Folksonomy Learning

Kristina Lerman
University of Southern California

This lecture is partly based on slides prepared by Anon Plangprasopchok
Formal vs Emergent Semantics

- **Ontology**: an explicit specification of the conceptualization of a domain
  - Challenges of formal ontologies
    - Complicated – Users are slow to adopt
    - Costly to produce
    - Ontology drift – do not keep up with evolving communities and user needs

- **Folksonomy**: emergent semantics arising out of interactions among many users
  - Advantages over formal ontologies
    - Created from collective agreement of many individuals;
    - Relatively inexpensive to obtain;
    - Can adapt to evolving vocabularies and community’s information needs;
Annotated according to a formal (Linnean) taxonomy or Scientific Classification System

<Kingdom>Animalia</Kingdom>
<Phylum>Chordata</Phylum>
<Class>Aves</Class>
<Genus>Merops</Genus>
<Species>M. ornatus</Species>
Classifying Entities using a Folksonomy

submitter

private
albums

public
groups

tags

Rainbow bee-eater
Merops ornatus
Australia
Queensland
Mackay Gardens

Tags
Rainbow bee-eater
Merops ornatus
Australia
Queensland
Mackay Gardens

Additional Information
© All rights reserved
Anyone can see this photo
- Taken with a 4/3
- More properties
- Taken on May 24, 2009
- 1 person adds this photo as a favorite
- Viewed 4 times

Comments
laRuth says:
Gorgeous colours on that bird!
Posted 10 months ago. (permalink)

aaardvark says:
If the bee-eaters won’t come to you, you go to the bee-eaters! It’s interesting to be where your local birds migrate to.
Posted 10 months ago. (permalink)

frangipanni IS HERE!! says:
Hi, I’m an admin for a group called Birds - Kingfishers, Pittas, Rollers & Bee-eaters, and we’d love to have this added to the group.
Posted 10 months ago. (permalink)
User-created Hierarchies

~Aquila~

Shapes, Textures and Patterns
3 sets

Places
4 sets

Natural Things
8 sets

People and Pets
4 sets

Flowers
17 photos

Birds
37 photos

Reptiles & Amphibians

Oz Fungi
2 photos

Up Close and Personal
6 photos
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]

Exploit user-specified hierarchies
- GiveALink [Markines06+]
- Constructing Folksonomies by Integrating Structured Metadata [Plangprasopchok09,+]
Users describe objects with metadata of their own choice

- Tags – keywords from uncontrolled personal vocabularies
- Structured metadata – user-specified hierarchies

Interactions between large numbers of users leads to a global consensus on semantics

- Consensus represents emergent semantics
  - Tags ~ Concepts
- Consensus emerges quickly [cf Golder & Huberman]
- Need a model of semantic-social networks [Mika, “Ontologies Are Us”, ISWC 2005]
Tripartite Model of Ontologies

- Resources (Instance)
- Users (Actors)
- Tags (Concepts)
Reduce tripartite hypergraph to three bipartite graphs

- User-Tag (Actor-Concept) graph
- Tag-Resource (Concept-Instance) graph
- User-Resource (Actor-Instance) graph
Fold bipartite graph to create two simple graphs

- CI graph represented by adjacency matrix $\mathbf{B} = \{b_{ij}\}$
  - Cf Document-Term matrix

1) social network that connects users based on shared tags $\mathbf{S} = \mathbf{BB}'$

2) lightweight ontology of concepts based on overlapping sets of docs $\mathbf{O} = \mathbf{B}'\mathbf{B}$
Tripartite Model of Ontologies – Step 3

• Bipartite CI graph leads to
  • A semantic network where links between tags are weighed by the number of resources they both tag
    • Cf text mining – terms are associated by their co-occurrence in documents

• AI graph leads to
  • A social network where links between users give the number of resources they both tagged
  • A graph where links between resources showing the number of people who tagged a given pair of resources
Folksonomy Learning as Network Analysis

- Learn concepts and broader → narrower relations between concept from semantic networks
  - Concept A is a superconcept of Concept B
    - If the set of entities classified under B is a subset of entities under A
    - Set of A is significantly larger than the set of B
- By applying network analysis tools to semantic networks
  - Clustering coefficient
  - Betweenness centrality
Empirical Validation

Delicious dataset

- 30,790 URLs (instances)
- 10198 users (actors)
- 29,476 tags (concepts)
Tag co-occurrence clusters

Main concept clusters in tag-resource network

<table>
<thead>
<tr>
<th>travel</th>
<th>cote, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>venture_capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, vc</td>
</tr>
<tr>
<td>free time</td>
<td>procrastination, info, advice, gtd, life, notes, planning, daily, reading, forums</td>
</tr>
<tr>
<td>sex</td>
<td>hot, to, street, pictures, on, photos, free, celeb, adult, lesbian</td>
</tr>
<tr>
<td>web design</td>
<td>design, designer, webdesign, premium, logo, logos, dreamweaver, templates, best, good</td>
</tr>
</tbody>
</table>
associations reflect overlapping communities of interest
### Broader $\rightarrow$ Narrower Relations

<table>
<thead>
<tr>
<th>Broader</th>
<th>Narrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>rss</td>
<td>atom</td>
</tr>
<tr>
<td>cmyk</td>
<td>rgb</td>
</tr>
<tr>
<td>cell</td>
<td>umts, wcdma, ev-do</td>
</tr>
<tr>
<td>phone</td>
<td>cell</td>
</tr>
<tr>
<td>ajax</td>
<td>json</td>
</tr>
<tr>
<td>xml</td>
<td>xslt</td>
</tr>
<tr>
<td>rdf</td>
<td>owl</td>
</tr>
<tr>
<td>flickr</td>
<td>gmail, picasa</td>
</tr>
<tr>
<td>ruby</td>
<td>rails</td>
</tr>
<tr>
<td>mac</td>
<td>iphoto</td>
</tr>
<tr>
<td>java</td>
<td>j2ee</td>
</tr>
<tr>
<td>google</td>
<td>gds</td>
</tr>
<tr>
<td>search</td>
<td>a9, engine</td>
</tr>
<tr>
<td>linux</td>
<td>ubuntu, gnome</td>
</tr>
<tr>
<td>flash</td>
<td>actionscript</td>
</tr>
<tr>
<td>flickr</td>
<td>licker, photoset</td>
</tr>
<tr>
<td>javascript</td>
<td>xmlhttprequest,</td>
</tr>
<tr>
<td></td>
<td>dom, sarissa</td>
</tr>
</tbody>
</table>

Relations in the Technology domain extracted from overlapping subcommunities on Delicious
Discussion

- Social tagging systems are effective, because they attract many like-minded people
- Community-based ontology extraction
  - Associations between concepts emerge as a consequence of social interactions
  - User graph-based tools to mine associations to create an ontology
- Limited quality
  - Associations are created from co-occurrence of objects
  - Problems of sparseness, ambiguity, synonymy
Subsumption approach applied to tag cooccurrence [Schmitz, 2006]

- Tag x subsumes y if
  \[ P(x|y) \geq t \] and \[ P(y|x) < t \]
- x is broader than y or \( x \rightarrow y \)
  - E.g., bird → finch

No. images tagged x

No. images tagged y
Some problems:

- **Generality vs Popularity**
  - Washington $\rightarrow$ United States
  - Car $\rightarrow$ Automobile

- **Mixing tags from different facets**
  - Insect $\rightarrow$ Hongkong
  - Color $\rightarrow$ Brazilian

Above relations induced using tag-based subsumption on Flickr data
Folksonomy learning: relational clustering approach

This material is based on “Growing a tree in the forest: constructing folksonomies by integrating structured metadata” by A. Plangprasopchok, K. Lerman & L. Getoor, 2010.
Can we recover the folksonomy from the observed personal hierarchies? → folksonomy learning!

Folksonomy = Communal Taxonomy

Folksonomy that users commonly have in their mind (*hidden*)

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

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

[shallow, noisy, sparse (incomplete) & inconsistent]
Structured Social Metadata in Flickr

Personal hierarchy of *maxi_millipede*

- **“collection”**
  - Plant Pests
  - Sap Suckers
  - Plant Eaters
  - Caterpillars

- **“set”**
  - Plant Parasites
  - Sap Suckers
  - Plant Eaters
  - Caterpillars

- **“photos”**
  - Plant Pests
  - Sap Suckers
  - Plant Eaters
  - Caterpillars

- **“tags”**

**Tags on each photo**

1) The set aggregates tags of all photos in the set
2) The collection aggregates all tags of all sets in the collection

Assume:

- Tags on each photo
- Collection aggregates tags of all photos in the set
- The collection aggregates all tags of all sets in the collection
Challenges in Folksonomy Learning

1.) Sparseness:

- most personal hierarchies contain very few child nodes

2.) Ambiguity:

3.) Conflict:

4.) Varying Granularity:
Basic idea: combine/aggregate personal hierarchies together in both horizontal and vertical directions.

**Horizontal aggregation expands folksonomy's width**

**Vertical aggregation extends folksonomy’s depth**
Basic idea: 2 nodes should be merged (clustered) if they are similar enough. Similarity is computed using *structural information*.

\[ \text{victoria1} \neq \text{victoria2} \quad \text{because:} \]

\[ \{\text{ChildNodes(victoria1)}\} \cap \{\text{ChildNodes(victoria2)}\} = \emptyset \]

\[ \{\text{TopTags(victoria1)}\} \cap \{\text{TopTags(victoria2)}\} = \emptyset \]
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*

Local similarity: \( \text{sim}(A,B) \)

Structural similarity: \( \text{sim}(\text{neighbor}(A), \text{neighbor}(B)) \)

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

We then merge nodes together if they are similar enough.

*see Bhattacharya & Getoor, 2007, Collective Entity Resolution in Relational Data, TKDD for more detail*
Structural Similarity

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

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

for normalizing \( K \)

\[
\text{structuralSim}(R_1, R_2) = K_{r_1,r_2} + (1 - K_{r_1,r_2}) \times \text{tagsim}(\text{leaf nodes of } A,B \text{ that do not have common name})
\]

**Leaf vs. Root:**

\[
\text{structuralSim}(L_1, R_2) = \text{structuralSim}(\text{root}(L_1), R_2)
\]

**Leaf vs. Leaf:**

If the parents of \( A \) and \( B \) are similar, we simply say that \( A \) and \( B \) are similar if they have the same name.
1.) A user specifies a root term, e.g., “canada”
2-4.) cluster personal hierarchies with “canada” as their root name

5.) pick a leaf node; cluster all personal hierarchies having their root name similar as the leaf; and attach the most similar merged hierarchy to it
Handling shortcuts and loops

Suppose we have the following clusters of hierarchies:

- Scotland
- London
- England
- Dockland
- B. Museum

Some users mistakenly put “England” under “London”

- shortcuts have to be removed to make the learned hierarchy consistent
- the order of attaching does matter – we would attach the England hierarchy before London one to the UK because England is “closer” (more similar) to UK than London.
Handling shortcuts and loops (2)

1) Attach "England"

2) Remove "London" shortcut

3) Attach "London"

4) Remove England loop
Evaluations

- Compare to the baseline approach
- Baseline*
  - Assumes that nodes with the same name refer to the same concept
  - Keeps the relations between two nodes that are statistically significant
  - Combines them together into a tree
  - Shown to produce better folksonomies than tag subsumption

* A. Plangprasopchok and K. Lerman, 2009, Constructing folksonomies from user-specified relations on flickr, WWW
1. Automatic evaluation
   • Compare against a reference hierarchy
   • Metrics: **Lexical Recall, Taxonomic Overlap**

2. Structural evaluation
   • How detailed is the learned tree?
   • Metrics: **Area Under Tree (AUT)**

3. Manual evaluation
   • Ask users whether portions of learned tree are correct: e.g., path from root to leaf of is correct
   • Metrics: **Accuracy**
Automatic Evaluation

Compare against a reference taxonomy, e.g. DMOZ

- **Taxonomic Overlap** [adapted from Maedche & Staab]
  - measures 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)

\[
LR(T_1, T_2) = \frac{|C_1 \cap C_2|}{|C_2|}
\]

*A. Maedche & S. Staab, 2002, Measuring Similarity between Ontologies, in EKAW*
Area Under Tree (AUT) combining bushiness and depth of the tree into a single number: the higher value, the bushier and deeper tree. (see the next slide for more intuition)
Plot the distribution on # of nodes at each depth

Then, compute the area here (trapezoids with height value = 1)
Area Under Tree (AUT)

Which tree is best in term of “bushiness” and “depth”?

a) AUT = 4.5

b) AUT = 4.5

c) AUT = 6

d) AUT = 5

The highest area we can get is from the tree that keeps spanning at each level.
Evaluate on 32 cases (seeds)

<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>Taxonimic 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>Accuracy (Manual)</td>
<td>5</td>
</tr>
</tbody>
</table>
Examples of Learned Folksonomies

Terms are stemmed
Advantages:

• Creates more accurate and detailed folksonomies than the current state-of-the-art approach, since it exploits structure information during the merging process
• More scalable: incrementally growing the folksonomies rather than using on an exhaustive search

Disadvantages:

• Many parameters are required to specify, e.g., (1) weights combination between local and structure information in similarity measures; (2) thresholds for deciding whether two nodes are similar or not. Small changes on parameter values can significantly change the quality of the result.
• Ad hoc – combining hierarchies and resolving their inconsistencies are independent processes
Summary

• Social annotation domain presents rich, interlinked data for analysis
  • Entities – users, documents, annotations (tags, ...)
  • Different links between entities
    • User→tag : tag is in user’s vocabulary
    • Document→tag : document annotated with the tag, ...
  • New types of data
    • Learning from relations (hierarchies), rather than flat tags

• Representation
  • As a network
  • Statistical representation

• Analysis
  • Graph-based methods
  • Statistical analysis methods
  • Probabilistic inference methods