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

Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices

Margaret E. I. Kipp
D. Grant Campbell

ASIST Annual Meeting, November 3-8, 2006, Austin, Texas

Background

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

What is Social Bookmarking?

- public sharing of links
  - association of tags (keywords) with links
- network of related links created by users
  - network of related tags created by users
- site for sharing bookmarks, articles, etc.
  - tags and articles are joined into networks of related terms
  - users are encouraged to share bookmarks and tags with others

What is Tagging?

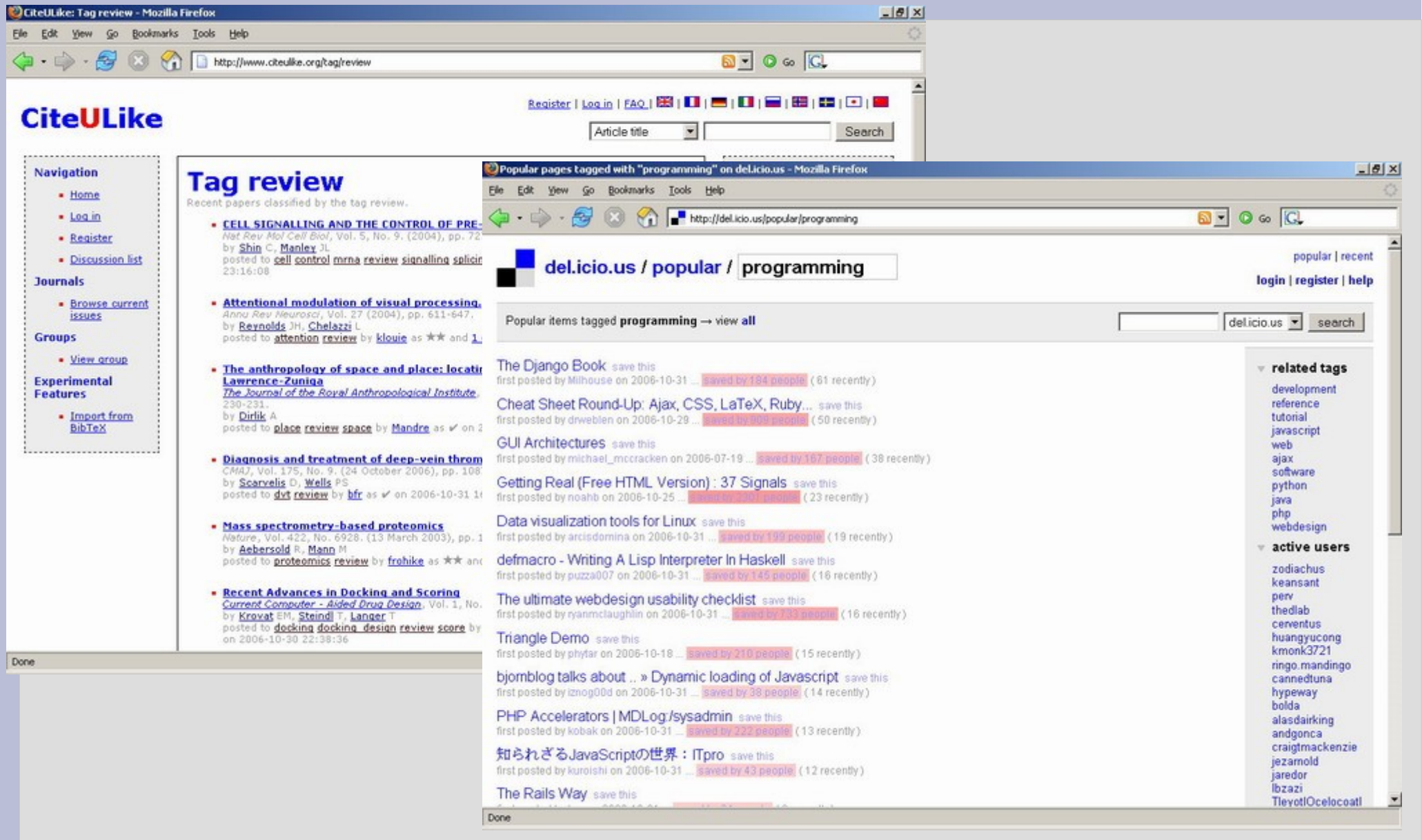
- the act of associating a term with a link or article
- labelling or classifying for personal use
- act of generating a dynamic taxonomy or folksonomy

- Related definitions:
  - folksonomy - user generated taxonomy of related tags
  - tag cloud - tag display where size equals popularity

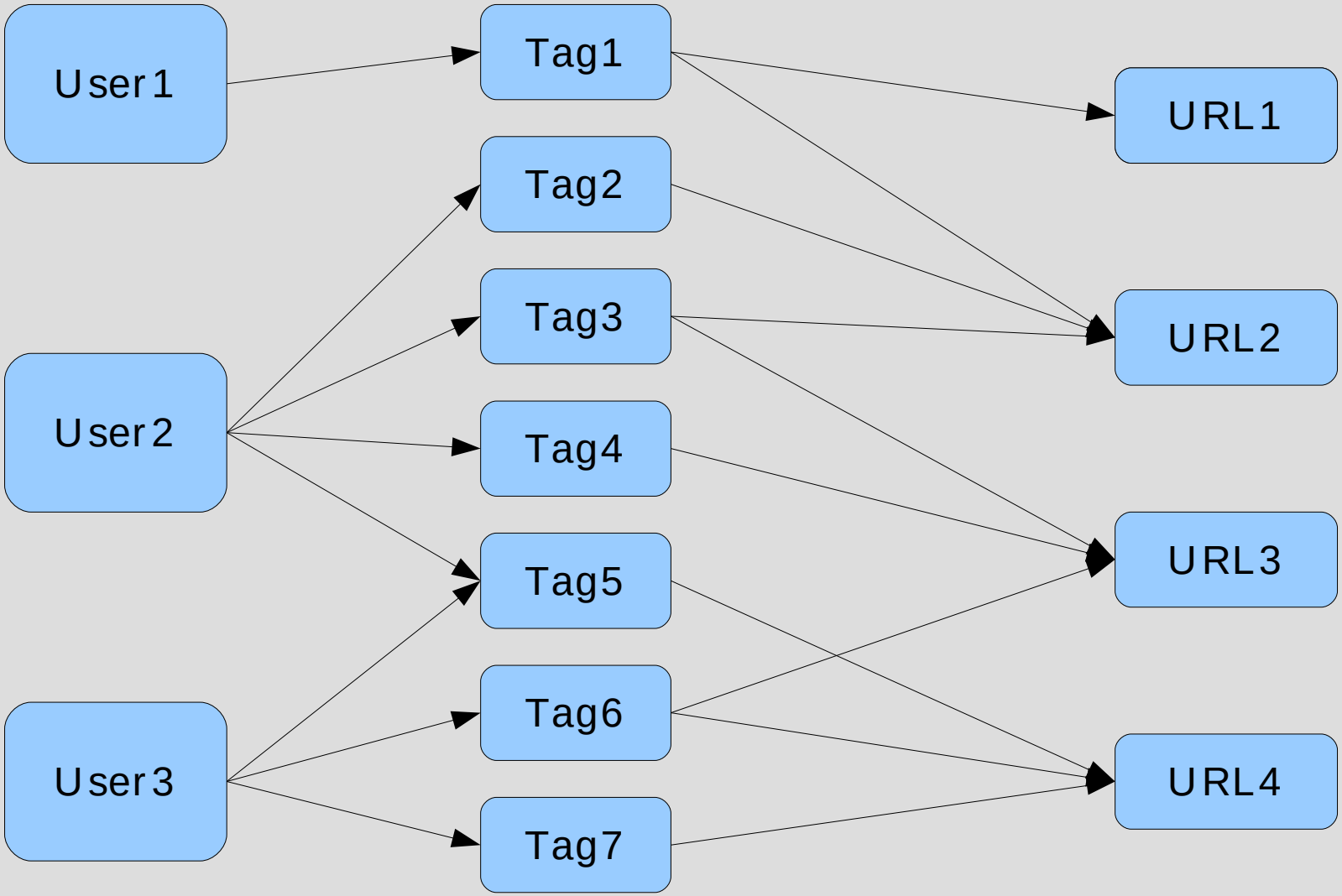
Social Bookmarking Sites

- citeulike
  - specialised for academic researchers
  - mainly journals and academic books
  - http://citeulike.org/
- del.icio.us
  - for anyone
  - bookmark anything
  - http://del.icio.us/

Screenshots - November 2006



Posting to Del.icio.us



The Controversy

Tagging is Good

- dynamic distributed classification
- related tag networks
- tag clouds show extent of collection
- user terminology
- diversity
- consensus by active users

Tagging is Bad

- mob indexing
- no controlled vocabulary
- poor browsing experience
- no thesaurus
- consensus by a mob or no consensus

The Study

- examine tags used by users of a social bookmarking service
- compare to traditional classification methods
  - examine similarities
  - examine differences
- analyse relationships
  - examine structures
  - examine related tags
  - frequency charts
  - coword analysis of tags in posts

Research Questions

1. What patterns of consistent user tagging activity emerge through analyses of tagging frequency and co-word analysis?
2. To what extent do these patterns of tagging support and enhance some of the other traditional ways of classifying documents?
3. To what extent do these patterns defy these traditional methods, suggesting viable and promising alternatives to traditional subject access tools?

Methodology

- Data source
  - Del.icio.us
- Collection times
  - January 30th-31st 2006
- Collection method:
  - python scripts
- Data collected
  - all posts for 64 URLs
  - posts for popular tags (http://del.icio.us/popular)
  - URLs posted by > 500 users with tags health, productivity or programming

Analysis Methods

- Descriptive Statistics
- Tag Frequency Charts
  - unique tag frequencies
  - a unique tag is alphabetically unique
- Tag Coword Analysis
  - examine frequency of occurrence of pairs of tags
  - if users A, B, and C had all tagged the same URL with tags X and Y, then X and Y would be a co-word pair with 3 occurrences

Frequency and Coword Analysis

- frequency graphs and coword graphs analysed for trends in tag usage
- analysis of frequency and coword graphs was qualitative not statistical
- coword graphs are a visual representation of tags clustered by similarity or commonness of co-occurrence

Descriptive Statistics

- number of posts: 58728
- number of tags: 165831
- number of unique tags: 18904
  - (per URL max: 1252, min: 23)
- average posts per URL: 917
  - (max: 5172, min: 53)
- average tags per URL: 295
  - (max: 13809, min: 49)

Descriptive Statistics

- users who did not tag: 6%
- users who used 1-3 tags: 65%

Tag	*Frequency*
productivity	6989
gtd	5782
lifehacks	4753
tools	4672
web2.0	4077
web	3876
organization	3469
css	3382
blog	3320
wiki	3187

Top 10 High Frequency Tags

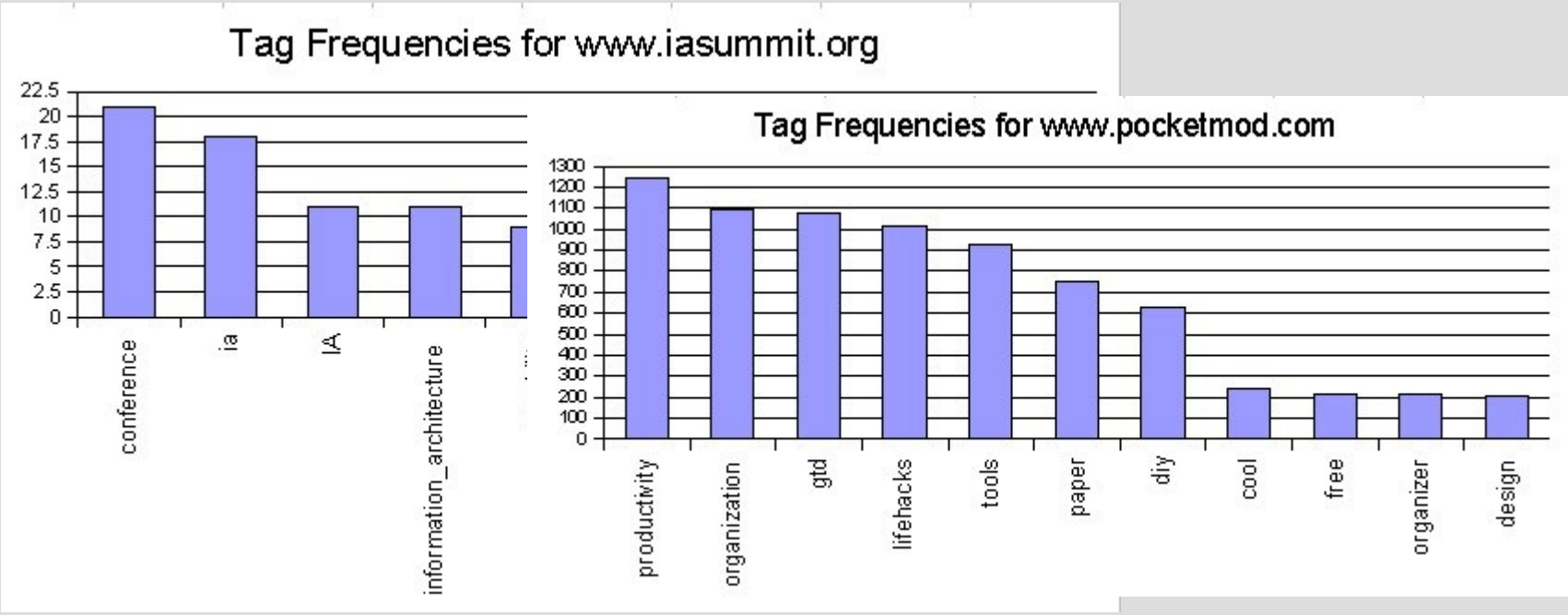
Number of uses of tag	*Number of tags at each frequency*
1	5593
2	1307
3	496
4	296
5	163
6	125
7	107
8	82
9	51
10	47

Infrequently Used Tags

Comparisons to Classification

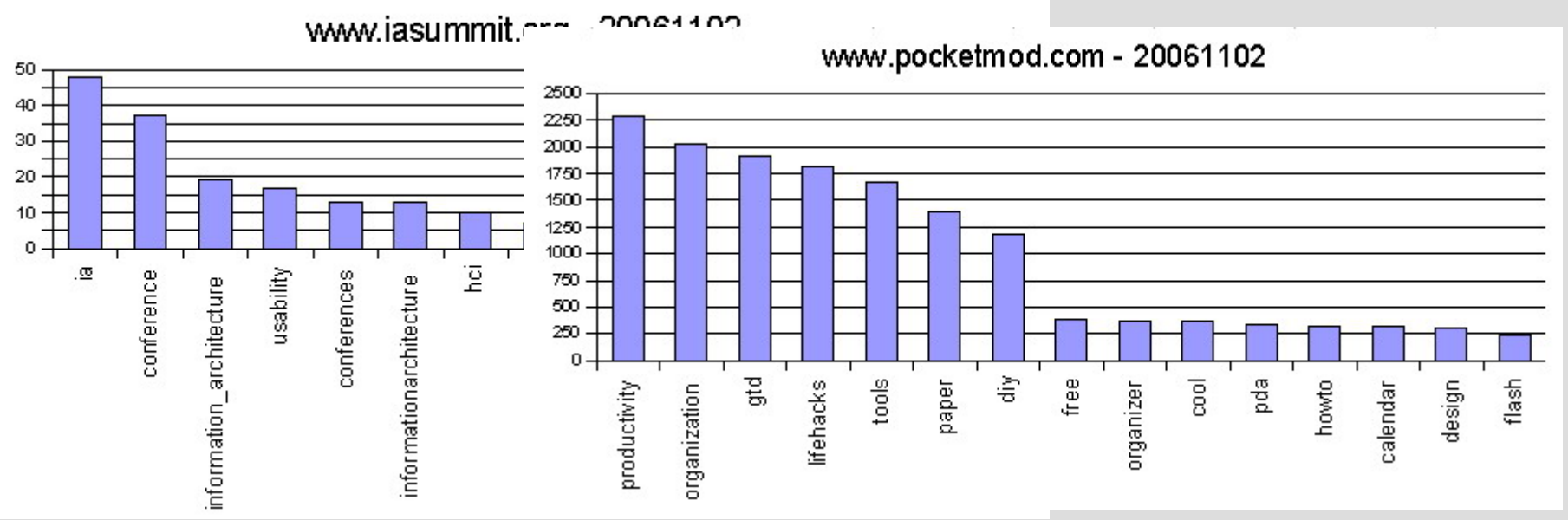
- spelling variations
  - British versus American spelling
  - singular or plural
  - conjugated versus stem
- synonyms or related terms
  - e.g. diet, nutrition, health, food, eating
- acronyms
  - e.g. www.iasummit.org
  - most common tags:
    - conference
    - ia
    - IA
    - information_architecture

Tag Frequency Graphs - January 2006



- frequency graph shows power law curve
- drop off is much shallower than expected
- pattern appears on highly tagged sites (4171 users)
- suggests users settle on cluster of terms

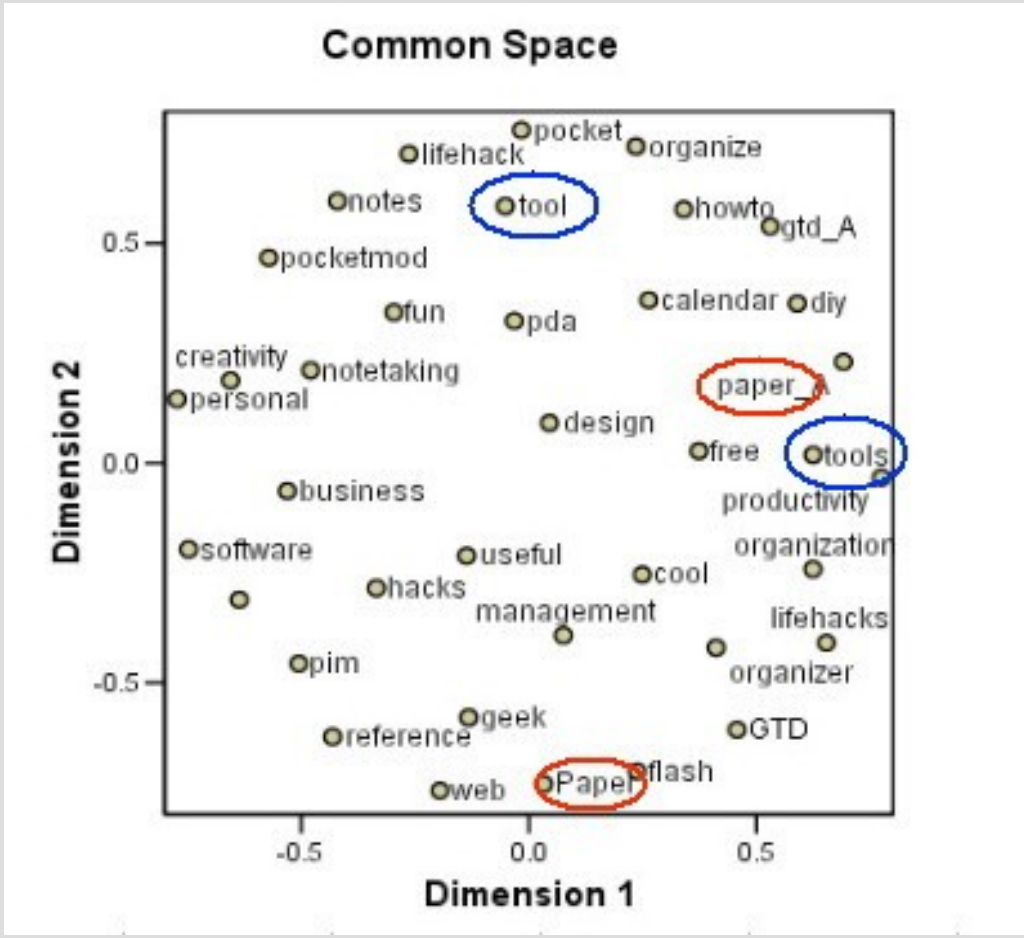
Tag Frequency Graphs - November 2006



- tag frequency graphs of recent data show the same patterns
- pocketmod (6754) has stabilised on a core set of terms while a tiny bit of shifting has occurred for the iasummit (63)

Cotag Graphs

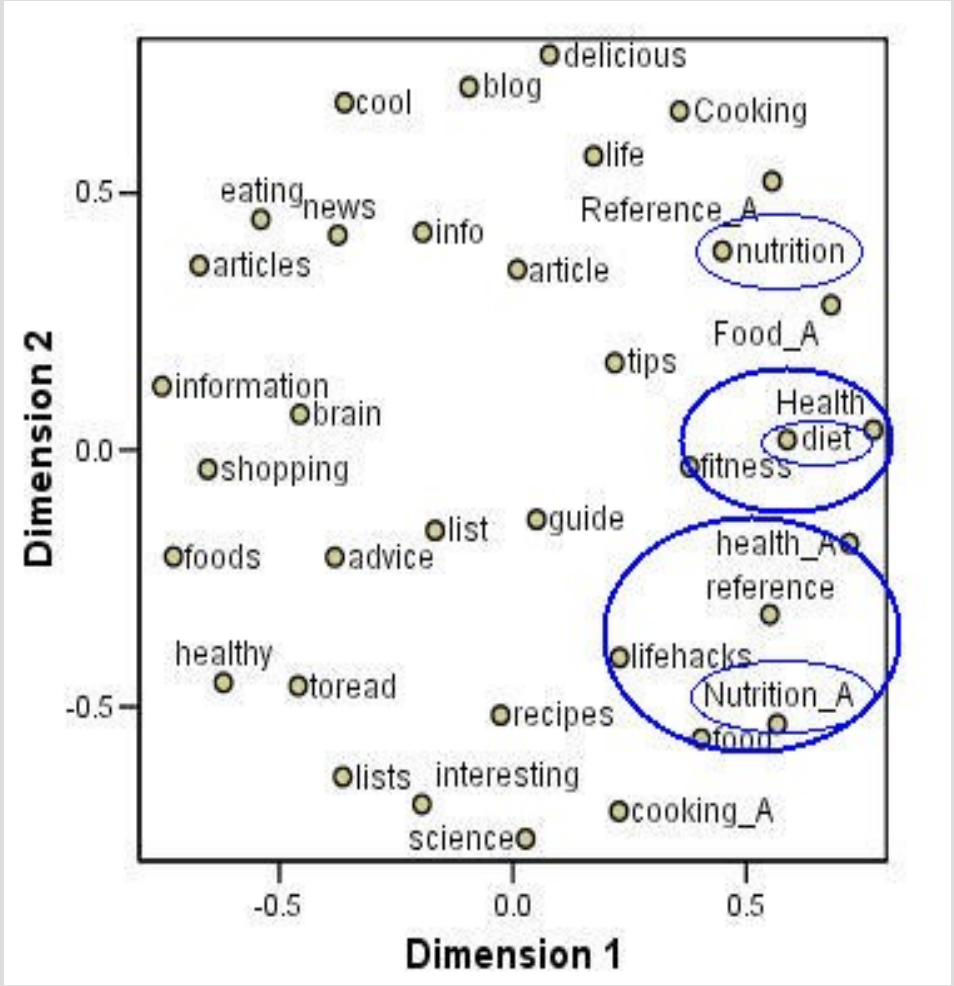
- coword results can be graphed using MDS (multi dimensional scaling)
- generally clusters show similarity
- note blue circled and red circled tags: similar but not clustered



Cotag graph
www.pocketmod.com

Cotag Graphs

- nutrition and diet do not cluster
- neither nutrition nor Nutrition cluster with diet (Nutrition clusters with food)
- perhaps evidence of different user groups in the tag clusters



Cotag graph www.bellybytes.com

Non Subject Tags

- Affective Tags
  - cool: 906 occurrences
- Time and Task Tags
  - toread: 939 occurrences
  - 3049 unique tags identified as time and task (16%)

Non Subject Tags

- intrinsically time-sensitive
- express response from user not subject of document
- suggest active engagement with the text
- show that user links perceived subject matter to:
  - specific task
  - specific set of interests
  - specific emotional reactions

Discussion and Conclusions

- closely-related terms are not necessarily revealed through co-occurrence
- users employ many conventions in constructing tags, but apply them inconsistently
- since the data collection period, del.icio.us has removed case sensitivity from tags

Discussions and Conclusions

- like indexing, tagging resorts to multiple terms to describe the aboutness of documents
- users demand finer grained indexing than is currently common
- users want to represent more than just the aboutness of a document

Future Directions

- continuities between tagging and indexing suggest the two may be complementary and that a combination would enrich both
- use of time and task or affective tags shows that tagging expresses a dynamic relationship between users and documents, suggesting possible new ways of modelling information access

Acknowledgements

- Faculty of Information and Media Studies, University of Western Ontario
- Government of Ontario
- Social Sciences and Humanities Research Council, Canada

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

Thank you!

Questions?

ASIST Annual Meeting, November 3-8, 2006, Austin, Texas