<td>16</td>
<td>Limousin</td>
<td>0.100</td>
</tr>
<tr>
<td>15</td>
<td>Rhône-Alpes</td>
<td>0.102</td>
</tr>
<tr>
<td>14</td>
<td>Auvergne</td>
<td>0.102</td>
</tr>
<tr>
<td>11</td>
<td>Centre</td>
<td>0.107</td>
</tr>
<tr>
<td>12</td>
<td>Basse-Norm</td>
<td>0.108</td>
</tr>
<tr>
<td>13</td>
<td>Bourgogne</td>
<td>0.108</td>
</tr>
<tr>
<td>21</td>
<td>Midi-Pyrénées</td>
<td>0.110</td>
</tr>
<tr>
<td>3</td>
<td>Franche-Cor</td>
<td>0.110</td>
</tr>
<tr>
<td>22</td>
<td>île-de-France</td>
<td>0.110</td>
</tr>
<tr>
<td>1</td>
<td>Alsace</td>
<td>0.112</td>
</tr>
<tr>
<td>23</td>
<td>Aquitaine</td>
<td>0.112</td>
</tr>
<tr>
<td>17</td>
<td>Poitou-Chant</td>
<td>0.113</td>
</tr>
<tr>
<td>25</td>
<td>Corse</td>
<td>0.114</td>
</tr>
<tr>
<td>2</td>
<td>Lorraine</td>
<td>0.122</td>
</tr>
<tr>
<td>10</td>
<td>Haute-Norm</td>
<td>0.124</td>
</tr>
<tr>
<td>6</td>
<td>Champagne</td>
<td>0.125</td>
</tr>
<tr>
<td>24</td>
<td>Provence-Alp</td>
<td>0.131</td>
</tr>
</tbody>
</table>
### Unemployment Table with additional Attributes

#### Input Table  |  Link  |  Additional Attributes

<table>
<thead>
<tr>
<th>Row No.</th>
<th>region</th>
<th>unempl...</th>
<th>region_uri</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
<th>name_uri__</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Limousin</td>
<td>0.100</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>15703.300</td>
<td>124.380</td>
<td>0.780</td>
<td>0.280</td>
<td>16510.400</td>
<td>3282</td>
<td>1248.700</td>
<td>1017.700</td>
<td>1528.700</td>
</tr>
<tr>
<td>15</td>
<td>Rhône-Alpes</td>
<td>0.102</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>158329</td>
<td>3896.363</td>
<td>1.880</td>
<td>0.390</td>
<td>17003</td>
<td>56881</td>
<td>28175.900</td>
<td>807.700</td>
<td>9428.700</td>
</tr>
<tr>
<td>14</td>
<td>Auvergne</td>
<td>0.102</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>29729.200</td>
<td>689.348</td>
<td>1.580</td>
<td>0.260</td>
<td>16396.800</td>
<td>6450</td>
<td>1009.400</td>
<td>951.900</td>
<td>2374.700</td>
</tr>
<tr>
<td>11</td>
<td>Centre</td>
<td>0.107</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>59502</td>
<td>868.893</td>
<td>1.140</td>
<td>0.340</td>
<td>16880.700</td>
<td>12928</td>
<td>13745.600</td>
<td>761.100</td>
<td>3838.700</td>
</tr>
<tr>
<td>12</td>
<td>Basse-Norme</td>
<td>0.108</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>32224.600</td>
<td>297.951</td>
<td>0.910</td>
<td>0.280</td>
<td>15678.300</td>
<td>7117</td>
<td>3085.400</td>
<td>750.100</td>
<td>2148.700</td>
</tr>
<tr>
<td>13</td>
<td>Bourgogne</td>
<td>0.108</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>37778.200</td>
<td>358.552</td>
<td>0.890</td>
<td>0.150</td>
<td>16654.500</td>
<td>8742</td>
<td>338.700</td>
<td>848.900</td>
<td>3206.700</td>
</tr>
<tr>
<td>21</td>
<td>Midi-Pyrénées</td>
<td>0.110</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>64525</td>
<td>228.610</td>
<td>2.180</td>
<td>1.160</td>
<td>15728.200</td>
<td>12598</td>
<td>7964</td>
<td>784.500</td>
<td>4820.700</td>
</tr>
<tr>
<td>25</td>
<td>Corse</td>
<td>0.114</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>26343.400</td>
<td>529.876</td>
<td>1.390</td>
<td>0.390</td>
<td>16445.900</td>
<td>6989</td>
<td>689.800</td>
<td>758.100</td>
<td>1649.700</td>
</tr>
<tr>
<td>22</td>
<td>Île-de-France</td>
<td>0.110</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>469046.700</td>
<td>14364.372</td>
<td>3.140</td>
<td>0.620</td>
<td>20912</td>
<td>51984</td>
<td>5914</td>
<td>650.200</td>
<td>24309.700</td>
</tr>
<tr>
<td>1</td>
<td>Alsace</td>
<td>0.112</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>45914</td>
<td>692.110</td>
<td>1.350</td>
<td>0.700</td>
<td>17112.900</td>
<td>11960</td>
<td>3563.700</td>
<td>883.100</td>
<td>2597.700</td>
</tr>
<tr>
<td>23</td>
<td>Aquitaine</td>
<td>0.112</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>74099.700</td>
<td>1146.621</td>
<td>1.070</td>
<td>0.930</td>
<td>16118.900</td>
<td>15200</td>
<td>5894</td>
<td>802.800</td>
<td>6088.700</td>
</tr>
<tr>
<td>17</td>
<td>Poitou-Charentes</td>
<td>0.113</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>37880.700</td>
<td>305.333</td>
<td>0.760</td>
<td>0.620</td>
<td>15766.100</td>
<td>7850</td>
<td>138.500</td>
<td>698.700</td>
<td>2490.700</td>
</tr>
<tr>
<td>25</td>
<td>Corse</td>
<td>0.114</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>5686.200</td>
<td>132.209</td>
<td>0.900</td>
<td>0.370</td>
<td>14775.800</td>
<td>8584</td>
<td>448</td>
<td>921.100</td>
<td>1115.700</td>
</tr>
<tr>
<td>2</td>
<td>Lorraine</td>
<td>0.122</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>51233.500</td>
<td>546.504</td>
<td>0.920</td>
<td>0.160</td>
<td>16035.400</td>
<td>17315</td>
<td>9565.700</td>
<td>824.400</td>
<td>3640.700</td>
</tr>
<tr>
<td>10</td>
<td>Haute-Normandie</td>
<td>0.124</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>43191.300</td>
<td>600.947</td>
<td>0.950</td>
<td>0.240</td>
<td>16464.700</td>
<td>14035</td>
<td>11481.400</td>
<td>650.400</td>
<td>2418.700</td>
</tr>
<tr>
<td>6</td>
<td>Champagne</td>
<td>0.125</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>33075</td>
<td>237.687</td>
<td>0.560</td>
<td>-0.100</td>
<td>15911.500</td>
<td>8134</td>
<td>5393.100</td>
<td>740.200</td>
<td>2065.700</td>
</tr>
<tr>
<td>24</td>
<td>Provence-Alpes</td>
<td>0.131</td>
<td><a href="http://wifo-5-0">http://wifo-5-0</a></td>
<td>117460.300</td>
<td>2099.396</td>
<td>1.650</td>
<td>0.870</td>
<td>16582.400</td>
<td>27335</td>
<td>5690</td>
<td>813.200</td>
<td>14913.700</td>
</tr>
</tbody>
</table>
RapidMiner Linked Open Data Extension

Allows you to

1. link local table to DBpedia and other LOD data sources
2. extend local table with additional attributes
3. mine extended tables using all Rapidminer features

http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/
Finding Correlations

- Use additional attributes to find interesting correlations

Example correlation for unemployment in France:

- African islands, Islands in the Indian Ocean, Outermost regions of the EU (positive)
- Population growth (positive)
- Disposable income (negative)
- Energy consumption (negative)
- Fast food restaurants (positive)
- Hospital beds/inhabitants (negative)
- Police stations (positive)
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
3. The Web Mining Process

Equal to the standard data mining process with the difference that data is gathered from the Web.
3.1 Gathering and Exploration

- **Gathering of Web Data**
  - 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
3.2 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
- Aggregation, sampling
- Dimensionality reduction / feature subset selection
- Attribute transformation / text to term vector
- Discretization and binarization

- Good data preparation is key to producing valid and reliable models.
- Data preparation estimated to take 70-80% of the time and effort of a data mining project!
3.3 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.
4. Course Outline and Organization
Course Organization

- **Lecture**
  - introduces the principle methods of Web mining
  - discusses how to evaluate the generated models
  - presents examples of Web mining applications

- **Exercise**
  - students experiment with given data sets

- **Project Work**
  - teams of two 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
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic Thursday</th>
<th>Topic Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.02.2014</td>
<td>Lecture: Introduction to Web Mining</td>
<td>Lecture: Web Usage Mining</td>
</tr>
<tr>
<td>20.02.2014</td>
<td>Lecture: Recommender Systems</td>
<td>Exercise: Recommender Systems</td>
</tr>
<tr>
<td>27.02.2014</td>
<td>Exercise: Recommender Systems</td>
<td>Lecture: Web Structure Mining</td>
</tr>
<tr>
<td>06.03.2014</td>
<td>Lecture: Social Network Analysis</td>
<td>Lecture: Introduction to Pajek</td>
</tr>
<tr>
<td>13.03.2014</td>
<td>Exercise: Social Network Analysis</td>
<td>Exercise: Social Network Analysis</td>
</tr>
<tr>
<td>20.03.2014</td>
<td>Lecture: Web Content Mining: Sentiment Analysis</td>
<td>Exercise: Sentiment Analysis</td>
</tr>
<tr>
<td></td>
<td>Easter break</td>
<td></td>
</tr>
<tr>
<td>01.05.2014</td>
<td>Holiday</td>
<td>Coaching</td>
</tr>
<tr>
<td>08.05.2014</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>15.05.2014</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>22.05.2014</td>
<td>Project work</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>30.05.2014</td>
<td>Holiday</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Course Organization

**Course Webpage**
- Contains general information and lecture slides

**Exercise Material**
- ILIAS eLearning System, [https://ilias.uni-mannheim.de/](https://ilias.uni-mannheim.de/)

**Time and Location**
- **Thursday, 15:30 to 17:00**, Room: B 6, A303
- **Friday, 12:00 to 13:30**, Room: B 6, A303
Literature


Software

- Powerful open-source data mining suite
- Download: http://rapid-i.com/
Software

- open-source machine learning library which can be deployed on a single machine as well as on hadoop clusters
- includes various recommendation algorithms
- Download: http://mahout.apache.org/
Software

- Tool for the analysis and visualization of large networks
- Download: http://pajek.imfm.si/doku.php?id=pajek
Software

- Stanford CoreNLP: A Suite of Core NLP Tools
Questions?