The Social Web: Social networks, tagging and what you can learn from them

Kristina Lerman
USC Information Sciences Institute
The Social Web

The Social Web is a collection of technologies, practices and services that turn the Web into a platform for users to create and use content in a social environment

- Authoring tools → blogs
- Collaboration tools → wikis, Wikipedia
- Tagging systems → del.icio.us, Flickr, CiteULike
- Social networking → Facebook, MySpace, Essembly
- Collaborative filtering → Digg, Amazon, Yahoo answers
Social Web Features

- **Users create content**
  - Articles, opinions, creative products

- **Users annotate content**
  - Metadata
    - *Tags* – *freely chosen labels*
    - *Geo-tags* – *location information*
  - Ratings

- **Users create connections**
  - Between content and metadata
  - Between content or metadata and users
  - Among users (social networks)

Users traverse these connections, creating new ones along the way

- **Allows users to interact**
  - See and discuss comment
  - Create new links between content, metadata and users
Flickr example

Reflections of a pretty sky
Compliments of a couple of my kids...Hope you enjoy!

Comments

Fpease says:

wonderful sentiment!

have a great day!

Posted 12 hours ago. (previews)

Grveone says:

Gorgeous pink and blue hues in this serenity, so lovely framed - wonderful shot!

Wishing you all the best and a happy day!

Posted 11 hours ago. (previews)

This photo also belongs to:

- Quality — (PLEASE Read the RULES)
- Spectacular Nature — invited images only
- reflections
- sky
- water
- pink
- blue
- trees
- reeds
- nature
- landscape
- seascape
- serene
- earth
- land
- QUALITY
- specnature

Additional information

- (All rights reserved)

- Taken on August 6, 2006
- 22 people call this photo Favorite
- Viewed 283 times
User’s social networks

User’s contacts (friends) and group membership
User’s tags

- Tags are keyword-based metadata added to content
  - Help users organize their own data
  - Facilitate searching and browsing for information
  - Freely chosen by user
User’s favorite images
(by other photographers)
Social Web is enormous and growing rapidly
• Some popular sites have >1 million users and >1 billion objects
• 2G/day of “authored” content
• 10-15G/day of user generated content [From Andrew Tomkins, Yahoo! Research]

Social Web is highly heterogeneous
• Different languages
• Different content types

Social Web is highly dynamic
• New users and content
• Links are created and destroyed
Social Web is interesting

- **Social Web as a complex dynamical system**
  - Collective behavior emerges from actions taken by many users
  - Interesting interactions between users: network-mediated, environment-mediated (e.g., popularity-based)

- **Social Web as a knowledge-generating system**
  - Users express personal knowledge (e.g., through tags)
  - Tailor information to user’s individual preferences ...
  - ... or combine users’ knowledge to create a *folksonomy* of concepts

- **Social Web as a problem-solving system**
  - By exposing human activity, Social Web allows users to harness the power of collective intelligence to solve problems

- **Lots of data for empirical studies**
  - Social Web is amenable to analysis
  - Design systems for optimal performance
I study how user-contributed metadata can be used to solve a variety of information processing problems, including information discovery and personalization.

- Dynamics of information spread on networks
  - Social browsing = social networks + recommendation
    - Patterns of information spread on networks indicative of content quality
  - Mathematical analysis of collaborative decision-making on Digg

- Learning from social tagging
  - Machine learning methods to extract information from tags created by distinct users
  - Better than Google: using Del.icio.us tags to find Web services
Dynamics of information spread on networks

with: Dipsy Kapoor, Aram Galstyan
Social news aggregator Digg

- Users submit stories
- Users vote on (digg) stories
  - Digg selects some stories for the front page based on users votes
- Users create social networks by adding other users as friends
  - Digg provides Friends Interface to track recent activity of friends
    - See stories friends submitted
    - See stories friends dugg
How the Friends interface works

submitter

‘see stories my friends submitted’

‘see stories my friends dugg’

fans of submitter

fans of voters

...
Top users

- **Digg ranks users**
  Based on how many of their stories were promoted to front page
  - *User with most stories is ranked #1, ...*

- **Top 1000 users data**
  - Usage statistics
    - *User rank*
    - *How many stories user submitted, dugg, commented on*
  - Social networks
    - *Friends: outgoing links*
      \[ A \to B := B \text{ is a friend of } A \]
    - *Reverse friends: incoming links*
      \[ A \to B := A \text{ is a fan of } B \]
• **Stories**
  Collected by scraping Digg ... now available through the API
  • Front pages stories: data about ~200 stories
    • *most recently promoted to the front page on June 30, 2006*
  • Newly submitted stories: data about ~900 stories
    • *Most recently submitted on June 30, 2006*
  • For each story extracted
    • *Submitter’s name*
    • *Title of story*
    • *Names of the first 215 users to vote on the story*
    • *Number of votes story received*
Dynamics of votes

story “interestingness”

number of votes (diggs)

0 1000 2000 3000 4000 5000
0 500 1000 1500 2000 2500

time (min)
Distribution of votes

~200 front page stories submitted in June 29-30, 2006

~30,000 front page stories submitted in 2006

Wu & Huberman, 2007
Dynamics of information spread

- How do producers promote their content?
- How do consumers find interesting new content?

How do stories become popular on Digg?
- Social networks play a major role in promoting stories on Digg
  - Patterns of information spread through networks can be used to predict how popular the story will become
- Mathematical model for collaborative decision-making on Digg
Stories spread through the network

Distribution of the numbers of users who can see the story through the Friends Interface

- At submission
- After 10 votes
- After 20 votes
And receive votes from within the network

Distribution of the number of in-network votes

Cascade = number of in-network votes (votes from fans of the previous voters)
But, **how** the story spreads through the network is different for different stories.
Correlation already after the first 10 votes!
Classification: Training

- **Decision tree classifier**
  - **Features**
    - *Number of in-network votes*
    - *Number of fans of submitter*
    - *Story interestingness*
      - **Yes if > 500 votes**
      - **No if < 500 votes**
  - **10-fold validation on 207 stories**
    - **Correctly classified 84% of instances**

```
<table>
<thead>
<tr>
<th>v10</th>
<th>&lt;=4</th>
<th>&gt;4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes(130/5)</td>
<td>v10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;=8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fans1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Predicting interestingness

Predict how interesting a story is based on the first 10 votes

• **Test dataset**
  • 900 stories submitted on June 30, 2006, but not yet promoted to the front page
  • 48 stories that were submitted by top users (rank<100) and received at least 10 votes
    • *Retrieve the final number of votes received by stories*

• **Classification**
  • Correctly classified 36 examples (TP=4, TN=32)
  • 12 errors (FP=11, FN=1)
  • Looking at the promoted stories only
    • *Digg prediction: 5 of 14 received more than 520 votes (Pr=0.36)*
    • *Our prediction: 4 of 7 received more than 520 votes (Pr=0.57)*
Mathematical analysis can help understand and predict the emergent behavior of collaborative information systems

- **Analysis of collective behavior on Digg**
  - Dynamics of collective voting
  - Dynamics of user rank
  - Analysis can aid the design of Digg
    - *Study the choice of the promotion algorithm before it is implemented*
    - *Study the effect of design choices on system behavior*
      - story timeliness, interestingness, user participation, incentives to join social networks, etc.
Dynamics of collective voting

Model characterizes a story by

- **Interestingness** $r$
  - probability a story will received a vote when seen by a user

- **Visibility**
  - Visibility on the upcoming stories page
    - *Decreases with time as new stories are submitted*
  - Visibility on the front page
    - *Decreases with time as new stories are promoted*
  - Visibility through the friends interface
    - *Stories friends submitted*
    - *Stories friends dugg (voted on)*
Mathematical model

- Mathematical model describes how the number of votes $m(t)$ changes in time

$$\Delta m(t) = r(v_f + v_u + v_i)\Delta t$$

- Solve equation
  - Solutions parametrized by $S, r$
  - Other parameters estimated from data
Dynamics of votes

Exploring the parameter space

Minimum $S$ required for the story to be promoted for a given $r$ for a fixed promotion threshold

Time taken for a story with $r$ and $S$ to be promoted to the front page for a fixed promotion threshold
Dynamics of user influence

- Digg ranked users according to how many front page stories they had

- Model of the dynamics of user influence
  - Number of stories promoted to the front page $F$
  - User’s social network growth $S$
Model of rank dynamics

- **Number of stories promoted to the front page** $F$
  - Number of stories $M$ submitted over $\Delta t$=week
  - User’s promotion success rate $\sim S(t)$
  
  $$\Delta F(t) = cS(t)M\Delta t$$

- **User’s social network $S$ grows as**
  - Others discover him through new front page stories $\sim \Delta F$
  - Others discover him through the Top Users list $\sim g(F)$

  $$\Delta S(t) = b\Delta F(t) + g(F)\Delta t$$

- **Solve equations**
  - Estimate $b$, $c$, $g(F)$ from data
Solutions 2

Learning from social tagging

with: Anon Plangrasopchok
**Metadata**

- Metadata (‘data about data’) used to “facilitate the understanding, characteristics, use and management of data.” [source: Wikipedia]
  - Terms from a formal taxonomy used to describe data
  - E.g., Linnean classification system describes living organisms

```
Animalia → Arthropoda → Insecta → Orthoptera → Caelifera → Tettigidae
```

[source: Linnean Classification System]
Semantic Web attempted to impose meaning on Web data to improve information access and usability

[Berners-Lee & Hendler in *Scientific American*, 2001]

- Web content annotated with machine-readable metadata (from a formal taxonomy) to aid automatic information retrieval and integration
- Still unrealized in 2008
  - Too complicated: specialized training to be used effectively
  - Costly and time-consuming to produce
  - Variety of specialized ontologies: ontology alignment problem
Tags serve a function similar to that of metadata from a formal taxonomy

- **Tags help users organize their data**
  - Facilitate searching and browsing for information
  - Describe the semantics (meaning) of content

- **Tags are keywords used to describe content**
  - Freely-chosen by user
  - No controlled vocabulary or formal taxonomy

Tags as alternative metadata
Collective tagging of content may lead to an emergent informal classification system

- **Folksonomy**
  - “user generated taxonomy used to categorize and retrieve web content using open-ended labels called tags.” [source: Wikipedia]
  - Advantages over formal taxonomies
    - *Simpler: easier cognitive process*
    - *Bottom-up: decentralized, emergent, scalable*
    - *Dynamic: adapts to changing needs and priorities*
  - But, since it comes from many different users,
    - *Noisy: need tools to extract meaning from data*
Tagging is simpler than categorization

Categorization = assigning object a single concept within a taxonomy

Cognitive process behind categorization

Stage 0: Object worth remembering (article, image, book...)
Stage 1: Multiple concepts are activated
Stage 2: Choose ONE of the activated concepts
Categorize it! Note the chosen concept

Analysis-Paralysis!

Cognitive process behind tagging

Stage 0: Object worth remembering (article, image, book...)
Stage 1: Multiple concepts are activated
Tag it! Write down activated concepts

Rashmi Sinha 2005
Collective tagging on Delicious

Web source

Welcome to FlyteComm
http://www.flytecomm.com/cgi-bin/trackflight
this url has been saved by 461 people.
save this to your bookmarks »

user notes

Tool for tracking flights

Track any flight anytime
http://www.flytecomm.com/cgi-bin/tracker
Welcome to FlyteComm

Track flights
greenjb

Flight Tracking
race/flight
Track you flights online

Flight Tracking Real Time Airport
woody7

popular tags

common tags
cloud | list:
204 travel
169 flight
140 tracking
84 flighttracker
83 airline
58 tracker
56 tools
33 reference
20 airlines
20 aviation
20 flights
15 tool
13 maps
13 web
11 airplanes
9 imported
7 flying
5 air
5 information
5 status
5 track
5 airplanes
5 airtravel
5 flight-tracking
5 internet

related items - show ↓

posting history

by adawson to travel/airlines
by ray_in_la to tracking/flights
by todd_2 to airplanes/travel/flighttracker

user tags
Given a collection of documents tagged by different users

- Use machine learning techniques to find ‘hidden’ or ‘latent’ topics in a collection of sources
  - Learn a hierarchy of hidden topics = folksonomy?
Alternative models

pLSA

MWA

ITM

Aggregate all tags from all users as if having only single user

Take into account individual difference

[Plangrasopchok & Lerman, in WWW’07]
The price we pay for aggregating data

- Keeping track of individual variations in tag usage, rather than aggregate tags over all users, can improve learning

Navarro+ 2006
Evaluation on synthetic data

Test how models perform when tags are ambiguous

- Created a synthetic data set with tunable noise parameter (tag ambiguity)
  - 40 resources tagged by 200 users
  - 10 topics
  - 25 unique tags
- Evaluate performance of the learning models on synthetic data
  - Find the topic distribution of resources using 3 models
    - $Distance_A = \text{distance between resources computed using the actual topic distribution}$
    - $Distance_L = \text{distance between resources computed using learned topic distribution}$
    - $\Delta = Distance_L - Distance_A$
  - Evaluation
    - *The better the learned topic distribution, the smaller the $\Delta$*
## Results on synthetic data

<table>
<thead>
<tr>
<th>Noise</th>
<th>Δ</th>
<th>Average(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pLSA</td>
<td>MWA</td>
</tr>
<tr>
<td>0</td>
<td>0.204(0.026)</td>
<td>0.216(0.030)</td>
</tr>
<tr>
<td>0.3</td>
<td>0.203(0.018)</td>
<td>0.224(0.031)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.205(0.018)</td>
<td>0.226(0.032)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.217(0.018)</td>
<td>0.233(0.034)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.216(0.018)</td>
<td>0.233(0.036)</td>
</tr>
<tr>
<td>1.1</td>
<td>0.224(0.019)</td>
<td>0.232(0.033)</td>
</tr>
<tr>
<td>1.3</td>
<td>0.227(0.019)</td>
<td>0.233(0.033)</td>
</tr>
<tr>
<td>1.5</td>
<td>0.242(0.022)</td>
<td>0.238(0.036)</td>
</tr>
</tbody>
</table>

### Significant < 0.01

Table 1: Deviations between distances computed from learned and actual topic distributions of resources of each model. Avg(sd) gives the average and standard deviation of difference between distances over all pairs of resources.

---

More noise, more tag ambiguity
Apply the learning framework to real data

- Leverage user-contributed tags to discover hidden topics in a collection of Web sources

- Find sources that provide some functionality
  - Simpler goal: find sources that provide the same functionality as the seed, e.g., http://flytecomm.com
    - Improve robustness of information integration applications
    - Increase coverage of the applications
Data sets for discovery task

• **Seeds**
  • *flytecomm*: flight status
  • *geocoder*: coordinates of an address
  • *wunderground*: weather conditions
  • *hotels*: hotel reservations
  • *whitepages*: phone book

• **Collected data by scraping del.icio.us**
  • For each seed, retrieve the 20 popular tags
  • For each tag, retrieve other *resources* annotated with same tag
  • For each resource, retrieve all *tags* added by *users*
• Find hidden topics in a collection of sources, using Probabilistic Generative Model

• Compute similarity between seed and source using hidden topics
Application to resource discovery

wunderground

whitepages

geocoder

flytecomm
Entrophy of user interests $p(i|u)$

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entropy (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>geocoder</td>
<td>1.419 (0.278)</td>
</tr>
<tr>
<td>wunderg</td>
<td>1.397 (0.227)</td>
</tr>
<tr>
<td>flytecomm</td>
<td>1.285 (0.279)</td>
</tr>
<tr>
<td>whitepage</td>
<td>1.157 (0.272)</td>
</tr>
<tr>
<td>online-res</td>
<td>0.629 (0.41)</td>
</tr>
</tbody>
</table>

Higher entropy => users seem to be interested in all topics equally likely
Conclusions

- In their every day use of Social Web sites, users create large quantity of data, which express their knowledge and opinions
  - Content
    - Articles, media content, opinion pieces, etc.
  - Metadata
    - Tags, ratings, discussion, social networks
  - Links between users, content, and metadata

- Social Web enables new problem solving approaches
  - Collective problem solving
    - Efficient, robust solutions beyond the scope of individual capabilities
  - Social information processing
    - Use knowledge, opinions of others for own information needs
Further reading

• To see more papers on Social Web
  http://www.citeulike.org/user/krisl/tag/socialweb

• To see my papers on the Social Web
  http://www.citeulike.org/user/krisl/tag/mysocialweb