A Case Study on Emergent Semantics in Communities

Elke Michlmayr,
Women's Postgraduate College for Internet Technologies (WIT),
Vienna University of Technology,
http://wit.tuwien.ac.at
Outline

- Folksonomies
  - What are they?
- Comparison to taxonomies
  - Methodology
  - On the data level
- Folksonomies and peer-to-peer networks
  - User behaviour
  - Usable as test data?
- Related work
- Summary
Folksonomies and (collaborative) tagging

- Multi-user web applications that provide a simple categorization system

- Items
  - Web pages (Deli.cio.us, Furl, ...)
  - Images (Flickr)
  - Citations (Connotea, CiteULike)

- Tags = keywords
  - Can be chosen freely

- Every user has a web page with a list of own items
  - Sorted in reverse-chronological order
  - Can be filtered by tag(s)

- Public access to item collections and meta-data
Example: del.icio.us user interface
“Bottom-up” approach to categorization

- No pre-defined model or hierarchy
- Inconsistencies
  - Synonyms, homonyms
  - Singular and plural versions of a tag
  - Keywords that consist of two terms
    - i.e., semantic web, semantic_web, semanticweb
- Relies on aggregation of meta-data
  - Tag frequency distribution
    - Tags most often used to annotate an item categorize it best
    - No need to reach consensus
  - Relationships between tags evolve from meta-data
- Amount of meta-data crucial!
  - Number of users, lifetime of folksonomy
Comparison of meta-data

- Lots of discussions about taxonomies vs. folksonomies, e.g., Clay Shirky 2005
- Experiment: compare meta-data from two big community projects that categorize Web pages to find out about the differences
  - DMOZ open directory project http://dmoz.org/
    - Taxonomy for Web pages
    - ~600000 concepts and ~5000000 instances
    - Available in RDF format (two big files)
  - Social bookmarking site http://del.icio.us/
    - No official numbers, ~100000 users
    - RDF file for each collection and for each item
Data gathering

■ Procedure
  ● Use only items from del.icio.us that were annotated by more than 100 users (= popular items)
  ● Download random popular items from del.icio.us
  ● Lookup if items are present in the DMOZ collection
    ○ ~25 % of the items were also present in DMOZ

■ 788 items with meta-data from both sources
  ● ~50 % of them are instances of DMOZ concept Top/Computers

URL http://arxiv.org/
DMOZ Top/Science/Physics/Publications
DMOZ Top/Science/Math/Publications
DMOZ Top/Science/Math/Publications/Online_Texts/Collections
DMOZ Top/Science/Publications/Archives/Free_Access_Online_Archives
ID 19aa8ff1e9c2a06677ab34f3f2a5b0c8
TITLE arXiv.org e-Print archive
TAGS physics:43; science:41; research:27; math:23; papers:19; reference:18; mathematics:15; journal:10; articles:10; archive:9; biology:8; eprint:7; library:7; preprint:6; books:6; programming:6; cs:5; article:5; academic:5; computer:4; arxiv:4; literature:4; toread:4; computer_science:4; ai:3; study:3;
Preparation of data

- Preparations
  - Convert to lower case, remove underscores and hyphens
  - Remove last character `s` because of singular/plural tags
  - Don’t consider Top/World (multi-lingual categories)
  - Remove all categories with one character only (/[A-Z]/)
  - Remove Top category
  - Sort category names in reverse to put most specific entry first
  - Rank tags by number

- Example
  - Top/Science/Math/Publication -> publication math science

- How to compare?
  - Avg. DMOZ hierarchy length: 4.67
  - Avg. deli.cio.us tags per item: 24.59
Comparison

- Lookup for each DMOZ category
  - Is it included in the del.icio.us tags?
- Take top 1, 3, 5, 10, 15, all tags into account
  - Top tag is included in ~50% of the cases
  - Top 5 is the fairest comparison
  - Top tags match more often than the less popular ones

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th to 11th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top tag</td>
<td>9.44 %</td>
<td>15.94 %</td>
<td>12.67 %</td>
<td>4.72 %</td>
<td>3.28 %</td>
<td>1.72 %</td>
<td>0.81 %</td>
</tr>
<tr>
<td>Top 3 tags</td>
<td>20.37 %</td>
<td>27.55 %</td>
<td>21.58 %</td>
<td>14.29 %</td>
<td>12.23 %</td>
<td>6.21 %</td>
<td>2.30 %</td>
</tr>
<tr>
<td>Top 5 tags</td>
<td>28.32 %</td>
<td>34.81 %</td>
<td>27.72 %</td>
<td>19.75 %</td>
<td>16.42 %</td>
<td>11.03 %</td>
<td>3.69 %</td>
</tr>
<tr>
<td>Top 10 tags</td>
<td>37.38 %</td>
<td>44.53 %</td>
<td>35.94 %</td>
<td>27.08 %</td>
<td>25.91 %</td>
<td>18.28 %</td>
<td>6.25 %</td>
</tr>
<tr>
<td>Top 15 tags</td>
<td>44.30 %</td>
<td>52.45 %</td>
<td>43.17 %</td>
<td>34.16 %</td>
<td>32.12 %</td>
<td>26.55 %</td>
<td>8.93 %</td>
</tr>
<tr>
<td>All tags</td>
<td>52.99 %</td>
<td>62.55 %</td>
<td>52.48 %</td>
<td>46.34 %</td>
<td>44.34 %</td>
<td>40.34 %</td>
<td>14.73 %</td>
</tr>
</tbody>
</table>
Folksonomies and peer-to-peer networks

- **Architectures are very different**
  - Folksonomies are centralized systems, aggregation is easy
  - Peer-to-peer networks are distributed, aggregation is hard

- **User behaviour is comparable**
  - Act autonomously
  - No central authority
  - Want to share information

- **Data from a folksonomy can be used to model peers and content distribution**
  - No data about queries available

- **Experiment**
  - Can subsets of the del.icio.us data be selected in such a way that the principle of interest-based locality be observed in these subsets?
Can interest-based locality be observed?

- **Interest-based locality**
  - “If peer A has a particular piece of content peer B is interested in, it is likely the case that the other information items stored by peer A are also of interest to peer B.”

- **Method**
  1. Retrieve all users from del.icio.us that store a random bookmark
  2. Retrieve all their collections

- **Retrieved 4 test sets**
  - 155, 248, 280, 551 users
  - Distribution of items among users nearly equal in the test sets
  - Avg.: 84% of items are not shared!

<table>
<thead>
<tr>
<th>Not shared</th>
<th>84 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>By 2 users</td>
<td>8.9 %</td>
</tr>
<tr>
<td>By 3 users</td>
<td>2.92 %</td>
</tr>
<tr>
<td>By 4 users</td>
<td>1.49 %</td>
</tr>
<tr>
<td>5-10 users</td>
<td>2.18 %</td>
</tr>
<tr>
<td>&gt; 10 users</td>
<td>0.51 %</td>
</tr>
</tbody>
</table>
Related work

- Adam Mathes, 2004: *Folksonomies – Cooperative Classification and Communication Through Shared Metadata*
  - Very good introduction

- Clay Shirky, 2005: *Ontology is Overrated: Categories, Links and Tags*
  - Controversial discussion of taxonomies vs. folksonomies

- Scott Golder and Bernardo Huberman, 2005: *The structure of Collaborative Tagging Systems*
  - Cognitive aspects
  - Data analysis: Tag frequency distribution for an item is stable over time
Summary

- Investigated the properties of meta-data provided by a folksonomy
- Compared it to DMOZ data collection
- Tried to find interest-based locality
- Paper contains some other experiments I did not have time to tell you about

Open questions
  - Is there a way to combine the bottom-up and top-down approach for creating metadata?
  - How much could the semantic web benefit from it?