Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems

Ciro Cattuto, Dominik Benz, Andreas Hotho, Gerd Stumme

Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, Torino, Italy
Research Unit Knowledge and Data Engineering (KDE), University of Kassel, Germany
Everybody is tagging...

- simple and intuitive way to organize resources, immediately useful
- uncontrolled vocabulary
- however: evidence for converging vocabulary / emergent semantics due to
  - shared implicit knowledge
  - mutual influence of users
  - underlying social networks

So how to pick the best from the tag soup?
- Extract concepts
- Harvest semantics
- Learn ontologies
Agenda

- Topic Definition
- Folksonomy Definition and Data
- Tag Similarity Measures
- Semantic Grounding
- Summary and Outlook
Topic Definition: Semantic Grounding of Tag similarity

- Final Goal: Understand “tag semantics” in a folksonomy, i.e.,
  - Which tags describe the same / a more specific / a more general concept?
- Two basic approaches:

Look up tags in external thesaurus:

+ semantically grounded metrics
- “folksonomy jargon” (misspellings, neologisms etc.) not present

Apply measures directly to folksonomy structure (e.g. cooccurrence statistics, ...)

+ inclusion of complete vocabulary
- semantic interpretation of measures is not clear

→ Understand characteristics of (distributional) measures
→ assess their applicability for concept extraction, ontology learning, ...
Agenda

■ Topic Definition
■ Folksonomy Definition and Data
■ Tag Similarity Measures
■ Semantic Grounding
■ Summary and Outlook
Folksonomy Model

- A folksonomy \( F = (U, T, R, Y) \) consists of the sets
  - Users \( U \)
  - Tags \( T \)
  - Resources \( R \)
  - Tag assignments \( Y \subseteq (U \times T \times R) \)

- Can be seen as
  - ternary relation
  - tripartite hypergraph \( G = ((U \cup T \cup R), Y) \)
Folksonomy Dataset

- **Del.icio.us crawl 2006**
  - $|U| = 667,128$  \  \  $|T| = 2,454,546$  \  \  $|R| = 18,782,132$
  - $|Y| = 140,333,714$

- **Excerpt: 10,000 most popular tags**
  - $|U| = 476,378$  \  \  $|T| = 10,000$  \  \  $|R| = 12,660,470$
  - $|Y| = 101,491,722$

- **In the following: tag rank = position in most-popular list:**
  - 1: design
  - 2: software
  - 3: blog
  - 4: web
  - ...
Agenda

- Topic Definition
- Folksonomy Definition and Data
- Tag Similarity Measures
- Semantic Grounding
- Summary and Outlook
Similarity Measures: Co-occurrence + Cosine

- Take Co-occurrence frequency as similarity measure (freq):

\[ freq(t_1, t_2) = |\{(u, r) \in U \times R : (u, t_1, r) \in Y \land (u, t_2, r) \in Y\}| \]

- Describe each tag as a vector, whereby each dimension of the vector space corresponds to another tag. Compute similar tags by cosine similarity (cosine).
  (The same can be done in the user space or the resource space and with TF-IDF.)
### Most related tags by cooccurrence / cosine similarity

<table>
<thead>
<tr>
<th>art</th>
<th>design</th>
<th>photography</th>
<th>illustration</th>
<th>blog</th>
<th>graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td>web2.0</td>
<td>ajax</td>
<td>web tools</td>
<td>blog</td>
<td>webdesign</td>
<td></td>
</tr>
<tr>
<td>news</td>
<td>blog technology</td>
<td>politics</td>
<td>media</td>
<td>daily</td>
<td></td>
</tr>
<tr>
<td>howto</td>
<td>tutorial</td>
<td>reference</td>
<td>tips</td>
<td>linux</td>
<td>programming</td>
</tr>
<tr>
<td>video</td>
<td>music</td>
<td>funny</td>
<td>tv</td>
<td>software</td>
<td>media</td>
</tr>
<tr>
<td>ajax</td>
<td>javascript</td>
<td>web2.0</td>
<td>web programming</td>
<td>webdesign</td>
<td></td>
</tr>
<tr>
<td>tutorial</td>
<td>howto</td>
<td>programming</td>
<td>reference</td>
<td>design</td>
<td>css</td>
</tr>
<tr>
<td>javascript</td>
<td>ajax</td>
<td>programming</td>
<td>css</td>
<td>web</td>
<td>webdesign</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>art</th>
<th>graphic</th>
<th>creative</th>
<th>print portfolios</th>
<th>nice</th>
</tr>
</thead>
<tbody>
<tr>
<td>web2.0</td>
<td>web2</td>
<td>web-2.0</td>
<td>webapp “web _2.0”</td>
<td></td>
</tr>
<tr>
<td>news</td>
<td>blogs</td>
<td>people</td>
<td>weblog culture</td>
<td>future</td>
</tr>
<tr>
<td>howto</td>
<td>how-to</td>
<td>guide tutorials</td>
<td>help</td>
<td>how_to</td>
</tr>
<tr>
<td>video</td>
<td>entertainment</td>
<td>awesome</td>
<td>fun</td>
<td>cool</td>
</tr>
<tr>
<td>ajax</td>
<td>dhtml</td>
<td>dom</td>
<td>js</td>
<td>ecmaScript</td>
</tr>
<tr>
<td>tutorial</td>
<td>tutorials</td>
<td>tips</td>
<td>coding</td>
<td>code</td>
</tr>
<tr>
<td>javascript</td>
<td>webdevelopment</td>
<td>webdev</td>
<td>example</td>
<td>examples</td>
</tr>
</tbody>
</table>
Example for cosine measure
Similarity Measures: FolkRank

- Take Co-occurrence frequency as similarity measure (freq).
- Cosine Similarity between tag vectors
- Use FolkRank to find related tags (folkrank).
  - Basic Idea: PageRank-like spreading of weights through folksonomy structure + high weights for a particular tag in the random surfer vector

Web graph

Folksonomy graph

### Examples of most related tags

<table>
<thead>
<tr>
<th>rank</th>
<th>tag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>web2.0</td>
<td>ajax</td>
<td>web</td>
<td>tools</td>
<td>blog</td>
<td>webdesign</td>
</tr>
<tr>
<td>15</td>
<td>howto</td>
<td>tutorial</td>
<td>reference</td>
<td>tips</td>
<td>linux</td>
<td>programming</td>
</tr>
<tr>
<td>28</td>
<td>games</td>
<td>fun</td>
<td>flash</td>
<td>game</td>
<td>free</td>
<td>software</td>
</tr>
<tr>
<td>30</td>
<td>java</td>
<td>programming</td>
<td>development</td>
<td>opensource</td>
<td>software</td>
<td>web</td>
</tr>
<tr>
<td>39</td>
<td>opensource</td>
<td>software</td>
<td>linux</td>
<td>programming</td>
<td>tools</td>
<td>free</td>
</tr>
<tr>
<td>1152</td>
<td>tobuy</td>
<td>shopping</td>
<td>books</td>
<td>book</td>
<td>design</td>
<td>toread</td>
</tr>
</tbody>
</table>

### FolkRank

<table>
<thead>
<tr>
<th>rank</th>
<th>tag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>web2.0</td>
<td>web</td>
<td>ajax</td>
<td>tools</td>
<td>design</td>
<td>blog</td>
</tr>
<tr>
<td>15</td>
<td>howto</td>
<td>reference</td>
<td>linux</td>
<td>tutorial</td>
<td>programming</td>
<td>software</td>
</tr>
<tr>
<td>28</td>
<td>games</td>
<td>game</td>
<td>fun</td>
<td>flash</td>
<td>software</td>
<td>programming</td>
</tr>
<tr>
<td>30</td>
<td>java</td>
<td>programming</td>
<td>development</td>
<td>software</td>
<td>ajax</td>
<td>web</td>
</tr>
<tr>
<td>39</td>
<td>opensource</td>
<td>software</td>
<td>linux</td>
<td>programming</td>
<td>tools</td>
<td>web</td>
</tr>
<tr>
<td>1152</td>
<td>tobuy</td>
<td>toread</td>
<td>shopping</td>
<td>design</td>
<td>books</td>
<td>music</td>
</tr>
</tbody>
</table>

### Cosine

<table>
<thead>
<tr>
<th>rank</th>
<th>tag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>web2.0</td>
<td>web2</td>
<td>web-2.0</td>
<td>webapp</td>
<td>“web”</td>
<td>web_2.0</td>
</tr>
<tr>
<td>15</td>
<td>howto</td>
<td>how-to</td>
<td>guide</td>
<td>tutorials</td>
<td>help</td>
<td>how_to</td>
</tr>
<tr>
<td>28</td>
<td>games</td>
<td>game</td>
<td>timewaster</td>
<td>spiel</td>
<td>jeu</td>
<td>bored</td>
</tr>
<tr>
<td>30</td>
<td>java</td>
<td>python</td>
<td>perl</td>
<td>code</td>
<td>c++</td>
<td>delphi</td>
</tr>
<tr>
<td>39</td>
<td>opensource</td>
<td>open_source</td>
<td>open-source</td>
<td>open.source</td>
<td>oss</td>
<td>foss</td>
</tr>
<tr>
<td>1152</td>
<td>tobuy</td>
<td>wishlist</td>
<td>to_buy</td>
<td>buyme</td>
<td>wish-list</td>
<td>iwant</td>
</tr>
</tbody>
</table>
Qualitative insights: Overlap & average rank of related tags

<table>
<thead>
<tr>
<th>Overlap</th>
<th>Average rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosine-freq</td>
<td>1.7</td>
</tr>
<tr>
<td>cosine-foolrank</td>
<td>1.1</td>
</tr>
<tr>
<td>freq-foolrank</td>
<td>6,7</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between tag rank and average rank for different overlap measures.](image)
First insights

- Freq / FolkRank show bias to high-frequency tags, i.e., to **hyperonyms**.

- Cosine seems to yield more **synonyms** and “siblings”.

→ Now: **grounding** of these observations in WordNet.
Agenda

- Topic Definition
- Folksonomy Definition and Data
- Tag Similarity Measures
- Semantic Grounding
- Summary and Outlook
Semantic Grounding in WordNet

- WordNet is a large lexical database for English.

- Words with same meaning are grouped in synsets, which are ordered by an is-a hierarchy.

- Introduction of single artificial root node enables application of graph-based similarity metrics between pairs of nouns / pairs of verbs.

- Inclusion of top $n$ del.icio.us tags in WordNet:
  - $100$: 82%
  - $1,000$: 79%
  - $5,000$: 69%
  - $10,000$: 61%
Example of Semantic Grounding

- Original tag:
  - „java“

- Most similar tag:
  - Freq, folkrank: „programming“
  - Cosine: „python“

Wordnet Synset Hierarchy:

- Map
- Grounded similarity

Diagram showing the hierarchy of Wordnet synsets with the terms “computers”, “programming”, “design_patterns”, “languages”, “java”, and “python”. The arrow indicates the grounded similarity between “java” and “python”.
Shortest path between original tag and most closely related one

Shown to be the semantically most adequate measure for similarity within WordNet [Budanitsky, Hirst, 2006].
shortest paths in WordNet

length of shortest path to most related tag
Agenda

- Topic Definition
- Folksonomy Definition and Data
- Tag Similarity Measures
- Semantic Grounding
- Summary and Outlook
Summary

- Analysis of tag similarity measures by mapping to WordNet

- Exposed clearly different characteristics:
  - freq measure and FolkRank tend to more general tags
  - Synonyms and siblings are the result of the cosine measure

- Implications for ontology learning:
  - Insights can inform the choice of an appropriate measure to extract semantic tag relations
  - e.g., FolkRank to find Hyperonyms, Cosine measure for Synonyms

- Now: Embed these measures in an ontology learning procedure
Thanks for your attention!

contact:
benz@cs.uni-kassel.de

Try it yourself:
www.bibsonomy.org
Appendix: Music Genre Taxonomy learned from last.fm