Social Tagging and Folksonomies

Kristina Lerman
University of Southern California

This lecture is partly based on slides prepared by Anon Plangprasopchok
Harvesting Knowledge from Social Annotations

Annotation / Metadata

Users

Web content

Categorical Knowledge

Organize, Search, Recommend, Leverage, Categorize

USC Information Sciences Institute
Formal vs Informal Taxonomies

- **Semantic Web**
  - Web content annotated with machine-readable metadata from a formal taxonomy, *ontology*, to aid information integration
  - Challenges of formal taxonomies
    - Very complicated - specialized training to be used effectively
    - Costly and time-consuming to produce
    - Agreement between data providers on a common ontology

- **Folksonomy**
  - “*user generated taxonomy* used to *categorize* and *retrieve* *web content* using open-ended labels called *tags*.” [source: Wikipedia]
    - Bottom-up: decentralized, emergent, scalable
    - Dynamic: adapts to changing needs and priorities
    - But, social annotations data is noisy: we need tools to extract meaning from his data
Rainbow bee-eater

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 on Flickr

submitter

private albums

public groups

tags
User-created Hierarchies

~Aquila~

Natural Things

Places

Shapes, Textures and Patterns
3 sets

Natural Things
8 sets

Places
4 sets

People and Pets
4 sets
User-created Hierarchies

~Aquila~

Natural Things

Places

Flowers

Birds

Reptiles

Rainbow bee-eater

Collection: Natural Things

Natural Things

Sunrise & Sunset
29 photos

Birds
37 photos

Reptiles & Amphibians
1 photo

Rainbow Bee-eater

Flowers
17 photos

Up Close and Personal
6 photos

Oz Fungi
2 photos

http://www.flickr.com/photos/cjhudson/cc
Overview

• Learning folksonomies, etc. from social annotation

• *Harvesting Social Knowledge from Folksonomies*
  • Links between users, tags and documents in collaborative tagging systems can be represented
    • Network of linked nodes
    • Adjacency matrix representation of the network
  • Analysis of the adjacency matrix can
    • Identify communities
    • Identify experts and recommend authoritative documents
    • Generate ontologies/ folksonomies

• *Constructing Folksonomies from User-specified Relations on Flickr*
  • Learning folksonomies from tags is problematic
  • Instead, learn folksonomies from user-specified hierarchical relations
Harvesting Social Knowledge from Folksonomies  Wu, Zubair, & Maly
3 Challenges of Collaborative Tagging

- **Community identification**
  - Users with common interests
  - Documents with common topics

- **Recommendation**
  - Identify high quality users (experts) and documents (authorities)

- **Ontology generation**
  - Create an efficient structure for navigation
  - Assist search and document retrieval

- **Proposed solution**
  - **Representation**
    - Network representation of collaborative tagging systems
  - **Analysis algorithms**
    - Matrix-based analysis of adjacency matrix representation
Network Representation

- Represent collaborative tagging system as a network
  - Entities
    - Users, documents, tags
  - Links

Users annotate documents
Documents are associated w/tags
Users create tags
Tags
Matrix Representation of a Network

- User-document adjacency matrix $D$ represents connections between users and documents
  - $D_{ij}=1$ if user $i$ annotated document $j$; otherwise $D_{ij}=0$
- User-tag adjacency matrix $T$ represents connections between users and tags

\[
D = \begin{bmatrix}
1 & 0 & 0 & 1 & \ldots \\
0 & 0 & 0 & 1 & \ldots \\
1 & 1 & 0 & 0 & \ldots \\
0 & 1 & 0 & 0 & \ldots \\
0 & 0 & 1 & 0 & \ldots \\
\vdots 
\end{bmatrix}
\]

\[
T = \begin{bmatrix}
1 & 0 & 0 & 0 & \ldots \\
0 & 1 & 1 & 1 & \ldots \\
1 & 1 & 0 & 1 & \ldots \\
1 & 1 & 0 & 0 & \ldots \\
0 & 0 & 1 & 1 & \ldots \\
\vdots 
\end{bmatrix}
\]
Analysis of Network-based Representation

- To find communities, decompose matrices to find their singular vectors

  (same technique employed by Latent Semantic Indexing (LSI)!)  

  - Singular vectors of $D$ correspond to
    - Communities of users
    - Key documents of interest to these communities
  
  - Singular vectors of $T$ correspond to
    - Communities of users
    - Most important tags in their vocabulary
Recommendation

- Identify experts and high quality documents
- HITS (Kleinberg 1999)
  - Iterative algorithm to find high quality sources in networks
    - Hubs – sources with many incoming links
    - Authorities – sources with many outgoing links
- Modify HITS to find
  - High-quality documents (authorities)
  - Experts (hubs) related to a given tag
  - Works on user-document network (matrix)
Ontology Generation

- Work on matrix representation of document-tag network
- Ontology generation as hierarchical clustering
  - Merge two similar documents into a group
  - Merge similar groups into another group
Discussion

- Framework for harvesting social knowledge from links between
  - Tags and documents
  - Users and documents

- Represents associations between users, tags and documents as an adjacency matrix
  - Algorithms manipulate matrices

- Approach addresses synonymy and polysemy challenges
  - Synonymous tags tend to include the same documents
  - Document associations with tags disambiguate polysemous tags
Construcing Folksonomies from User-specified Relations on Flickr

Plangprasopchok and Lerman
Learning Concept Hierarchies
Hierarchical Relations in Social Web

- Appear Implicitly

- Appear Explicitly

Tags:
Insect
Grasshopper
Australian
Macro Orthoptera

Goal: to induce deeper hierarchies from this metadata
Learning Hierarchy from Tags

- Subsumption approach applied to tag cooccurrence [Schmitz, 2006]
  - Tag x subsumes y, if $P(x|y) \geq t$ and $P(y|x) < t$

![Diagram showing a subsumption relationship between tags x and y.](image)
Learning Hierarchy from Tags

Some problems:

- Washington $\rightarrow$ United States
- Car $\rightarrow$ Automobile
- Insect $\rightarrow$ Hongkong
- Color $\rightarrow$ Brazilian

Notation: $A \rightarrow B$
(A is broader than B)

Generality vs Popularity

Mixing tags from different facets

Above relations induced using tag-based subsumption on Flickr data
User-specified hierarchical relations?

- User specified relations, e.g.,
  - Flickr’s Collection-Set,
  - Delicious’ Bundle-Tag,
  - Bibsonomy’s Relation-Tag

- Key intuition: Not so many people specify odd relations like
  - “car” → “automobile”, or
  - “Washington” → “United States”
Simple Strategy

Collection

Set

The Netherlands - Holanda

Blijdorp - Rotterdam

Concept relations

netherland

holanda

blijdorp

netherland

blijdorp

holanda

rotterdam

countri

holland

china

blijdorp

Assume that same terms refer to the same concept

Remove “noise” Link concepts

Tokenize + Stem
Remove noisy relations

- Conflict Resolution (when both $a \rightarrow b$ and $b \rightarrow a$ appear)
  - Relation conflicts occur because of noise
  - Voting scheme:

Keep $a \rightarrow b$ (and discard $b \rightarrow a$)

If $N_u(a \rightarrow b) > 1$ and $N_u(a \rightarrow b) > N_u(b \rightarrow a)$
Path Selection

- Link relations from many users can cause a *spaghetti* graph

4 possible paths from anim $\rightarrow$ moth:
1) $a \rightarrow b \rightarrow i \rightarrow m$
2) $a \rightarrow i \rightarrow m$
3) $a \rightarrow m$
4) $a \rightarrow b \rightarrow m$

Network Bottleneck idea:

"the flow bottleneck is a minimum flow capacity among all relations in the path"

1) $a \rightarrow b \rightarrow i \rightarrow m$ [BN score = $\min(26,1,18) = 1$]
2) $a \rightarrow i \rightarrow m$ [BN score = $\min(72,18) = 18$]
3) $a \rightarrow m$ [BN score = $\min(10) = 10$]
4) $a \rightarrow b \rightarrow m$ [BN score = $\min(26,4) = 4$]
Evaluation & Data Set

- Hypothesis: *the approach that takes explicit relations into account can induce better hierarchies.*
  - more consistent with hand-built hierarchies (ODP)
- Compare to *subsumption* approach
  - Collection and set terms are used instead of tags, making it comparable.

**Data Set:**
- Data from 17 user groups, devoted to wildlife and naturalist photography
- 21,792 of 39,922 users specify at least one collection
- 110,543 unique terms (c.f. 166,153 unique terms in ODP), 15,495 terms in common.
Metrics

- **Taxonomic Overlap**
  - measuring *structure similarity* between two trees
  - for each node, determining how many ancestor and descendant nodes overlap to those in the reference tree. (detail in the proposal)

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

- Manually selected 32 trees

- Taxonomic Overlap:
  - 27 of them are better than those by subsumption
  - 3 of them get zero score in both approaches

- Lexical Recall:
  - 28 of them are better than those by subsumption
  - 2 of them get similar score on both approaches
  - the rest, by subsumption, only induce the root node.

- *Proposed approach induces deeper trees than subsumption*
Sport hierarchy

- sunset
- basketball
- soccer
- golf
- summer
- travel
- nate
- diver sport
- volleyball
- foot
- vtt
- tour
- hocke
- handball
- tennis
- rugby
- other sport
- American football
- motocross
- BMX
- water sport
- roger cup
- lacrosse
- softball
- kite surf
- gymnastics
- football
- motorsport
- flag football
- bike
- ride
- kayak
- snowboard
Discussion

• 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, rather than flat tags

• Representation
  • As a network
  • Hidden topic model

• Analysis
  • Matrix manipulation of adjacency matrix representation of networks
  • Probabilistic inference methods