Influences on Tag Choices in del.icio.us

Emilee Rader and Rick Wash
School of Information, University of Michigan
ejrader, rwash @ umich.edu
introduction
Folksonomy

The practice and method of collaboratively creating and managing tags to annotate and categorize content.

http://en.wikipedia.org/wiki/Folksonomy
Why do people choose some tags over others?
Individual choices  ➔  Internet-scale patterns
Users’ future tag choices are heavily influenced by tags they have chosen in the past.
background
Exact word choices matter for organizing, re-finding and navigating in tagging systems, but there are competing hypotheses in the literature for how people choose tags.
Explore Everyone's Tags
See what's popular, or delve into your own interests.

Tag

type a tag

Click a tag below or type in a word above, and we'll show you the latest bookmarks saved with that tag.

Tag Cloud: Popular

Sort: Alphabetically | By size

blog, blogs, design, development, entertainment, flash, fun, games, google, howto, html, inspiration, interactive, interesting, internet, javascript, jobs, jquery, kids, language, learning, library, linux, list, lists, literature, mac, management, maps, marketing, math, media, microsoft, mobile, money, movie, movies, music, network, networking, news, online, opensource, osx, people, phone, photography, photos, photoshop, php, plugin, podcast, politics, portfolio, privacy, productivity, programming, psychology, python, radio, rails, real estate, recipe, recipes, reference, religion, research, resources, reviews, rss, ruby, ruby on rails, school, science, search, security, seo, shopping, social, social networking, software, statistics, streaming, teaching, tech, technology, tips, todo, tool, tools, toread, travel, tutorial, tutorials, tv, twitter, typography, ubuntu, useability, video, videos, vim, visualization, web, web 2.0, web design, webdev, wiki, wikipedia, windows, wishlist, wordpress, work, writing, youtube
Save a Bookmark on Delicious

URL: http://www.loc.gov/index.html

TITLE: Library of Congress Home

NOTES:

TAGS:
- ref
- reference

Do Not Share

Save  Cancel

Tags
- Recommended
  - research
  - reference
  - tools
  - library
  - usa
  - book
  - database
  - information
  - libraries
  - art
  - education
  - history
  - images
  - search

- Popular
  - government
  - books
  - congress
  - resources
  - loc
  - archive
  - literature
Social Hypothesis
Users’ tag choices are influenced by the tag choices of others
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)

Cumulative Frequency in Percent

Time in Units of Bookmarks Added
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)

Cumulative Frequency in Percent

Time in Units of Bookmarks Added
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)

Cumulative Frequency in Percent

Time in Units of Bookmarks Added
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)

- reference
- library
- books
- government
Tag Distribution for del.icio.us ID:
3b3cb60120b337fa373871347e15418c (libcongress)

Cumulative Frequency in Percent

Time in Units of Bookmarks Added

- Reference
- Library
- Books
- Government
Sliding Window Tag Distribution, for del.icio.us ID: 3b3cb60120b337fa373871347e15418c (libcongress)

Time in Units of Bookmarks Added

Frequency in Percent

reference

library

books

government
Organizing Hypothesis

Users’ tag choices are personal and idiosyncratic, not influenced by others’ tag choices
Wash and Rader (2007)

- Respondents generally used one or more heuristics for choosing tags:
  - Reuse tags they have applied before to other web pages
  - Create and adhere to mental rules or definitions for specific tags
  - Choose terms they imagine using to re-find bookmarks in the future
We set out to look for a connection between the small scale (individual tag choices) and the large scale (aggregate patterns) for tags on del.icio.us.
logistic regression
Hypotheses
Hypotheses

• **Imitation**: Users imitate tags that previous users have applied to a web page
Hypotheses

- **Imitation**: Users imitate tags that previous users have applied to a web page
- **Organizing**: Users re-use tags that they have applied to other web pages
Hypotheses

• **Imitation**: Users imitate tags that previous users have applied to a web page

• **Organizing**: Users re-use tags that they have applied to other web pages

• **Recommended**: Users choose tags that are suggested via the del.icio.us posting interface
\[
tag_{\text{chosen}} = f(\text{used.onSite}, \text{used.byUser}, \text{interaction}, \tag_{\text{dummies}}, \text{random\_effect}(\text{user}))\]
\[
\text{tag\_chosen} = f(\text{used\_onSite}, \text{used\_byUser}, \text{interaction}, \\
\text{tag\_dummys}, \text{random\_effect(user)})
\]
tag_chosen = f(used.onSite, used.byUser, interaction, tag_dummies, random_effect(user))
tag_chosen = f(used.onSite, used.byUser, interaction,
tag_dummys, random_effect(user))
\( \text{tag\_chosen} = f(\text{used\_onSite, used\_byUser, interaction, tag\_dummies, random\_effect(user)}) \)
\[ \text{tag\_chosen} = f(\text{used.onSite}, \text{used.byUser}, \text{interaction}, \text{tag\_dummies}, \text{random\_effect(user)}) \]
## Summary of Logistic Regression Results:

<table>
<thead>
<tr>
<th>Mean Coefficient</th>
<th>onSite (Mean)</th>
<th>byUser (Mean)</th>
<th>Interaction (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.245</td>
<td>3.439</td>
<td>-0.280</td>
</tr>
</tbody>
</table>
Summary of Logistic Regression Results:

<table>
<thead>
<tr>
<th>Mean Coefficient</th>
<th>onSite</th>
<th>byUser</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.245</td>
<td>3.439</td>
<td>-0.280</td>
</tr>
</tbody>
</table>
Summary of Logistic Regression Results:

<table>
<thead>
<tr>
<th>Mean Coefficient</th>
<th>onSite</th>
<th>byUser</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.245</td>
<td>3.439</td>
<td>-0.280</td>
</tr>
</tbody>
</table>
Summary of Logistic Regression Results:

<table>
<thead>
<tr>
<th>Mean Coefficient</th>
<th>onsite</th>
<th>byUser</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.245</td>
<td>3.439</td>
<td>-0.280</td>
<td></td>
</tr>
</tbody>
</table>

**Organizing hypothesis** is strongly supported
computer model
Computer model
Computer model

- Imitation-Popular: Users might select popular tags presented to them in the del.icio.us interface
Computer model

- **Imitation-Popular**: Users might select popular tags presented to them in the del.icio.us interface

- **Imitation-Urn**: Imitation of other users’ tag choices might be achieved using a path-dependent process
Computer model

- **Imitation-Popular**: Users might select popular tags presented to them in the del.icio.us interface.

- **Imitation-Urn**: Imitation of other users’ tag choices might be achieved using a path-dependent process.

- **Organizing**: Users might favor tags that they had used previously.
Two complimentary measures
Two complimentary measures

- **Distributional Equivalence**: Fit to a powerlaw distribution (KS-test)
Two complimentary measures

- **Distributional Equivalence**: Fit to a powerlaw distribution (KS-test)

- **Numerical Identity**: Inter-user Agreement (Furnas et. al, 1983)
Imitation − Random
Tag Index $i$
Tags with index $> i$
$10^{-2.0}$
$10^{-1.5}$
$10^{-1.0}$
$10^{-0.5}$
$10^{0.0}$
$10^{0.5}$ $10^{1.0}$ $10^{1.5}$

$KS = 0.409$
Summary of Computer Model Results for **Distributional Equivalence**

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean KS statistic</td>
<td>0.069</td>
<td>0.223</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>
Summary of Computer Model Results for **Distributional Equivalence**

<table>
<thead>
<tr>
<th>Mean KS statistic</th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.069</td>
<td>0.223</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>
# Summary of Computer Model Results for Distributional Equivalence

<table>
<thead>
<tr>
<th>Mean KS statistic</th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.069</td>
<td>0.223</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>

**Mean KS statistic**: Measures the maximum difference between the empirical cumulative distribution functions of the real world and the model-generated data.
Summary of Computer Model Results for Distributional Equivalence

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean KS statistic</td>
<td>0.069</td>
<td>0.223</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>
Summary of Computer Model Results for *Distributional Equivalence*

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean KS statistic</td>
<td>0.069</td>
<td>0.223</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>

Rule out *Imitation-Popular* and *Imitation-Urn*
Summary of Computer Model Results for Numerical Identity
Summary of Computer Model Results for **Numerical Identity**

<table>
<thead>
<tr>
<th>Mean IUA</th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.144</td>
<td>0.317</td>
<td>0.184</td>
<td>0.182</td>
</tr>
</tbody>
</table>
### Summary of Computer Model Results for Numerical Identity

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean IUA</strong></td>
<td>0.144</td>
<td>0.317</td>
<td>0.184</td>
<td>0.182</td>
</tr>
</tbody>
</table>
# Summary of Computer Model Results for Numerical Identity

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean IUA</td>
<td>0.144</td>
<td>0.317</td>
<td>0.184</td>
<td>0.182</td>
</tr>
</tbody>
</table>
# Summary of Computer Model Results for Numerical Identity

<table>
<thead>
<tr>
<th>Mean IUA</th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.144</td>
<td>0.317</td>
<td>0.184</td>
<td>0.182</td>
</tr>
</tbody>
</table>
Summary of Computer Model Results for **Numerical Identity**

<table>
<thead>
<tr>
<th></th>
<th>Real World</th>
<th>Imitation-Popular</th>
<th>Imitation-Urn</th>
<th>Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean IUA</td>
<td>0.144</td>
<td>0.317</td>
<td>0.184</td>
<td>0.182</td>
</tr>
</tbody>
</table>

Rule out **Imitation-Popular**
1. **Imitation-Popular**: rule out based on both numerical identity and distributional equivalence

2. **Imitation-Urn**: rule out based on distributional equivalence

3. **Organizing**: cannot rule out (similar to “real world” on all measures)
1. **Imitation-Popular**: rule out based on both numerical identity and distributional equivalence

2. **Imitation-Urn**: rule out based on distributional equivalence

3. **Organizing**: cannot rule out (similar to “real world” on all measures)
1. **Imitation-Popular**: *rule out* based on both numerical identity and distributional equivalence

2. **Imitation-Urn**: *rule out* based on distributional equivalence

3. **Organizing**: *cannot rule out* (similar to “real world” on all measures)
The **Organizing hypothesis** was **strongly supported** by logistic regression, and we could **not rule it out** based on our computer model.
Users’ future tag choices are heavily influenced by tags they have chosen in the past.
conclusion
• **Exact word choices** matter for re-finding and navigation in tagging system, but there are **competing hypotheses** in the literature for how people choose tags.

• We set out to **establish a connection** between the small scale (individual tag choices) and the large scale (aggregate patterns) for tags on del.icio.us.

• **Organizing hypothesis** was **strongly supported** by logistic regression, and we could not rule it out based on our computer model.
Thank You!

Emilee Rader and Rick Wash
School of Information, University of Michigan
ejrader, rwash @ umich.edu

Online Appendix: http://bierdoctor.com/papers/cscw08/