Evaluating Similarity Measures for Emergent Semantics of Social Tagging

Ben Markines, Ciro Cattuto, Fil Menczer, Dominik Benz, Andreas Hotho, Gerd Stumme
Social Applications

BibSonomy

givealink.org
I donated my bookmarks to science.

myspace.com
a place for friends...

LinkedIn

Facebook

YouTube

flickr

Broadcast Yourself
Related URLs sorted by Related URL List

1-10 of 1000 results for http://www.cnn.com

BBC NEWS | News Front Page
Find Similar Results

Los Angeles, California, national and world news, jobs, real estate, cars - Los Angeles Times
Find Similar Results

Find Similar Results

CNET News.com -- Technology news and business reports
Find Similar Results
Goals

• Tag-tag, resource-resource, user-user similarity

• Capture relationships
  – Effectively and Efficiently
  – Shannon information of annotations
Folksonomy Model

F = (U, T, R, Y), Y ⊆ U × T × R (the triples)

- hyper-graph
- complex
- user-driven
- large-scale
- many-projections
- literature
Agenda

• Design
  – Aggregation
  – Similarity Measures

• Evaluation
Aggregation Methods

**Projection**

<table>
<thead>
<tr>
<th></th>
<th>news</th>
<th>web</th>
<th>tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnn.com</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>www2009.org</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>wired.com</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Distributional**

<table>
<thead>
<tr>
<th></th>
<th>news</th>
<th>web</th>
<th>tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnn.com</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>www2009.org</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>wired.com</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Aggregation Methods: Incremental

\[ \sigma(x, y) = \sum_u \sigma_u(x, y) \]

Macro

<table>
<thead>
<tr>
<th></th>
<th>news</th>
<th>web</th>
<th>tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnn.com</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>www2009.org</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

alice

bob
Aggregation Methods: Incremental (2)

Collaborative

<table>
<thead>
<tr>
<th></th>
<th>news</th>
<th>web</th>
<th>tech</th>
<th>alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnn.com</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>www2009.org</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \sigma(x, y) = \sum_u \sigma_u(x, y) \]
Similarity Measures

- **Jaccard**
  \[ \sigma(x_1, x_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \]

- **Matching**
  \[ \sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1 \cup X_2} \log p(y)} \]

- **Overlap**
  \[ \sigma_u(x_1, x_2) = \frac{\sum_{y \in X_1^u \cap X_2^u} \log p(y|u)}{\sum_{y \in X_1^u \cup X_2^u} \log p(y|u)} \]

- **Dice**
  \[ p(y|u) = \frac{N(u, y)}{N(u) + \delta} \]
More Similarity Measures

- **Cosine**
  \[
  \sigma(x_1, x_2) = \frac{X_1 \cdot X_2}{||X_1|| \cdot ||X_2||}
  \]

- **Mutual Information**
  \[
  \sigma(x_1, x_2) = \sum_{y_1 \in X_1} \sum_{y_2 \in X_2} p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)}
  \]

- **Maximum Information Path**
  \[
  \sigma(x_1, x_2) = \frac{2 \times \log(\min_{y \in X_1 \cap X_2}[p(y)])}{\log(\min_{y \in X_1}[p(y)]) + \log(\min_{y \in X_2}[p(y)])}
  \]
Agenda

• Design
• Evaluation
  – Efficiency
  – Predicting tag relations
  – Semantic Grounding
Predicting User-defined Tag Relations
Predicting User-defined Tag Relations Area Under ROC Curve
ROC limitations

- Data is sparse: 2,000 tags
  - 142 user tag relations
- Similarity values are broadly distributed
- Tag relations rely on hierarchical relationships
- Only available with tags (not resources)
Semantic Grounding

**WordNet**

- 17,041 tags
  - Overlap between WordNet and Bibsonomy tags
- Limited to the top 2,000 resources
- Relationships established with Jiang-Conrath
  - user-validated

**dmoz**

- 3,323 resources
  - Overlap between ODP and Bibsonomy resources
- Relationships established with Maguitman’s graph based similarity
  - user-validated

Jiang ROCLING 1997

Maguitman WWW 2005
Kendall’s $\tau$

<table>
<thead>
<tr>
<th>Rank</th>
<th>Reference</th>
<th>Measure A</th>
<th>Measure B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tech-web</td>
<td>news-tech</td>
<td>news-web</td>
</tr>
<tr>
<td>2</td>
<td>news-web</td>
<td>tech-web</td>
<td>tech-web</td>
</tr>
<tr>
<td>3</td>
<td>news-tech</td>
<td>news-web</td>
<td>news-tech</td>
</tr>
<tr>
<td>$\tau$</td>
<td>1</td>
<td>1/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>

\[ \tau = \frac{|\text{agreed ranked pairs}|}{|\text{total number of ranked pairs}|} \]

Kendall Biometrika 1938
Tag Similarity

random $\tau = 10^{-4}$
Resource Similarity

projection

Distributive

Macro

Collab. $\delta = 1$

Collab. $\delta = 10^6$

$\tau / \tau_{\text{random}}$

random $\tau = 8 \times 10^{-5}$

HT 2009
Related Work

• User, tag and resource similarity

• Ranking
  – Hotho et al. 2006

• Organization

• Link Prediction
  – Liben-Nowell and Kleinberg 2003

• Recommendation
Conclusion

• Similarity framework
  – Folksonomy-based tag/resource similarity measures
  – Aggregation methods

• Evaluation
  – Efficiency/performance tradeoffs
  – Direct vs. semantic grounding
  – Distributional Mutual Information performs well, but is inefficient
  – Collaborative aggregation is both efficient and effective, especially Maximum Information Path

• Techniques presented here can immediately support Social Web applications
Thank You!

- Similarity framework
  - Folksonomy-based tag/resource similarity measures
  - Aggregation methods
- Evaluation
  - Efficiency/performance tradeoffs
  - Direct vs. semantic grounding
  - Distributional Mutual Information performs well, but is inefficient
  - Collaborative aggregation is both efficient and effective, especially Maximum Information Path
- Techniques presented here can immediately support Social Web applications

Ben Markines
Ciro Cattuto
Fil Menczer
Dominik Benz
Andreas Hotho
Gerd Stumme