

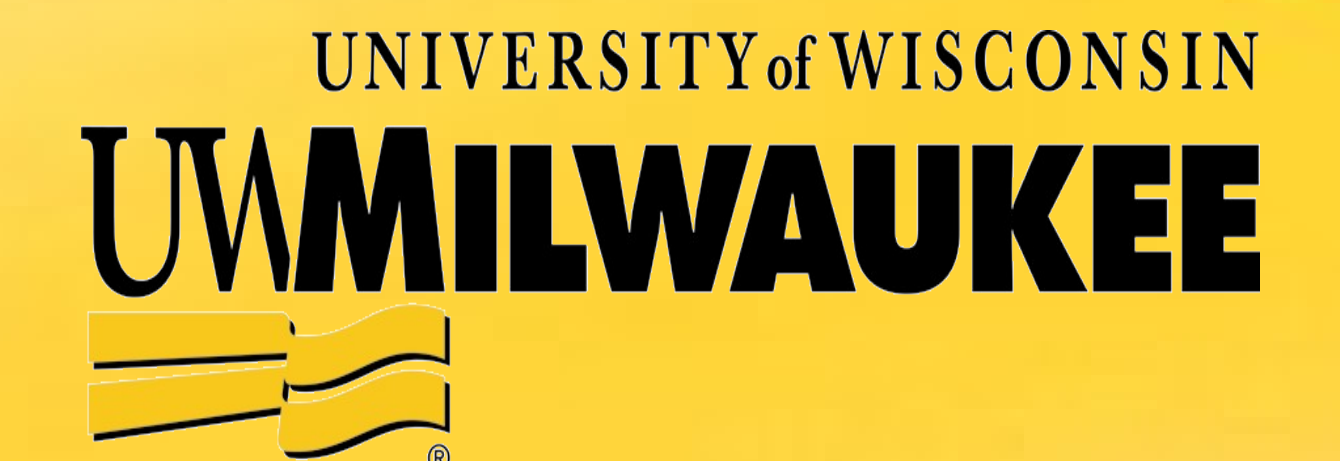
# Convergence and Divergence in Tagging Systems: An Examination of Tagging Practices Over a Four Year Period

Margaret E.I. Kipp <kipp@uwm.edu>

Information Organization Research Group (IOrg)

School of Information Studies, University of Wisconsin Milwaukee

https://pantherfile.uwm.edu/kipp/public/



## INTRODUCTION

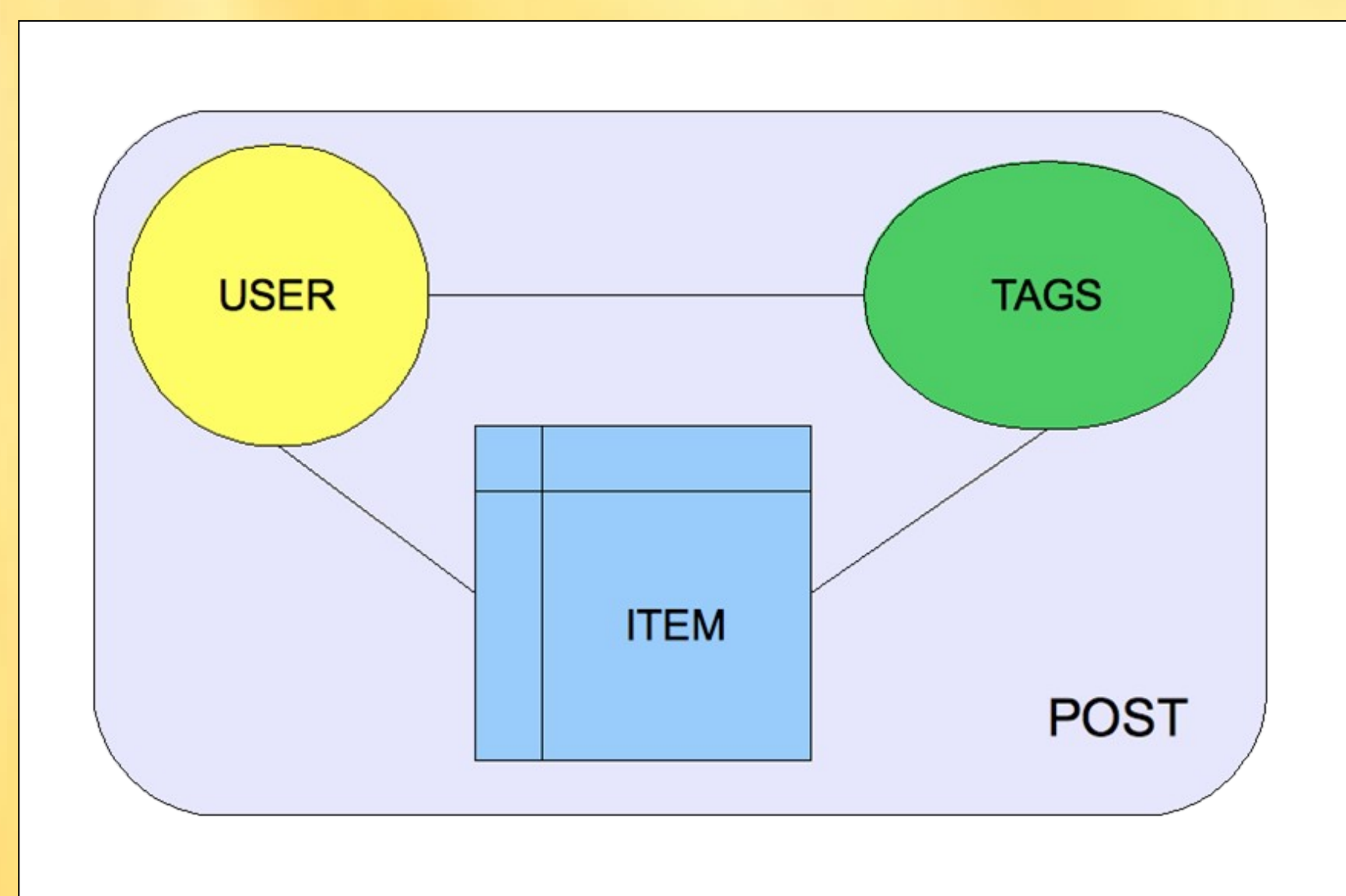
- early tagging research showed strong similarities to conventional indexing as well as substantial differences
- patterns in tagging show both divergence and convergence over time
- this study examines patterns in tagging of URLs over a 4 year period using the following questions:
  - What patterns of user tagging activity emerge through analyses of tags applied to a set of documents (URLs)?
  - To what extent do these patterns converge and diverge over time?

## BACKGROUND

- Hammond et al. 2005, Sen et al. 2006: tag frequencies show convergence of terminology
- Kipp 2005, Kipp NASKO2007: tags are often related terms (in a thesaural sense) or similar terms which are not yet in the thesaurus
- Golder and Huberman 2006: tagging activity has predictable patterns and proportions stabilise after as few as 100 taggers
- Kipp and Campbell ASIST2006: tagging has many similarities to conventional indexing and substantial differences
- Kipp IASummit2007: non subject tags may be a sign of emotional involvement, reviewing function, deeper relationship with text

## METHODOLOGY

- This study analyses changes over time in tags applied to 63 documents with up to 4 years data in order to track fluctuations in tagging.
- data source:** delicious.com
- crawler:** delicious.py
- post:** single item posted by a user, may have associated tags
  - each URL may have been posted by many users; each user may post many URLs
- analysis:** delicious-analyse.py
  - descriptive statistics
  - frequency graphs and proportion graphs
  - cotag analysis with MDS
  - intertagger consistency measures

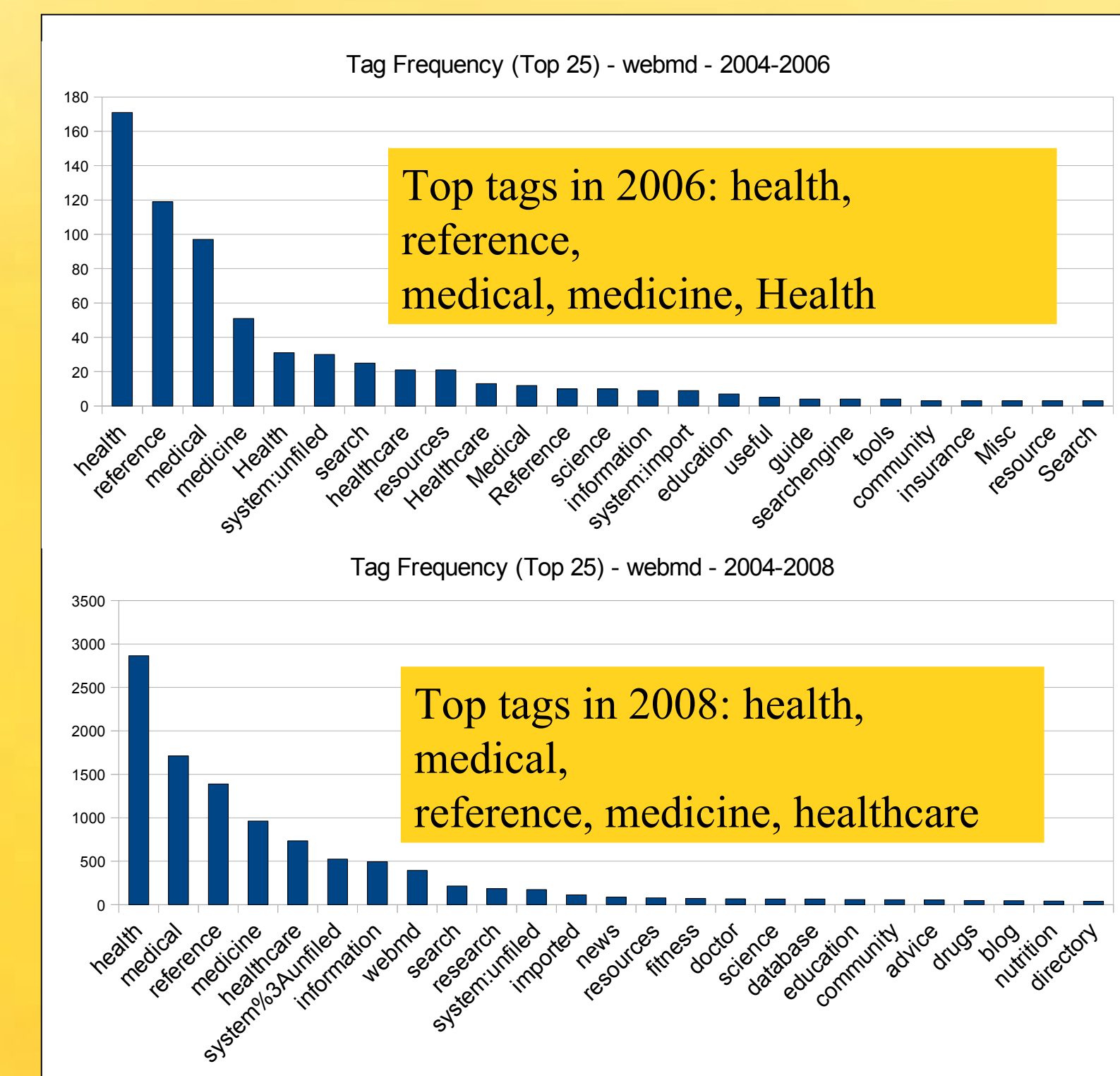
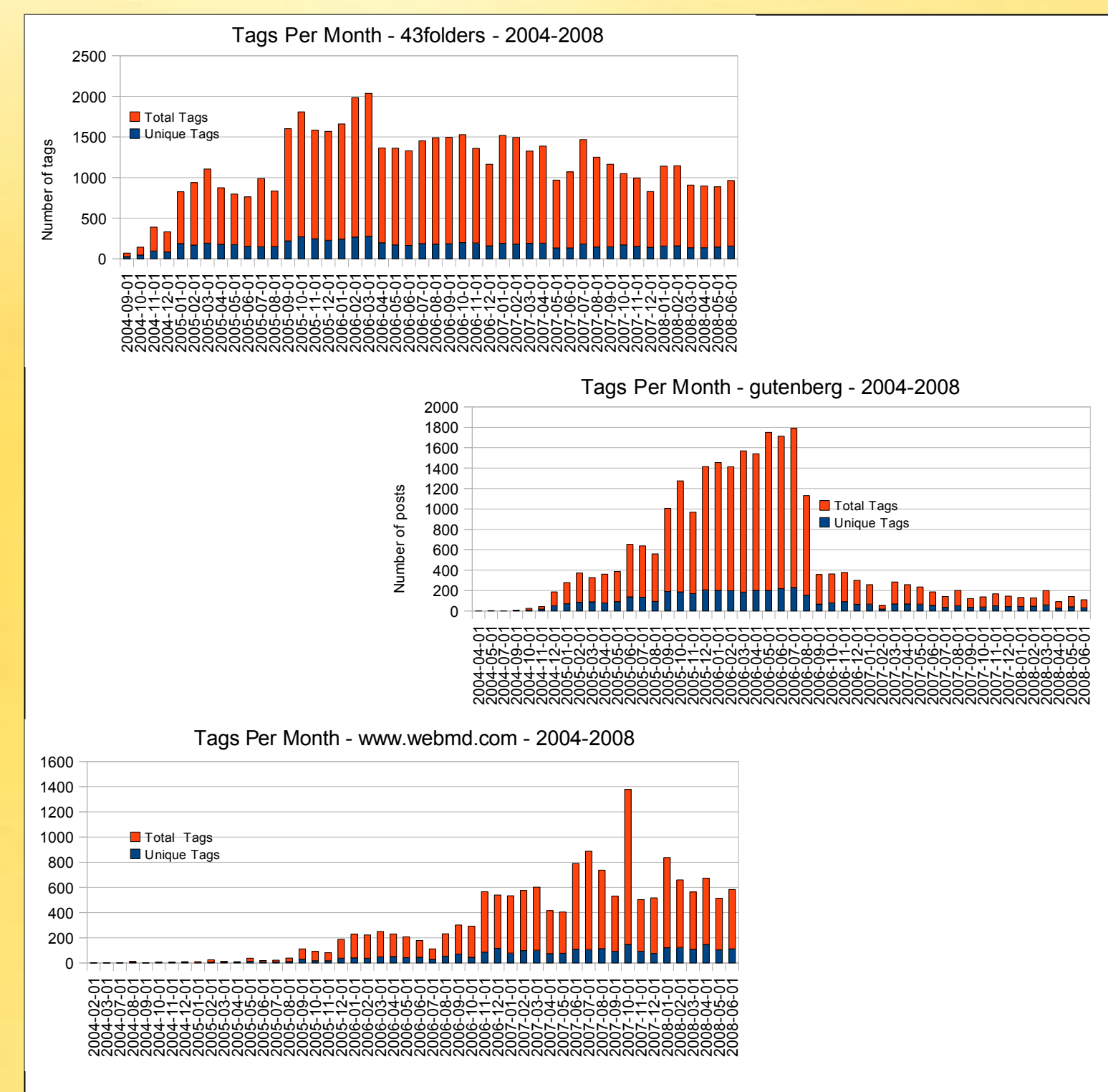


## RESULTS

- total number of posts:** 56646 (2006) and 136656 (2008) increased by 2.88 times
  - popular URLs experienced a greater growth (maximum 15.36)
- total number of tags:** 163761 (2006) and 430451 (2008)
- unique tags:** 62% in 2006, 68% in 2008
- user vocabulary length:** 10 in 2006 (median 2), 24 in 2008 (median 2)
- tagless entries:** 6% in 2006, 7% in 2008
  - system:unfiled retains its popularity (system assigned tag for tagless entries)
- popularity:** 9 of 10 most popular items from 2006 sample still in top 10 in sample in 2008 (webmd was a newcomer to the list in 2008)
  - popular tags from 2006 in sample still popular in 2008 (health was a new popular tag in 2008)
- 3 Sample URLs:** with different growth patterns
  - <http://www.43folders.com/> (43folders - personal productivity site/blog)
  - <http://www.gutenberg.org/> (gutenberg - public domain ebooks)
  - <http://www.webmd.com/> (webmd - health information)

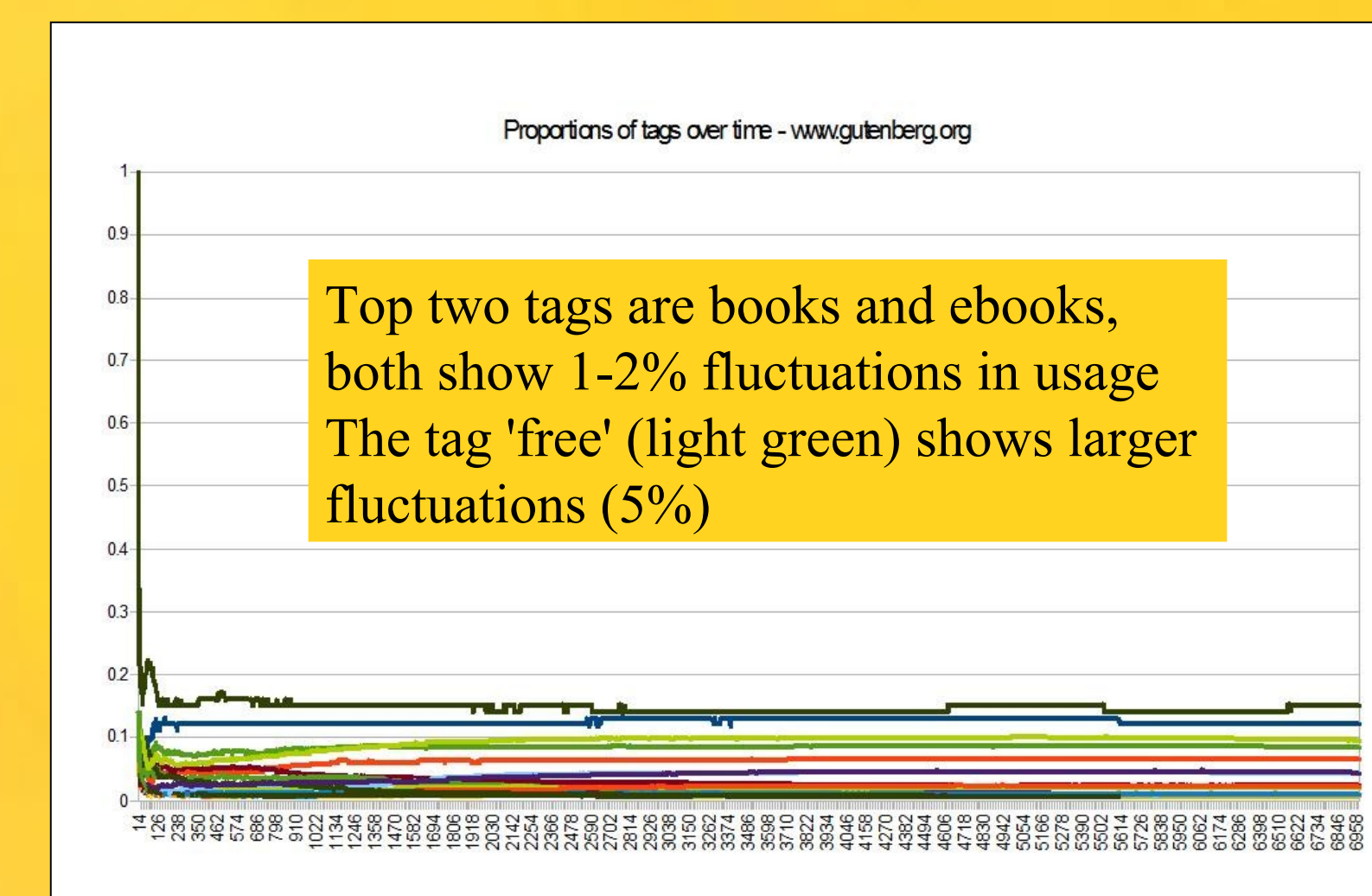
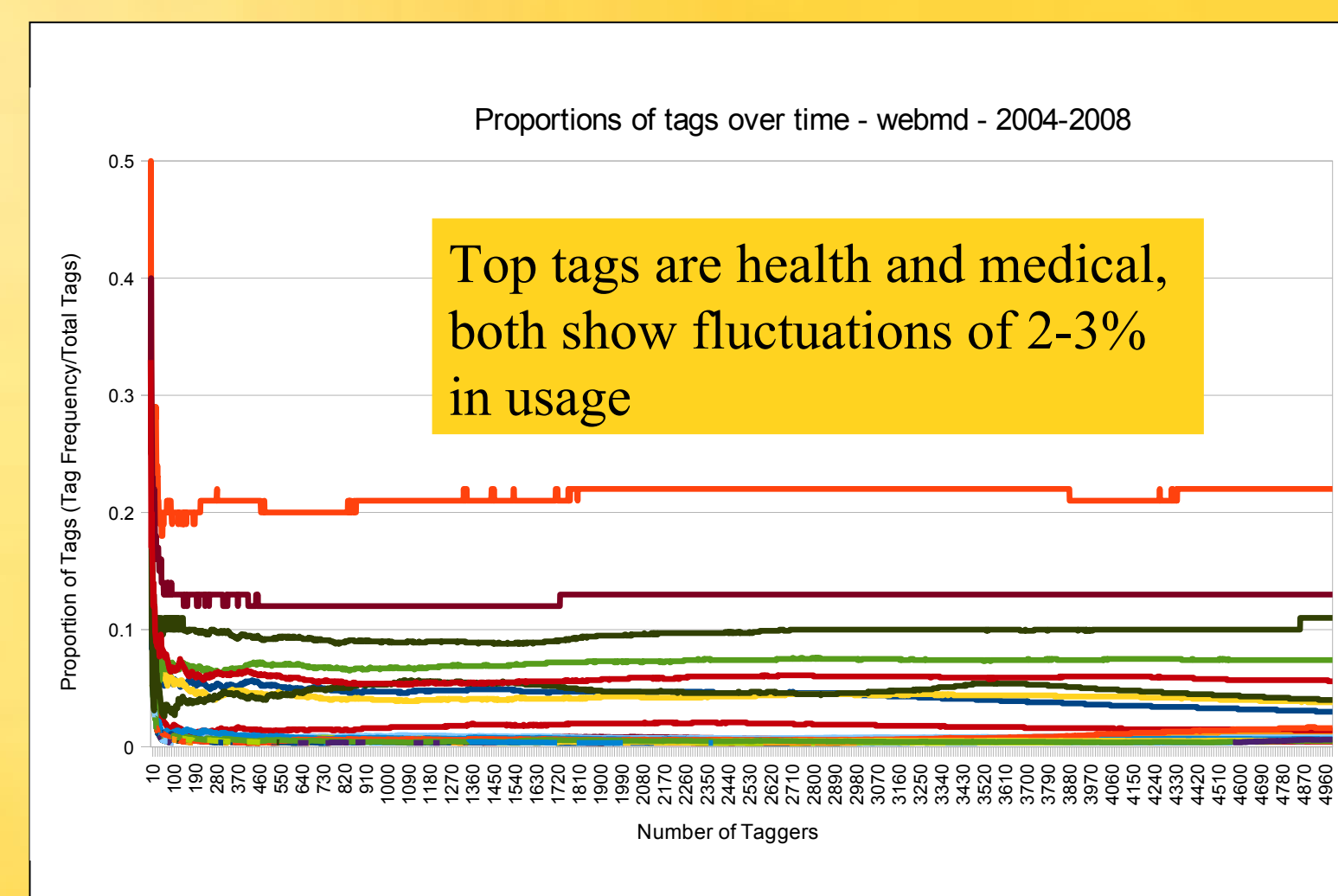
## Fluctuations in Taggers and Tags

- significant positive correlation between # of users and # of tags
- significant positive correlation between # of tags and # of unique tags



## Stability and Proportions of Tags

- many frequency graphs show top 6-7 items remaining relatively stable
- proportions of each tag as compared to total number of tags fluctuate, but have islands of stability
  - tag decay (Russell ASIST2007), changes in perspectives, changes in terminology use over time

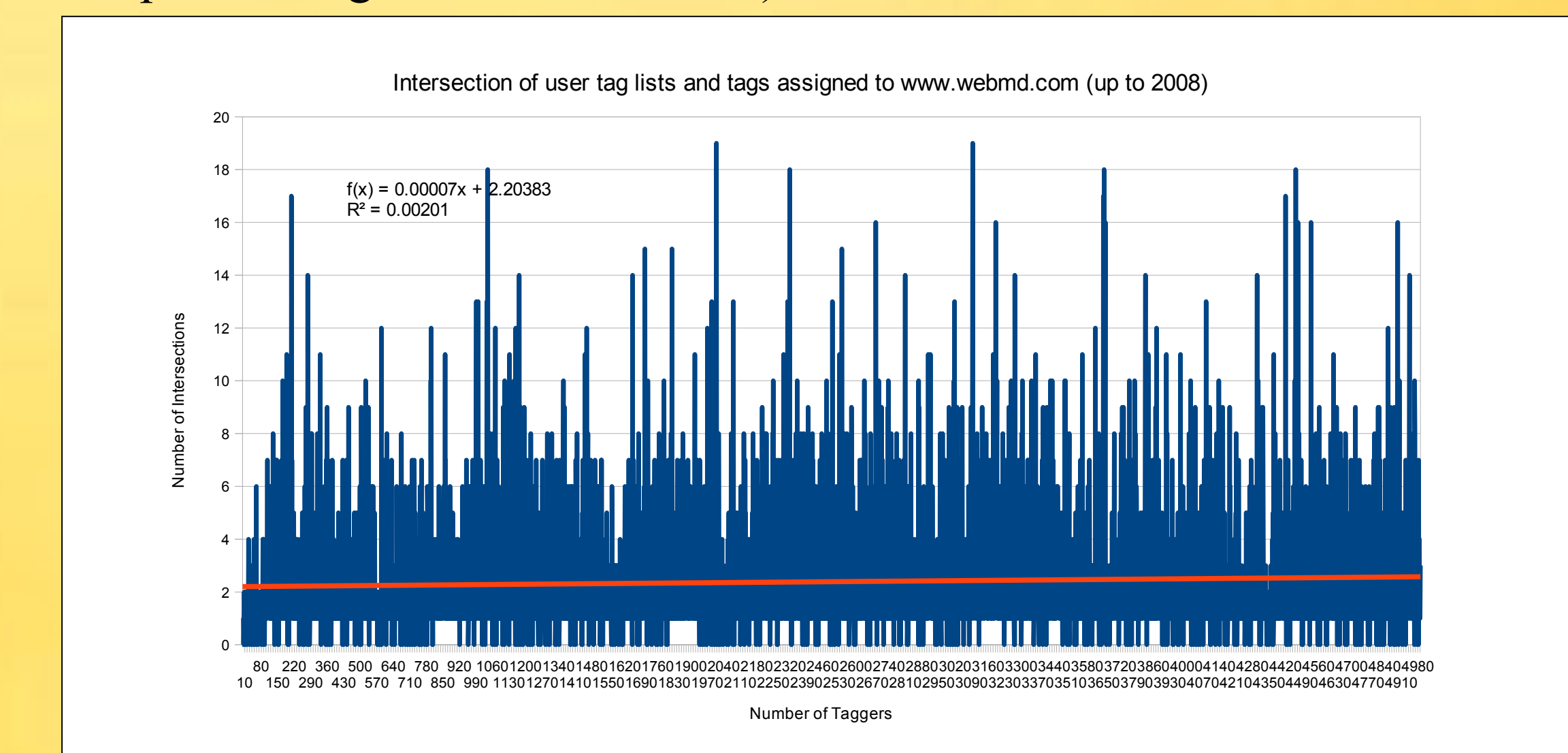


## Inter Tagger Consistency

- inter indexer consistency studies show that indexers are not consistent (from 4% to 82% in Markey 1984): varies based on expertise of indexers, type of vocabulary used
- various methods: with centroid, centroid represents good indexing; without centroid: pairwise comparisons
  - e.g.: webmd
    - indexer1: 'health', 'medical', 'medicine', 'information'
    - indexer2: 'biology', 'diet', 'healthcare', 'webmd', 'medicine', 'medical', 'health', 'information', 'food'
- with centroid: intertagger values ranged from 1% to 25%
- pairwise: intertagger values ranged from 3% to 30%

## Influence in Tagging

- positive correlation between time since initial posting and user vocab length ( $R^2 0.198$   $p < 0.05$ )
- comparing number of intersections between user tag lists and the tag cloud for an item shows an increase in intersections (most show small positive significant correlation)



## DISCUSSION AND CONCLUSIONS

- early studies of tagging reported that convergence and stability were simply a matter of having sufficient users tag an item
- this study suggests tagging may always show fluctuations in term usage based on fluctuations in user interests in everyday life and emergent terminology
- librarians and information specialists can use these fluctuations to enhance search and browse by allowing terminology to change over time while still providing stable controlled vocabularies for long term information organisation and retrieval

## REFERENCES

Golder, S.A., and Huberman, B.A. 2006. The structure of collaborative tagging systems. *Journal of Information Science* 32, no. 2:198-208.

Hammond, T., Hannay, T., Lund, B., and Scott, J. 2005. Social bookmarking tools (I): A general review. *D-Lib Magazine* 11, no. 4. <http://www.dlib.org/dlib/april05/hammond/04hammond.html>

Kipp, M.E.I. 2005. Complementary or discrete contexts in on-line indexing: A comparison of user, creator and intermediary keywords. *Canadian Journal of Information and Library Science* 29, no. 4:419-436. <http://eprints.rclis.org/8379/>

Kipp, M.E.I. 2007. @toread and cool: Tagging for time, task and emotion. *Proceedings of the 8th Information Architecture Summit, Las Vegas, USA, March 22-26*. <http://eprints.rclis.org/10445/>

Kipp, M.E.I. 2007. Tagging for Health Information Organization and Retrieval. *North American Symposium on Knowledge Organization (NASKO), Toronto, June 14-15, 2007*. <http://eprints.rclis.org/11412/>

Kipp, M.E.I., and Campbell, D.G. 2006. Patterns and inconsistencies in collaborative tagging practices: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology, Austin, TX, USA, November 3-8, 2006*. <http://eprints.rclis.org/8315/>

Markey, K. 1984. Inter-indexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research* 6, no. 2:155-177.

Russell, T. 2007. Tag decay: A view into aging folksonomies. *Proceedings of the American Society for Information Science and Technology*, 44: 1-5. <http://www.terrellrussell.com/projects/ifon8-tagdecay.pdf>

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, M. F., and Riedl, J. 2006. Tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 CSCW Conference*, New York, NY, USA.