Web Mining

Introduction and Course Outline

Simone Paolo Ponzetto
Anne Lauscher
Chia-Chien Hung

FSS 2020
Hi!

- Since 2016 full professor of Information Systems (Wifo) in Mannheim

- Main research areas
  - Knowledge Acquisition
  - Natural Language Processing
  - Computational Social Science

- TAs: Anne Lauscher, Chia-Chien Hung
- {anne, chia-chien, simone}@informatik.uni-mannheim.de
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
1. Course Organization

- **Lecture (IE 671)**
  - covers different types of Web Mining methods
  - presents examples of Web Mining applications
  - discusses how to evaluate learned models

- **Labs**
  - students get their hand dirty with tools and code

- **Evaluation**
  - 60 min final exam

- **Project (IE 684)**
  - 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 a summary about the project, present the project results

- **Evaluation**
  - Implementation + report + presentation
Course Organization

- **Materials**
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/

- **Time and Location**
  - **Lecture:**
    Tuesday, 10:15 to 11:45,
    Room: B 6, A104
  - **Labs:**
    Tuesday, 15:30 to 17:00,
    Room: B 6, A104
<table>
<thead>
<tr>
<th>Morning session (10:15-11:45)</th>
<th>Afternoon session (15:30-15:00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/2/20 Lecture: Introduction to Web Mining</td>
<td>11/2/20 -</td>
</tr>
<tr>
<td>18/2/20 Lecture: Recommender Systems</td>
<td>18/2/20 Lab: Recommender Systems</td>
</tr>
<tr>
<td>25/2/20 Lecture: Recommender Systems</td>
<td>25/2/20 Lab: Recommender Systems</td>
</tr>
<tr>
<td>10/3/20 Lecture: Web Content Mining</td>
<td>10/3/20 Lab: Web Content Mining</td>
</tr>
<tr>
<td>17/3/20 Lecture: Web Content Mining</td>
<td>17/3/20 Lab: Web Content Mining</td>
</tr>
<tr>
<td>24/3/20 Lecture: Web Content Mining</td>
<td>24/3/20 Lab: Web Content Mining</td>
</tr>
<tr>
<td>31/3/20 Introduction to Student Projects</td>
<td>-</td>
</tr>
<tr>
<td>7/4/20</td>
<td>Osterferien</td>
</tr>
<tr>
<td>14/4/20</td>
<td>Osterferien</td>
</tr>
<tr>
<td>21/4/20 Feedback about Project Outlines</td>
<td>21/4/20 -</td>
</tr>
<tr>
<td>28/4/20 -</td>
<td>28/4/20 -</td>
</tr>
<tr>
<td>5/5/20 Coaching session (optional)</td>
<td>5/5/20 -</td>
</tr>
<tr>
<td>12/5/20 Coaching session (optional)</td>
<td>12/5/20 -</td>
</tr>
<tr>
<td>19/5/20 Coaching session (optional)</td>
<td>19/5/20 -</td>
</tr>
<tr>
<td>26/5/20 Presentation of the projects</td>
<td>26/5/20 Presentation of the projects</td>
</tr>
</tbody>
</table>
Lecture Videos

Video recordings of past lectures (outdated but useful as extra complementary materials)

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

- You will not succeed in this module if you do not know how to program (we assume some fluency in Python)...

- spaCy
- SNAP
- surprise
- scikit-learn

[Links to websites and resources for software and libraries]
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
2.1 The Classic Document Web

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

- Size of the Indexed Web: approx. 60 billion pages
- [http://www.worldwidewebssize.com/](http://www.worldwidewebssize.com/)

  - Estimated on the basis of a method that combines word frequencies from a corpus and search counts returned by the engines
  - 50 words are sent to all four search engines
  - The number of webpages found for these words are recorded
  - Multiple extrapolated estimations are made of the size of the engine's index, which are subsequently averaged
  - Example: The word „the“ appears in 67% of all English pages and has 25.2 billion hits on Google
Link Structure of the Web: In-Degree

- 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%

**Probability 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.
From a Web of Documents to a Web of Data

Web of documents

Key characteristics:

1. **Names** (URIs)
2. **Documents** (Resources) described by HTML, XML, etc.
3. **Interactions** via HTTP
4. **(Hyper)Links** between documents or anchors in these documents
Web of Documents vs. Web of Data

"Documents"

Hyperlinks

"Things"

Typed Links
Web of data

Key characteristics:

- Links between arbitrary things (e.g., persons, locations, events, buildings)
- Structure of data on Web pages is made explicit
- Things described on Web pages are named and get URIs
- Links between things are made explicit and are typed
Available approaches

1. semantically markup the content of their HTML pages
2. 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

- Key idea
  - use existing HTML attributes to embed structured data types

- Shortcoming
  - Fixed formats means that it can not represent any kind of data
Microformats: example hCard

- **hCard** is a simple, open format for publishing people, companies, organizations on the web, using a 1:1 representation of vCard

```html
<div class="vcard">
  <a class="fn org url" href="http://www.commerce.net/">CommerceNet</a>
  <div class="adr">
    <span class="type">Work</span>: 169 University Avenue
    <span class="locality">Palo Alto</span>,  
    <abbr class="region" title="California">CA</abbr>&nbsp;&nbsp;
    <span class="postal-code">94301</span>
    <span class="country-name">USA</span>
  </div>
  <div class="tel">
    <span class="type">Work</span> +1-650-289-4040
  </div>
  <div class="tel">
    <span class="type">Fax</span> +1-650-289-4041
  </div>
  <div>Email:
    <span class="email">info@commerce.net</span>
  </div>
</div>
```
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>
```
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>
```
Problem: which vocabulary to use?

**Schema.org** provides a collection of shared vocabularies

Launched in June 2011 by Bing, Google and Yahoo

Create a common set of schemas for webmasters to mark-up with structured data their websites.

- 200+ Types: Event, Organization, Person, Place, Product, Review
- Encoding: Microdata, RDFa, JSON-LD
## Schema.org: example schema

**Person**

A person (alive, dead, undead, or fictional).

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties from Person</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>additionalName</td>
<td>Text</td>
<td>An additional name for a Person, can be used for a middle name.</td>
</tr>
<tr>
<td>address</td>
<td>PostalAddress, Text</td>
<td>Physical address of the item.</td>
</tr>
<tr>
<td>affiliation</td>
<td>Organization</td>
<td>An organization that this person is affiliated with. For example, a school/university, a club, or a team.</td>
</tr>
<tr>
<td>alumniOf</td>
<td>EducationalOrganization, Organization</td>
<td>An organization that the person is an alumni of. Inverse property: alumni.</td>
</tr>
<tr>
<td>award</td>
<td>Text</td>
<td>An award won by or for this item. Supersedes awards.</td>
</tr>
<tr>
<td>birthDate</td>
<td>Date</td>
<td>Date of birth.</td>
</tr>
<tr>
<td>birthPlace</td>
<td>Place</td>
<td>The place where the person was born.</td>
</tr>
</tbody>
</table>
Schema.org: examples

■ RDFa

```html
<div vocab="http://schema.org/" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <div property="director" typeof="Person">
    Director: <span property="name">James Cameron</span>
    (born <time property="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span property="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" property="trailer">Trailer</a>
</div>
```

■ Microdata

```html
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
    Director: <span itemprop="name">James Cameron</span>
    (born <time itemprop="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```
Usage of Schema.org Data @ Google

Data snippets within search results

Data snippets within info boxes

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>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun 20 Fri</td>
<td>The Black Keys</td>
</tr>
<tr>
<td></td>
<td>Neuhausen ob Eck (near you)</td>
</tr>
<tr>
<td>May 16 Fri</td>
<td>The Black Keys</td>
</tr>
<tr>
<td></td>
<td>Gulf Shores, AL</td>
</tr>
<tr>
<td>Jun 22 Sun</td>
<td>The Black Keys</td>
</tr>
<tr>
<td></td>
<td>Schelfel</td>
</tr>
</tbody>
</table>
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.

Diagram:
- RDF
- RDF link
- RDF links
- A
- B
- C
- D
- E
The Web Today as a Multitude of Data Silos
From Data Silos vs. the Web of Data

1. **Use URIs as names** for things.
2. **Use HTTP URIs** so that people can look up those names.
3. When someone looks up a URI, **provide useful RDF information**.
4. Include RDF **statements that link to other URIs** so that they can discover related things.
Entities are identified with HTTP URIs

HTTP URIs take the role of global primary keys.

\[
\begin{align*}
\text{pd:cygri} &= \text{http://richard.cyganiak.de/foaf.rdf#cygri} \\
\text{dbpedia:Berlin} &= \text{http://dbpedia.org/resource/Berlin}
\end{align*}
\]
URIs can be looked up on the Web

By following RDF links applications can

- navigate the global data graph
URIs can be linked to navigate the data space

- By following RDF links applications can:
  - navigate the global data graph
  - discover new data sources
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
HTML Interfaces

- Allows browser-based access to profile, communication, etc.
Web APIs

- Access the data programmatically

```python
import pandas as pd

# Search tweets
dict_ = {'user': [], 'date': [], 'text': [], 'favorite_count': []}
for status in python_tweets.search(**query)['statuses']:
    dict_['user'].append(status['user']['screen_name'])
    dict_['date'].append(status['created_at'])
    dict_['text'].append(status['text'])
    dict_['favorite_count'].append(status['favorite_count'])

# Structure data in a pandas DataFrame for easier manipulation
df = pd.DataFrame(dict_)
df.sort_values(by='favorite_count', inplace=True, ascending=False)
df.head(5)
```
Web APIs slice the Web into Data Silos

3. What is Web Mining?

- **Definition**
  - Acquiring useful information from
    - Web *content*
    - Web *structure*
    - Web *usage data*

- **Web data pose unique challenges:**
  1. Large volume
  2. Semi-structured
  3. Heterogeneous
  4. Distributed
Web Mining is a Multi-Disciplinary Field

- Draws ideas and techniques from

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

- **Definition**

  Discovery of patterns in 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.
The Web Usage Mining Process
Example: product recommendation
Example: product recommendation
Example: personalized search
Example: personalized search
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.
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
Link Prediction

Question: Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? (Liben-Nowell & Kleinberg, 2007)

Applications

- Facebook: recommending possible friends
- Tinder: recommending potential matches

3.3 Web Content Mining

- Definition

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

- Content Mining Tasks
  - Content classification and clustering on Web content
  - Applications of NLP techniques to social network content
Content Classification

- **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**
  - News categorization
  - Topic classification
  - Spam detection
  - Product categorization

- **Classification methods commonly used for**
  - Naive Bayes, Support Vector Machines, Deep Neural Nets
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

Applications

- Topic discovery
- Search result clustering

Techniques

- Algorithms: K-Means, Expectation-Maximization (EM)
- Similarity measures: Cosine, Jaccard
Mixture of Document Clustering and Classification
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
  - Stars

- **Applications**
  - Document-level: poll prediction from tweets
  - Feature/Aspect-level: analysis of product reviews
Applications of NLP to social media data

---

**Trends for you**

- **Politics** - Trending
  - #Merkel
    - 13.5K Tweets
  - Frankfurter Allgemeine is Tweeting about this

- **Politics** - Trending
  - #AKKRuecktritt
    - 9,327 Tweets
  - tagesschau, Frankfurter Allgemeine, and 1 more are Tweeting about this

- **Celebrities** - Trending
  - Parasite
    - 2.12M Tweets

- **Politics** - Trending
  - #BlackRock
    - 15.7K Tweets

- Trending in Germany
  - #Merz
    - 6,429 Tweets
    - Der Postillon is Tweeting about this

---

**Twitter**

<table>
<thead>
<tr>
<th></th>
<th>% posts</th>
<th>Whisper</th>
</tr>
</thead>
<tbody>
<tr>
<td>I hate</td>
<td>70.5</td>
<td>I hate</td>
</tr>
<tr>
<td>I can't stand</td>
<td>7.7</td>
<td>I don't like</td>
</tr>
<tr>
<td>I don't like</td>
<td>7.2</td>
<td>I can't stand</td>
</tr>
<tr>
<td>I really hate</td>
<td>4.9</td>
<td>I really hate</td>
</tr>
<tr>
<td>I fucking hate</td>
<td>1.8</td>
<td>I fucking hate</td>
</tr>
<tr>
<td>I'm sick of</td>
<td>0.8</td>
<td>I'm sick of</td>
</tr>
<tr>
<td>I can't stand</td>
<td>0.7</td>
<td>I'm so sick of</td>
</tr>
<tr>
<td>I fuckin hate</td>
<td>0.6</td>
<td>I just hate</td>
</tr>
<tr>
<td>I just hate</td>
<td>0.6</td>
<td>I really don't like</td>
</tr>
<tr>
<td>I'm so sick of</td>
<td>0.6</td>
<td>I secretly hate</td>
</tr>
</tbody>
</table>

**Whisper**

<table>
<thead>
<tr>
<th></th>
<th>% posts</th>
<th>Whisper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigga</td>
<td>31.11</td>
<td>Black people</td>
</tr>
<tr>
<td>White people</td>
<td>9.76</td>
<td>Fake people</td>
</tr>
<tr>
<td>Fake people</td>
<td>5.07</td>
<td>Fat people</td>
</tr>
<tr>
<td>Black people</td>
<td>4.91</td>
<td>Stupid people</td>
</tr>
<tr>
<td>Stupid people</td>
<td>2.62</td>
<td>Gay people</td>
</tr>
<tr>
<td>Rude people</td>
<td>2.60</td>
<td>White people</td>
</tr>
<tr>
<td>Negative people</td>
<td>2.53</td>
<td>Racist people</td>
</tr>
<tr>
<td>Ignorant people</td>
<td>2.13</td>
<td>Ignorant people</td>
</tr>
<tr>
<td>Nigger</td>
<td>1.84</td>
<td>Rude people</td>
</tr>
<tr>
<td>Ungrateful people</td>
<td>1.80</td>
<td>Old people</td>
</tr>
</tbody>
</table>

---
Questions?