WIT Wissenschafterinnenkolleg Internettechnologien

TU WIEN Technische Universität Wien
Vienna University of Technology

A Case Study on Emergent Semantics in Communities

Elke Michlmayr,
Women's Postgraduate College for Internet Technologies (WIT),
Vienna University of Technology,
http://wit.tuwien.ac.at

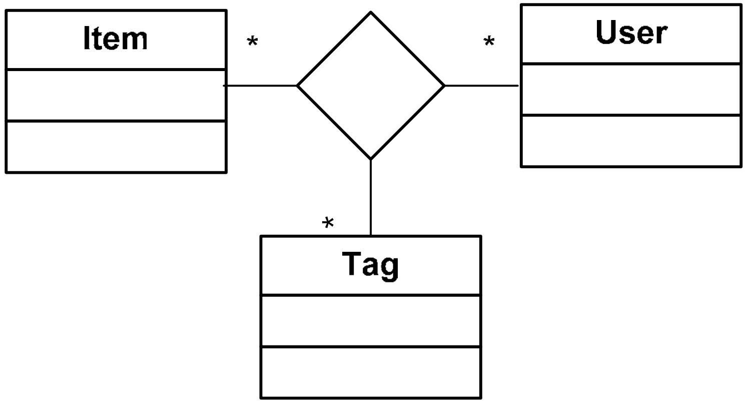
fFORTE bm:bwk esf

Outline

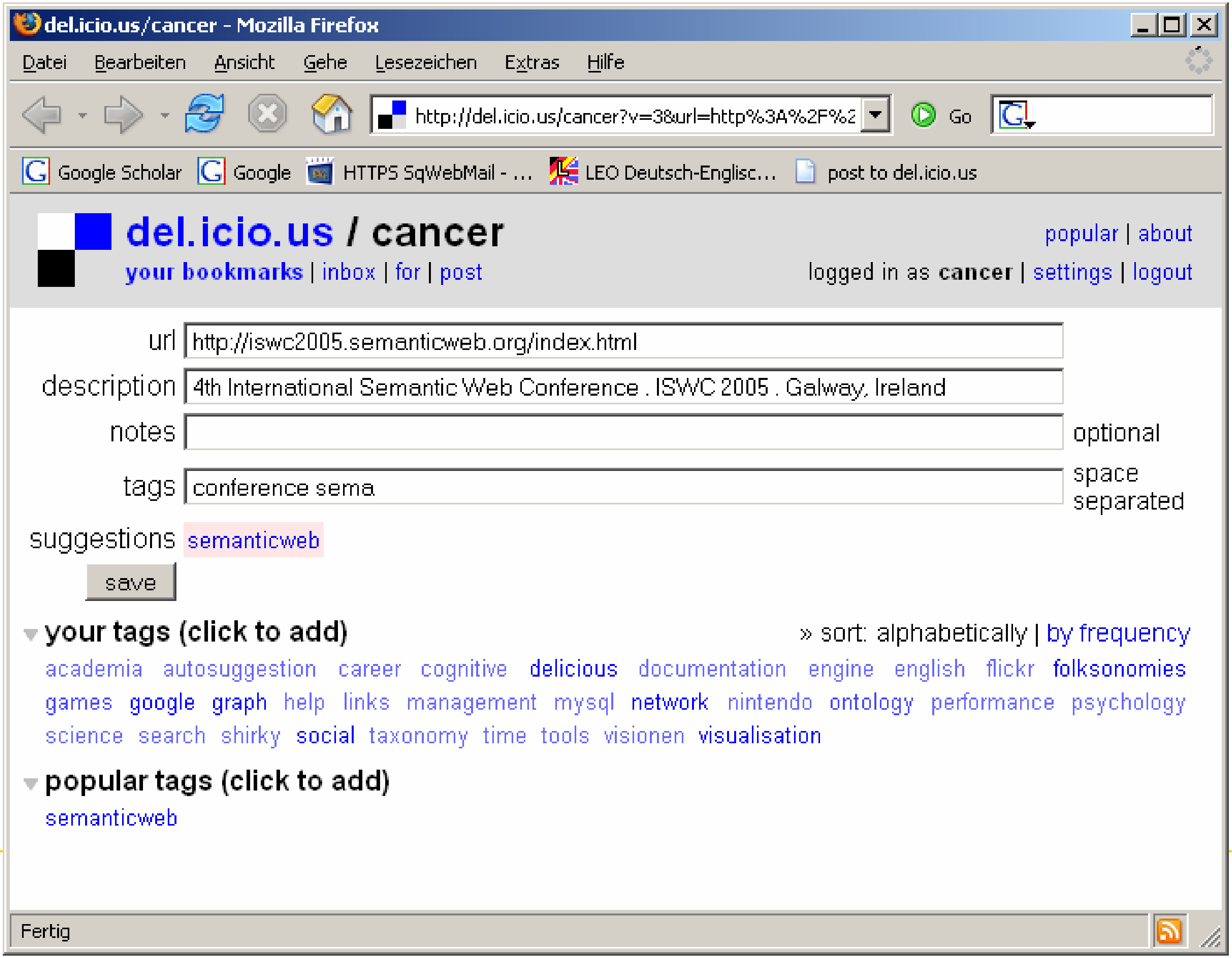
- Folksonomies
  - What are they?
- Comparison to taxonomies
  - Methodology
  - On the data level
- Folksonomies and peer-to-peer networks
  - User behaviour
  - Usable as test data?
- Related work
- Summary

Folksonomies and (collaborative) tagging

- Multi-user web applications that provide a simple categorization system
- Items
  - Web pages (Deli.cio.us, Furl, ...)
  - Images (Flickr)
  - Citations (Connotea, CiteULike)
- Tags = keywords
  - Can be chosen freely
- Every user has a web page with a list of own items
  - Sorted in reverse-chronological order
  - Can be filtered by tag(s)
- Public access to item collections and meta-data



Example: del.icio.us user interface



del.icio.us/cancer - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://del.icio.us/cancer?v=3&url=http%3A%2F%2
Go
Google Scholar
Google
HTTPS SqWebMail - ...
LEO Deutsch-Englisc...
post to del.icio.us

del.icio.us / cancer

your bookmarks | inbox | for | post

popular | about

logged in as **cancer** | settings | logout

url: http://iswc2005.semanticweb.org/index.html

description: 4th International Semantic Web Conference . ISWC 2005 . Galway, Ireland

notes: optional

tags: conference sema (space separated)

suggestions: semanticweb

save

your tags (click to add)

» sort: alphabetically | by frequency

academia autosuggestion career cognitive delicious documentation engine english flickr folksonomies games google graph help links management mysql network nintendo ontology performance psychology science search shirky social taxonomy time tools visionen visualisation

popular tags (click to add)

semanticweb

“Bottom-up” approach to categorization

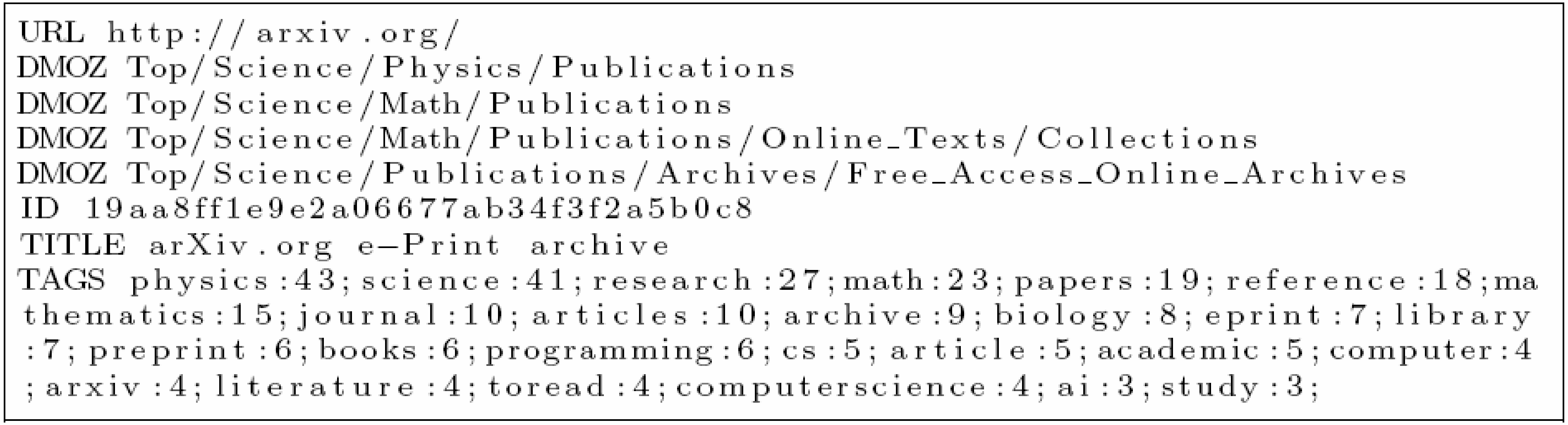
- No pre-defined model or hierarchy
- Inconsistencies
  - Synonyms, homonyms
  - Singular and plural versions of a tag
  - Keywords that consist of two terms
    - i.e., semantic web, semantic_web, semanticweb
- Relies on aggregation of meta-data
  - Tag frequency distribution
    - Tags most often used to annotate an item categorize it best
    - No need to reach consensus
  - Relationships between tags evolve from meta-data
- Amount of meta-data crucial!
  - Number of users, lifetime of folksonomy

Comparison of meta-data

- Lots of discussions about taxonomies vs. folksonomies, e.g., Clay Shirky 2005
- Experiment: compare meta-data from two big community projects that categorize Web pages to find out about the differences
  - DMOZ open directory project http://dmoz.org/
    - Taxonomy for Web pages
    - ~600000 concepts and ~5000000 instances
    - Available in RDF format (two big files)
  - Social bookmarking site http://del.icio.us/
    - No official numbers, ~100000 users
    - RDF file for each collection and for each item

Data gathering

- Procedure
  - Use only items from del.icio.us that were annotated by more than 100 users (= popular items)
  - Download random popular items from del.icio.us
  - Lookup if items are present in the DMOZ collection
    - ~25 % of the items were also present in DMOZ
- 788 items with meta-data from both sources
  - ~50 % of them are instances of DMOZ concept `Top/Computers`



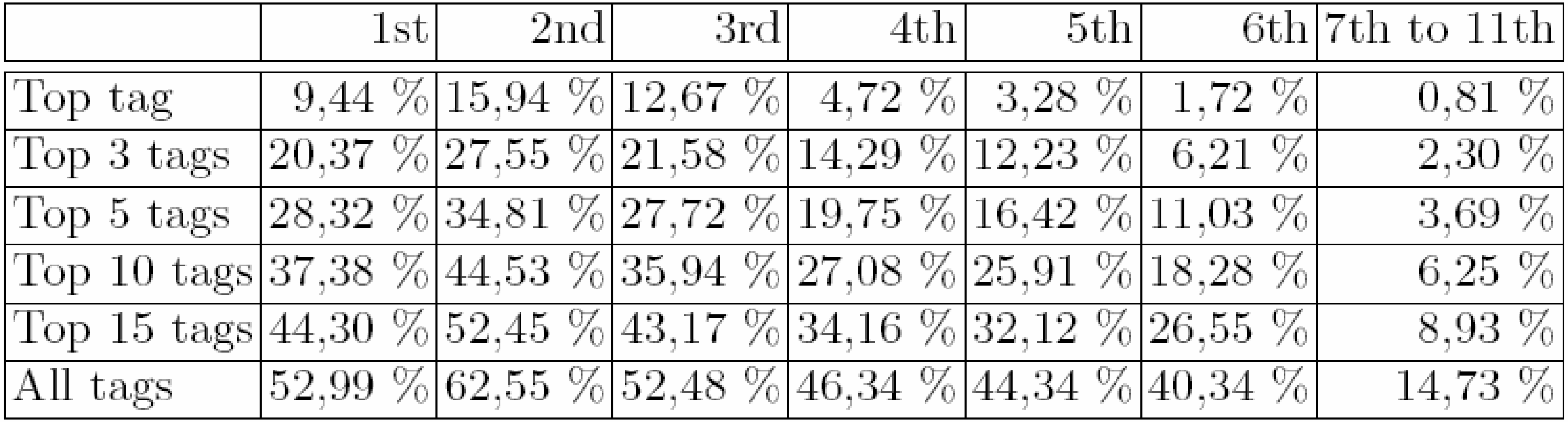
```
URL http://arxiv.org/
DMOZ Top/Science/Physics/Publications
DMOZ Top/Science/Math/Publications
DMOZ Top/Science/Math/Publications/Online_Texts/Collections
DMOZ Top/Science/Publications/Archives/Free_Access_Online_Archives
ID 19aa8ff1e9e2a06677ab34f3f2a5b0c8
TITLE arXiv.org e-Print archive
TAGS physics:43;science:41;research:27;math:23;papers:19;reference:18;ma
thematics:15;journal:10;articles:10;archive:9;biology:8;eprint:7;library
:7;preprint:6;books:6;programming:6;cs:5;article:5;academic:5;computer:4
;arxiv:4;literature:4;toread:4;computerscience:4;ai:3;study:3;
```

Preparation of data

- Preparations
  - Convert to lower case, remove underscores and hyphens
  - Remove last character s because of singular/plural tags
  - Don’t consider Top/World (multi-lingual categories)
  - Remove all categories with one character only (/A - /Z)
  - Remove Top category
  - Sort category names in reverse to put most specific entry first
  - Rank tags by number
- Example
  - Top/Science/Math/Publication -> publication math science
- How to compare?
  - Avg. DMOZ hierarchy length: 4,67
  - Avg. deli.cio.us tags per item: 24,59

Comparison

- Lookup for each DMOZ category
  - Is it included in the del.icio.us tags?
- Take top 1, 3, 5, 10, 15, all tags into account
  - Top tag is included in ~50% of the cases
  - Top 5 is the fairest comparison
  - Top tags match more often than the less popular ones



	1st	2nd	3rd	4th	5th	6th	7th to 11th
Top tag	9,44 %	15,94 %	12,67 %	4,72 %	3,28 %	1,72 %	0,81 %
Top 3 tags	20,37 %	27,55 %	21,58 %	14,29 %	12,23 %	6,21 %	2,30 %
Top 5 tags	28,32 %	34,81 %	27,72 %	19,75 %	16,42 %	11,03 %	3,69 %
Top 10 tags	37,38 %	44,53 %	35,94 %	27,08 %	25,91 %	18,28 %	6,25 %
Top 15 tags	44,30 %	52,45 %	43,17 %	34,16 %	32,12 %	26,55 %	8,93 %
All tags	52,99 %	62,55 %	52,48 %	46,34 %	44,34 %	40,34 %	14,73 %

Folksonomies and peer-to-peer networks

- Architectures are very different
  - Folksonomies are centralized systems, aggregation is easy
  - Peer-to-peer networks are distributed, aggregation is hard
- User behaviour is comparable
  - Act autonomously
  - No central authority
  - Want to share information
- Data from a folksonomy can be used to model peers and content distribution
  - No data about queries available
- Experiment
  - Can subsets of the del.icio.us data be selected in such a way that the principle of interest-based locality be observed in these subsets?

Can interest-based locality be observed?

- Interest-based locality
  - "If peer A has a particular piece of content peer B is interested in, it is likely the case that the other information items stored by peer A are also of interest to peer B."
- Method
  1. Retrieve all users from del.icio.us that store a random bookmark
  2. Retrieve all their collections
- Retrieved 4 test sets
  - 155, 248, 280, 551 users
  - Distribution of items among users nearly equal in the test sets
  - Avg.: 84% of items are not shared!

Not shared	84 %
By 2 users	8.9 %
By 3 users	2.92 %
By 4 users	1.49 %
5-10 users	2.18 %
> 10 users	0.51 %

Related work

- Adam Mathes, 2004: *Folksonomies – Cooperative Classification and Communication Through Shared Metadata*
    - Very good introduction
- Clay Shirky, 2005: *Ontology is Overrated: Categories, Links and Tags*
    - Controversial discussion of taxonomies vs. folksonomies
- Scott Golder and Bernardo Huberman, 2005: *The structure of Collaborative Tagging Systems*
    - Cognitive aspects
    - Data analysis: Tag frequency distribution for an item is stable over time

Summary

- Investigated the properties of meta-data provided by a folksonomy
- Compared it to DMOZ data collection
- Tried to find interest-based locality
- Paper contains some other experiments I did not have time to tell you about
- Open questions
  - Is there a way to combine the bottom-up and top-down approach for creating metadata?
  - How much could the semantic web benefit from it?