Tagging Practices on Research Oriented Social Bookmarking Sites

Margaret E. I. Kipp
Faculty of Information and Media Studies
University of Western Ontario
margaret.kipp@gmail.com

CAIS/ACSI 10-12 Mai 2007, Montréal, QC

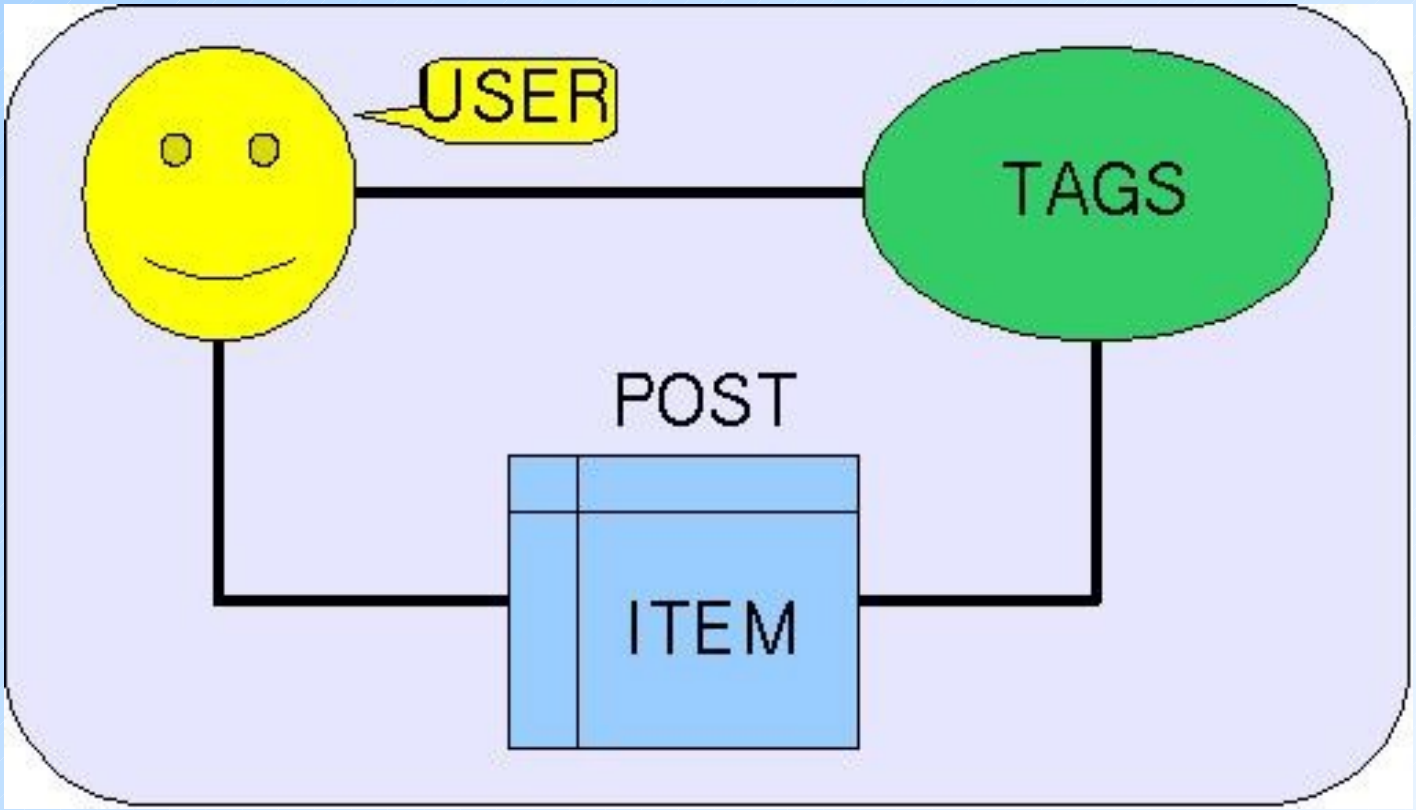
Background

- My research examines:
  - how people organise things on the web
  - how this compares to traditional library classification techniques
- Specific points of interest:
  - structures and the creation of structures in classification systems
  - relationship between personal information management and classification

Social Bookmarking and Tagging

- Social Bookmarking:
  - site for sharing bookmarks, articles, etc.
  - association of tags (keywords) with links
  - tags and articles are joined into networks of related terms
  - users are encouraged to share bookmarks and tags with others
- Tagging:
  - the act of associating a term with a link or article
  - labelling or classifying for personal use

Social Bookmarking Post



A post is a relationship between a user, an item and a set of tags.

citeulike

Navigation

- Home
- Log in
- Register
- Discussion list

Journals

- Browse current issues

Groups

- View group

Experimental Features

- Import from BibTeX

Everyone's library

Some recent papers posted to CiteULike - all mixed together.

- **Supramolecular Structure and Dynamics Special Feature: Crystalline molecular machines: Encoding supramolecular dynamics into molecular structure**
  PNAS, Vol. 102, No. 31. (2 August 2005), pp. 10771-10776.
  by Garcia-Garibay MA
  posted to review superamolecular by barry as ★★ on 2007-05-09 00:07:02

- **The fabrication of the translucent ZnO by sintering**
  Journal of Materials Science, Vol. 12, No. 11. (1 November 1977), pp. 2347-2349.
  by Moriyoshi Y, Isobe M, Hasegawa Y, Komatsu W
  posted to ceramics transparent_ceramic zno by polyparadigm as ★★★ on 2007-05-09 00:02:28

- **Measuring Coexisting Densities from a Two-Phase Molecular Dynamics Simulation by Voronoi Tessellations**
  J. Phys. Chem. B, Vol. 111, No. 13. (5 April 2007), pp. 3469-3475.
  by Fern JT, Keffer DJ, Steele WV
  posted to int by jwagoner as ★★ on 2007-05-08 23:59:06

- **Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations.**
  Proc Natl Acad Sci U S A, Vol. 100, No. 26. (23 December 2003), pp. 15310-15315.
  by Xayaphoummine A, Bucher T, Thalmann F, Isambert H
  posted to pseudoknots rna rna-secondary-structure by Kiarostami as ★★ and 5 others ... on 2007-05-08 23:43:09

- **Psychology in Canada.**
  Annu Rev Psychol, Vol. 47 (1996), pp. 341-370.
  by Adair JG, Paivio A, Ritchie P
  posted to canada memory-history by khm as ★★ on 2007-05-08 23:42:09

- **A comparison of RNA folding measures.**
  BMC Bioinformatics, Vol. 6, No. 1. (3 October 2005)
  by Freyhult E, Gardner PP, Moulton V
  posted to ncrna rna-folding by Kiarostami as ★★ and 5 others ... on 2007-05-08 23:39:40

- **Subcubic Time Algorithms for RNA Secondary Structure Prediction**

Everyone's Tags

Most active tags on CiteULike

Filter:

adaptation agents algorithm algorithms analysis annotation attention bayesian bioinformatics biology book cancer classification clustering coding collaboration communication community complexity cscw culture data database design detection development digital disease distributed dynamics economics education engineering evaluation evolution experiment expression fmri function gene genetics genome genomics geometry graph hci health history human information interaction internet ir

Previous Studies

- Study 1: Del.icio.us
- study of Del.icio.us tag usage on highly tagged sites
- examination of convergence of tag usage
- co-occurrence analysis for co-used tags

- Study 2: CiteULike
- study of CiteULike tag usage in comparison to author keywords and subject headings from on-line journal databases
- examine types of tags and more traditional index terms

Commonalities Between Studies

- Study 3: Del.icio.us, CiteULike, Connotea
- use of affective tags (e.g. cool, interesting) and time and task related tags (e.g. @toread, todo) in both studies
- > 16% of tags in Del.icio.us study
- average of 1-3 tags per article in CiteULike study were not directly subject related
- categories: time and task, affective, geographic, methodology, emergent vocabulary, other

Motivations

- Builds on study 2 of CiteULike
  - Kipp (2006): users do use words from thesaurus as tags, but often use similar or related terms from other fields
- Examine use of indexing terms by users, authors and intermediaries
- Do they appear to provide a similar context?

Related Studies

- Mathes (2004): suggested examination of user, author and intermediary terms
- Voorbij (1998), Ansari (2005): relatively high degree of match between descriptors (intermediary terms) and title keywords
- Kipp (2006): found differences in term usage between users, authors and intermediaries

Research Question

- To what extent do term usage patterns of user tags, author keywords and intermediary descriptors suggest a similar context between users, authors and intermediaries?

Data Collection

- Tag data collected from CiteULike
- Journals used: Proteins and Journal of Molecular Biology
- Data collected included DOI or URL for collection of keywords and descriptors
- Author Keywords and Pubmed Descriptors from journal sites and Pubmed respectively
- 1083 articles (1588 posts) collected

Data Analysis

- Informetric analysis using SQL (see Wolfram 2005)
  - standard informetric measures: frequency of occurrence of unique tags
- Thesaural analysis (see Voorbij 1998, Kipp 2006)
  - comparison of terms using Pubmed thesaurus (range from SAME, SYN, NT, BT, RT, related and Not related)

Authors, Users and Journals

- Authors:
  - 80% of articles had between 2 and 5 authors
  - one article had 48 authors
- Users:
  - 239 unique users, 1588 posts
  - most prolific user had posted 94 posts
  - 42 users posted 10 or more articles

Tags, Keywords and Descriptors

	Tags	Keywords	Descriptors
Unique	1136	3181	2746
Total	3788	4866	12473

- ratio of unique terms to total terms highest for author keywords
- supports findings from previous study in which author keywords were found to be more diverse than tags or descriptors

Popular Tags, Keywords and Descriptors

Tags	Frequency
protein_structure	140
no-tag	114
protein	114
structure	103
docking	97

Author Keywords	Frequency
protein folding	58
protein structure	49
molecular dynamics	46
protein structure predictio	38
docking	31

Descriptors	Frequency
Models, Molecular	649
Protein Conformation	511
Proteins	388
Amino Acid Sequence	306
Binding Sites	280

- 645 tags were used only once in the data set
- 2548 keywords were used only once
- 731 descriptors were used only once
- keywords are more diverse

Tags, Keywords and Descriptors by Article

- maximum number of tags per article was 29, minimum 1 and median 2
  - article with 29 tags was tagged by 14 users (most users still use 1-3 tags to an article)
- maximum number of tags per post was 15, minimum 1 and median 2
- maximum number of keywords per article in the data set was 13, minimum 1, median 5
- maximum number of descriptors per article was 36, minimum 2, median 11

User Vocabulary Length

User	Max tag list length	Min tag list length	Number of articles posted
3109	7	2	15
3063	6	1	73
4068	15	2	9

- user vocabulary length: the number of unique terms used by a user
- largest user vocabulary length was 62 (min. 1, median 2)
- most users use a small number of tags

Thesaural Analysis

Tags	Keywords	Descriptors
3d	16 S RNA	Base Sequence
algorithms	ribosome	Computer Simulation
prediction	computer modeling	Cross-Linking Reagents
rna	distance geometry	Escherichia coli
16s		Models, Molecular
distance_geometry		Molecular Sequence Data
bioinformatics		Nucleic Acid Conformation
structure		RNA, Ribosomal, 16S
structure_prediction		

- Article 788: Computer modeling 16 S ribosomal RNA
- across all three sets of terms are variants on RNA and 16s
- use of term bioinformatics versus computer modelling/simulation

Discussion

- results from the previous study (Kipp 2006) using a smaller data set from library science are relevant to other fields and to larger data sets
- users use some terminology which is rare or completely absent from author keyword lists or descriptor lists (e.g. time and task tags)
- user terms often not part of formal thesaurus
  - 'protein' and 'structures' as separate tags were linked in the thesaurus
  - abbreviations such as 'PDB' for 'Databases, Protein'

Discussion 2

- tags: 'human', 'animal', and 'family-studies' showed users tagging biology related articles are extremely interested in methodology and user groups associated with articles, this did not occur in the previous study

Margaret E. I. Kipp
Faculty of Information and Media Studies
University of Western Ontario
margaret.kipp@gmail.com
http://publish.uwo.ca/~mkipp/

Thank you/Merci!

Questions?