Aspects of Broad Folksonomies

Mathias Lux

Alpen Adria Universität Klagenfurt

Michael Granitzer

Know-Center Graz

Roman Kern

Know-Center Graz

Content

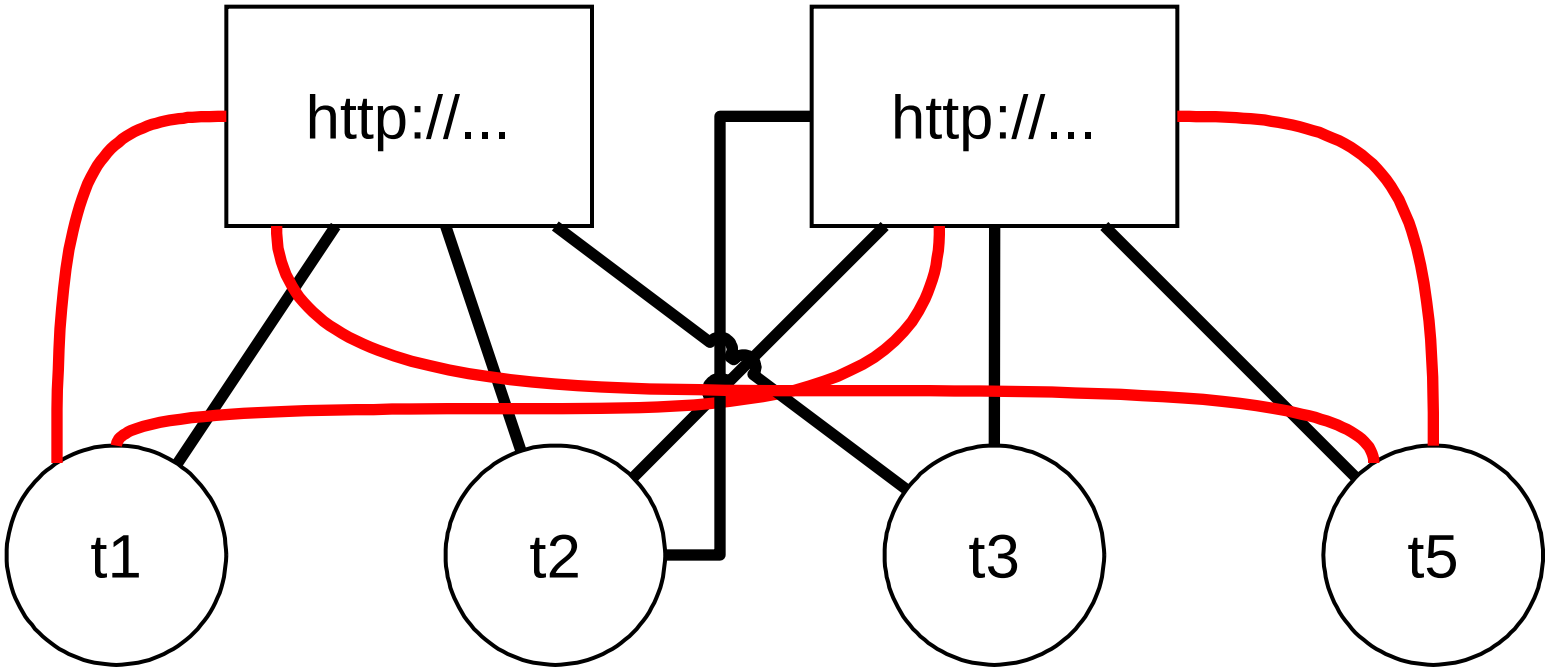
- What is a broad folksonomy?
- Motivation & related work
- Methodology
- Results
- Conclusion

Folksonomy

- Term Coined by **Thomas Vander Wal**
  - folk + taxonomy
- Definition is not clear
  - Web 2.0: Everyone makes up his own definition
- Definition of T. Vander Wal as base
  - Users add tags (keywords) to resources
  - F. emerge from this (mostly personal) organization
  - F. is hypergraph: agents, tags & resources (cp. P. Mika, 2005, ‘Ontologies Are Us’)

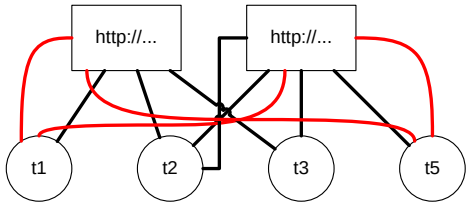
Folksonomy - Example Create Bookmark

Folksonomy



Motivation

- F. is a **complex & huge** graph
- F. represents **metadata**
- F. represents **relations**
  - between users, tags & resources
- F. might be **utilized for retrieval**
  - Some problems already identified
    - e.g. ambiguity, scope and misspellings



Research Questions

- Does a F. provide (good) metadata for retrieval?
- Does a F. (or parts of a F.) stabilize over time?
- Is there a structure that emerges from a F. and what does it look like?

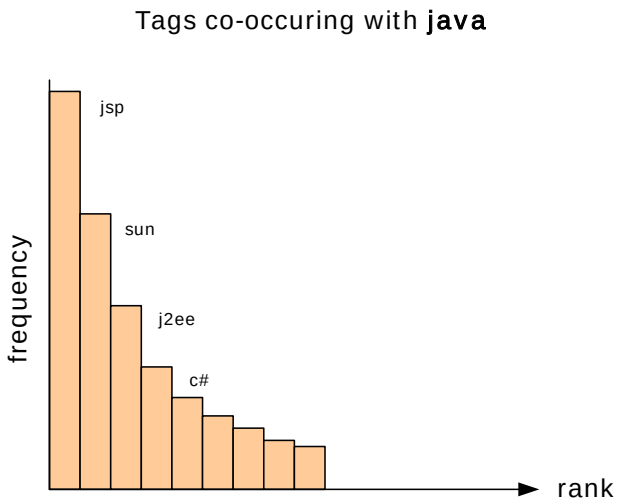
Assumptions

- Tags are co-assigned to resources
- Frequent co-assignment means:
  - “Tags are related semantically”
- If tags are semantically related:
  - There are few tags highly related
  - Some tags somewhat related
  - Many tags not related

Related Work

Cattuto, Loretto & Pietronero (2007)

- Investigated Frequency-Rank distribution of co-occurrence of tags.
- Empirical evidence that power law applies
- Shown for 4 tags
  - Blog, Ajax, Xml, H5N1



Further Assumptions

- Analyzing co-occuring tags of 4 tags is not enough to infer global emergence.
  - What about broader tags like ‘funny’?
  - Wu, Zhang & Yu (2006) use an entropy function to identify such broad tags ...
- Broad tags might not follow a power law.
  - They are associated to many other tags
    - e.g. video, image, page, joke, photo

Test Data Set: A Quasi Random Sample

- Social Bookmarking: del.icio.us
  - Investigated e.g. by Cattuto et al., Mika
  - One of the biggest available
- Continuous aggregation of bookmarks
  - Recent additions every 7th minute
  - Only bookmarks used at least 2 times
  - URL, user, description, note, date and tags

Test Data Set: A Quasi Random Sample

- Sample size
  - 3.234.956 bookmarks
  - 9.241.878 tag associations of
  - 356.838 different tags by
  - 84.121 different users
- Sub sample (due to computation issues)
  - 838.804 bookmarks having
  - 2.408.935 tag associations of
  - 135.473 different tags by
  - 26.919 different users

Methodology

What is a *power law*?

- Heavy-tail distributions, Pareto distributions, Zipfian distributions, etc.
- Much heavier tails than others (e.g. exponential distributions)
- Not characterized well by mean and variance
- Log-log plot is a straight line
- Examples: Size of cities, sizes of solar flares

cf. Clauset, Shalizi & Newman (2007) “Power-law distributions in empirical data” and Mitzenbacher (2002) “A Brief History of Generative Models for Power Law and Lognormal Distributions”

Methodology

- Simple empirical test
  - Plot a sample on a logarithmic scale
  - If it resembles a ‘straight line’ a power law might apply
- Statistical tests: χ^2 (chi square) test
  - Estimate constant and exponential parameter
  - Calculate χ^2 statistic for each rank & estimate significance

$$y = \alpha \cdot x^{\beta}$$

Tag Co-Occurence

- What tags are co-occuring to Tag t?
  - R_t set of resources it has been assigned to
  - co-occuring tags are all tags that are assigned to resources in R_t
- Frequency of a co-occuring tag
  - Number of overall assignments in R_t

Tag Co-Occurence

- Does the frequency-rank distribution for co-occuring tags follow a power law?
  - cp. Cattutos finding for a few tags

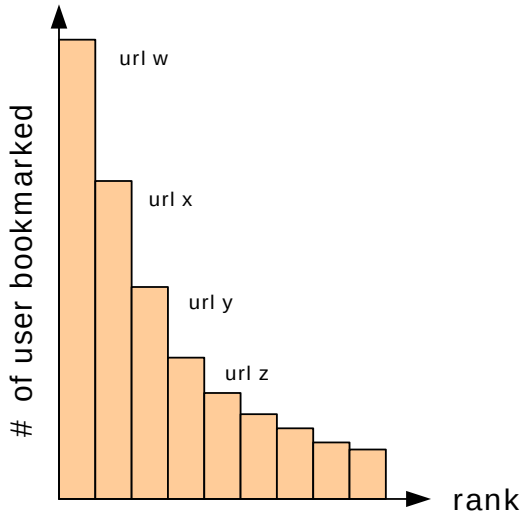
- We found that
  - 80% of the tags the co-occuring tags have a Zipf's frequency-rank distribution.
  - For 90% of those β is in [-1.5, -0.5]

Conclusions ...

- Tag Co-Occurence
  - Power law does not apply to whole folksonomy
  - In our results power law applies to co-occuring tags of 4 out of 5 tags.
  - Assumptions:
    - Data set too small
    - Tags too ambigous

Resource based Tagging Characteristics

- What is the distribution of users vs. the rank of the resource w.r.t. a tag?
  - Are there few resources where many users assign the tag and
  - Many resources where few users assign the tag?

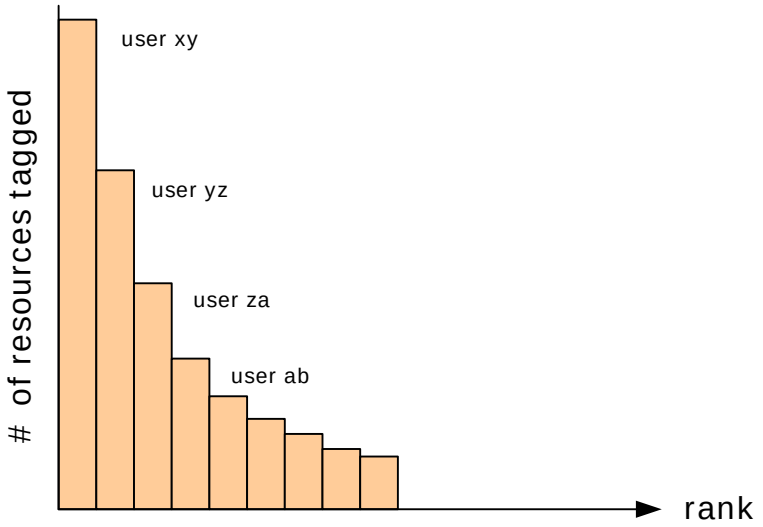


Resource based Tagging Characteristics

- Restricted to tags having been assigned 30+ times
- Around 18.4 % of the analyzed tags had a Zipfian user count to resource rank distribution.

User based Tagging Characteristics

- What is the distribution of resource count vs. user rank for tags?
  - Are there many users who assign the tag to few resources and
  - Few users who assign it to many resources?



User based Tagging Characteristics

- Restricted to tags having been assigned 30+ times
- Around 13 % of the analyzed tags had a Zipfian user count to resource rank distribution.

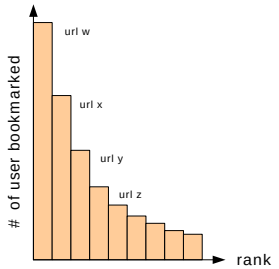
Conclusions ...

- Tagging Characteristics
  - Power law does not apply to most the tags in this respect.
  - We think that tags that for that the power law applies
    - are mostly unambiguous
    - have ‘narrow’ semantics (cp. ‘C3PO’ to ‘funny’)

Semantically Different Sub Communities?

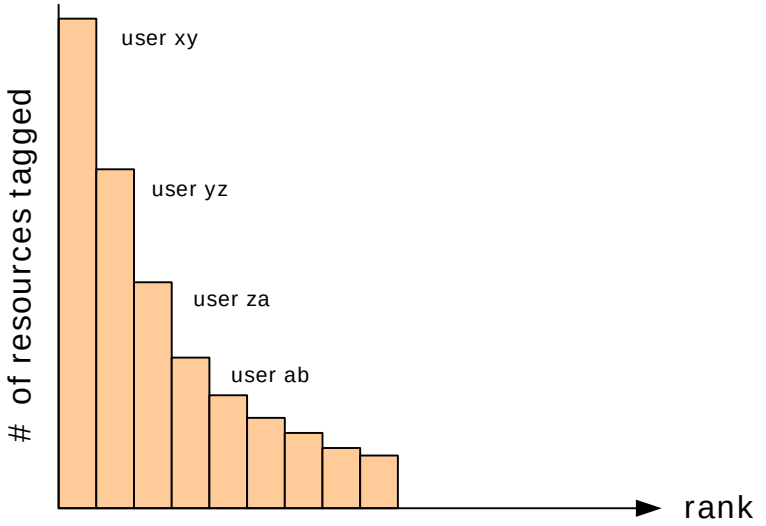
Analyzing resource based tagging characteristics
18.4 % of the tags showed a power law distribution of user frequency.

- Is there a disagreement upon tag assignment between users in the tail?
- Splitting to three groups (high, medium and low ranked resources, each 1/3) showed:
  - There is only a small overlap between the users in these groups.



Semantically Different Sub Communities?

- Also only a small overlap could be found in the user based tagging characteristics
  - High ranked users do not tag the same resources as low ranked users.



Tags not following a power law ..

- w.r.t. to user and resource based tagging characteristics
- Applies to more than 80%

Tags not following a power law ..

- D1: Tags used 30+ times
- D2: Tags used less than 30 times

	D1	D2
Tag only used once (e.g. typos)	-	57.0%
Tag used by single user (personal vocabulary)	3.9%	19.0%
Tag used once per user (unpopular tags)	12.0%	38.7%

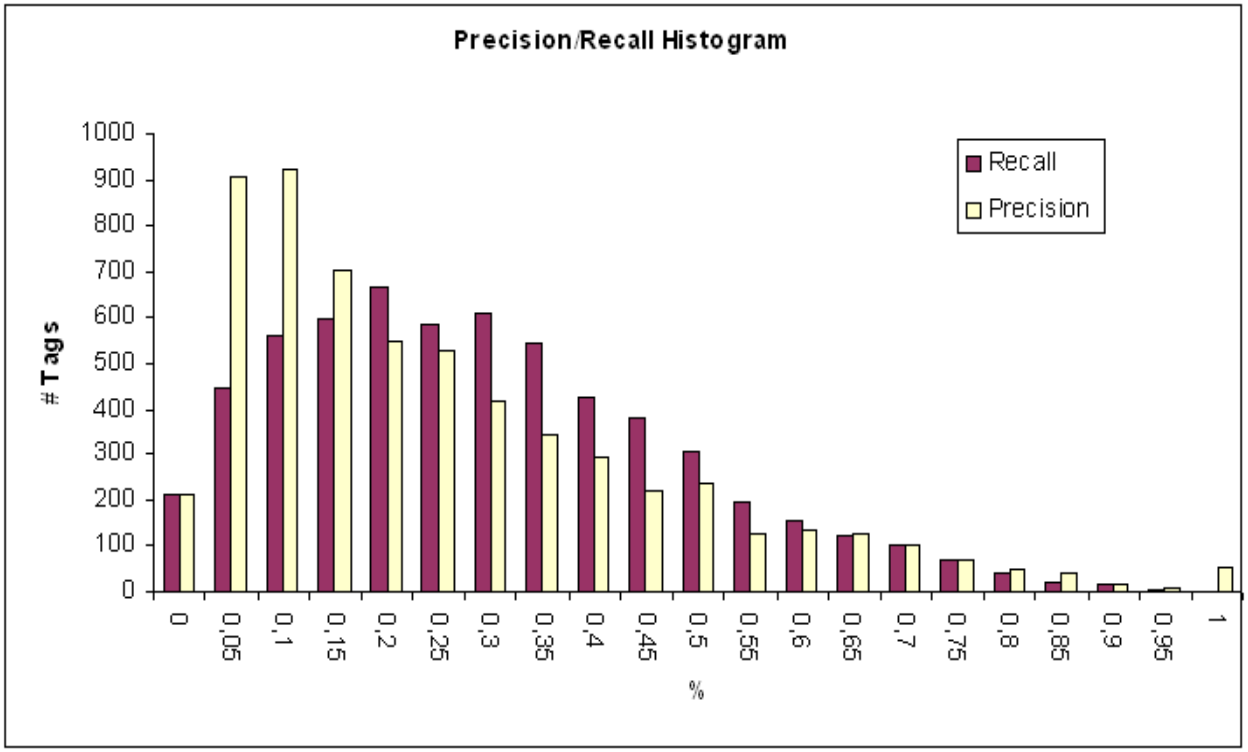
Conclusion

- Large number of tags are specific to users or groups of users.
- Personal vocabulary is integrated in larger structure
  - perhaps even (intermediate) community vocabulary
- Sub communities have to be taken into account for query expansion, etc.

Retrieval based on Folksonomies

- Research question: Does a folksonomy provide added value?
- Approach:
  - Tags assignment provides ‘ground truth’
  - Title (and description) get searched
  - Done for the 6000 most frequent tags

Retrieval based on Folksonomies



Precision & Recall for title only search

Conclusions

- Precision and recall mostly remain below 0.5 in this test
- Adding the description performance even decreases
  - Only 20% of the bookmarks have a description assigned
- But it shows: Tags are not redundant and provide ‘added value’ for retrieval

Overall Conclusions

- Power law for co-occuring tags applies to ~ 80% of the tags
  - Open question: Which 80%?
- User and resource based tagging statistics indicate a ‘more complex’ underlying structure in folksonomies
  - Open question: Are there sub communites and how can we identify them?
- Tags are not redundant
  - Retrieval has ‘added value’
  - Open question: Does this added value increase retrieval performance?

Questions?

Are there any questions left?

Contact:

- Mathias Lux, mlux@itec.uni-klu.ac.at