Growing a Tree in the Forest: Constructing Folksonomies by Integrating Structured Metadata

Anon Plangprasopchok (USC/ISI)
Kristina Lerman (USC/ISI)
Lise Getoor (UMD)
Explosion of user-generated content

- Images: Flickr, Picasa,..
- Videos: YouTube, Vimeo,..
- Maps: WikiMapia,..
- Stories: Blogs, Twitter,..
- Relational Data: Metaweb, Google Base,..

User-generated semantics: annotation/metadata

- Tags, Geotags
- Personal Hierarchies

Goal: extract users’ knowledge (folk knowledge) from this metadata
Folksonomy that users commonly have in their mind (hidden)

Users select a portion of the hierarchy to organize their content.

Personal hierarchies from various users (observed) such as users’ folder-sub folders

[shallow, noisy, sparse(incomplete) & inconsistent]

Can we recover the folksonomy back from many observed hierarchies? → folksonomy learning
Outline

• Personal Hierarchies in Social Web
• Challenges in Folksonomy Learning
• Integrate Personal Hierarchies to Folksonomy
  • Relational Clustering for Learning Folksonomy
• Evaluations
  • Metrics
  • Results
• Related Work
• Discussion & Conclusions
Structured Social Metadata in Flickr

Shallow Hierarchy by user “maxi millipede”

Assume:
1) A set of tags on a set is an aggregation of all tags of all photos in the set
2) A set of tags on a collection is an aggregation of all tags of all sets in the collection
Challenges in Folksonomy Learning

1.) Sparseness:
- Animal
  - Bird
  - Cat

 mostly personal hierarchies contain very few child nodes

2.) Ambiguity:
- Bird
  - Duck
  - Geese
  - Turkey
- Turkey
  - Istanbul
  - Antalya
  - Amasya

3.) Conflict:
- Travel
  - Japan
  - China
  - USA
- China
  - Travel
  - Food
  - People

4.) Varying Granularity:
- UK
  - Scotland
  - London
  - Glasgow
  - Edinburgh
- Scotland
  - Shetland
  - London

ubiquitous

very rare!
Sketched idea: *combine/aggregate personal hierarchies together in both horizontal and vertical directions.*

Horizontal aggregation: expanding folksonomy’s width

Vertical aggregation: extending folksonomy’s depth
**Sketched idea:** 2 nodes should be clustered if they are similar enough
- similarity is computed using **contextual & relational information**

Check common tags & child nodes
Relational Clustering for Folksonomy Learning

Formally, two nodes are considered similar if:

1. their features are similar, i.e., have similar names, have many common tags – *local similarity*
2. their neighbors are similar – *structural similarity*

![Diagram](image)

\[
\text{Sim}(A,B) = (1-\alpha) \times \text{localsim}(A,B) + \alpha \times \text{structuralSim}(A,B)
\]

We then merge nodes together if they are similar enough.

Note that we use naïve version of relational clustering by simply using neighbors' local features to compute structural similarity, instead of neighbors' class labels*.

*see Bhattacharya & Getoor, 2007, Collective Entity Resolution in Relational Data, TKDD for more detail
- Pick a seed (root) term, e.g., “canada”

Making it broader

Making it deeper

& remove structural inconsistencies that may appear

“noisy” nodes are removed (using # users)
Suppose we have the following clusters of hierarchies:

- Shortcuts have to be removed to make the learned hierarchy consistent.
1) attach “England”
2) remove “London” shortcut
3) Attach “London”

- Keep the longer path since it captures more specific knowledge
- Growing trees from 32 seed terms & uses personal hierarchies from Flickr as in the previous work.*

**Evaluation Methodologies:**

1) **Against the reference hierarchy** (DMOZ)
2) **Structural evaluation**
3) **Manual evaluation**

**Baseline Approach**
- Assume nodes having the same name refer to the same concept
- Keep the relations between node pairs if they are not generated at random (using significance test)
- Then, combine all relations into a tree

* A. Plangprasopchok and K. Lerman, 2009, Constructing folksonomies from user-specified relations on flickr, WWW
1.) an automatic comparison to the reference hierarchy

- **Taxonomic Overlap** [adapted from Maedche & Staab] measuring structure similarity between two trees. For each node, determining how many ancestor and descendant nodes overlap to those in the reference tree.

- **Lexical Recall** measuring how well an approach can discover concepts, existing in the reference hierarchy (coverage)

2.) Structural evaluation

- **Area Under Tree (AUT)** combining bushiness and depth of the tree into a single number: the higher value, the bushier and deeper tree.

3.) Manual evaluation

- **Accuracy**: simply asking users whether a path from root to leaf of is correct: if there are some nodes misplaced in the wrong order, users will judge the whole path incorrect

Maedche & S. Staab, 2002, Measuring Similarity between Ontologies, in EKAW
Area Under Tree (AUT)

Which structures are the best in term of “bushiness” and “depth”?

a)

b)

c)

d)

Plot the distribution on # of nodes at each depth

Then, compute the area here (trapezoids with height value = 1)
The largest area we can get is from the tree that keeps spanning at each level.
Evaluate on 32 cases

<table>
<thead>
<tr>
<th>Metrics</th>
<th># of cases that are superior to the other approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
</tr>
<tr>
<td>Taxonomic Overlap</td>
<td>7</td>
</tr>
<tr>
<td>Lexical Recall</td>
<td>6</td>
</tr>
<tr>
<td>AUT</td>
<td>3</td>
</tr>
<tr>
<td><strong>TO+LR+AUT</strong></td>
<td>0</td>
</tr>
<tr>
<td>Accuracy (Manual)</td>
<td>5</td>
</tr>
</tbody>
</table>
Some of the Learned Folkonomies

Terms are stemmed
Some of the Learned Folkonomies
Related Work

- Learning concept hierarchy from text data
  - Syntactic based [Hearst92, Caraballo99, Pasca04, Cimiano+05, Snow+06]
  - Word clustering [e.g., Segal+02, Blei+03]

- Induce concept hierarchy from tags
  - Graph-based & clustering based [Mika05, Brooks+06, Heymann+06, Zhou07+]
  - Probabilistic subsumption [Schmitz06]

- Ontology alignment [e.g., Udrea+07]

- Exploit user-specified hierarchy for recommendation
  - GiveALink [Markines06+]
Discussion & Conclusions

• The present work can create more accurate and more detailed folksonomies than the current state-of-the-art approach, since it exploits structural information during the merging process.

• The present work is more scalable: incrementally growing the folksonomies rather than using on an exhaustive search.

• Future work:
  • **Automatically separate broader/narrower from related-to relations (facets)**
  • **Combining more sources of evidence such as geographical information**
  • **Apply on different data sets: e.g., personal workspaces, semantic network**
• Thank you..
• Our poster is #21
Outline

• Personal Hierarchies in Social Web
• Challenges in Folksonomy Learning
• Integrate Personal Hierarchies to Folksonomy
  • Relational Clustering for Learning Folksonomy
• Evaluations
  • Metrics
  • Results
• Discussion & Conclusions
• Related Work
AUT (2)

A Learned Tree

A Distribution of Nodes at Each Depth

# of nodes

AUT = 5.5

@ depth

0 1 1st 2nd 3rd 4th
Structural Similarity

Depends on the roles (root or leaf) of two nodes to be compared:

Root - Root:
Let \( K_{A,B} = \frac{|\text{name(leaves(A))} \cap \text{name(leaves(B))}|}{\text{min}(|\text{leaves(A)}|, |\text{leaves(B)}|)} \)

\[
\text{structuralSim}(R1,R2) = K_{r1,r2} + (1 - K_{r1,r2}) \times \text{tagsim( leaf nodes of A,B that do not have common name)}
\]

Leaf - Root:
\[
\text{structuralSim}(L1,R2) = \text{structuralSim(root(L1), R2)}
\]

Leaf - Leaf:
If the roots of node A and B are considered similar, we simply say A and B are similar if they have the same name.
Plot the distribution on # of nodes at each depth

Then, compute the area here (trapezoids with height value = 1)