13. Social / Distributed Categorization

INFO 202 - 10 October 2007

Bob Glushko
Plan for Today's Lecture

"Social/distributed categorization" -- not "classification"

del.icio.us and flickr

The Tradeoffs for "Metadata-makers"

Themes and memes about social / distributed information organization and categorization
Varieties of Categorization Systems

Cultural Categorization Systems

Individual Categorization ("Tagging")

Institutional Categorization ("Business Semantics")
Individual Categorization Systems

A system developed by an individual for organizing a personal domain to aid memory, retrieval, or usage

Can serve social goals to convey information, develop a community, manage reputation

Have exploded with the advent of cyberspace, especially in applications based on "tagging"
Why This is "Social"
Why It's Not "Classification"

"Social' or "Cooperative" "Classification" are more common than S or C "Categorization" -- that's wrong!

- CATEGORIZATION (from September 10) - Categories are equivalence classes - sets of material and abstract things, processes, and events that we treat the same

- CLASSIFICATION (from September 19) - A Classification (noun) is a system of categories, ordered according to a PRE-DETERMINED SET OF PRINCIPLES and used to organize a set of instances or entities; Classification (verb) is the process of assigning instances or entities to the categories in a classification system

Most "end user tagging" systems don't impose any pre-existing system of categories -- indeed, that's the point!
Tagging Taxonomy

- Tag User
  - Self
    - Technorati HTML Meta Tags
    - Flickr
  - Others
    - CiteULike
    - Connotea
    - del.icio.us
    - Frassle
    - Furl
    - Simpy
    - Spurl
    - unalog

- Others
  - (Wikipedia)
del.icio.us -- Shared Bookmarks
del.icio.us -- User Interface for Tagging
flickr - Photo Collections
flickr - Search for Tag "Glushko"
flickr -- Uploading Photos
flickr -- Uploading with "Batch Tagging"
Tagging Is...

Creative and dynamic
Unconstrained, open-ended
Interpretive
Statistical
Types of Tags

Subject /Taxonomic or Keyword Tags (most common, but rarely from a controlled vocabulary)

Property or Attribute Tags ("red," "expensive")

"Purpose" Tags (e.g, "toread" or "buythis" or "tagthis")

Evaluative Tags ("interesting," "good")
Why Tag?

To organize for your own future use
  - Content-based organization
  - Task-based organization

To enable sharing and communication to known audiences

To express opinions or to entertain
Tagging Functionality / User Interfaces

Context is recorded automatically (tagger, time, date, resource name)

Share/Don't Share (or Private/Public): enable both personal organization and group organization (default is{"public"})

Tag suggestion (tagging precedents) -- might be before or after your own tags are applied

Tag organization into groups or categories

Batch uploading and tagging

Tag Visualization ("tag clouds")
del.icio.us "Tag Cloud" for all Tags
del.icio.us "Tag Cloud" for "Doc Or Die" Blog
Tag Quality / Correctness?

The del.icio.us instructions say:

*Tagging is intuitive*  
*A tag can be anything you want*  
*There are no wrong tags*
Tag Me "Stanford" and "BarryBonds"
"Tag Soup"

Users are free to assign any number of labels or tags they choose

No vocabulary control
Responses to Tag Soup

Some people consider the unstructured, uncontrolled nature of "tag soup" to be its great strength, just as faceted classification overcomes some of the limitations of strict hierarchies.

Others adopt personal conventions to encode hierarchical and derivational relationships (e.g. using CamelCase; basic and specific level categories).

Using multiple accounts for the same application is another approach for organizing tags and the resources they describe.

Some systems are introducing "tag bundles" to enable more hierarchy; it might also be possible to infer the hierarchy using dictionaries or thesauri.
Geotagging and Taxonomic Tagging

Most tags don't come from controlled vocabularies, but geotagging and biological tagging are the exceptions that prove the rule.

Map interfaces in flickr and google earth can be used for geotagging but any GPS will do - by convention 3 tags are used:

- geotagged
- geo:lat=latitude e.g. geo:lat=51.4989
- geo:lon=longitude e.g. geo:lon=-0.1786
Combined Geo and Bio Tagging
Tag Convergence?

Some systems (like del.icio.us) don't allow users to see the tags assigned by other users when they are tagging a resource.

But once a user tags a resource, most systems reveal the tags applied by other users.

If your tag(s) don't match what others are using, do you?

- Change your tag to adapt to the group norm (maybe you'd look at the other resources with that tag to compare "senses")
- Keep your tag to influence the group norm
- Add the group tag but keep yours as well
Semiotic Dynamics, or Tagging Over Time

http://slashdot.org

http://engadget.com

time

rank
The Long Tail
Golder and Huberman Study

"The Structure of Collaborative Tagging Systems" studies tagging patterns for individuals and the most popular resources tagged on del.icio.us.

They observe "tension between tags that may be useful to the Delicious community at large and those useful only to oneself"

The diversity of tags for many resources and tags whose meaning is intrinsic to the tagger demonstrates that a significant amount of tagging, if not all, is done for personal use rather than public benefit.

Nonetheless...
Divergence, Stabilization, or Convergence?

Will individuals’ varying tag collections and personal preferences, compounded by an ever-increasing number of users, yield a chaotic pattern of tags?

Or will the combined tags of many users converge?

Or will a stable pattern emerge in which the proportions of each tag are nearly fixed?
Tag Stabilization in Golder and Huberman Study
Fringe [1]
Fringe [2]
Who Are the Metadata-Makers or Taggers?

Professionals (the emphasis of traditional library science); we might also add "publishers," "literary critics," and "program committees" for scholarly publications

Authors

Users

Machines (either via computational or contextual processes)
Authors {and,or} Users

We need to distinguish authors from users because only authors can be assumed to know some of the metadata about the object and the intent.

In del.icio.us, the taggers are users who are categorizing web sites by making bookmark lists.

In flickr, the taggers are most often categorizing their own photos.
End User Tagging

ADVANTAGES:

- There are lots of them
- They understand their intent

DISADVANTAGES:

- No training - no consistency or standardization
- They have diverse intent
Tagging by Professionals

ADVANTAGES:

- Consistent quality
- Conformance to standards

DISADVANTAGES:

- Expensive; steep learning curve to acquire expertise
- They can't tag very much
- They make assumptions about user intent that may not be correct
Professional Tags {and, or, vs} User Tags

Contrasting "professionals" and "users" this way assumes that they are different people.

But we can also view them as two "roles" or choices a single person can make about how much effort to put into categorizing information.

What are the incentives or tradeoffs that influence your decision?
Can We Have The Best of Both Worlds?

A number of systems combine "authoritative" or "professional" metadata with user-generated metadata - examples are CiteULike and Connotea.

The former is reliable for retrieving the specific item, while the latter might be more useful in exploratory browsing to find related information.

These systems appear to be targeted to academic researchers, whose "authority" as user-taggers might lessen the usual concerns and enable the discovery of "invisible colleges" of researchers working in related fields.
The Underlying Philosophies / Assumptions -- Why Social / Distributed Efforts (Supposedly) Work

"Architectures of Participation"

"Given Enough Eyeballs, All Bugs Are Shallow"

"Harnessing Collective Intelligence"
"Web 2.0"

**Web 2.0 Meme Map**

- **Strategic Positioning:**
  - The Web as Platform

- **User Positioning:**
  - You control your own data

- **Core Competencies:**
  - Services, not packaged software
  - Architecture of Participation
  - Cost-effective scalability
  - Removable data source and data transformations
  - Software above the level of a single device
  - Harnessing collective intelligence

**Key Concepts:**

- Flickr, del.icio.us: Tagging, not taxonomy
- PageRank, eBay reputation, Amazon reviews: user as contributor
- Blogs: Participation, Not publishing
- BitTorrent: Radical Decentralization
- Gmail, Google Maps and AJAX: Rich User Experiences
- Google AdSense: customer self-service enabling the long tail
- Wikipedia: Radical Trust

**Additional Points:**

- "An attitude, not a technology"
- The Long Tail
- Data as the "Intel Inside"
- The perpetual beta
- Hackability
- The Right to Remix: "Some rights reserved"
- Granular Addressability of content
- Trust your users
- Play
- Rich User Experience
- Small Pieces Loosely Joined (web as components)
- Emergent: User behavior not predetermined
Readings for INFO Lecture #14


