

10. Social / Distributed Classification

IS 202 - 28 September 2006

Bob Glushko

Plan for Today's Lecture

Themes and memes about social information organization

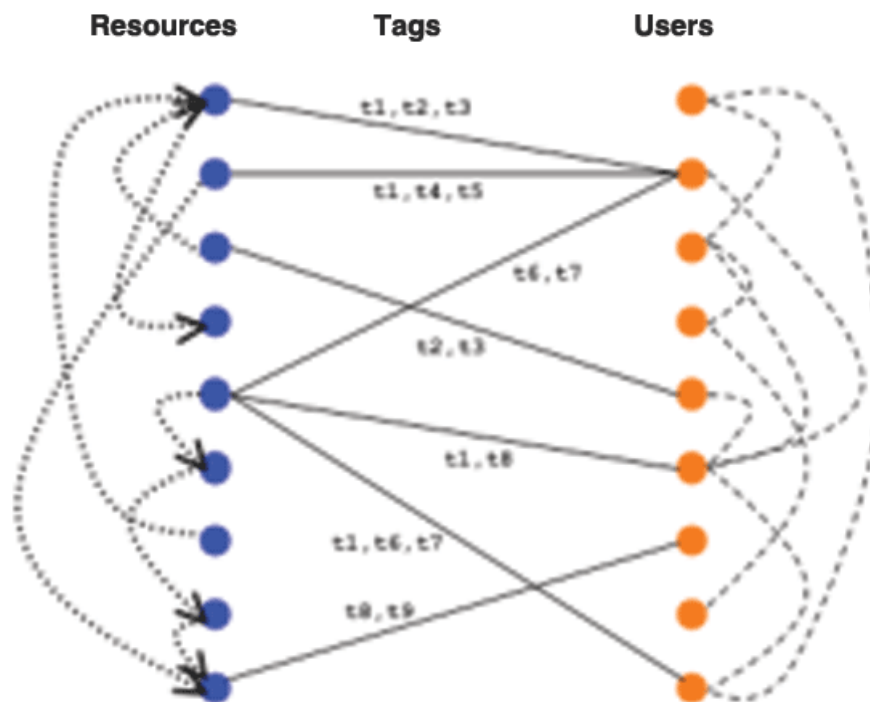
A brief history of "social software" and "social network" software

Social Classification

Collaborative Filtering / Recommendation Systems

The Tradeoffs for "Metadata-makers"

The Underlying Model



"Anyone Can Be A Publisher"

A few years ago it was popular to say that "the Web enables anyone to be a publisher"

If "Publish on the Web" means putting HTML on a Web server then this statement isn't very meaningful

If "Publish on the Web" means carrying out the functions of a publisher using Web- based technology as a distribution channel then the statement is false

But there is definitely a shift underway from "host-provided content" to "user-provided content"

"Architectures of Participation"

Successful "grassroots" communities embody an "architecture of participation" with a self-regulating free market of ideas

Anyone can propose a solution to a problem; it becomes adopted, if at all, by acclamation and the organic spread of its usefulness.

It is not only possible but encouraged to create a custom solution to a problem and share your results.

It's important to have an attitude that doesn't treat "things that don't come from the center" as second class citizens.

"Given Enough Eyeballs, All Bugs Are Shallow"

Eric Raymond's essay "[The Cathedral and the Bazaar](#)" is a manifesto of the Open Source "movement"

It explains why open source software development can produce high-quality software -- the diversity of contexts and experiences that people bring to testing software makes it easy to find bugs that a sole programmer might never find

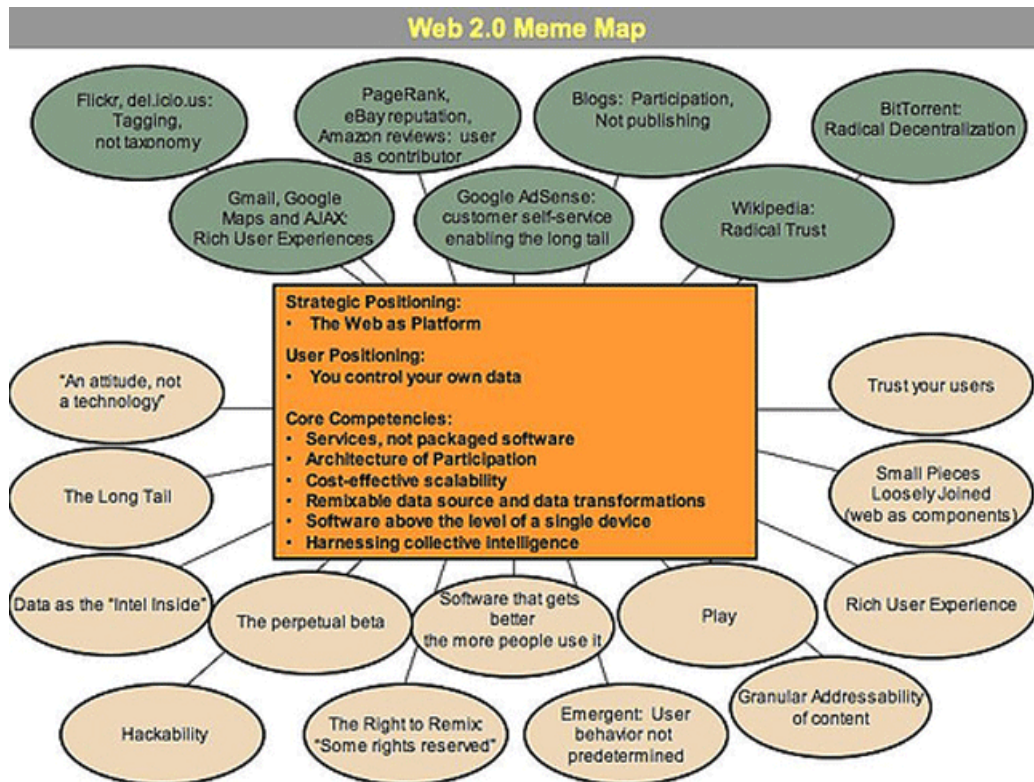
"Harnessing Collective Intelligence"

This idea that harnessing collective intelligence can yield knowledge that can't be created by individuals is also a fundamental theme in today's lecture

This collective intelligence doesn't have to be collected in an explicit way - it can be collected incidentally or implicitly as a by-product of what people do in their course of using and creating information

- In Google, collective link structure increases precision
- In eBay, collective actions of buyers and sellers set prices
- In flickr, de.licio.us, technorati or other sites appropriate descriptions of content or information sources can emerge through the aggregation of "tags" applied by individuals
- In Myspace.com and similar sites, collective preferences for music, products, events, etc. creates categories of "winners" and "losers" in popular culture

"Web 2.0"



Social Software (A Partial History)

[1]

1970s

- PLATO educational timesharing
- Email with CCs and distribution lists
- 1978 - [first Bulletin Board System](#) uses 300- baud dial-up line

1980s

- 100,000 BBS arise
- USENET news groups
- CSCW research groupware

Social Software (A Partial History)

[2]

1990s

- 100,000 BBS disappear without a trace
- Internet Relay Chat (IRC)
- Web chat rooms
- Exportable bookmarks in web browsers

Social Software (A Partial History)

[3]

2000s

- Friendster (May 2002)
- LinkedIn
- MySpace.com (July 2003, acquired by News Corp in Sept 2005)
- del.icio.us (late 2003, acquired by Yahoo! December 2005)
- flickr (Feb 2004, acquired by Yahoo! March 2005)
- *Social Software becomes overnight success!*

Social Network Software

"Software that supports group interaction" (Clay Shirky)

Social network software is software that aids in the formation, maintenance, or securing of one or more social networks.

Social networks can be created when users explicitly make choices to exchange information with others or participate in interactions that create networks; this can be thought of as "classifying people"

Social networks can also be created implicitly through the ordinary information exchange actions that users perform with software applications

Explicit vs Implicit Networks

You can send an email message to anyone without asking them permission, and you can link to or bookmark web pages without getting permission

The "inbound" links or tags applied to web resources can be viewed as votes that raise the value or prominence of the resource

In contrast, most social network applications that involve people rather than web resources require the recipient to acknowledge the request to create a connection

Vocabularies for Social Network Analysis

Wouldn't it be great to be able to compare or merge social networks created by different social network applications?

But this is an information integration problem that requires that we have a common or interoperable data model that describes each network and its members

What kinds of concepts or metadata should this vocabulary contain?

Social Classification

From upcoming workshop: **SOCIAL CLASSIFICATION: PANACEA OR PANDORA?** (Austin TX 4 November 2006):

- ... Any of a number of broadly related processes by which the resources in some collection are categorized by multiple people over an ongoing period
- ... With the potential result that any given resource will come to be represented by a set of labels or descriptors that have been generated by different people.

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 Taxonomy

Tag User	Others	<i>Technorati</i> <i>HTML Meta Tags</i>	<i>(Wikipedia)</i>
	Self	Flickr	CiteULike Connotea del.icio.us Frassle Furl Simpy Spurl unalog
		Self	Others
		Content Creator	

"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 (Examples [1](#) and [2](#))

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

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

Many resources have a "long tail" tag distribution

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?

[Golder and Huberman results](#)

Collaborative Filtering / Recommendation Systems

User preferences or relationships to resources (or people) can be used to facilitate identification of relevant resources (or people)

Finding "good" or "bad" ratings isn't enough and average ratings can be misleading

What you want is to "find good things" and "keep bad things away" -- where "good" and "bad" reflect YOUR preferences

Many applications - but web shopping for products and services is the most obvious one; also spam detection (cloudmark.com)

The technical challenges here are:

- How to state your preferences
- How to determine "similar" preferences

Explicit Preferences

Fill out a form to state preferences

Rate items on some scale to facilitate statistical processing

Are expert preferences worth more than those by ordinary users?

Does the effort required to make explicit ratings create free-rider problems?

Implicit Preferences

Collecting implicit ratings eliminates some of the problems with explicit ones

Actions like "reply," "save," "copy," "bookmark," "link to" etc indicate interest in a message or document












Buying something indicates you like it

Other implicit preference data?

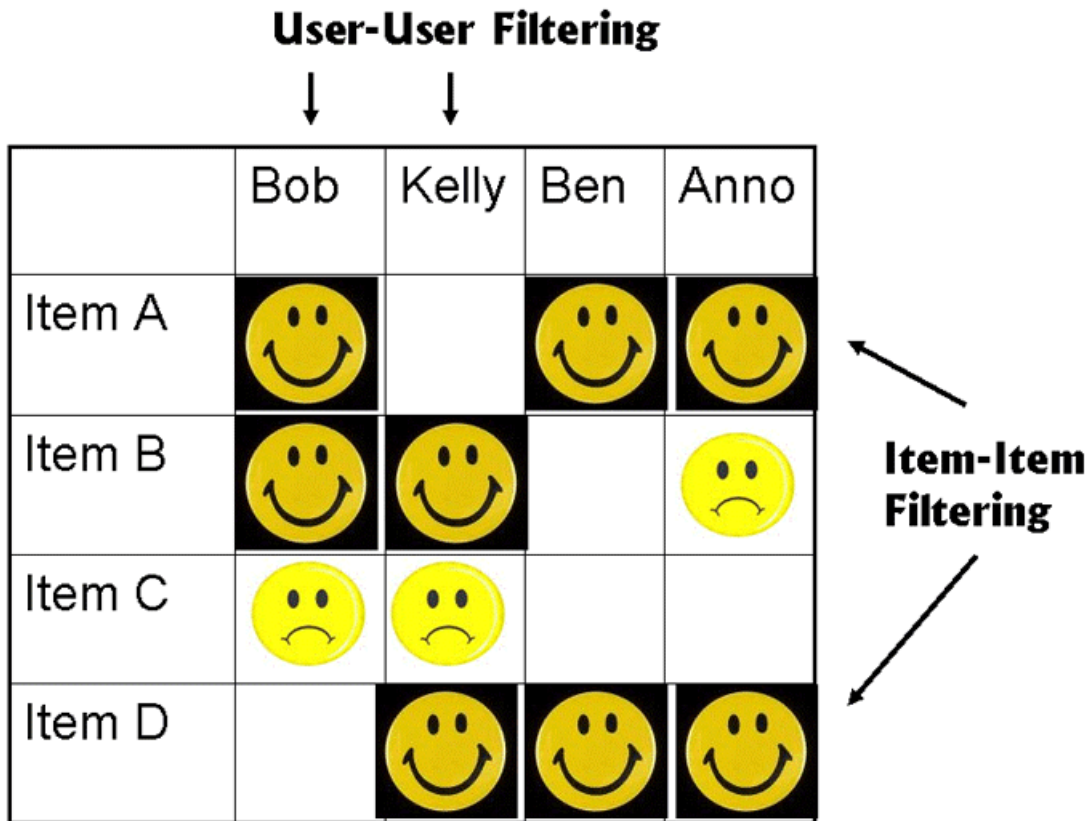
User-User and Item-Item Filtering

User-User Filtering

↓ ↓

	Bob	Kelly	Ben	Anno
Item A				
Item B				
Item C				
Item D				

Item-Item Filtering



User-User Collaborative Filtering

Principle: Find users with similar preferences and listen to their "word of mouth"

Bob and Kelly agree on Item B and C

So Bob's preference for A gets recommended to Kelly,
and Kelly's recommendation for D gets recommended to
Bob

Item-Item Collaborative Filtering

Principle: Find items with similar appeal

Item A and Item D are both preferred by Ben and Anno

So if people who like D also like A, then A can be recommended to Kelly, who likes D

Limitations on Collaborative Filtering

Privacy concerns

Recommendation "spam" and dishonest ratings

Variability and preference change

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

Readings for IO & IR Lecture #19

"Ontology 101 (1-20, through section 4)" Natalya Noy and Deborah McGuinness

"Ontology is Overrated: Categories, Links, and Tags"
Clay Shirky