Harvesting Knowledge from Social Annotations

Kristina Lerman
Information Sciences Institute
University of Southern California
Knowledge production is in the hands of the masses

Lay people help scientists collect and analyze data
Library of Congress mystery pictures

LOC asked amateur historians on Flickr to identify old travel photochroms
Community mapping of Haiti after the earthquake

Data created by OpenStreetMap after the earthquake

Sean Groman, Fortius One
Crowd-sourcing knowledge

Collaborative knowledge creation

• Citizen science projects
• Collaborative mapping
• Wikipedia, etc.

Passive knowledge creation: user-generated content and annotations on the Social Web

• People create content
  • Images, videos, bookmarks, locations, ...
• People annotate content
  • Tags: descriptive labels
  • Geo-tags, ...
Harvesting Knowledge from User-Generated Content

User-generated content and annotations ...

... aggregated from many people ...

photos tagged ‘los angeles’

... to extract social knowledge
Formal vs social knowledge

Ontology: an explicit specification of domain knowledge
- Created by a (small) group of experts
- Precise concepts and relations
- 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 people
- Represents social (implicit) knowledge
- Relatively inexpensive to obtain
- Can adapt to evolving vocabularies and community’s information needs
- May conflict with formal knowledge
Outline

• Emergent knowledge in social annotations
  • Folksonomy: Taxonomic knowledge of a community

• Learning folksonomies from structured metadata
  • Novel data source for taxonomic knowledge
  • Relational Affinity Propagation
    • New probabilistic algorithm for clustering structured data
    • Validation on Flickr data

• Leveraging diversity
  • Identifying experts
  • Integrating expert knowledge in folksonomy learning

• Geospatial knowledge
  • Learning about places and relations between them
Emergent knowledge in social annotation

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 emerges quickly [Golder & Huberman, 2006]
- Consensus represents emergent semantics
  - Tags ~ Concepts
Latent Hierarchical Structure

Can we exploit tag co-occurrence to learn taxonomies?

Probabilistic subsumption

A is broader than B if:

\[ P(A|B) = \frac{Freq(A,B)}{Freq(B)} \geq \text{threshold} \]
and

\[ P(B|A) = \frac{Freq(A,B)}{Freq(A)} < \text{threshold} \]

... e.g., bird → toucan

Latent Hierarchical Structure

Can we exploit tag co-occurrence structure to learn taxonomies?

Tag co-occurrence graph

Tag co-occurrence graph of Del.icio.us (circa 2005)

Some problems with tag co-occurrence

Popularity vs generality

Tag co-occurrence methods may erroneously conclude car is more general than automobile
Structured annotation

Novel source of evidence for broader/narrower relations

Users create structured metadata to organize content

- Personal hierarchies or directories
- Tag bundles
- Relations between tags
Structured Metadata on Flickr: Sets and Collections

Flickr users organize photos hierarchically

- Users group related photos in sets (e.g., “Caterpillars”)
- Users group related sets in collections (e.g., “Plant Pests”)

Users may also group related collections within other collections, ...
Personal taxonomies as saplings

**Personal hierarchy/taxonomy**

- **Plant Pests**
  - **Plant Parasites**: 10 photos
  - **Sap Suckers**: 114 photos
  - **Plant Eaters**: 59 photos
  - **Caterpillars**: 59 photos

**Tags on each photo**

**Assume:**
- Set aggregates tags of all constituent photos
- Collection aggregates all tags of all sets in the collection
Saplings created by maxi_millipede
Social knowledge and structured annotations

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

Folksonomy (Folk + Taxonomy) or ‘folk knowledge’ in hierarchical form

Folksonomy Learning

Personal taxonomies/saplings
Challenges in Folksonomy Learning

Sparseness

Ambiguity

Conflict

Varying Granularity
Folksonomy learning as clustering

Goal: cluster saplings into a common hierarchy (folksonomy)

Definitions:
1. Node = instance
2. Concept = a group of “similar” nodes
3. Relation = a link from an individual node to another node
Similarity function for folksonomy learning

Two nodes are similar if

- Their features are similar (similar names, many common tags, etc.) \(\rightarrow\) \textit{local similarity}
- Their neighbors are similar \(\rightarrow\) \textit{structural similarity}
  - Simply use neighbor’s features, rather than their class labels*

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

\(\alpha\) is a weight on how much we rely on structural information

*see Bhattacharya & Getoor, 2007, Collective Entity Resolution in Relational Data, TKDD for more detail
Illustration: merging ‘victoria’s’

Two nodes are merged if they are similar enough

- Similarity is based on local (contextual) & structural (relational) information

Root nodes will not be merged, since their tags and children are not similar
→ Two ‘victoria’s refer to different concepts
**Clustering structured data**

**Definitions:**
1. Node = instance
2. Concept = a group of “similar” nodes
3. Relation = a link from an individual node to another node
Clustering structured data

Constrain the clustering process to ensure resulting structure is a hierarchy

**Definitions:**
1. Node = instance
2. Concept = a group of “similar” nodes
3. Relation = a link from an individual node to another node

**Constraint 0:**
Similar nodes should be merged together into a cluster or “concept”

**Constraint 1:**
No incoming links to the root concept

**Constraint 2:**
All incoming links must be from the same cluster → a concept cannot have multiple parents.
Clustering as Distributed Inference: Affinity Propagation

Affinity Propagation (AP) [Frey & Dueck, Science, 2007]

- Clustering algorithm that identifies a set of exemplars that are representative of all data points
- Exemplar set that maximizes the overall sum of similarity between all exemplars and their data points gives the clusters

AP provides a framework to use structural information during clustering
Binary AP [Givoni & Frey, 2009]

Mapping data points to Binary AP’s hidden variables

Data points

Hidden binary variable matrix

\( c_{ij} \) Hidden binary variable \( c_{ij} \) represents node \( i \) choosing \( j \) as its exemplar

\[
\begin{array}{cccccc}
  & n & i & k & j & h \\
 n & 0 & 0 & 0 & 0 & 0 \\
 i & 0 & 0 & 0 & 1 & 0 \\
 k & 0 & 0 & 0 & 1 & 0 \\
 j & 0 & 0 & 0 & 1 & 0 \\
 h & 0 & 0 & 0 & 0 & 0 \\
 m & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

\( \circ = 0 \)

\( \bullet = 1 \)
Constrained inference in AP

Similarity:

- Similarity between node $i$ and its potential exemplar $j$

Cluster consistency constraints:

1. A node must belong to only one exemplar
   - $l_i$

2. If a node has been chosen by some other nodes as their exemplar, then the node must be its own exemplar.
   - $E_j$

Goal: find assignments of $c$ which maximizes the overall similarity while not violating cluster consistency constraints

$$S(c_{11},...,c_{NN}) = \sum_{i,j} S_{ij}(c_{ij}) + \sum_{i} I_i(c_{i1},...,c_{iN}) + \sum_{j} E_j(c_{1j},...,c_{1N})$$
Extending AP to structured data

Clustering structured data

- Each data point also carries parent-child relations
- Need to take structural information into account to find good integration

→ similar structures should be combined into a tree

*Structured* data points

![Diagram of a tree structure with nodes h, m, n, i, j, k, ...]
Relational Affinity Propagation (RAP)

Introduce new “single parent constraint” (F-constraint) to avoid bad merges

*Structured* data points

Hidden binary variable matrix

Plangprasopchok, et al., *A Probabilistic Approach for Learning Folksonomies from Structured Data*. In WSDM11
F-constraint

“Single parent” constraint: Nodes cannot be assigned to the same exemplar if their parents belong to different exemplars

\[
F_j(...)= \begin{cases} 
-\infty & \exists i,k : c_{ij} = 1; c_{kj} = 1; \text{exlr}(pa(i)) \neq \text{exlr}(pa(k)) \\
0 & \text{otherwise}
\end{cases}
\]

Goal: find assignments of \( c \) which maximizes the overall similarity \( S \) while not violating cluster consistency and structure constraints

\[
S(c_{11},...,c_{NN}) = \sum_{i,j} S_{ij}(c_{ij}) + \sum_{i} I_i(c_{i1},...,c_{iN}) + \sum_{j} E_j(c_{1j},...,c_{1N}) + \sum_{j} F_j(c_{1j},...,c_{1N})
\]
Learning folksonomies from saplings

• Data set
  • Saplings from 7K Flickr users
  • Available on my web site
    http://www.isi.edu/~lerman/downloads/flickr/flickr_taxonomies.html

• Use RAP to learn folksonomies for 32 seed terms
  • africa, amphibian, animal, asia, australia, bird, building, canada, central america, cat, city, country, craft, dog, europe, fauna, fish, flora, flower, invertebrate, insect, north america, plant, reptile, south africa, south america, sport, united kingdom, united states, urban, vertebrate, world
Illustration: ‘africa’ folksonomy

Some of the saplings
Evaluation

Methodologies

• Comparison to a reference hierarchy (ODP)
  • Taxonomic Overlap (mTO) [WWW09]: correctness of node ordering in the learn folksonomy relative to the reference hierarchy
  • #Opaths: number of overlapping paths
  • Lexical Recall (LR): how well an approach can discover concepts, existing in the reference hierarchy (coverage)

• Structural evaluation
  • Area Under Tree (AUT): heuristic for combining bushiness and depth of the tree into a single number

• Baseline Approach: SAP [KDD10]
  • Uses relational clustering (local+structural similarity) to incrementally grow a tree

Plangprasopchok & Lerman, *Constructing folksonomies from user-specified relations on flickr*, WWW09
Plangprasopchok et al., *Growing a Tree in the Forest: Constructing Folksonomies by Integrating Structured Metadata*, KDD10
Results

Number of cases (out of 32) where one approach outperforms the other

Why AUT & LR are worse in RAP: SAP grows one tree at a time → no competition between different trees as in RAP. So SAP tree is larger, with more coverage (and mistakes) than RAP.
How much does structural information help?

Additional Metrics:

- **#Conflicts**: the number of structural conflicts (the number of nodes whose parents belong to different clusters) – gauge how consistent the output structure is.
- **NetSim**: the overall similarity of all points to their exemplars.
Results

Count how many cases (out of 32) one approach outperforms the other

→ RAP substantially reduces the number of conflicts, producing better structure
→ Less stringent merging criteria allow AP to merge more nodes, resulting in higher NetSim
Leveraging user diversity

All users are not created equal

• Expert and novice users vary in
  • Degree of expertise and knowledge
  • Degree of expressiveness
  • Degree of enthusiasm for the subject
  → higher quality annotations from expert users

• Integrate knowledge from expert and novice users
  • Automatically identify expert users
  • Use experts’ knowledge to guide folksonomy learning
    • Better folksonomies?
    • Comprehensive?
Novice users

Most users create shallow saplings
Expert users

Few users create deep, detailed saplings
Identifying experts

Structure and semantic consistency are most important factors

- Generally creates a variety of saplings
- Creates deep or broad saplings
  - 3-5 levels deep
  - e.g., ‘bird’ sapling contains 20 bird species
- Provides generalizable knowledge
  - Specifies categories that are meaningful to others
  - Overly broad top-level categories imply non-specialist, e.g., `life', `things', `misc', `all sets'.
- Consistent granularity level
  - NOT world → los angeles
  - NOT vacation → china & vacation → disneyland
- Does not create conflicts
  - NOT journey → los angeles & los angeles → journey
- Does not create multiple child concepts with same name in the same sapling
Features

User-based features
• Number of saplings N
• Number of relations T
• Balance – uniformity of sapling size (measured by entropy)
• Disparity – similarity of concepts expressed by saplings (measured by JS divergence of tag distributions of saplings)

Sapling-based features
• Complexity – depth and breadth of sapling
• Balance – how balanced sapling is at each level (measured by entropy of the distribution of child nodes)
• Consistency – number of conflicts, number of users who expressed the same relation, ...
• Generalizability – how general or specific sapling’s root is
Generalizability

More general

- life
  - friends
  - kids
  - animals
  - travel

Other roots: stuff, out and about, random, home, misc, thing, life, ...

More specific

- europe
  - france
  - austria

Other roots: nature, animal, europe, travel, flower, bird, holiday, usa, ...

prob of occurrence vs. distinct child nodes

30%
**Automatically identifying experts**

<table>
<thead>
<tr>
<th>iteration</th>
<th>Training set cross validation</th>
<th>Training set size</th>
<th>Num positives (experts)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pr</td>
<td>Re</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>0.76</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.88</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>0.80</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>6</td>
<td>0.84</td>
<td>0.50</td>
<td>0.63</td>
</tr>
<tr>
<td>7</td>
<td>0.88</td>
<td>0.66</td>
<td>0.75</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>0.88</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Start w/200 labeled users

Self-train on unlabeled users with libSVM

→ Final model identifies 60 experts among 7K users
Integrating expert and folk knowledge

Self-similarity, or preference, in RAP: $S(c_{jj})$

- Likelihood node $j$ will become an exemplar
- Same preference values $\rightarrow$ all nodes equally likely to become exemplars
  - Usually set to *mean* of all similarities

Using expert knowledge in folksonomy learning

- Assign higher preference values to nodes from saplings created by experts
  - More likely to become exemplars
  - Expert knowledge guides folksonomy learning with RAP
Illustration: ‘africa’ folksonomy
Illustration: ‘africa’ folksonomy

- africa
  - south africa
  - christmas
    - south africa
    - cape town
      - south africa
      - cape town
      - muizenberg
      - kalk bay
      - table mountain
      - cape town
      - fish hoek
      - city bowl
  - cape town
  - xmas
    - family
    - card
  - africa
  - krusger national park
  - cape town

- africa
  - kenya
  - south africa
  - christmas
    - cape town
  - holiday
    - christmas
  - bloemfontein
  - cape town

... Saplings from experts
Evaluation

• Compare folksonomies learned with and without expert knowledge
  • Model M1 (no expert knowledge)
    • all preferences set to the same value
  • Model M2 (using expert knowledge)
    • preference values of expert nodes twice non-expert preference values

• Automatic evaluation
  • Depth
  • Lexical precision (LP)
    • with respect to reference taxonomy (ODP)
  • Taxonomic Overlap (TO)
    • with respected to reference taxonomy
Results: Automatic Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Model M1</th>
<th></th>
<th>Model M2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>LP</td>
<td>TO</td>
<td>Depth</td>
<td>LP</td>
</tr>
<tr>
<td>Ave. over</td>
<td>3.61</td>
<td>0.13</td>
<td>0.47</td>
<td>4.58</td>
<td>0.19</td>
</tr>
<tr>
<td>32 seeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kang & Lerman, *Integrating Specialist and Folk Knowledge with Affinity Propagation*, NIPS workshop 2010
Without experts
With experts
Without experts
Without experts
Without experts
With experts
Manual evaluation: reduced trees

Five annotators judged reduced trees (common nodes taken out)

M1

45.30% of trees correct

M2

68.24% of trees correct
Summary

• RAP: Novel extension to affinity propagation algorithm
  • Integrates many small structures by using structural constraints
  • Merging and removing inconsistencies is jointly determined
  • RAP produces folksonomies more consistent with the reference taxonomy than those produced by relational clustering

• RAP allows expert knowledge to be integrated into folksonomy learning
  • Folksonomies learned with expert knowledge are more detailed and accurate
  • But, need novice knowledge to learn comprehensive folksonomies
Related Work

Social annotation characteristics

• Tags are of multiple aspects [Mathes04, Rashmi05]
• Tag occurrences are not at random [Golder+06, Marlow+06, Heymann+08]
• Relationships between user groups and tags [Mika05, Marlow+06]
• Different levels of specificity [Golder+06]
• Synonymy/Polysemy, Acronym [Mathes04, Golder+06]

Folksonomy learning

• Inducing concept hierarchies
  – Syntax based [e.g., Hearst92, Caraballo99, Pasca04, Cimiano+05, Snow+06, Yang&Callan09]
  – Hierarchical clustering [e.g., Segal+02, Blei+03]
  – Induce hierarchy from tags [Mika05, Brooks+06, Heymann+06, Schmitz06]

• Ontology Alignment [e.g., Udrea07+, Euzenat&Shvaiko07]

• Structure Learning
  – Probabilistic Network [e.g., Heckerman95, Kok&Domingos10]
Thanks

• Collaborators
  • Anon Plangprasopchok – now at the National Electronics and Computer Technology Center, Thailand
  • Jeon-Hyung Kang (USC)
  • Suradej Intagorn (USC)
  • Lise Getoor (UMD)

• Support

  National Science Foundation
  • CMMI-0753124
  • IIS-0812677