Ontology Learning from Folksonomies

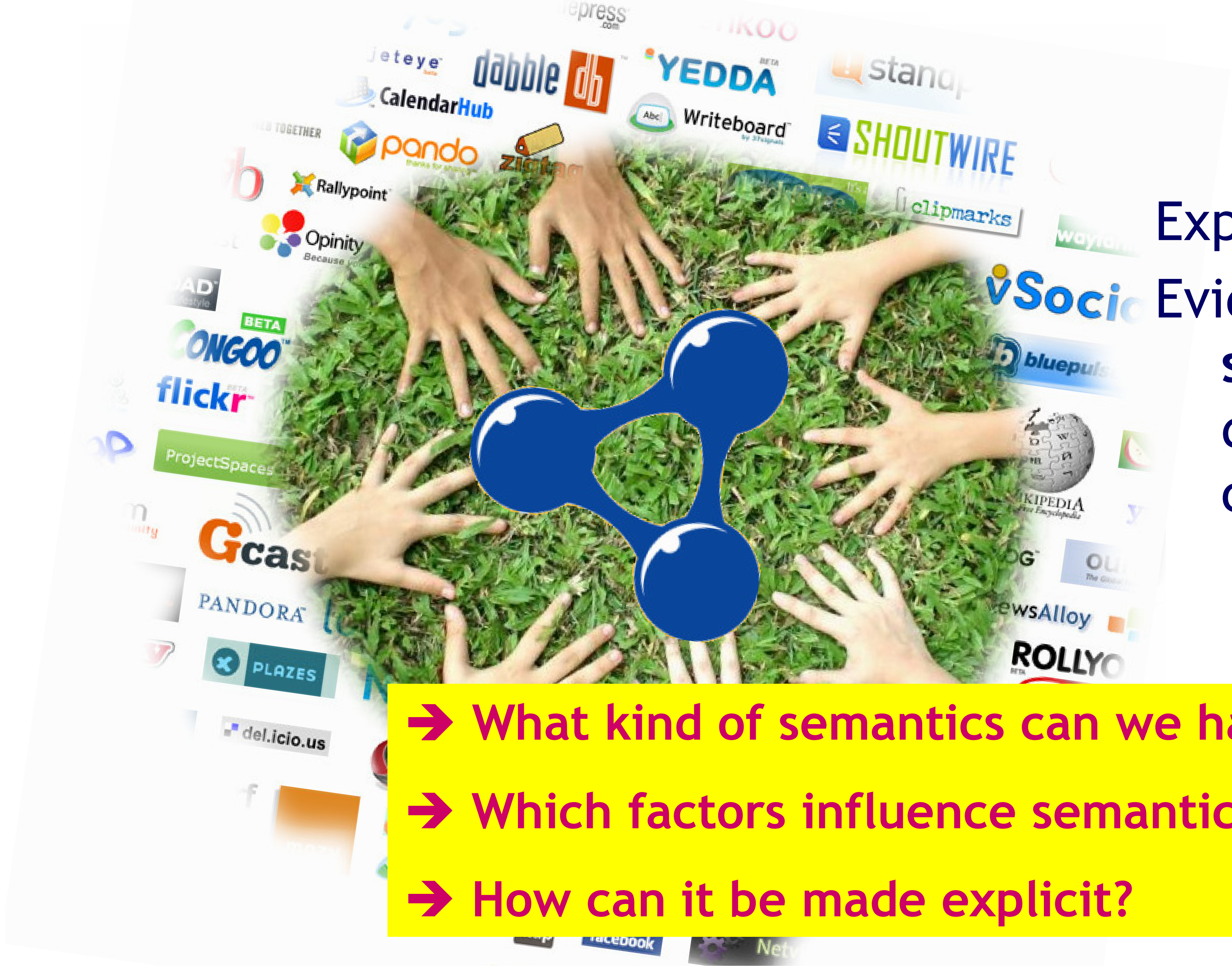
Tutorial at ICFCA 2011, Nicosia

*Andreas Hotho[1], **Robert Jäschke[2]***

[1]Data Mining and Information Retrieval Group, University of Würzburg
[2]Knowledge & Data Engineering Group, University of Kassel

Where do Semantics come from?

Semantically annotated content is the „fuel" of the **next generation** World Wide Web – but where is the petrol station?



Expert-built → expensive

Evidence for **emergent semantics** in Web2.0 data → Built by the crowd!

→ What kind of semantics can we harvest?

→ Which factors influence semantics?

→ How can it be made explicit?

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

Research and Topic Tag Cloud

Tag Recommender

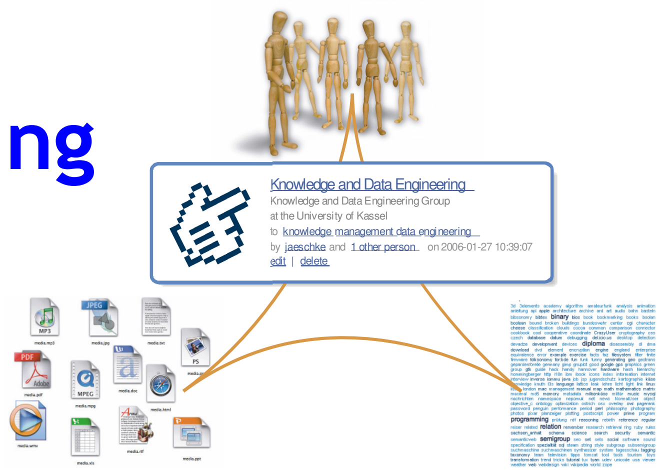
Semantic

Formal Concept Analysis

Semantic Web

Web 2.0

Data Mining



LogSonomies

Ontology Learning

Tag Similarity

Community detection

Trend detection

Ranking

Spam

Information Retrieval

Graph Structures

Social Network Analysis

Statistical Physics

Definition: Web 2.0

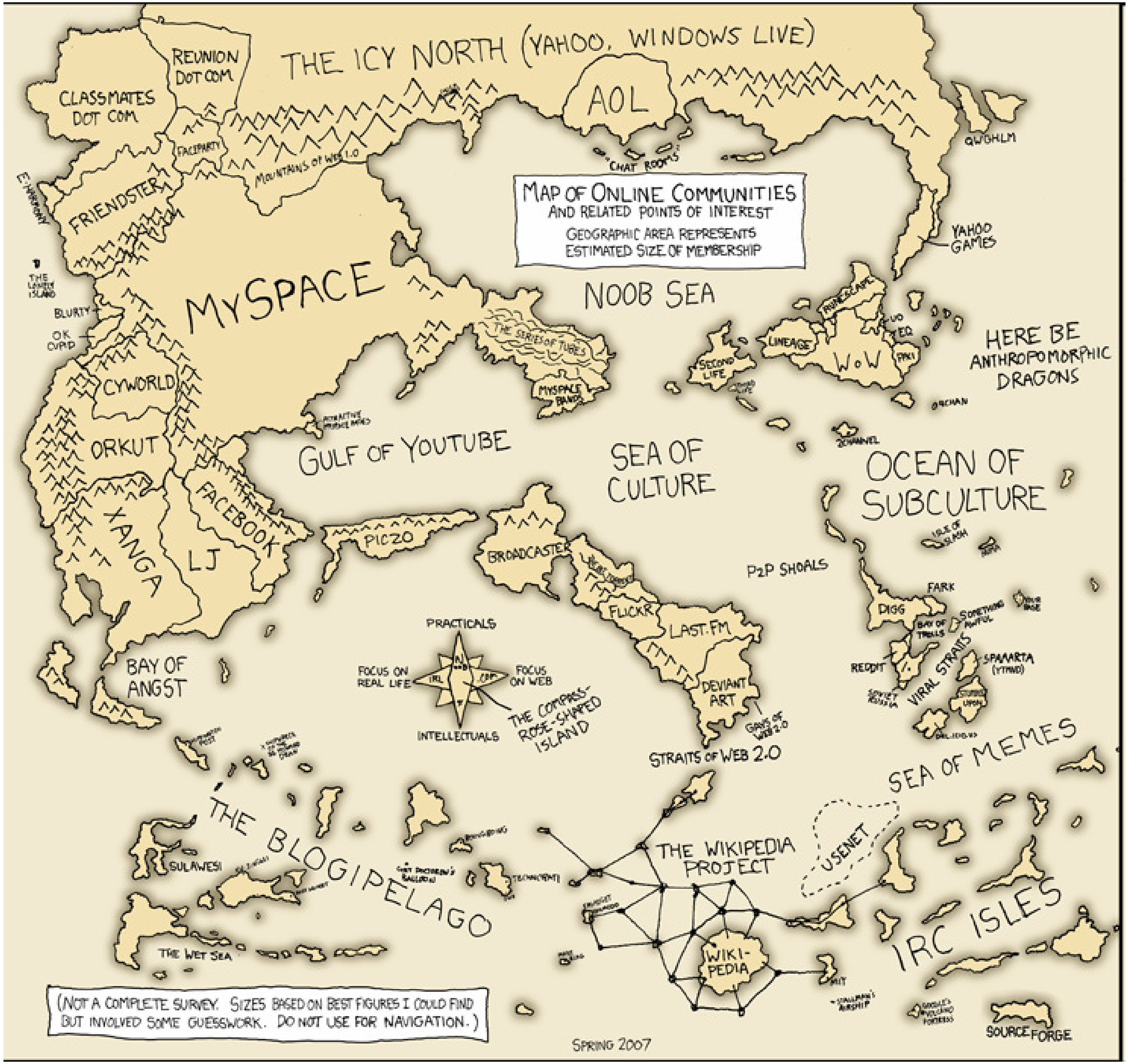
“The term **Web 2.0** is commonly associated with web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web.

Although the term suggests a new version of the World Wide Web, it does not refer to an update to any technical specifications, but rather to cumulative changes in the ways software developers and end-users use the Web.“

Wikipedia
http://en.wikipedia.org/wiki/Web_2.0

- The term was coined in 1999 by Darcy DiNucci in her article „Fragmented Future“.
- Tim O'Reilly shaped it by his work „What is Web 2.0“ (Sep. 2005) and the Web 2.0 conference in 2004.

A Map of the Web 2.0



- Blogs
- Wikis
- Bookmarking
- Youtube
- Flickr
- 43Things
- MySpace
- Facebook
- …

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

In this tutorial we will focus on collaborative tagging, in particular on social bookmarking:

- everybody knows (web) bookmarks
- has them in his/her own browser
- uses them on a daily basis
- bookmark repositories emerge totally independent

Interesting source of data which can be analyzed by using data mining and machine learning methods for (semi-)automatically learning ontologies

Web Bookmarks

del.icio.us / cahnmedia / hornsby

popular | recent

login | **register** | **help**

cahnmedia's items tagged **hornsby** → view **all**, **popular**

del.icio.us search

« earlier | later » showing the only item

BRUCE HORNSBY: Looking Back, Moving Forward save this

New one-hour music intensive radio special features Bruce Hornsby's exclusive comments on his favorite live performances from throughout his 20-year career.

to cahnmedia radio specials aaa americana npr noncom eclectic hornsby brucehornsby piano treshombres ... on nov 15

« earlier | later » showing the only item

» showing **10**, 25, 50, 100 items per page

del.icio.us | about | blog | terms of service | privacy policy | copyright policy | contact us | RSS feed for this page

related tags

1 + aaa
1 + americana
1 + brucehornsby
1 + cahnmedia
1 + eclectic
1 + noncom
1 + npr
1 + piano
1 + radio
1 + specials
1 + treshombres

FeaturedArtists

1 bobwills
1 brucehornsby
2 byrds
5 FooFighters
1 gramparsons
1 johnnycash
19 rdm
18 rythmsdelmundo
1 waylonjennings

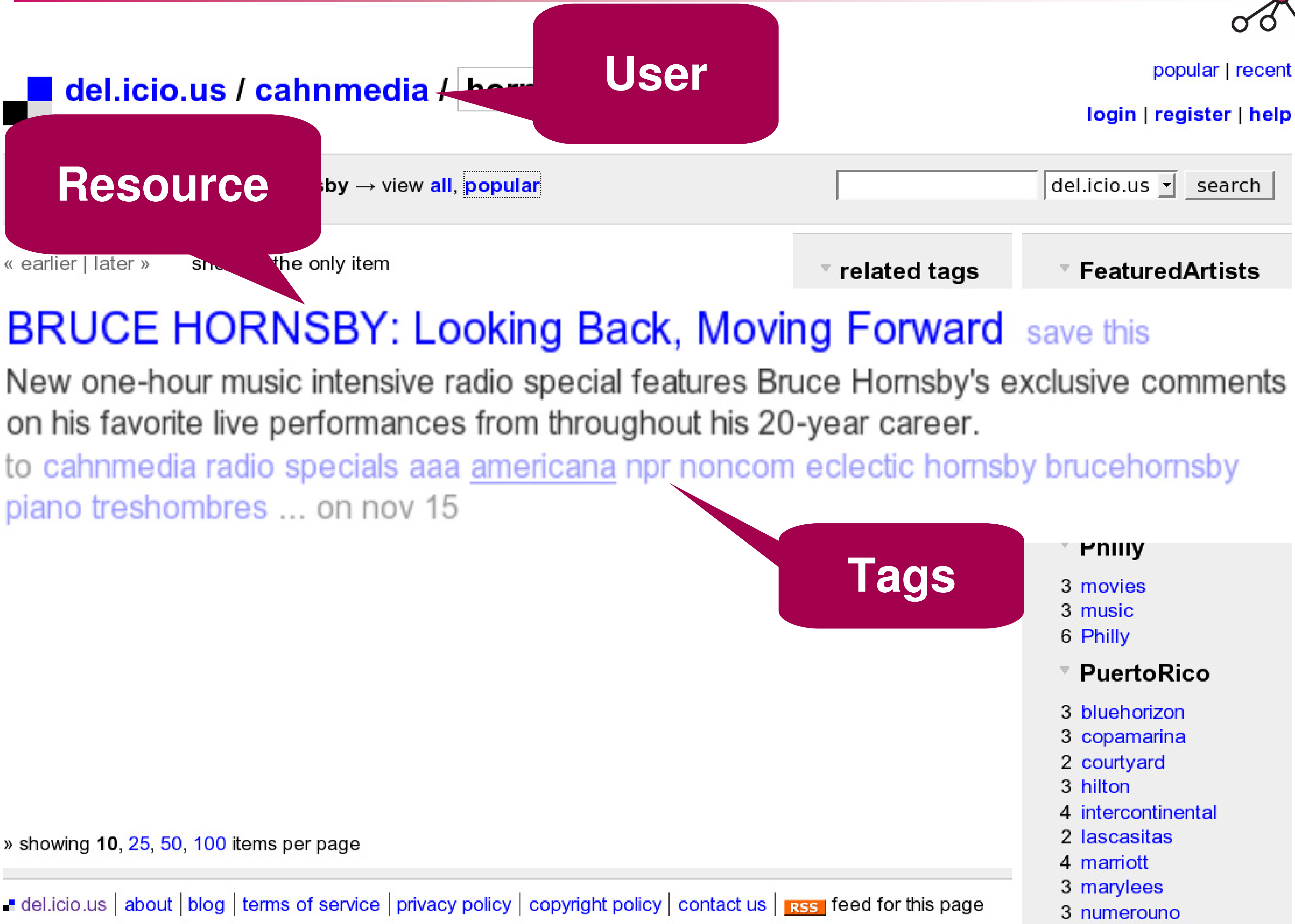
Philly

3 movies
3 music
6 Philly

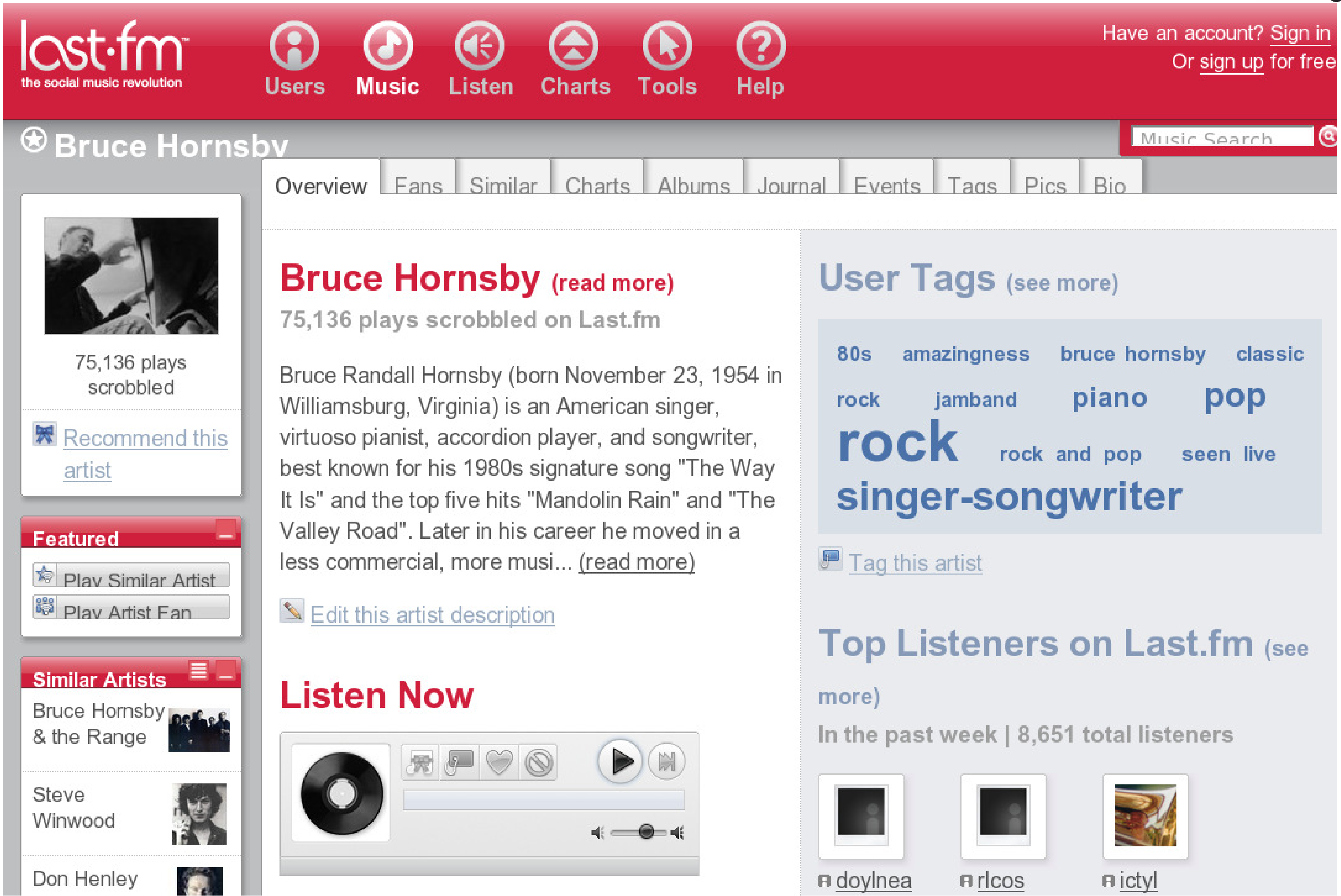
PuertoRico

3 bluehorizon
3 copamarina
2 courtyard
3 hilton
4 intercontinental
2 lascasitas
4 marriott
3 marylees
3 numerouno

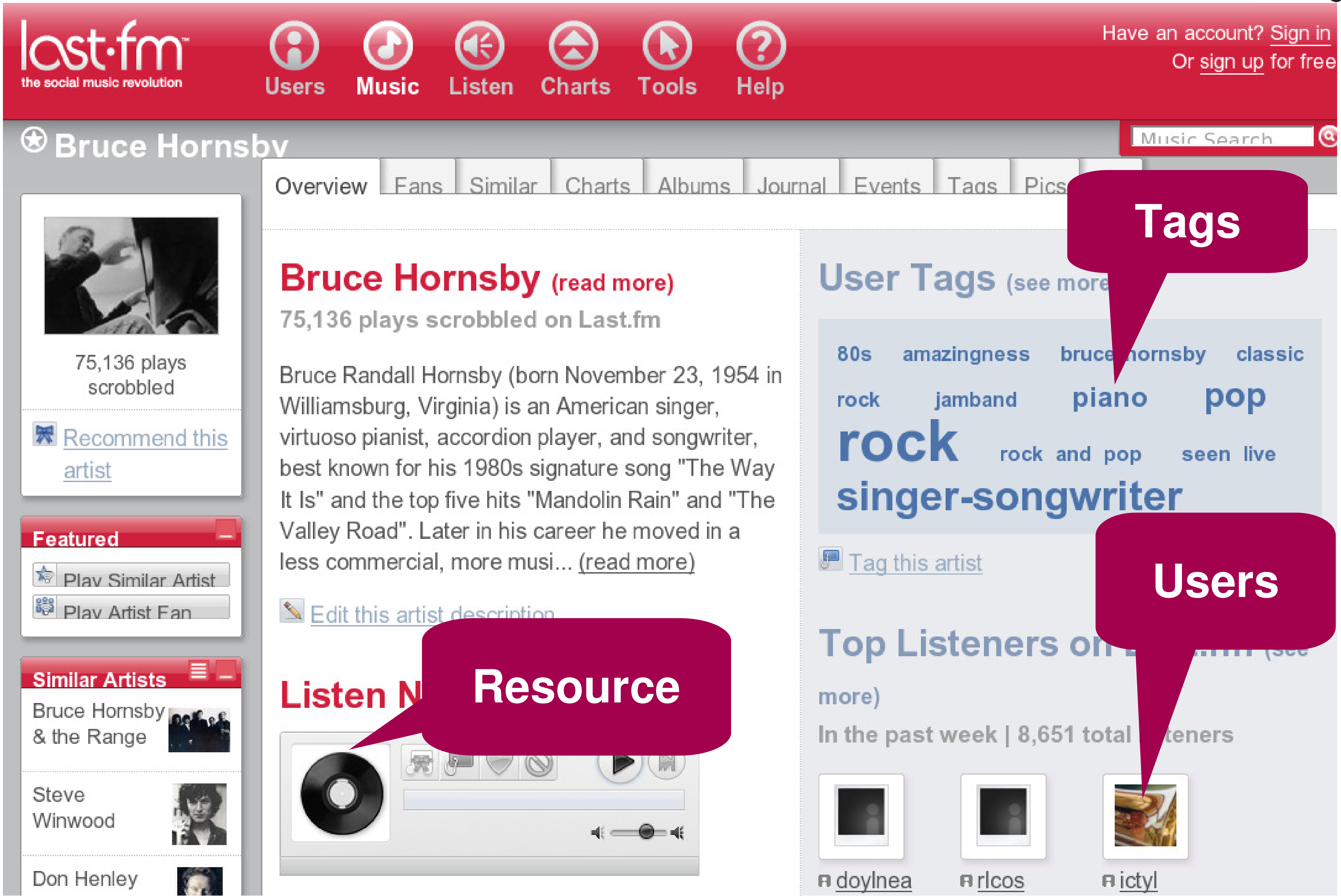
Web Bookmarks



Audio Streams



Audio Streams



Photos

flickr GAMMA

You aren't signed in Sign In Help

Home Learn More Sign Up! Explore

Search everyone's phot Search

Bruce Hornsby

Bruce Hornsby

Would you like to comment?

Sign up for a free account, or sign in (if you're already a member).

Uploaded on July 22, 2005 by **Shotlivephoto**

Shotlivephoto's photostream

Bruce Hornsby (Set)

more browse more

4 photos View as slideshow

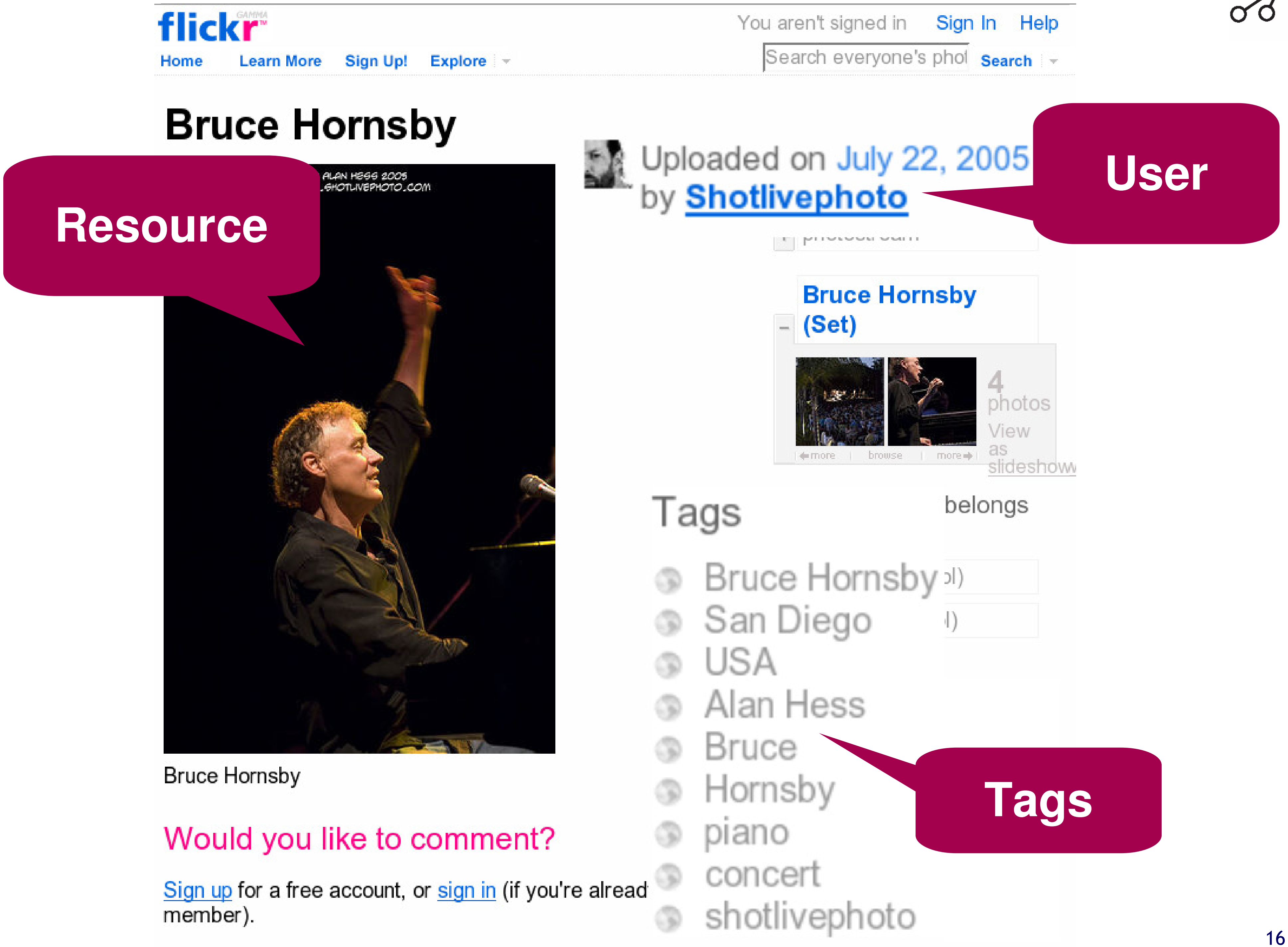
This photo also belongs to:

- San Diego (Pool)
- musicians (Pool)

Tags

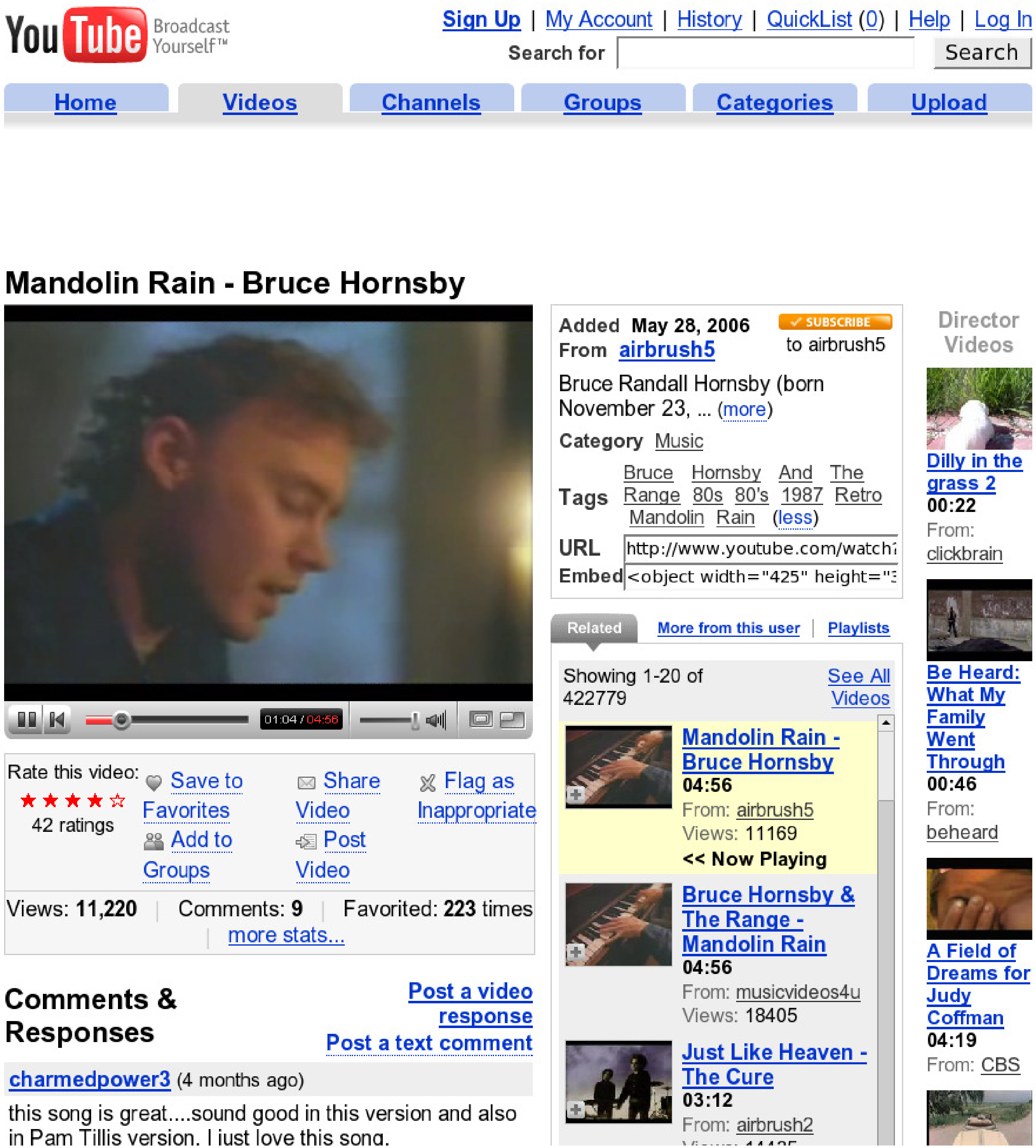
- Bruce Hornsby
- San Diego
- USA
- Alan Hess
- Bruce
- Hornsby
- piano
- concert
- shotlivephoto

Photos



flickr
You aren't signed in
Sign In
Help
Home
Learn More
Sign Up!
Explore
Search everyone's phot
Search
Bruce Hornsby
Uploaded on July 22, 2005
by Shotlivephoto
Resource
User
ALAN HESS 2005
SHOTLIVEPHOTO.COM
Bruce Hornsby
(Set)
4
photos
View
as
slideshow
more
browse
more
Tags
belongs
Bruce Hornsby
San Diego
USA
Alan Hess
Bruce
Hornsby
piano
concert
shotlivephoto
Tags
Bruce Hornsby
Would you like to comment?
Sign up for a free account, or sign in (if you're alread
member).

Videos



Broadcast Yourself™
Sign Up | My Account | History | QuickList (0) | Help | Log In
Search for
Search
Home
Videos
Channels
Groups
Categories
Upload
Mandolin Rain - Bruce Hornsby
01:04 / 04:56
Rate this video:
42 ratings
Save to Favorites
Share Video
Flag as Inappropriate
Add to Groups
Post Video
Views: 11,220 | Comments: 9 | Favorited: 223 times
more stats...
Comments & Responses
Post a video response
Post a text comment
charmedpower3 (4 months ago)
this song is great....sound good in this version and also in Pam Tillis version. I just love this song.
Added May 28, 2006
From airbrush5
SUBSCRIBE to airbrush5
Bruce Randall Hornsby (born November 23, ... (more)
Category Music
Tags Bruce Hornsby And The Range 80s 80's 1987 Retro Mandolin Rain (less)
URL http://www.youtube.com/watch?
Embed <object width="425" height="3
Related
More from this user
Playlists
Showing 1-20 of 422779
See All Videos
Mandolin Rain - Bruce Hornsby
04:56
From: airbrush5
Views: 11169
<< Now Playing
Bruce Hornsby & The Range - Mandolin Rain
04:56
From: musicvideos4u
Views: 18405
Just Like Heaven - The Cure
03:12
From: airbrush2
Director Videos
Dilly in the grass 2
00:22
From: clickbrain
Be Heard: What My Family Went Through
00:46
From: beheard
A Field of Dreams for Judy Coffman
04:19
From: CBS

Videos

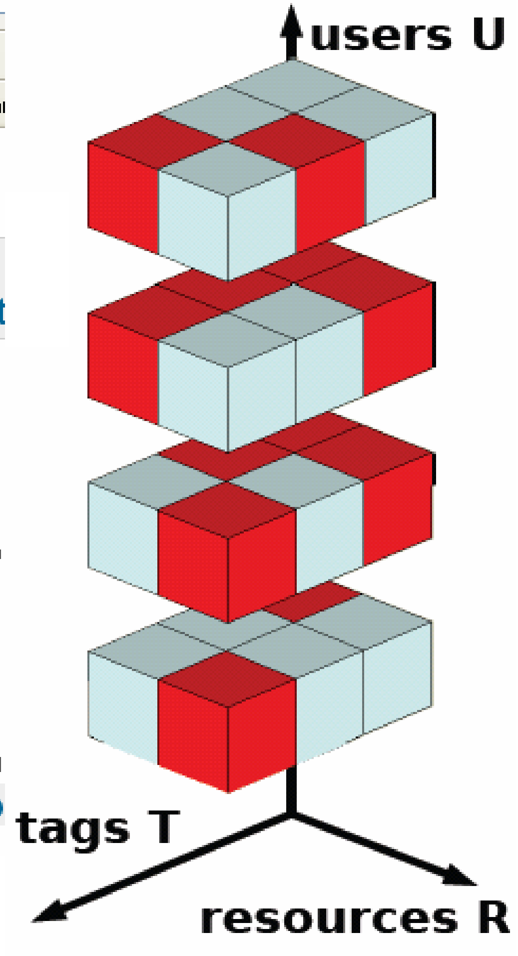
You Tube Broadcast Yourself™
Sign Up | My Account | History | QuickList (0) | Help | Log In
Search for
Search
Home
Videos
Channels
Groups
Categories
Upload
Resource
User
Rain - Bruce Hornsby
From airbrush5
Bruce Randall Hornsby (born November 23, ... (more)
Category Music
Tags Bruce Hornsby And The Range 80s 80's 1987 Retro Mandolin Rain (less)
Tags
Rate this video:
42 ratings
Save to Favorites
Share Video
Flag as Inappropriate
Add to Groups
Post Video
Views: 11,220 | Comments: 9 | Favorited: 223 times | more stats...
Comments & Responses
Post a video response
Post a text comment
charmedpower3 (4 months ago)
this song is great....sound good in this version and also in Pam Tillis version. I just love this song.
Related
More from this user
Playlists
Showing 1-20 of 422779
See All Videos
Mandolin Rain - Bruce Hornsby
04:56
From: airbrush5
Views: 11169
<< Now Playing
Bruce Hornsby & The Range - Mandolin Rain
04:56
From: musicvideos4u
Views: 18405
Just Like Heaven - The Cure
03:12
From: airbrush2
Be Heard: What My Family Went Through
00:46
From: beheard
A Field of Dreams for Judy Coffman
04:19
From: CBS

Folksonomies

Folksonomies allow **users**

to assign **tags**

to **resources.**

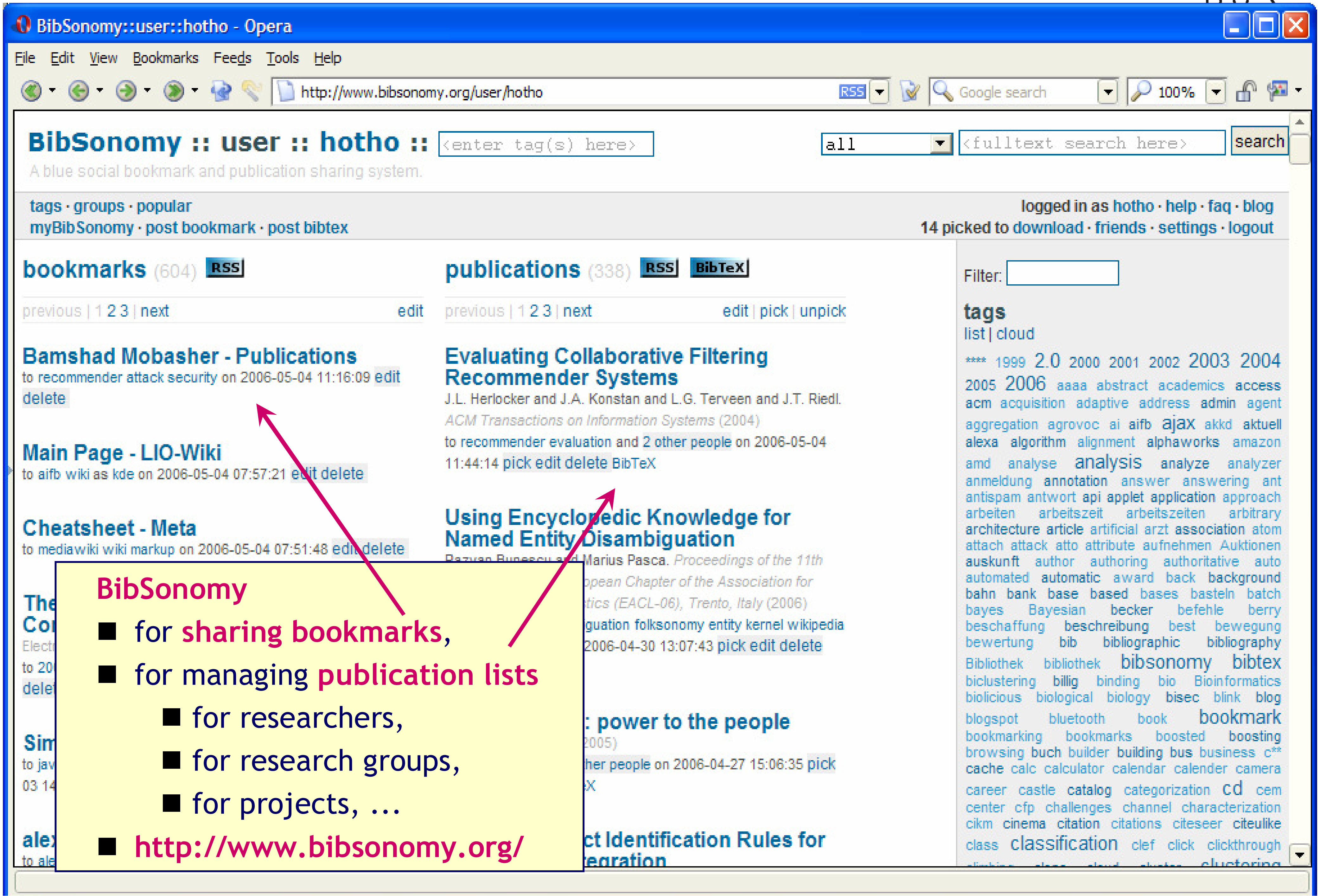


A *folksonomy* is a tuple $\mathbf{F} := (U, T, R, Y, \prec)$ where

- U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*,
- $Y \subseteq U \times T \times R$, called set of *tag assignments*,
- $\prec \subseteq U \times T \times T$ is a user-specific sub-tag/super-tag relation.

→ Without $\prec$ relation: tripartite hypergraph, triadic formal context, 3-dim. tensor

Our system: BibSonomy



BibSonomy

- for **sharing bookmarks**,
- for managing **publication lists**
  - for researchers,
  - for research groups,
  - for projects, …
- **http://www.bibsonomy.org/**

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

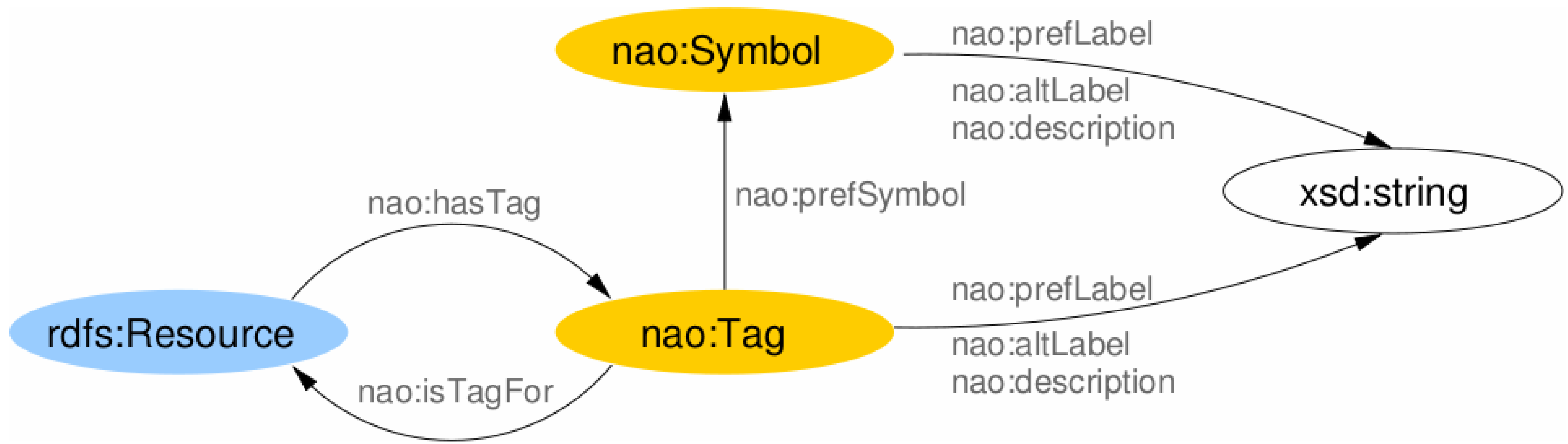
Folksonomies and Ontologies: Folksonomies are Ontologies

- Viewpoint: Folksonomies ***are*** lightweight ontologies

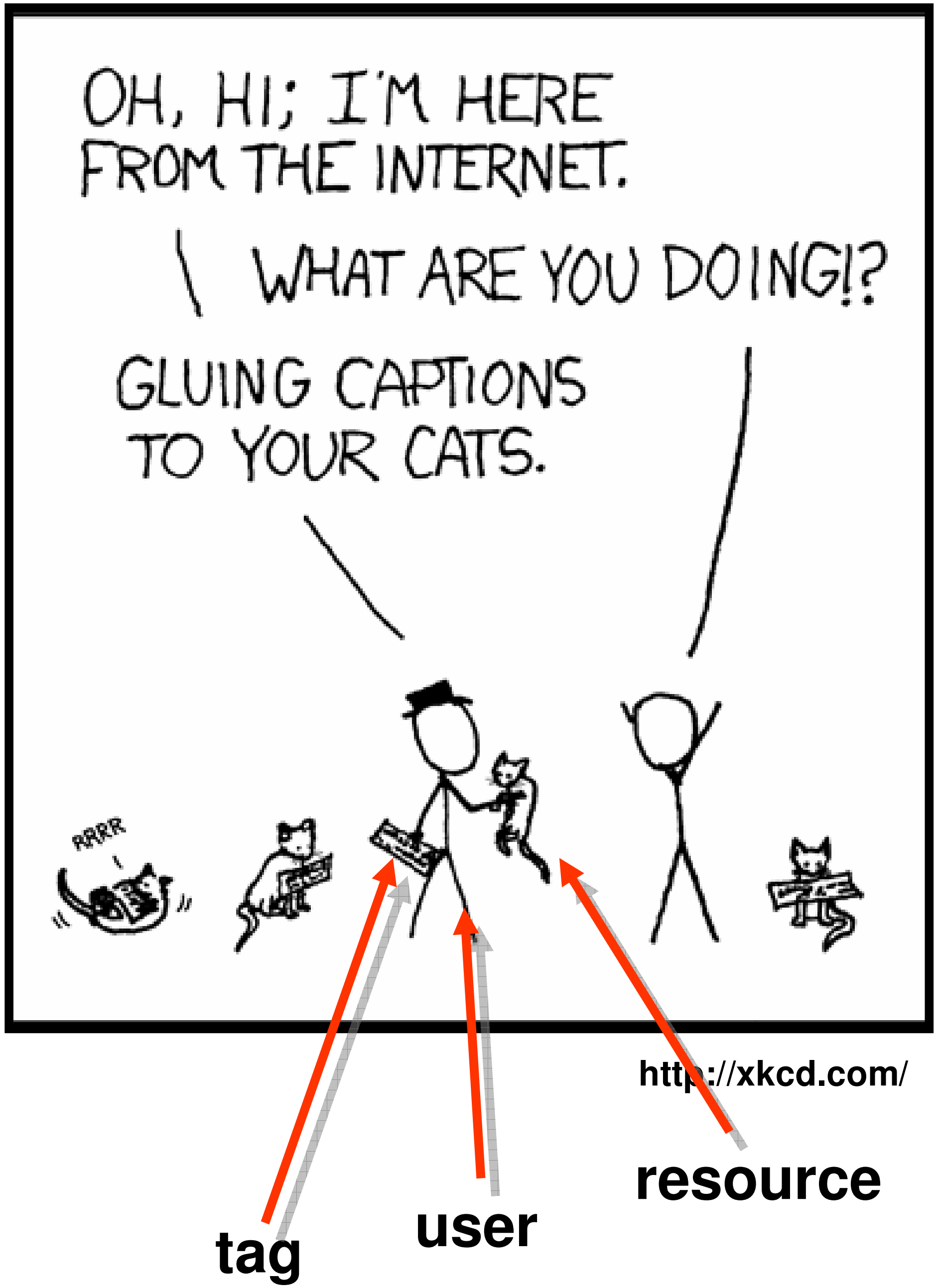
Overview (Java 2 Platform SE v1.4.2)
to java manual api programming reference by jaeschke and 3 other users on
Aug 19, 2005, 12:33 PM

- E.g., posts represent concepts
  - Resource = instance
  - Tags = terms

- Folksonomies can be represented using ontologies
- Several such ontologies available
- Overview:
  - The state of the art in tag ontologies: a semantic model for tagging and folksonomies by: Hak Lae Kim, Simon Scerri, John G. Breslin, Stefan Decker and Hong Gee Kim
- Example: representing a tagging in NAO (Nepomuk):



Folksonomies and Ontologies: Aspects of Tagging



- Tagging is a distributed process
- Tagging has a small cognitive overhead
- System contents can be browsed by tag
- The systems evolves in time: new resources, new users, new tags
- There may be an underlying social network, explicitly exposed or not
- The behavior of users is “selfish”
- Users are exposed to each other’s activity
- Users share implicit knowledge (language, cultural background)

- Emergence of the data happens in a ubiquitous way
- Data emergence in a distributed and independent way (no central control) - users are distributed

- Folksonomies:
  - Lightweight conceptualization
  - Shared vocabulary
  - Rather implicit

- Ontology learning methods extract knowledge and make it explicit

- Goals:
  - Benefit from huge amounts of data
  - Improve navigation, search, recommendation
  - Bridge the gap to the Semantic Web
  - Feed back semantics to improve folksonomies

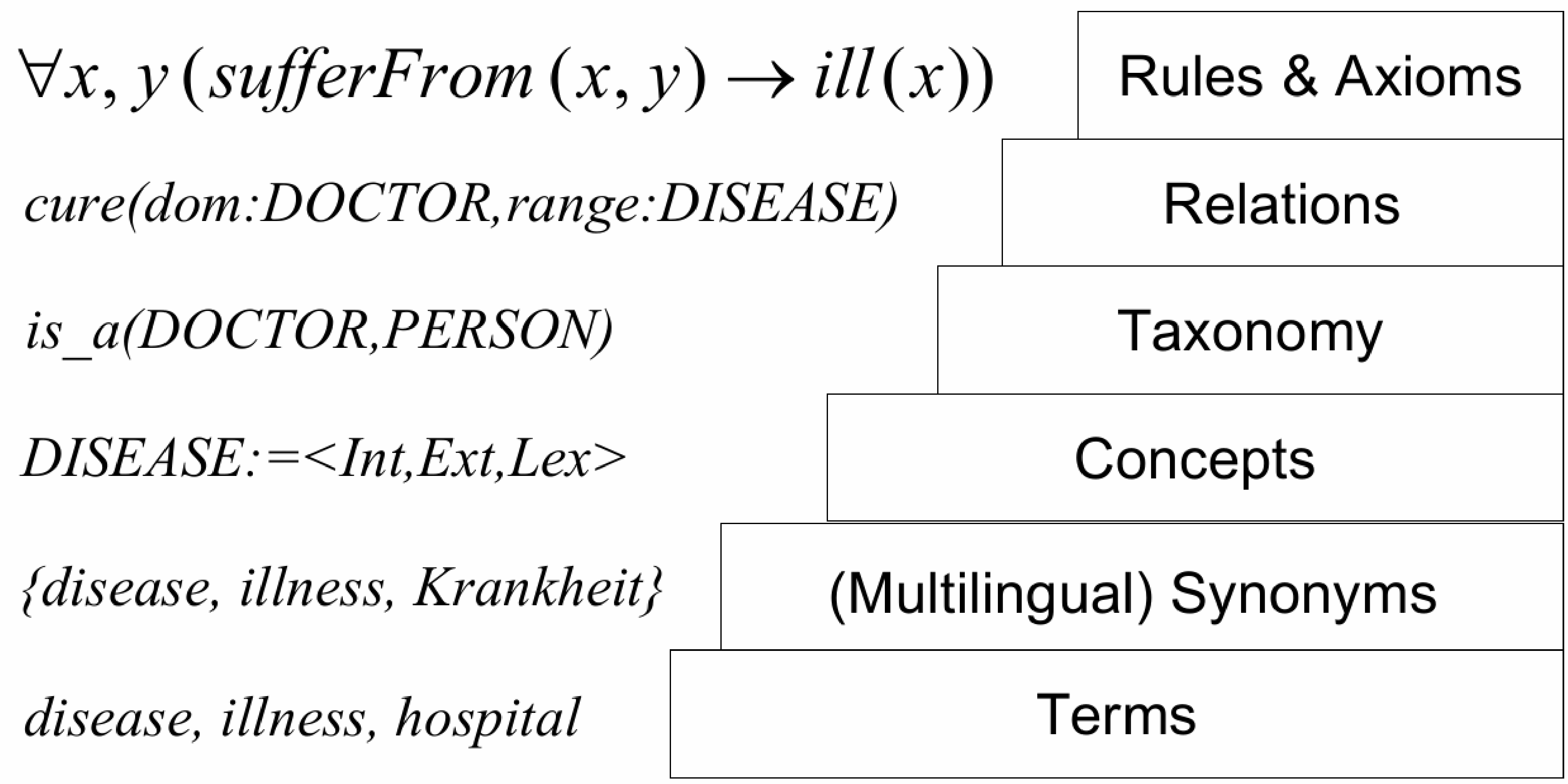
Ontology Learning from Text

- Techniques:
  - Linguistic analysis (NLP)
  - Data mining / machine learning
  - Statistics
  - Googling (i.e., asking the web)

- Overview:
  - P. Cimiano: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer, New York, 2006.

- OL methods for texts often don‘t fit in Folksonomies, because the sentence structure is missing

Ontology Learning Layer Cake



[Buitelaar, P., Cimiano, P. & Magnini, B.: Ontology Learning from Text: An Overview, IOS Press, 2005.]

Ontology Learning Layer Cake for Folksonomies

$\forall x, y\,(sufferFrom\,(x, y) \rightarrow ill(x))$	Rules & Axioms	
cure(dom:DOCTOR,range:DISEASE)	Relations	(✓)
is_a(teaching, education)	Taxonomy	✓
TEACHING := <Int, Ext, Lex>	Concepts	✓
{howto, how-to, guide}	(Multilingual) Synonyms	✓
howto guide programming	Tags	✓

Ontology Learning from the Web 2.0

Learning ontologies from …

- Wikis,
- Blogs,
- Micro blogging,
- Social networks,
- Social software

- … any other kind of Web 2.0 data except of tagging data is **not** the topic of this tutorial …

Ontology Learning from the Web 2.0

… but there is plenty of work dealing with Wikipedia

- S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. In ISWC/ASWC, LNCS 4825, Springer, 2007.
- R. Studer, M. Krötsch, D. Vrandecic, M. Völkel, H. Haller. Semantic Wikipedia. *Journal of Web Semantics*, *5*, 2007.
- M. Ruiz-Casado, E. Alfonseca and P. Castells, Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia. Proceedings of NLDB-2005. In *Natural Language Processing and Information Systems*, LNCS 3513, Springer, 2005.
- Simone Paolo Ponzetto , Michael Strube, Deriving a large scale taxonomy from Wikipedia, Proceedings of the 22nd National Conference on Artificial Intelligence, Vancouver, 2007.
- Suchanek, F. M., Kasneci, G., and Weikum, G. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semant*. 6, 3, 2008.

… or (micro) blogs

- S. Narayan, S. Prodanovic, M.F. Elahi, Z. Bogart. Population and Enrichment of Event Ontology using Twitter, Proceedings of the 1st Workshop on Semantic Personalized Information Management (SPIM 2010), Malta, 2010.
- M. Hepp. HyperTwitter: Collaborative Knowledge Engineering via Twitter Messages, Technical Report, 2010.
- C. Wagner, M. Strohmaier. The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams, Semantic Search 2010 Workshop (SemSearch2010), Raleigh, NC, USA, ACM, 2010.

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

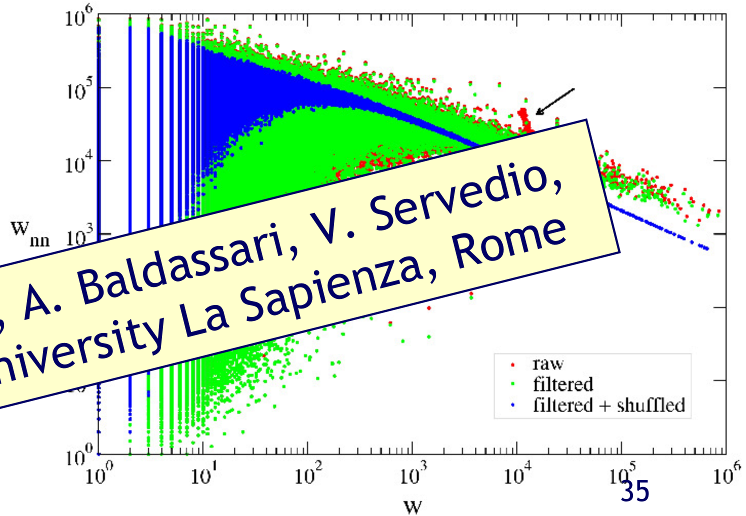
Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook



with C. Cattuto, A. Baldassari, V. Servedio,
V. Loreto, University La Sapienza, Rome

Data from the Delicious folksonomy site

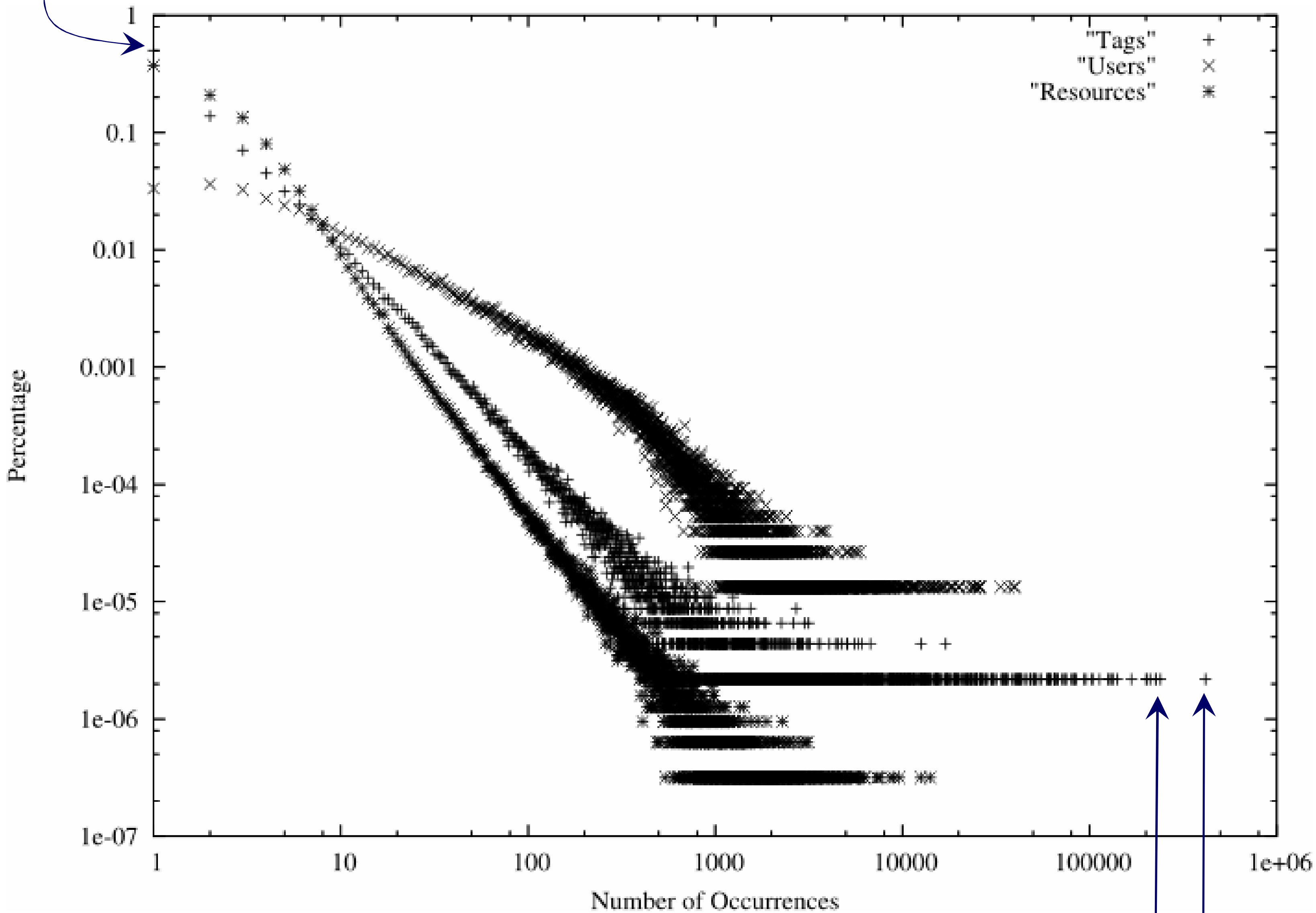
- Obtained in July 2005 (monthly dumps (14) June 2004 – July 2005)
- Consists of
  - |U| = 75,242 users
  - |T| = 533,191 tags
  - |R| = 3,158,297 resources
  - |Y| = 17,362,212 triples

Data from BibSonomy

- Latest obtained in July 2006 (20 monthly dumps)
- Consists of
  - |U| = 428 users
  - |T| = 13,108 tags
  - |R| = 47,538 resources
  - |Y| = 161,438 triples

Power Law Distribution in Delicious

approx. 40% of the tags occur only once



tag “web” occurs 238,891 times

tag “unlabeled” occurs 415,950 times

Milgram introduced the notion of a „small world“:

(Stanley Milgram. The small world problem. Psychology Today, 67(1):61–67, 1967.)

- Practical experiment in the US
- Any two person in the US are connected by a very short chain: six degrees of separation

Formal definition of the small world property for graphs:

- (Erdös) random graph
- Large clustering degree

Folksonomies exhibit small world properties:

- Small characteristic path lengths
- Large clustering degree (connectedness and cliquishness)

Networks of Tag Co-Occurrence

Consider tag-tag co-occurrences

- Link weight = number of common posts:

$$w(t_1, t_2) := |\{(u, r) \in U \times R \mid (t_1, u, r) \in Y, (t_2, u, r) \in Y\}|$$

Strength of a node t: total weight of its edges

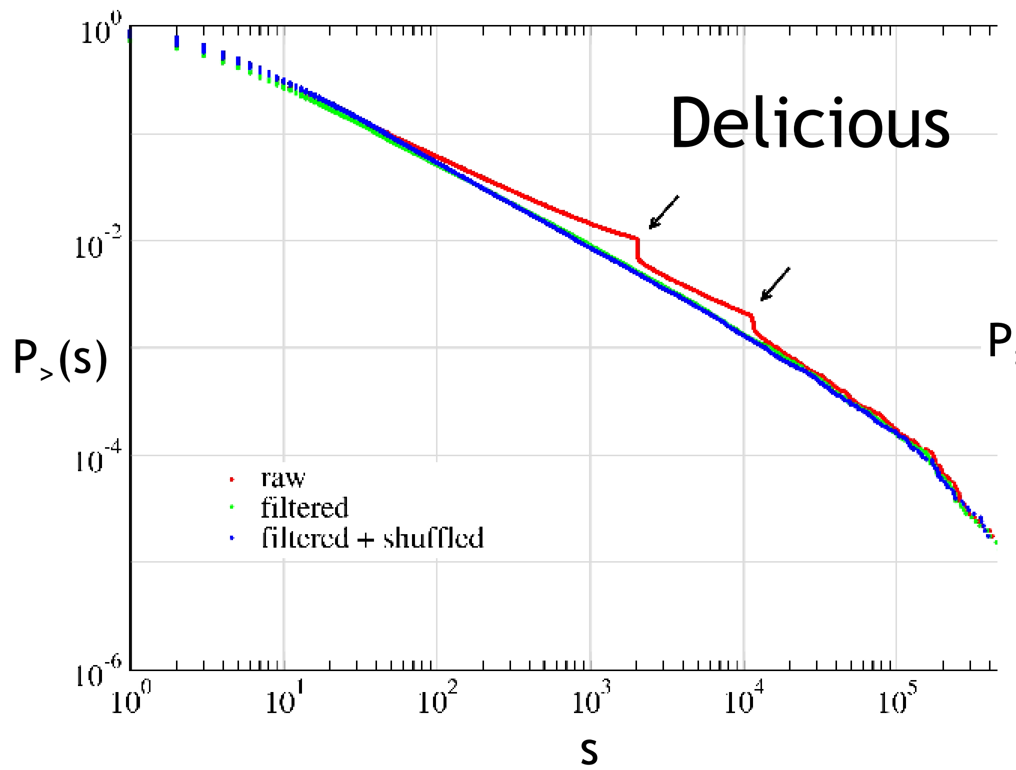
$$s_t := \sum_{t \neq t'} w(t, t')$$

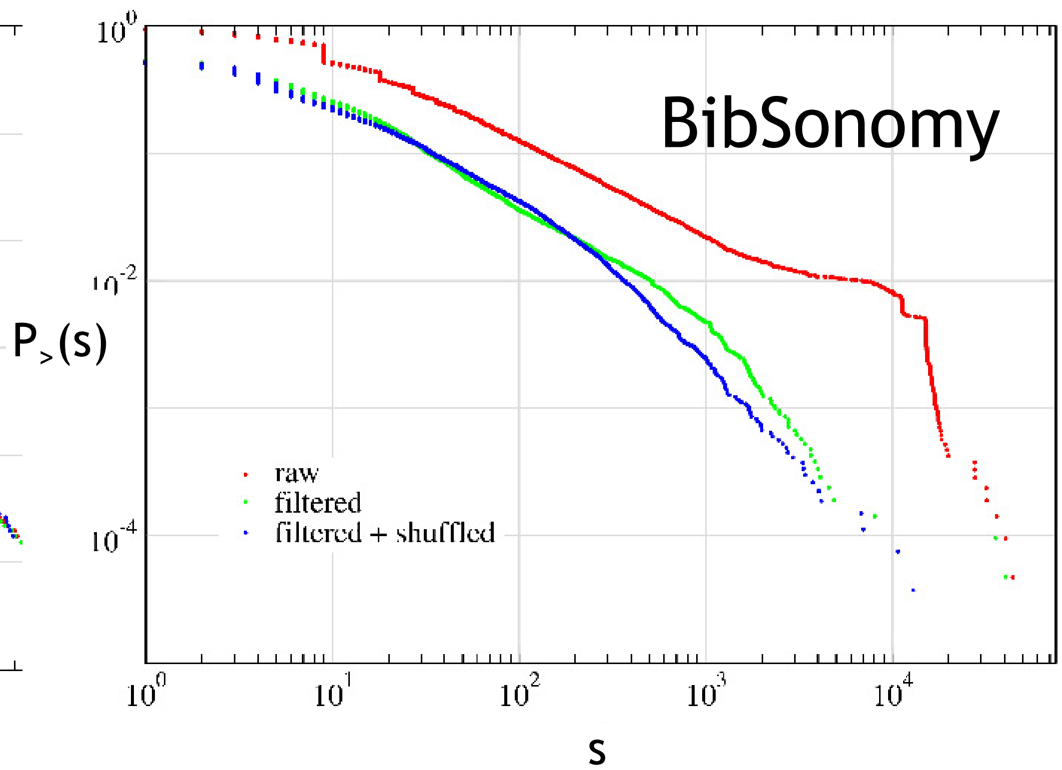
Examine cumulative strength distribution [Vazquez 2005]

$P_>(s)$:= probability of node strength exceeding s

Compare with shuffled graph: tags exchanged randomly between posts

Cumulative Strength Distribution





- Fat-tailed distribution
- Irregularities due to spamming activity, e.g.
  - Large number of tags per post
  - Regular number of tags (10, 50) per post
- Same distribution for shuffled tags
  - Behaviour determined solely by tag frequencies

Nearest-Neighbor Strength

Examine strength correlations between neighbors

Average nearest-neighbor strength for node *i*:

$$S_{nn}(s_i) := \frac{1}{k_i} \sum_{j=1}^{k_i} s_j$$

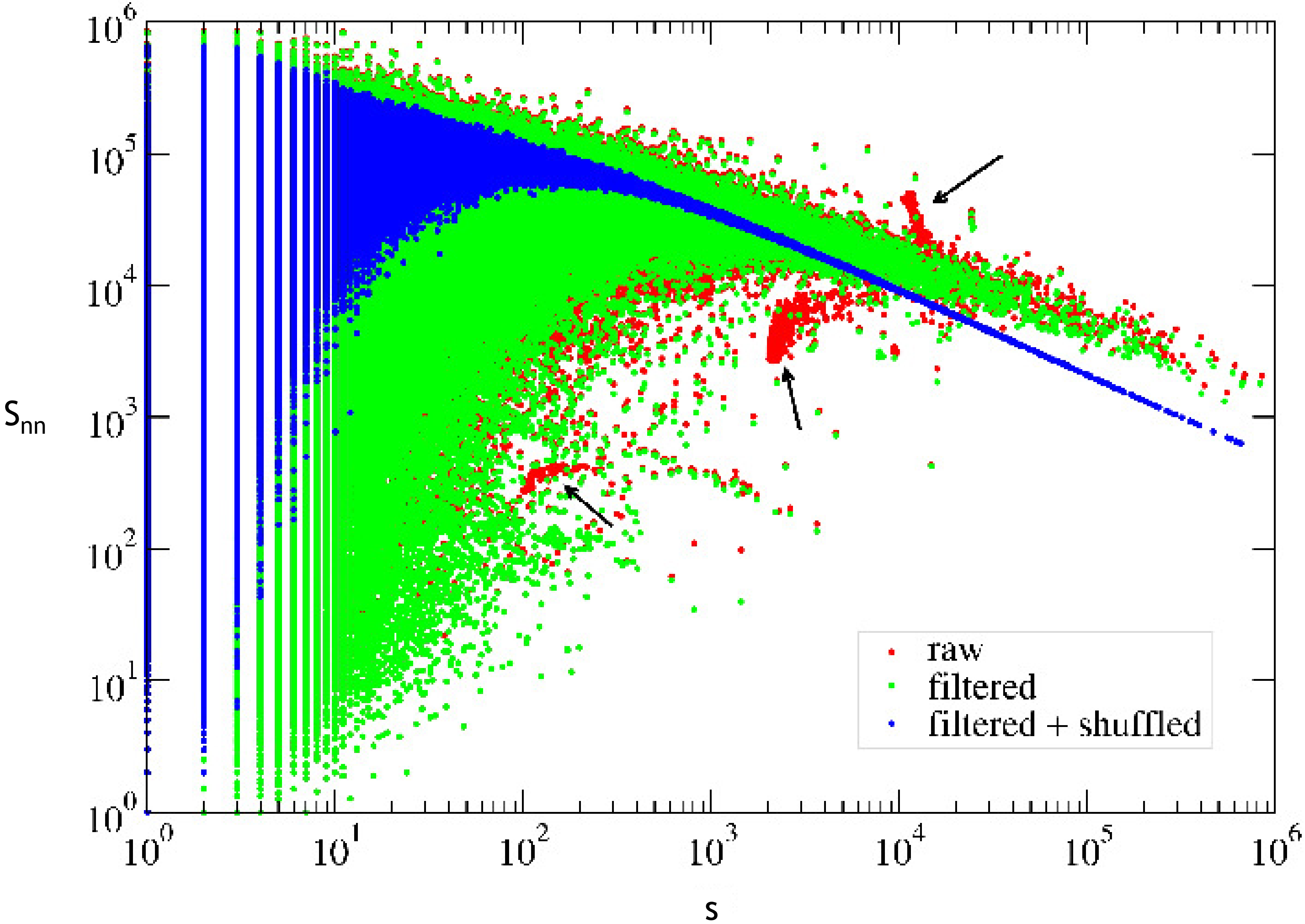
Assortative mixing: S_{nn} positively correlated to s

- E.g. social networks

Disassortative mixing: S_{nn} negatively correlated to s

- E.g. man-made, hierarchical networks

Nearest-Neighbor Strength: Delicious



Nearest-Neighbor Strength: Interpretation

Similar structure for BibSonomy and Delicious

- General pattern

Assortative as well as disassortative regions

Spamming activity: outliers

- Use for semi-automatic spam detection (work in progress)

Shuffling tag affects distribution

- Change of nearest-neighbor strength indicates semantic relations of tags

Related Work – Analysis of the Folksonomy Graph

Network properties of Web 2.0 applications

- K. Shen, L. Wu. Folksonomy as a Complex Network, 2005.
- R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network. 2005.
- P. Kolari, T. Finin, Y. Yesha, Y. Yesha, K. Lyons, S. Perelgut and J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007), 2007.
- A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. Phys. Rev. E, 74:036116, 2006.

Introduction into tagging systems

- S. Golder and B. A. Huberman. The Structure of Collaborative Tagging Systems cs/0508082 (2005)
- A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computermediated-communication/folksonomies.html.

Analysis of tagging behaviour

- C. Cattuto, A. Baldassarri,V. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems, 2007.
- C. Cattuto, V. Loreto and L. Pietronero. Collaborative Tagging and Semiotic Dynamics, PNAS, 2007.
- E. Santos-Neto, M. Ripeanu, A. Iamnitchi. Tracking User Attention in Collaborative Tagging Communities, 2007.

More under: http://www.bibsonomy.org/tag/network+folksonomy

Agenda

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

Types of Tags [Golder & Huberman, 2006]

Golder & Hubermann identified seven types of tags:

- Identifying what (or who) it is about, e.g., *ontology*, *learning*
- Identifying what it is, e.g., *article*, *blog*
- Identifiying who owns it, e.g., *apple*, *google*
- Refining categories, e.g., *2010*
- Identifying qualities or characteristics, e.g., *interesting*, *cool* (also called sentiment tags)
- Self reference, e.g., *myown*
- Task organization, e.g., *toread*, *tobuy* (also called intent or purpose tags)

Additionally, we can find

- Category of a resource
- System tags, e.g., *for:andrea*

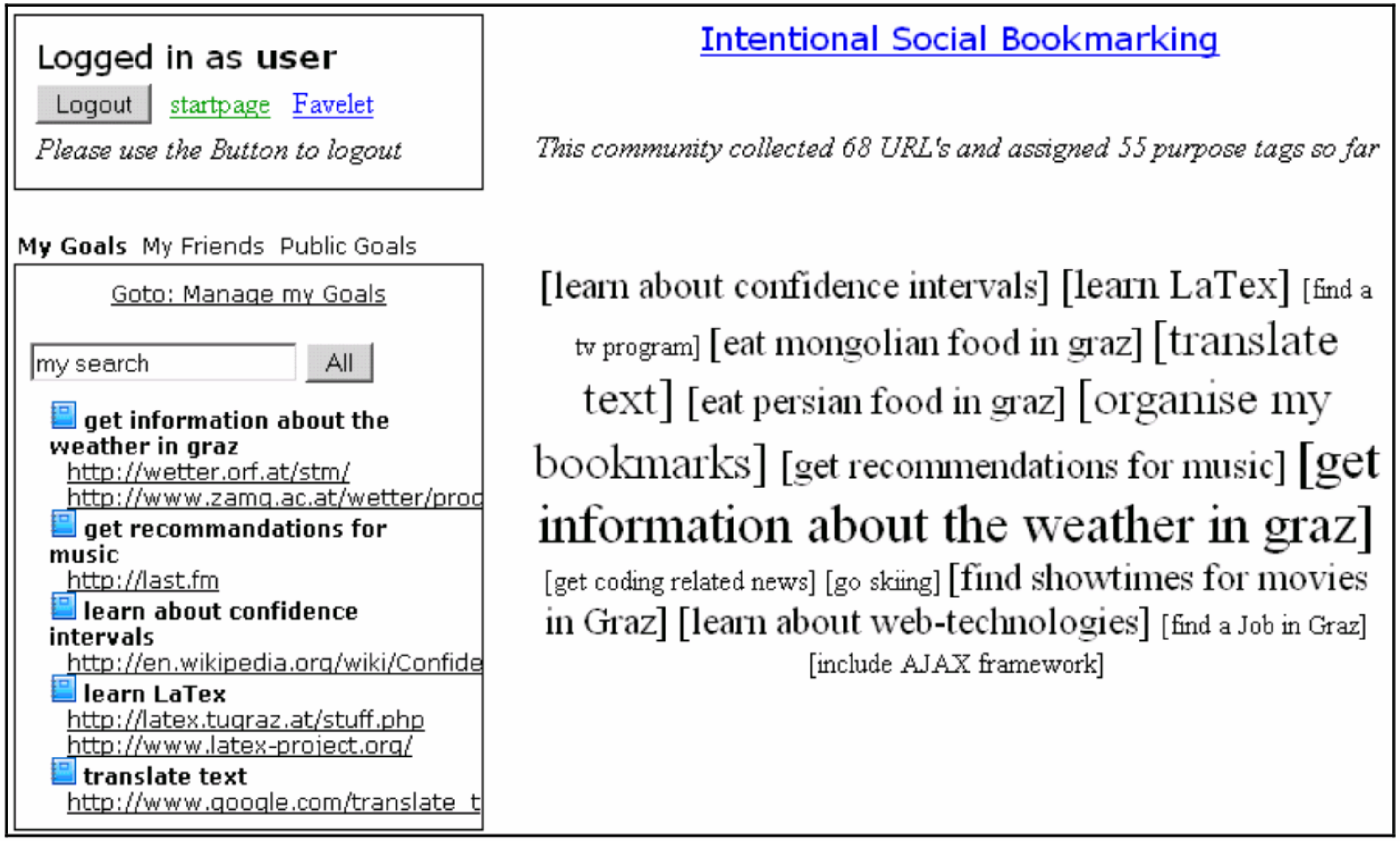
Types of Tags

- A tag can have several types (e.g., *ontology* can mean an actual ontology or an article about ontologies)
- Depending on the user, a tag can have different types
- Knowledge discovery methods should pay respect to the different types of tags
  - E.g., recommendation, ontology learning
  - not addressed so far (?)

Types of Tags - Purpose Tags [Strohmaier, 2008]

Goal: „find a physician in Seattle“

- Delicious tags for www.yellowpages.com would not help
- Most tags describe the content, not the intent



Tags

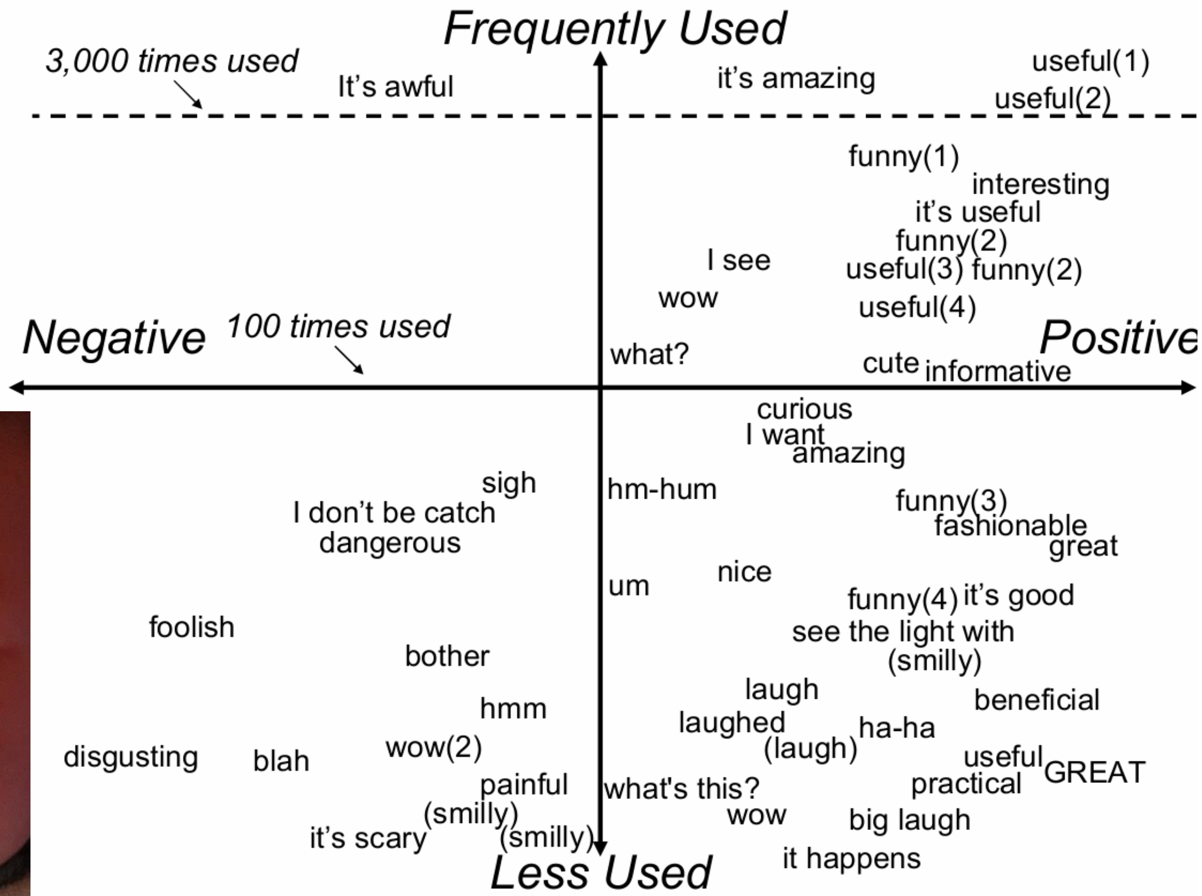
▾ Top Tags

directory	203
yellowpages	184
reference	167
search	153
phone	142
telephone	97
business	82
directories	65
yellow	59
pages	48
information	44
tools	37
local	35
research	25
imported	22
people	21
phonebook	19
resources	17
address	14
searchengine	13
book	12
seo	11
maps	10
usa	10
design	9
it	8
bit200w07	8
businesses	7
yellow_pages	7
kgb	7

Types of Tags - Sentiment Tags [Yanbe et al., 2007]

Tag Name	N
Web	16,633
google	15,674
troll	14,453
javascript	11,840
youtube	10,858
tips	10,784
css	9,411
design	8,423
2ch (huge BBS)	8,381
society	7,412

Tag Name	N
useful (1)	5,381
it's amazing	5,046
it's awful	4,123
useful (2)	3,041
interesting	638
funny (1)	
it's useful (3)	
funny (2)	
useful (4)	
I see	



Types of Tags - Event/Place Tags [Rattenbury et al., 2007]

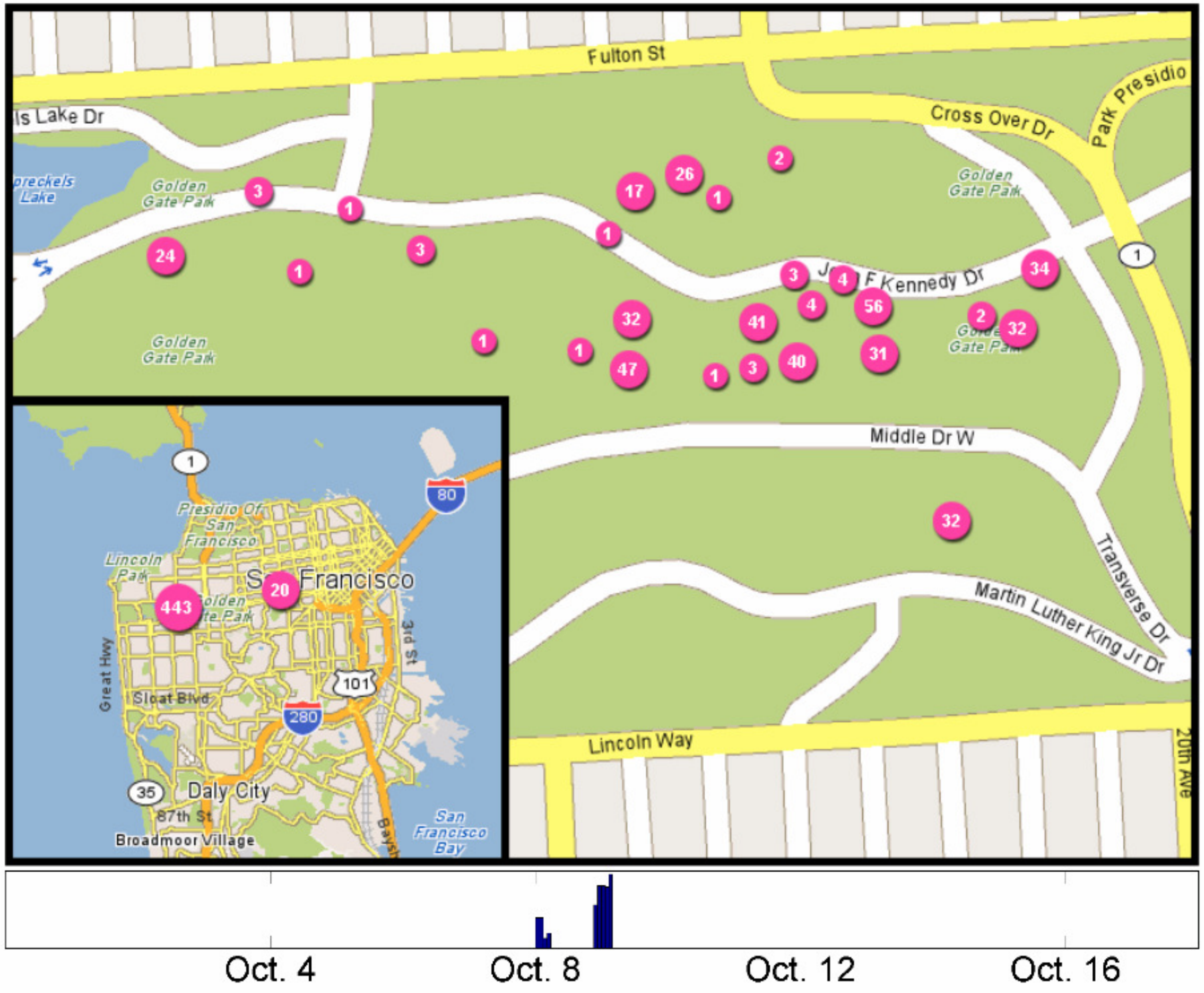


Figure 2: Location (top) and time (bottom) usage distributions for the tag `Hardly Strictly Bluegrass` in the San Francisco Bay Area. The zoomed in map view shows the details of the larger location cluster from the zoomed out view.

Based on time + location information, automatically extract event/place tags

Extracted place tags:

pet cemetary, Revision3, Ruby Red, Dahlias, *MashPitSF2*, VS Hoe Down, Red Devil Lounge, Club Neon, Future of Web Apps, Bottom of the Hill

Extracted event tags:

zombiemob, Bay to Breakers 2006, valleyschwag, zombie, zombiemob2006, eatbrains, VS Hoe Down, eatbrains2006, zombies, *air race*

(italics: false positives)

Types of Tags - Related Work

- Usage patterns of collaborative tagging systems. S. Golder and B. Huberman, Journal of Information Science 32, 2006.
- M. Strohmaier, Purpose Tagging - Capturing User Intent to Assist Goal-Oriented Social Search, SSM'08 Workshop on Search in Social Media, in conjunction with CIKM'08, Napa Valley, USA, 2008.
- M. Strohmaier, C. Körner, and R. Kern, Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems, 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26, 2010.
- Yanbe, Y.; Jatowt, A.; Nakamura, S. & Tanaka, K. (2007), Can social bookmarking enhance search in the web?, in 'JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries' , ACM, New York, NY, USA , pp. 107--116 .
- Rattenbury, T.; Good, N. & Naaman, M. (2007), Towards automatic extraction of event and place semantics from flickr tags, in 'SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval' , ACM Press, New York, NY, USA , pp. 103--110 .

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data

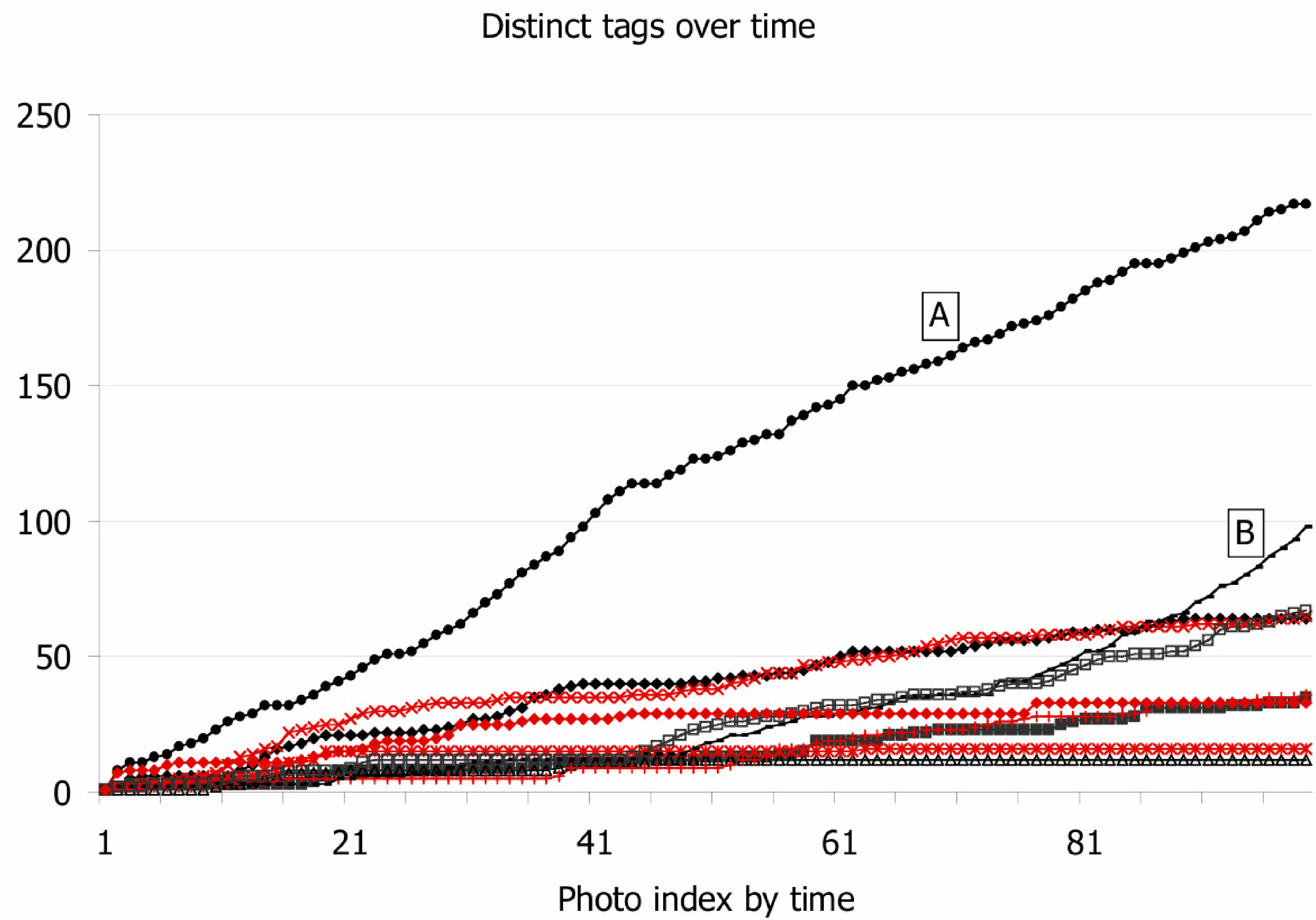
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

Types of Users [Marlow et al., 2006]

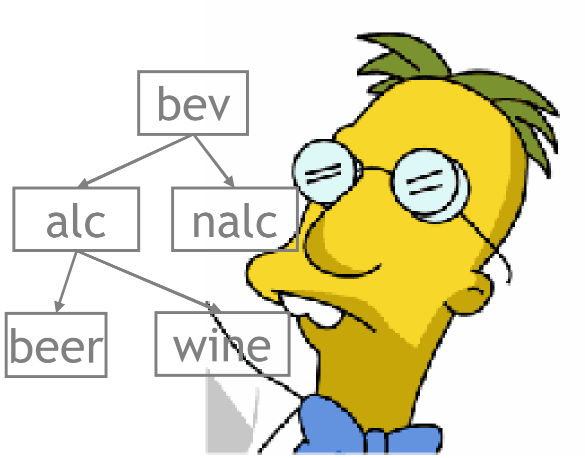


A: consistently new tags as new photos are uploaded

B: few tags, sudden growth later

Evidence of different ways **HOW** users tag (Tagging **Pragmatics**)

Broad distinction by tagging **motivation** [Strohmaier 2009]:



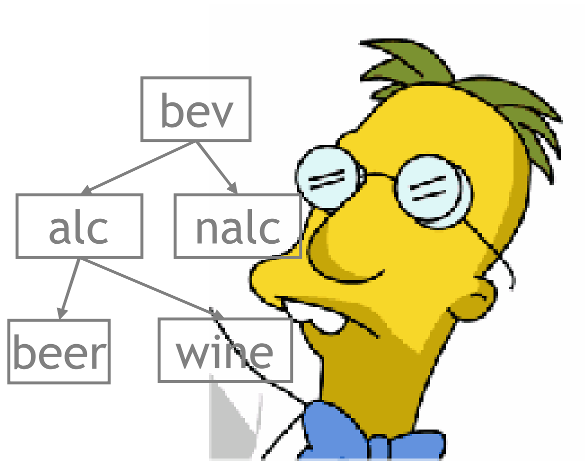
„Describers"…

- tag „verbously" with freely chosen words
- vocabulary not necessarily consistent (synomyms, spelling variants, …)
- goal: describe content, ease retrieval

„Categorizers"…

- use a small controlled tag vocabulary
- goal: „ontology-like" categorization by tags, for later browsing
- tags as replacement for folders

Types of Users [Strohmaier et al., 2010]



	Categorizer	**Describer**
Goal	Later Browsing	Later Retrieval
Change of Vocabulary	costly	cheap
Size of Vocabulary	limited	open
Tags	subjective	objective
Tag Reuse	frequent	rare
Tag Purpose	mimicking taxonomy	descriptive labels

We will come back to describers and categorizers later …

Types of Users – Related Work

- Marlow, C.; Naaman, M.; Boyd, D. & Davis, M. (2006), HT06, tagging paper, taxonomy, Flickr, academic article, to read, in 'HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia' , ACM, New York, NY, USA , pp. 31--40 .
- C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier: Of categorizers and describers: an evaluation of quantitative measures for tagging motivation, HT '10: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, New York, NY, USA, ACM, 2010.
- Strohmaier, M.; Körner, C. & Kern, R. (2010), Why do users tag? Detecting users' motivation for tagging in social tagging systems, in 'International AAAI Conference on Weblogs and Social Media (ICWSM2010)' .
- http://src.acm.org/2010/ChristianKoerner/understanding_the_motivation_behind_tagging/index.html

Agenda

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

Types of Resources

Basically, there are systems to tag *anything* …

photos 43Things MISTER WONG BibSonomy

faves.com Sites you'll love, from people like you.

videos flickr givealink.org Share your links!

vimeo

Knowledge Plaza HOWL digg contacts reddit

publication references

news StumbleUpon newsvine

delicious social bookmarking

bookmarks simpy goals in life

Slashdot citeulike gnolia

… to name just a few.

Types of Resources

- Specialized methods for certain types
  - E.g., NLP for web pages, blog articles, publications, etc.
  - Information extraction for documents
  - Image recognition/analysis techniques
  - Social network analysis for contacts/people
- Goal: disregard type, focus on type-independent techniques

Agenda

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

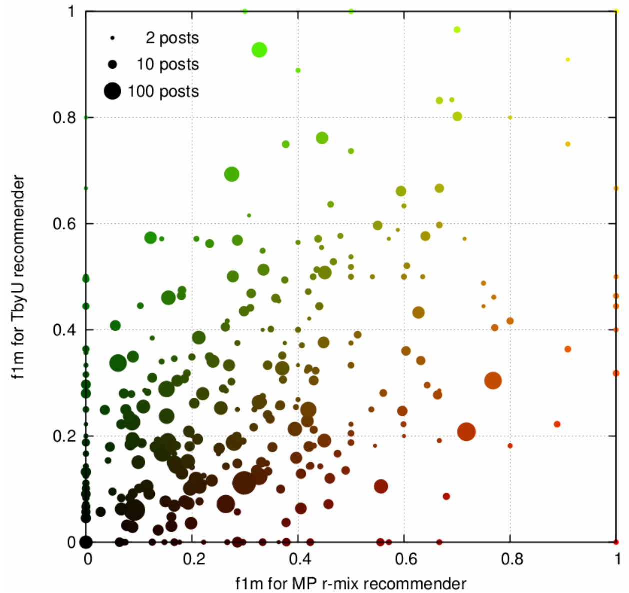
Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

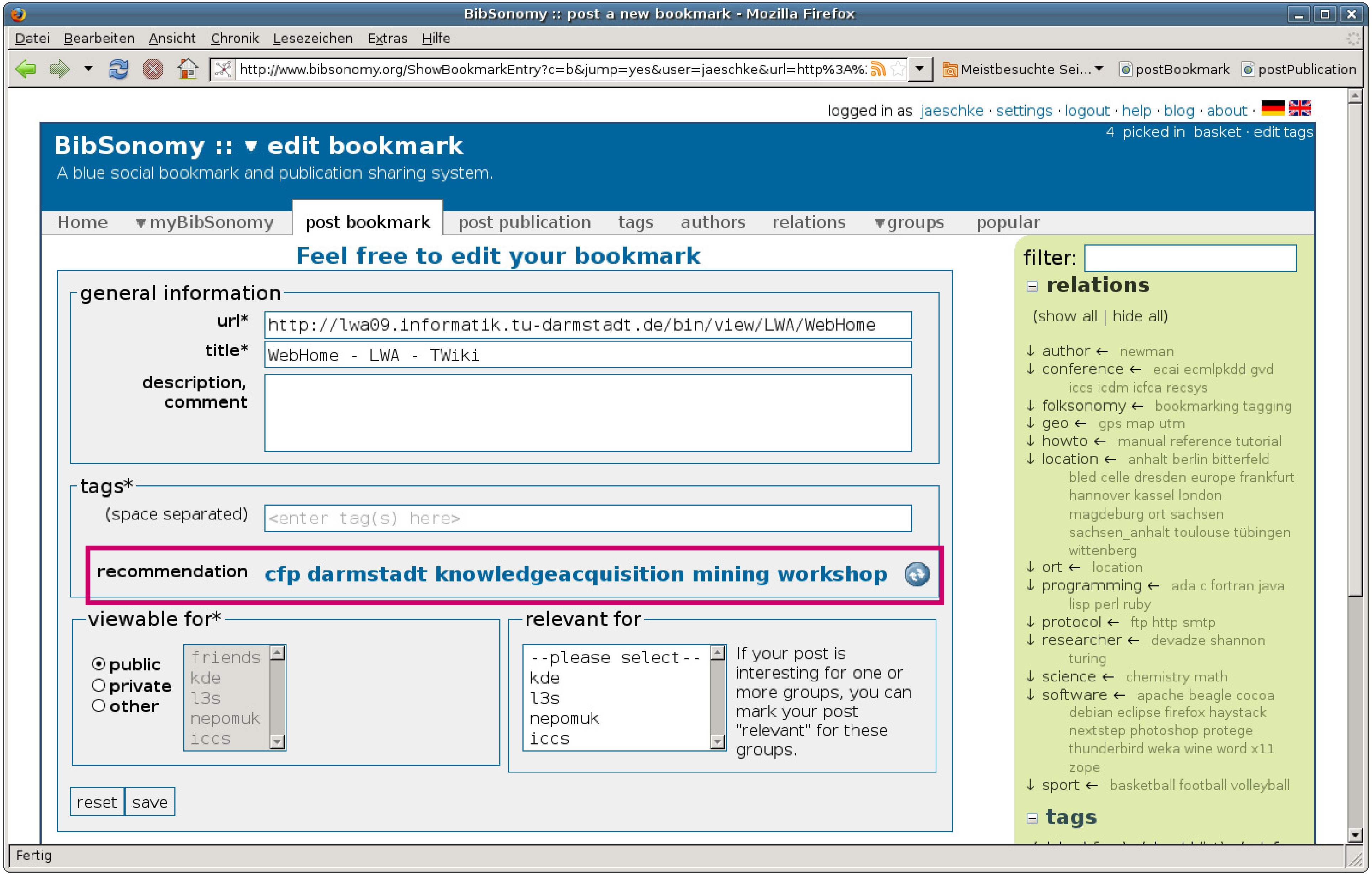
Summary and Outlook

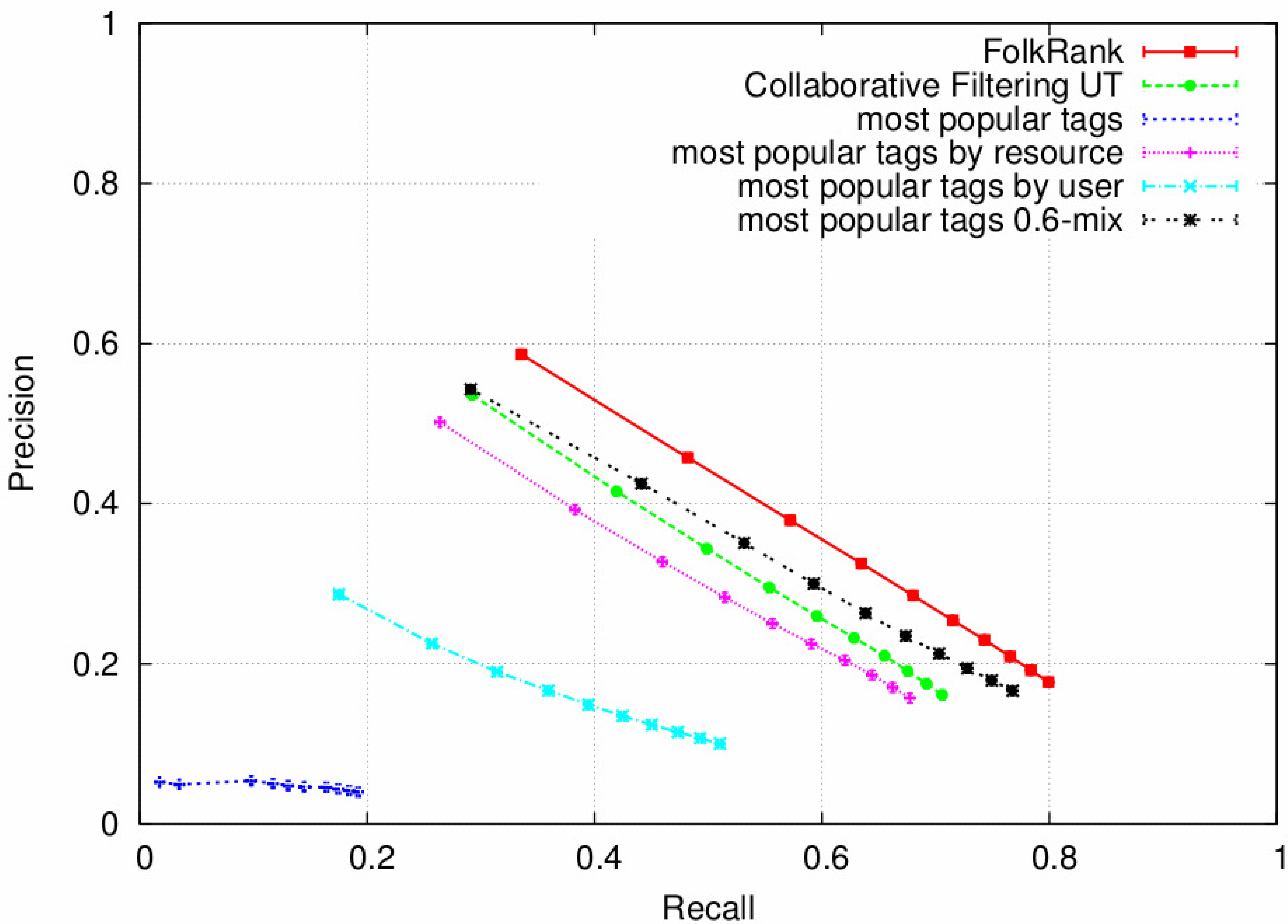


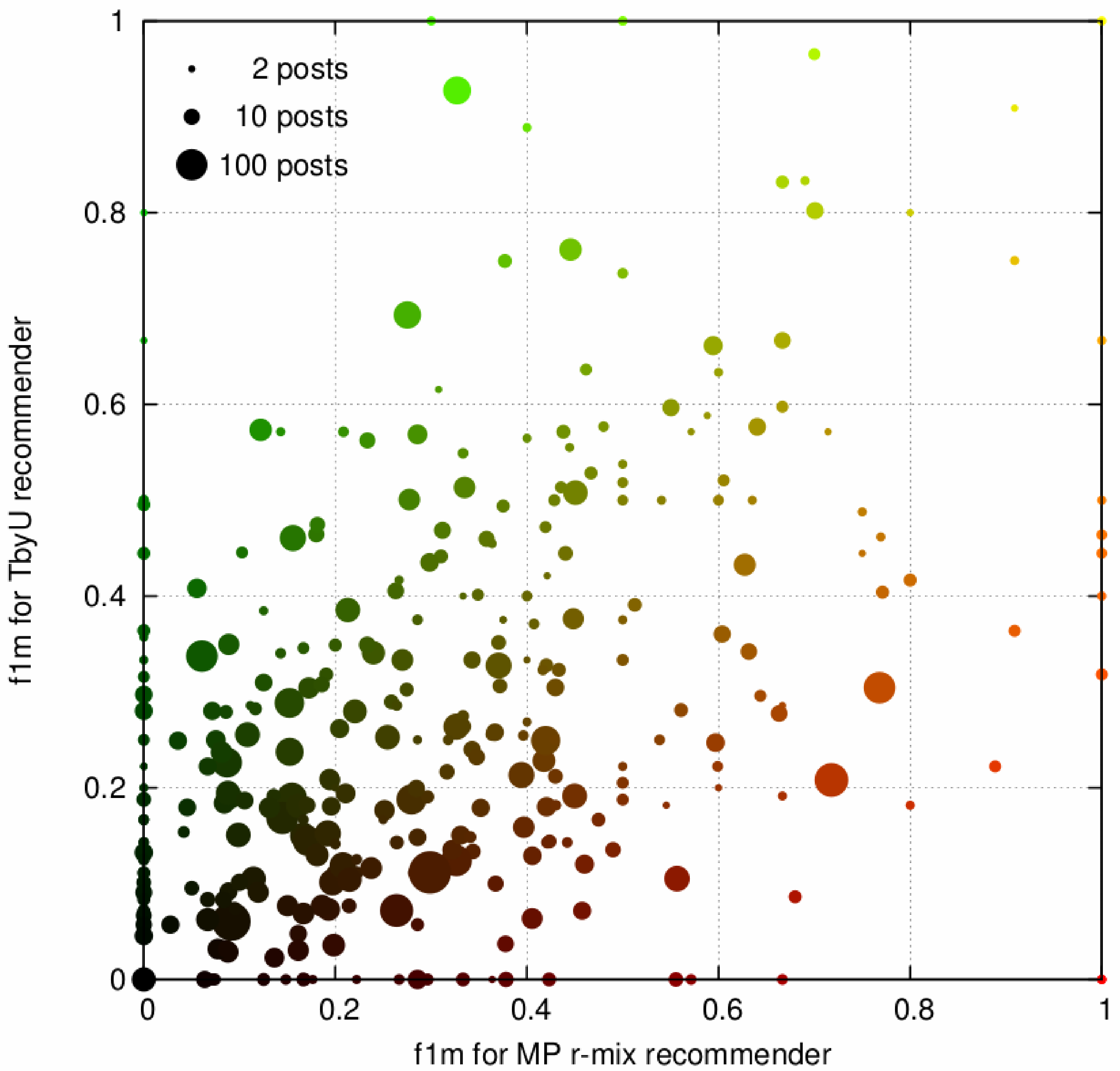
Factors influencing the Development of Folksonomies

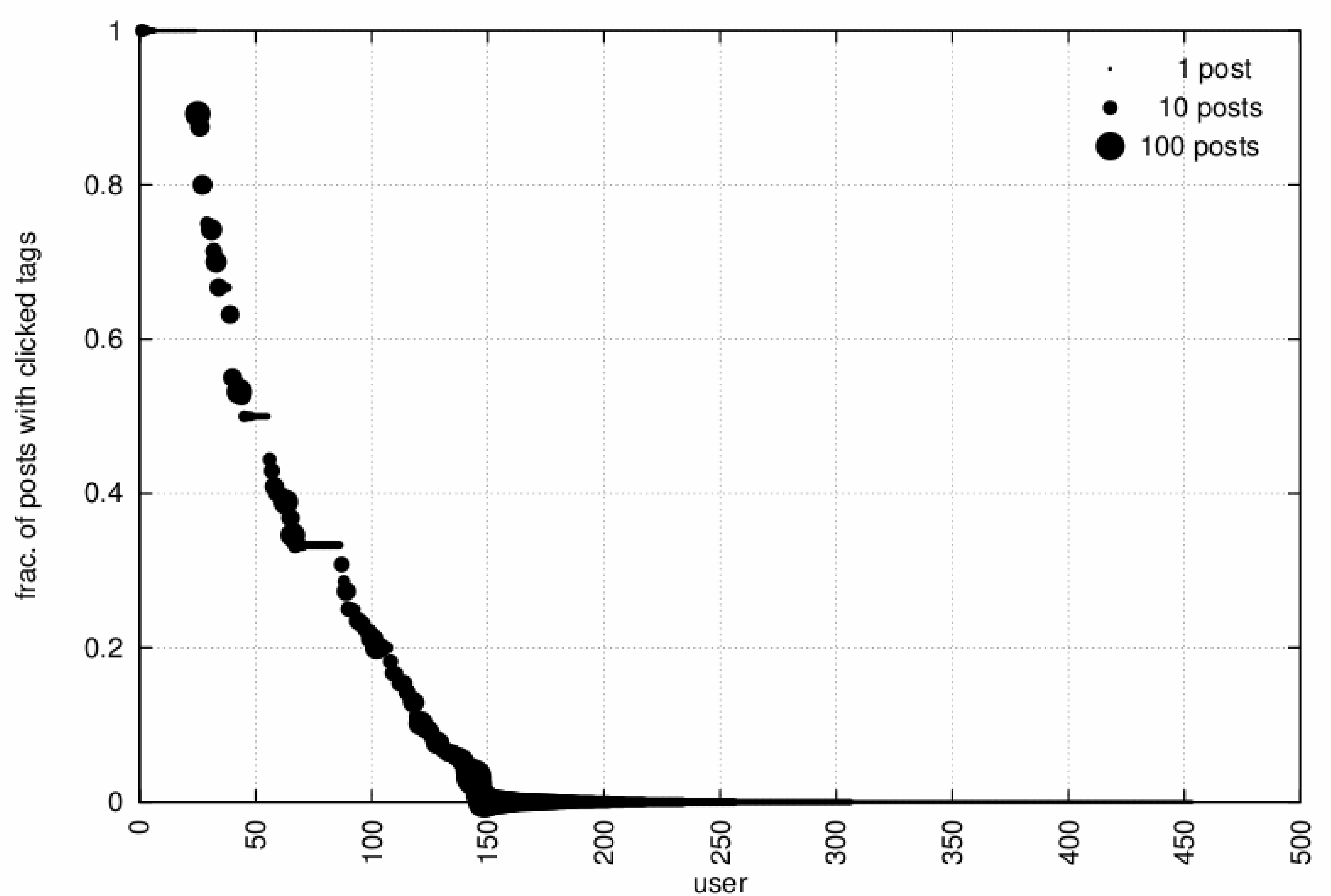
- Presentation/layout
- Systems
- Search Engines
- Trends
- Tools (e.g., automatic posting)
- **(Tag) Recommender**
- **Spam**
- Social Components
- Types of users, resources, tags
- …

Factors influencing the Development of Folksonomies




FolkRank
Collaborative Filtering UT
most popular tags
most popular tags by resource
most popular tags by user
most popular tags 0.6-mix
Precision
Recall
0
0.2
0.4
0.6
0.8
1


2 posts
10 posts
100 posts
f1m for TbyU recommender
f1m for MP r-mix recommender
0
0.2
0.4
0.6
0.8
1


1 post
10 posts
100 posts
frac. of posts with clicked tags
user
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250
300
350
400
450
500

Recommender

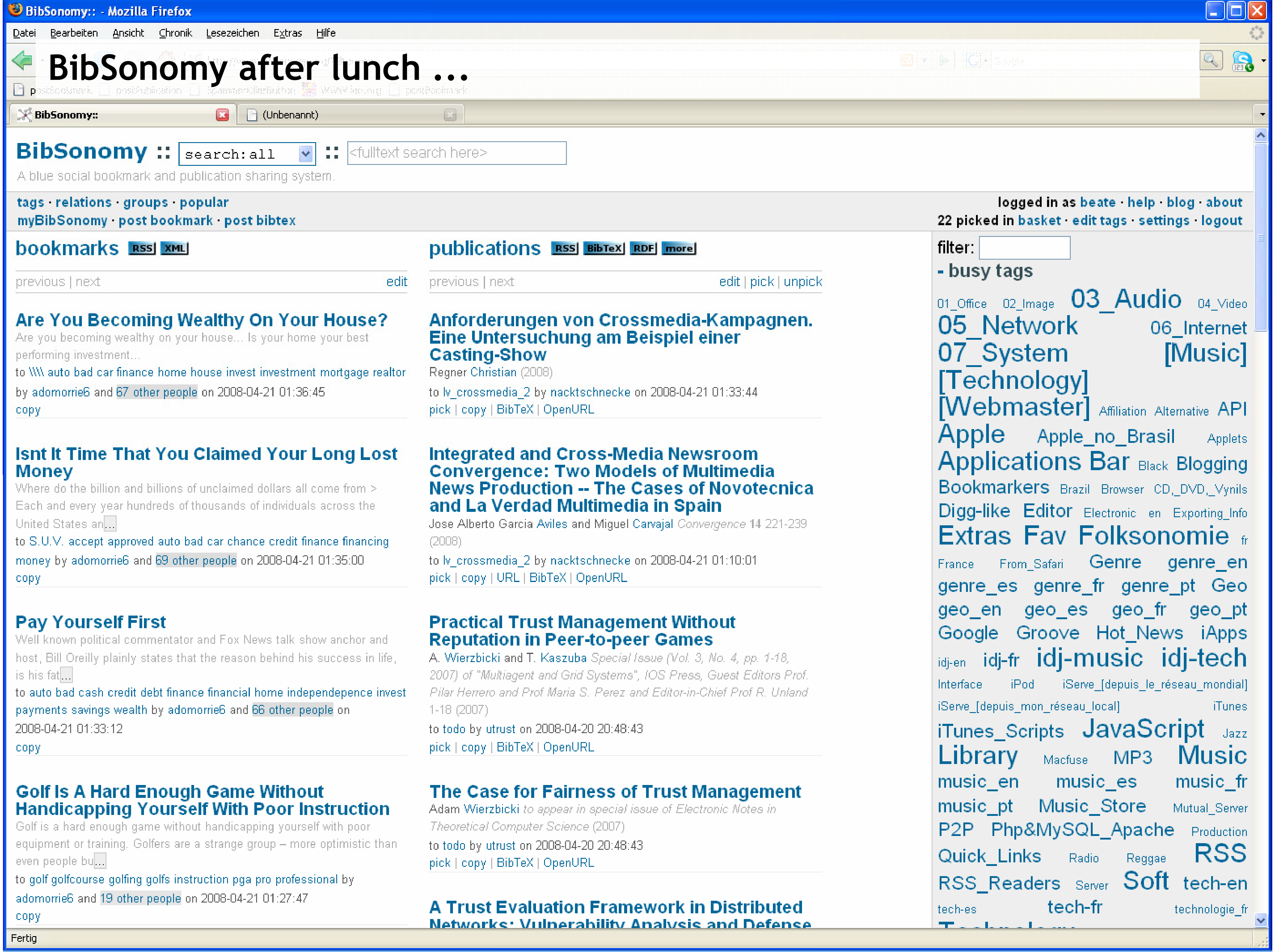
- G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. Knowledge and Data Engineering, IEEE Transactions on, (17)6:734--749, 2005.

Tag Recommender

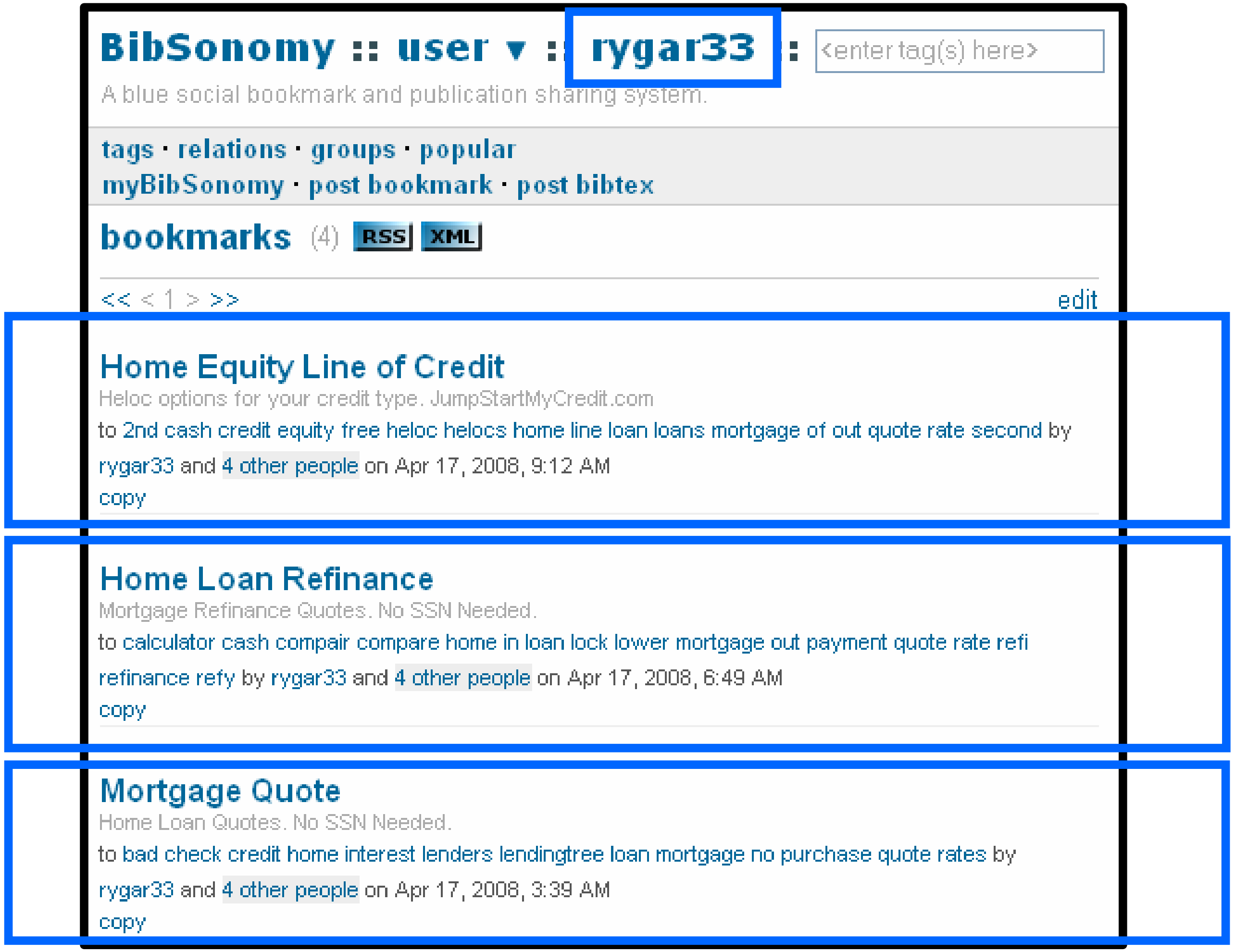
- Z. Xu and Y. Fu and J. Mao and D. Su. Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May, 2006.
- Yanfei Xu and Liang Zhang and Wei Liu. Cubic Analysis of Social Bookmarking for Personalized Recommendation. Frontiers of WWW Research and Development - APWeb 2006, 733--738, 2006.
- Jäschke, R.; Eisterlehner, F.; Hotho, A. & Stumme, G. (2009), Testing and Evaluating Tag Recommenders in a Live System, in Dominik Benz & Frederik Janssen, ed., 'Workshop on Knowledge Discovery, Data Mining, and Machine Learning' , pp. 44--51 .
- Jäschke, R.; Marinho, L.; Hotho, A.; Schmidt-Thieme, L. & Stumme, G. (2008), 'Tag Recommendations in Social Bookmarking Systems', AI Communications 21 (4) , 231-247 .

More under: http://www.bibsonomy.org/tag/recommender+folksonomy

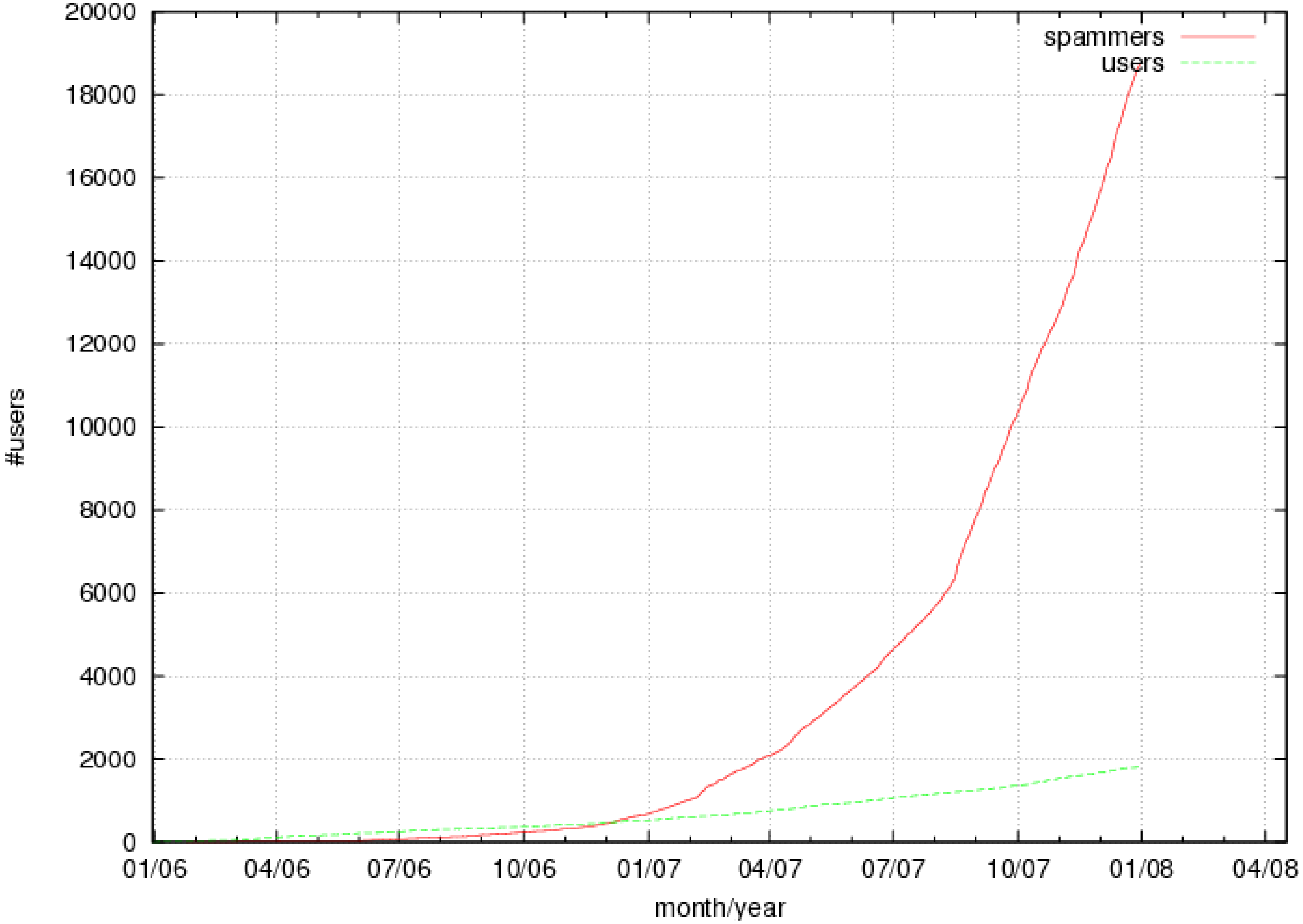
BibSonomy after lunch ...



Spam – User Level vs. Post Level



BibSonomy “active” user accounts over time …



BibSonomy admins and developers flag users as spammers

Decision is based on

- Links (websites) of posts
- Added tags
- Also influenced by personal information:
  - E-mail
  - Choice of name
  - Registration IP
  - …

Dataset creation process

Dataset - Figures

	Users	Spammer	Tags	Resources	TAS
All	1,411	18,681	306,993	920,176	8,709,417
Training	1,306	15,891	282,473	774,678	7,904,735
Test	100	2,790	49,644	153,512	804,682

- Time frame: until end of 2007
- Only users with at least one post
- No consideration of private posts
- Tags not normalized

Features

- 25 features
- 4 different categories
- Normalization of each user's feature vector

Profile Information

- Realname with 2 or 3 words
- lenght of the user name, email, realname
- digits in user name

Activity Information

- time between registration and first post
- number of tags per post
- average number of TAS
  - 470 for spammers, 334 for users

Location Information

- number of users in the same domain
- number of users in the same top level domain
- number of spam users with this IP

Semantic Information

- blacklist of tags
- Co-Occurrence information of the graph, e.g. spammer shares resources with other spammers

Frequency ROC Area: 0.80

TFIDF ROC Area: 0.79

Table 4: Baseline with all tags as features (frequency)

	Spam	Non-Spam
Spam	466	2324
Non-Spam	0	100

Table 5: Baseline with all tags as features (tfidf)

	Spam	Non-Spam
Spam	530	2260
Non-Spam	0	100

Tool:

- Weka

Classification Algorithms:

- SVM
- J48
- Logistic Regression
- Naïve Bayes

All features

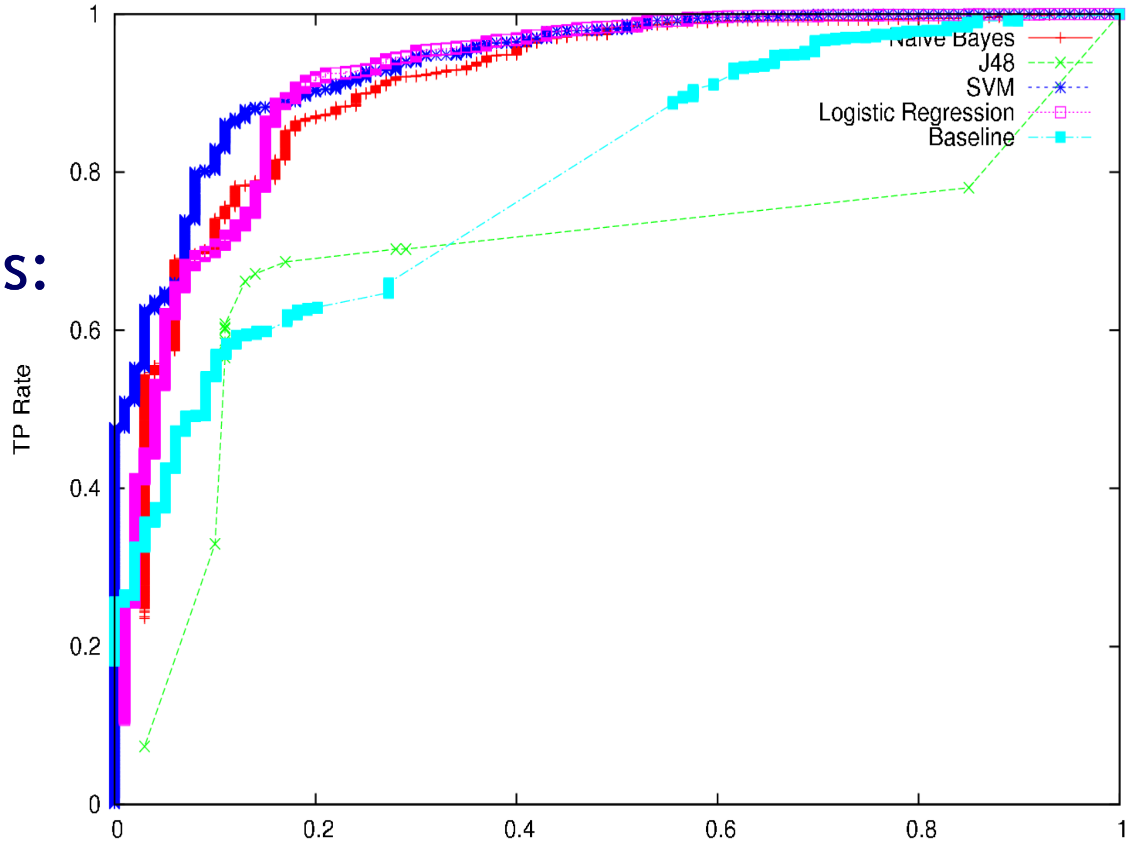


Table 10: Evaluation values all features

Classifier	ROC area	F1	FP	FN
Naive Bayes	0.906	0.876	14	603
SVM	0.936	0.986	53	23
Logistic Regression	0.918	0.968	30	144
J48	0.692	0.749	11	1112

Spam Detection - Related Work

Paul Heymann and Georgia Koutrika and Hector Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. IEEE Internet Computing, (11)6:36-45, 2007.

Georgia Koutrika and Frans Adjie Effendi and Zoltan Gyöngyi and Paul Heymann and Hector Garcia-Molina. Combating spam in tagging systems. AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, 57--64, ACM Press, New York, NY, USA, 2007.

Benjamin Markines and Ciro Cattuto and Filippo Menczer. Social spam detection.. In Dennis Fetterly and Zoltán Gyöngyi, editor(s), AIRWeb, 41-48, 2009.

Zoltán Gyöngyi and Hector Garcia-Molina and Jan Pedersen. Combating Web Spam with TrustRank.. VLDB, 576-587, 2004.

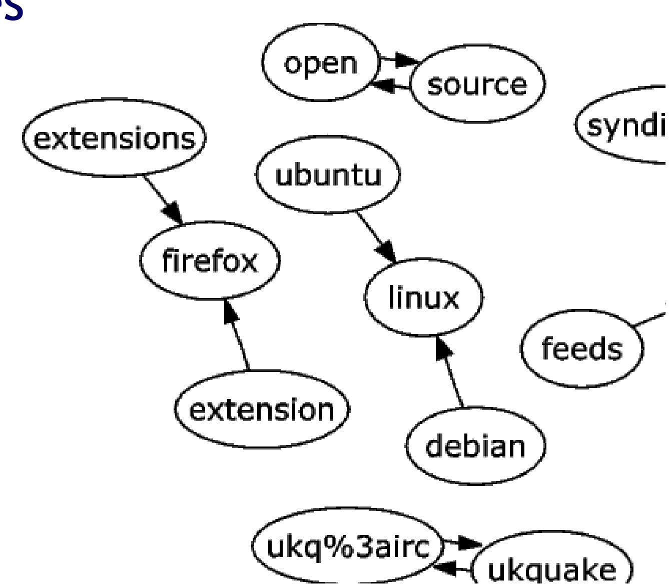
Agenda

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

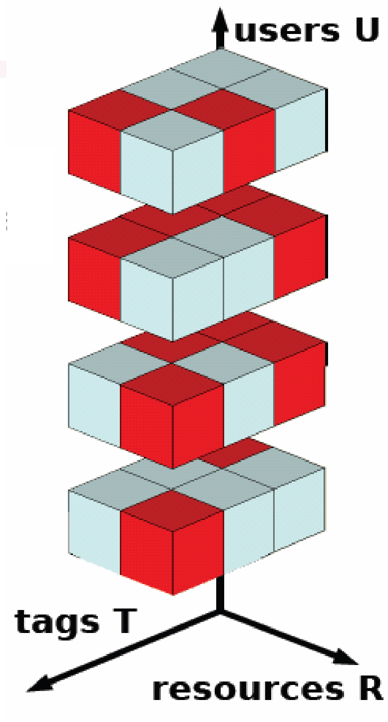
Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook



Mining Association Rules in Folksonomies



Task: Find all rules of the form:

Many people who buy $i_1, \ldots, i_n$ also buy $j_1, \ldots, j_m$.

Problem: folksonomies are of triadic nature:

- Cube Y instead of matrix I
- Tripartite hypergraph instead of bipartite graph

Straightforward Solution:

- ternary relation $\rightarrow$ projection on dyadic context $\rightarrow$ Apriori algorithm

$\cong$ *transactions*

$\cong$ *items*

Dimension reduction:

convert $\mathbf{F} = (U, T, R, Y)$ to some $\mathbf{K} = (G, M, I)$ by

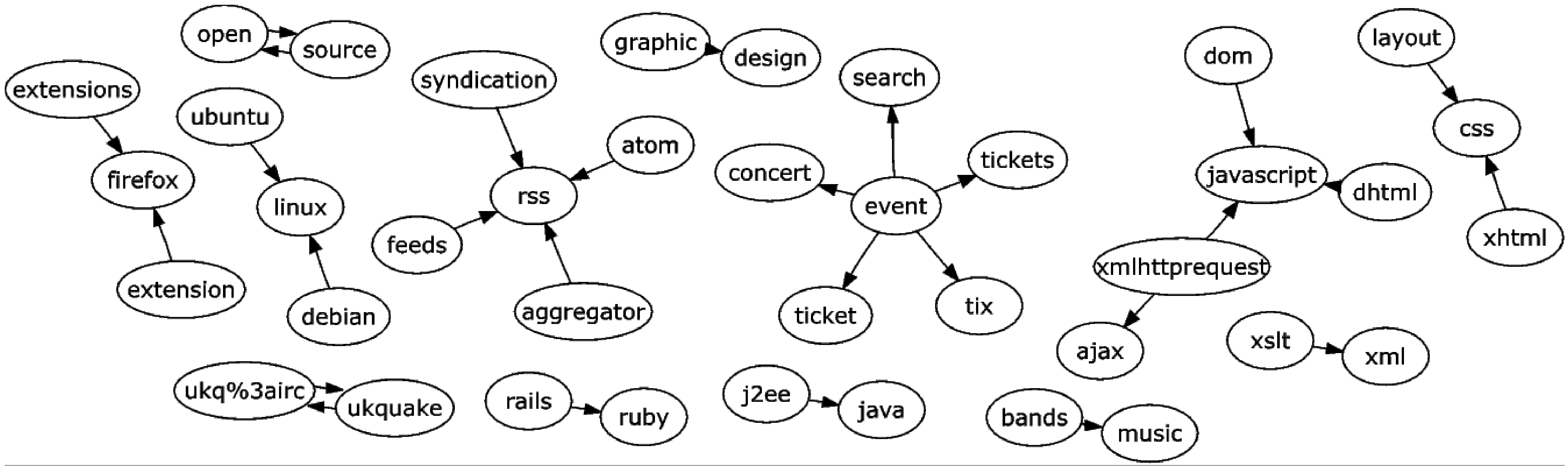
- slicing along one dimension
- projection & aggregation

Association Rules

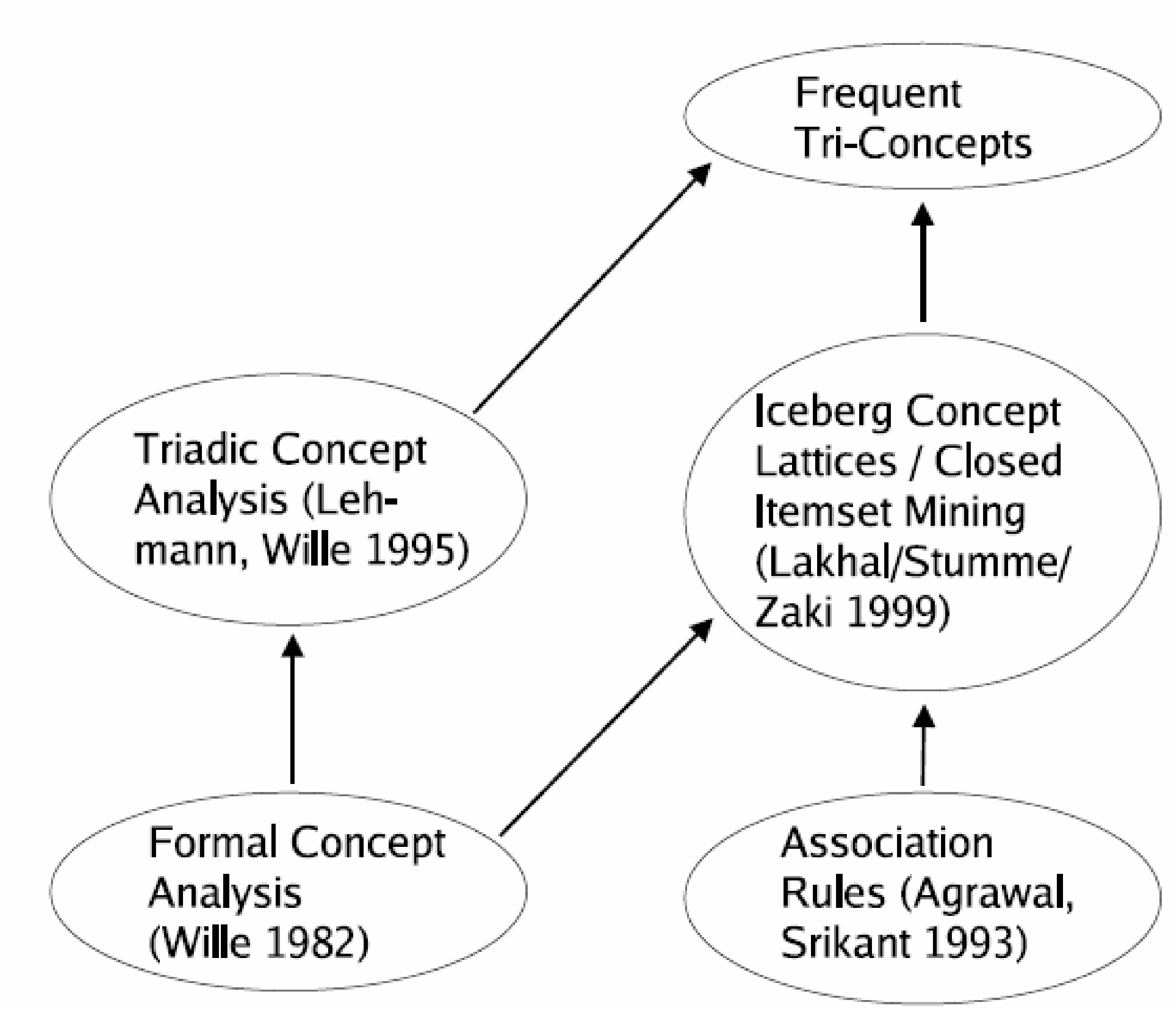
$\cong$ *transactions*

$\cong$ *items*

- $\mathbf{K}_1 = (U \times R, T, I_1)$
- If users tag some resource with tag t_i, they frequently also use t_j for it.
- Usage:
  - tag recommendations
  - learning implications (tag hierarchy)



Triadic Concept Analysis



with B. Ganter, TU Dresden

Recall: Closed Itemsets / Formal Concept Analysis

The itemset {Horseback Riding, Fishing} has the same „customers“ as the itemset *B*.

⇒ It is sufficient to consider one of them for association rules!

The maximal sets with this property are called **closed itemsets**.

Def.: *(A,B)* is called a **formal concept** if *A* and *B* are maximal with $A \times B \subseteq I$.

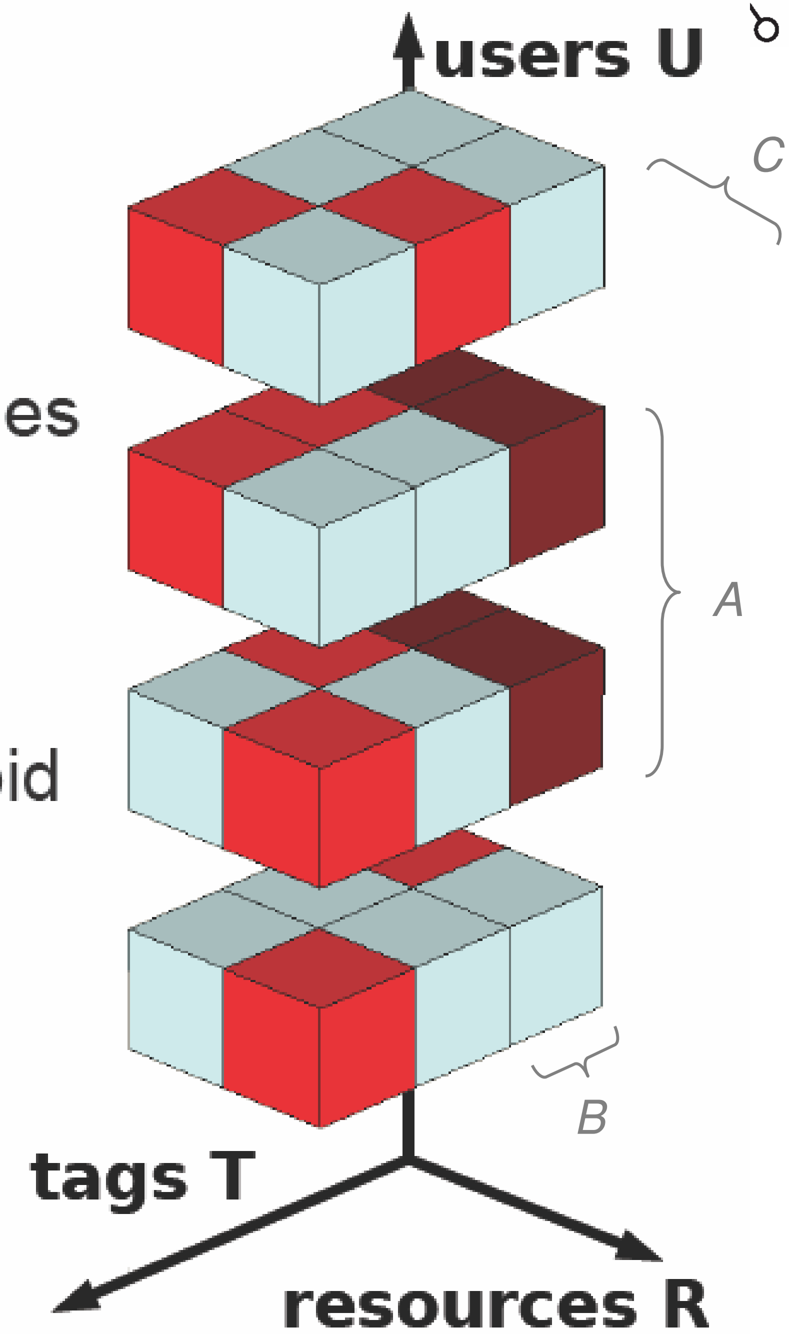
Intent *B*

Extent *A*

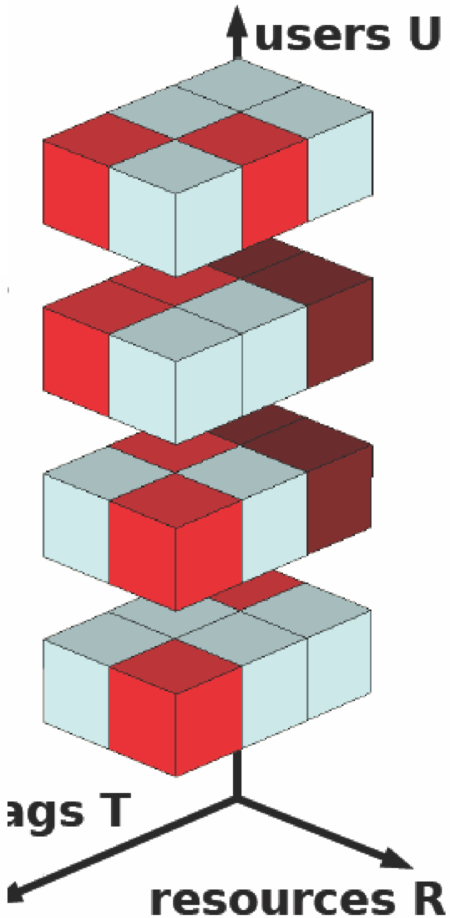
National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Triadic Concept Analysis

- conceptual clustering of folksonomies
  - find interesting concepts/clusters
  - support browsing, community detection, recommendation
- tri-concept (A, B, C): maximal cuboid where each user in A tagged each resource in C with each tag from B

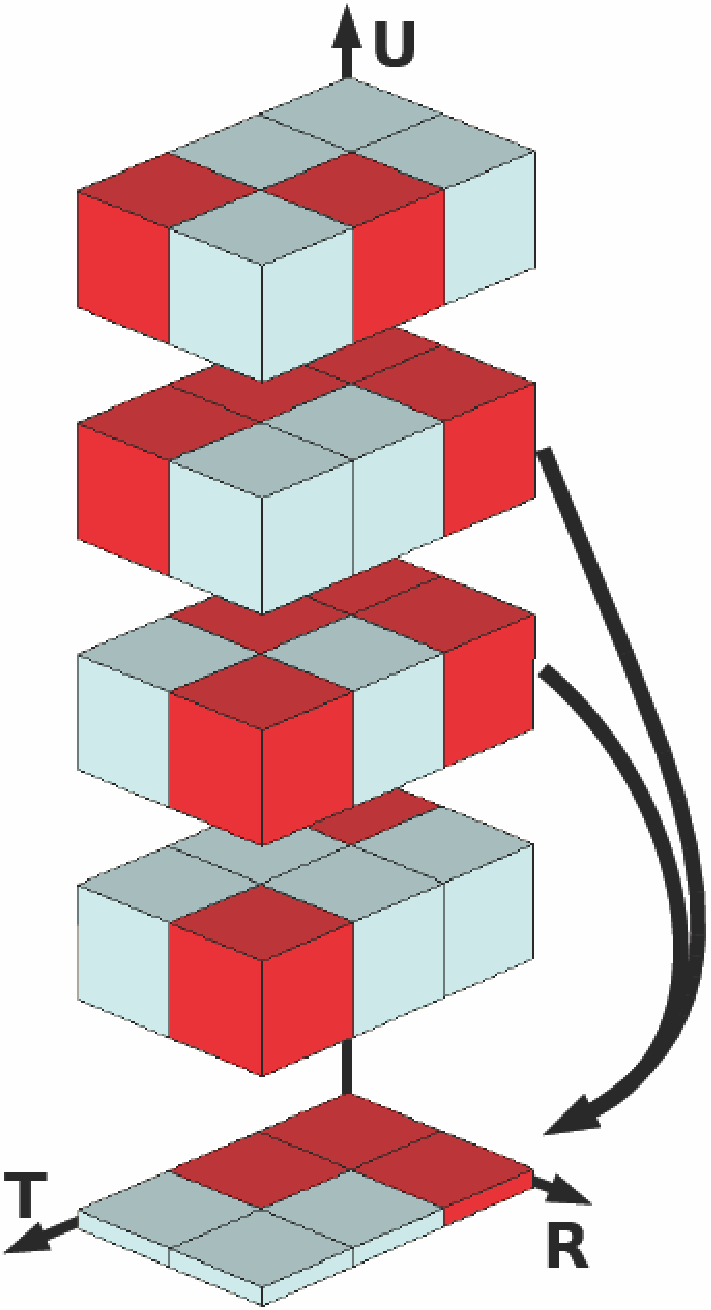


Triadic Concept Analysis: formal definition of problem



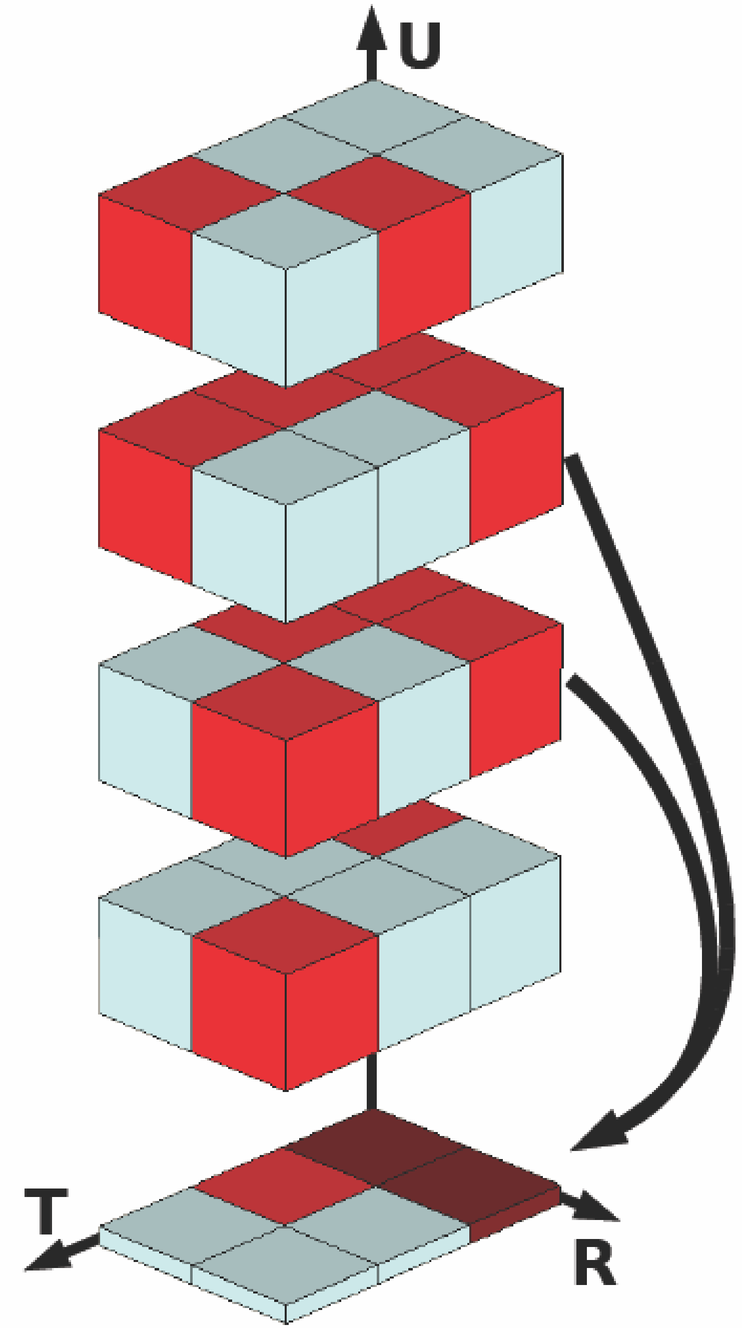
- given
  - sets U, T, R
  - ternary relation $Y \subseteq U \times T \times R$
  - minimal support constraints τ_u, τ_t, τ_r
- find (A, B, C) with
  - $A \subseteq U$, $B \subseteq T$, $C \subseteq R$
  - $|A| \geq \tau_u$, $|B| \geq \tau_t$, $|C| \geq \tau_r$
  - $A \times B \times C \subseteq Y$
  - none of A, B or C can be enlarged without violating last condition

TRIAS algorithm



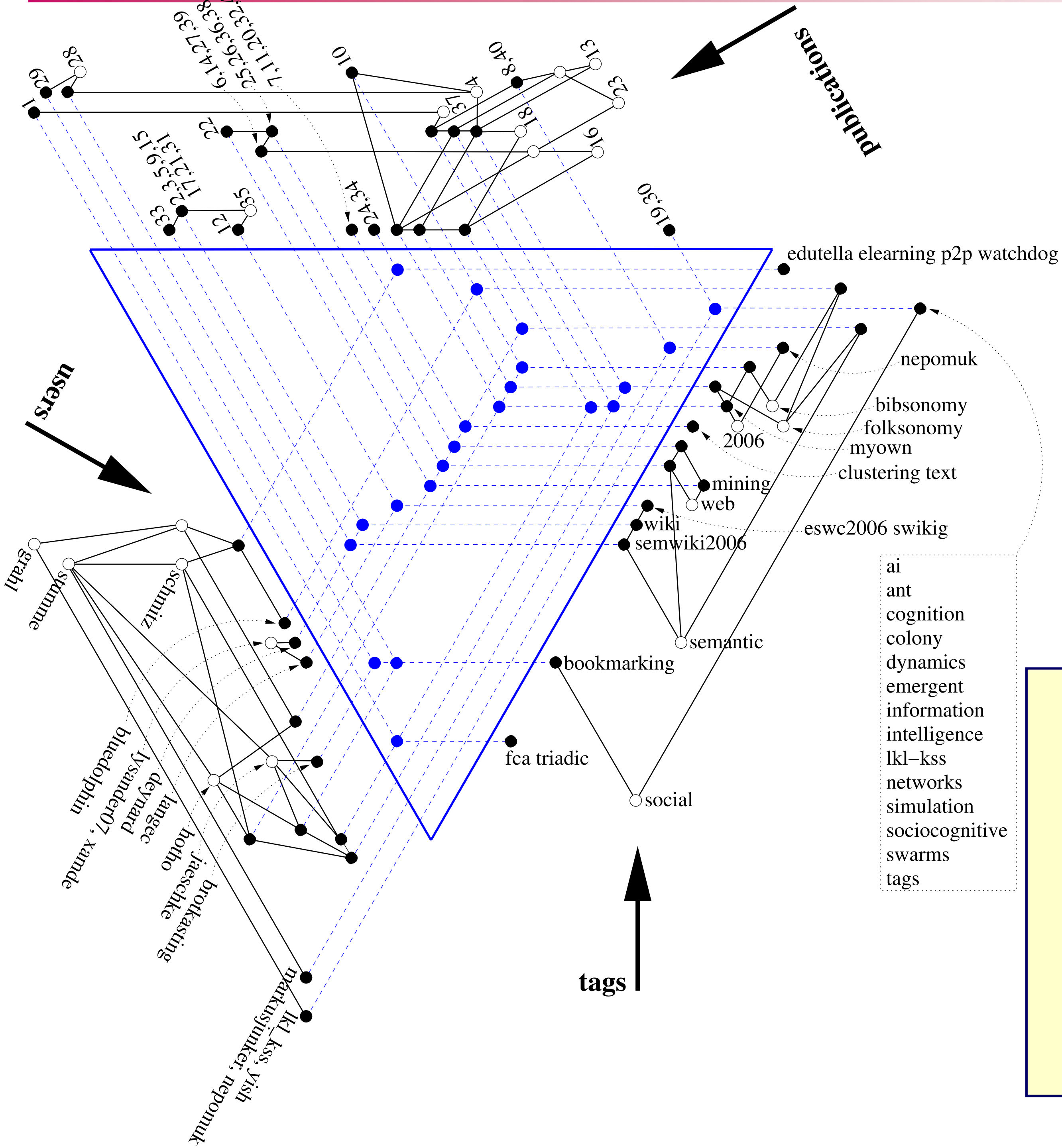
- let $\tilde{Y} := \{(u,(t,r)) \mid (u,t,r) \in Y\}$
- outer loop: find frequent concepts (A, I) in $(U, T \times R, \tilde{Y})$

TRIAS algorithm



- let $\tilde{Y} := \{(u,(t,r)) \mid (u,t,r) \in Y\}$
- outer loop: find frequent concepts (A, I) in $(U, T \times R, \tilde{Y})$
- inner loop: find frequent concepts (B, C) in (T, R, I)
- if $A = (B \times C)^{\tilde{Y}}$ output (A, B, C)

Frequent tri-concepts for BibSonomy publications



BibSonomy Publications

- |U| = 262 users
- |R| = 11,101 publications
- |T| = 5,954 distinct tags
- |Y| = 44,944 tag assignments

- 13,992 tri-concepts in total
- 21 frequent ones for 3x2x2 threshold

Related Work

Association rule mining

- Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, 487-499, Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA,1994.

Formal Concept Analysis

- Rudolf Wille: Restructuring lattice theory: An approach based on hierarchies of concepts. *Ordered Sets*, page 445-470. Reidel, Dordrecht-Boston, 1982.
- Ganter, Bernhard; Wille, Rudolf: *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag, Berlin 1998.

Triadic extension of association rule mining

- Fritz Lehmann and Rudolf Wille. A triadic approach to formal concept analysis. In G. Ellis and R. Levinson and W. Rich and J. F. Sowa, editor(s), Conceptual structures: applications, implementation and theory, LNAI 954, Springer1995, 32-43.
- Bernhard Ganter, Sergei A. Obiedkov: Implications in Triadic Formal Contexts. Proc. Intl. Conf. on Conceptual Structures 2004, LNAI 3127, Springer 2004, 186-195.
- Gerd Stumme. A Finite State Model for On-Line Analytical Processing in Triadic Contexts. In Bernhard Ganter and Robert Godin, editor(s), Proceedings of the 3rd International Conference on Formal Concept Analysis, LNAI 3403, Springer, 2005, 315-328.
- Robert Jäschke and Andreas Hotho and Christoph Schmitz and Bernhard Ganter and Gerd Stumme. TRIAS - An Algorithm for Mining Iceberg Tri-Lattices. Proc. 6th ICDM conference, Hong Kong,2006.

More under: http://www.bibsonomy.org/tag/association+rule

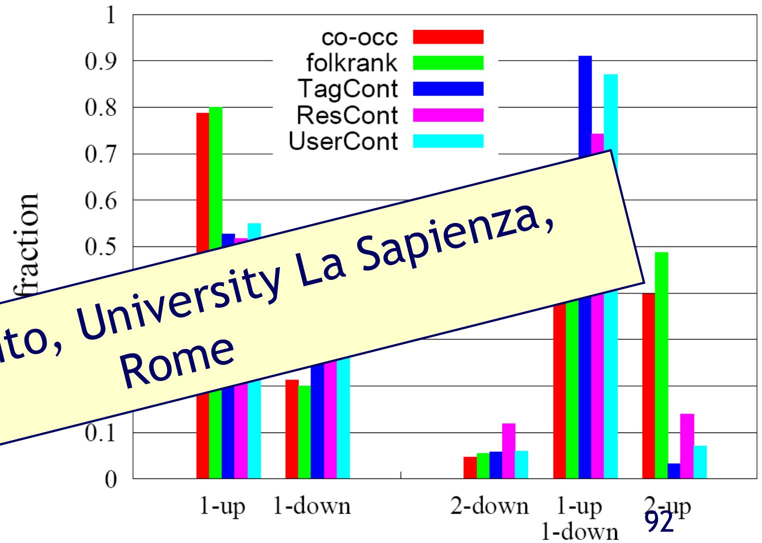
Agenda

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

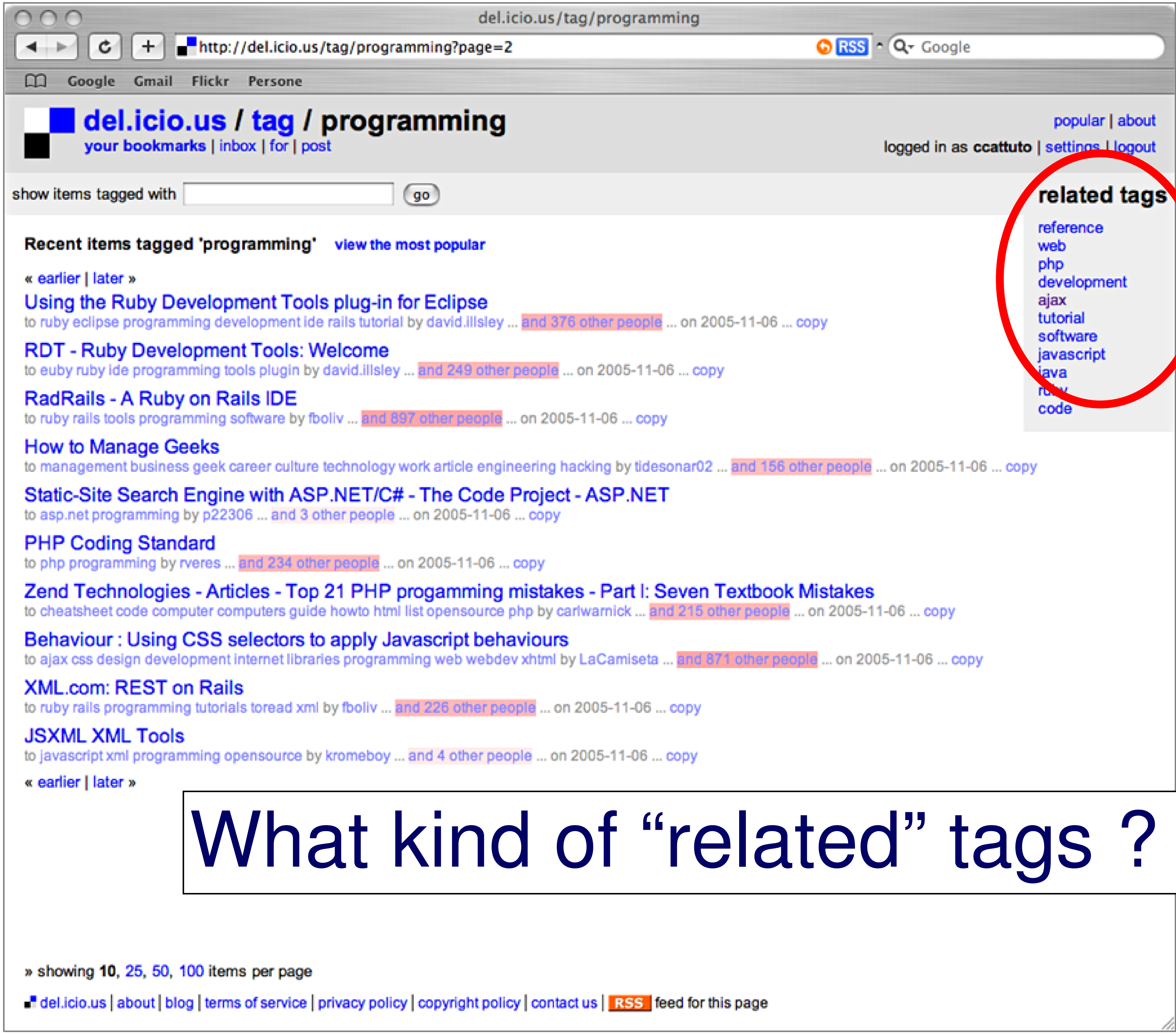
Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook



with C. Cattuto, University La Sapienza, Rome



What kind of “related” tags ?

Social Annotations

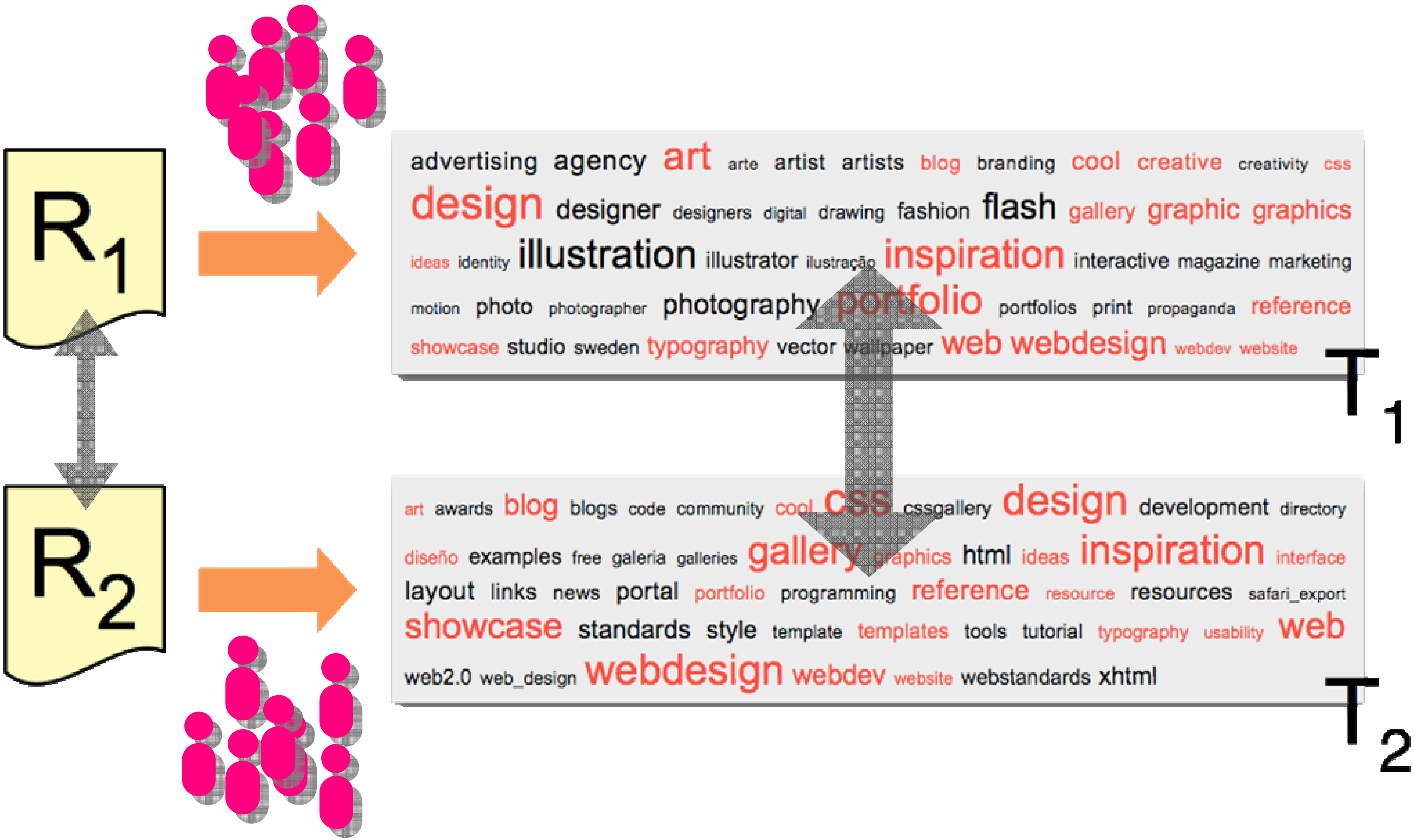
folksonomy level (macro)

semantics behavior

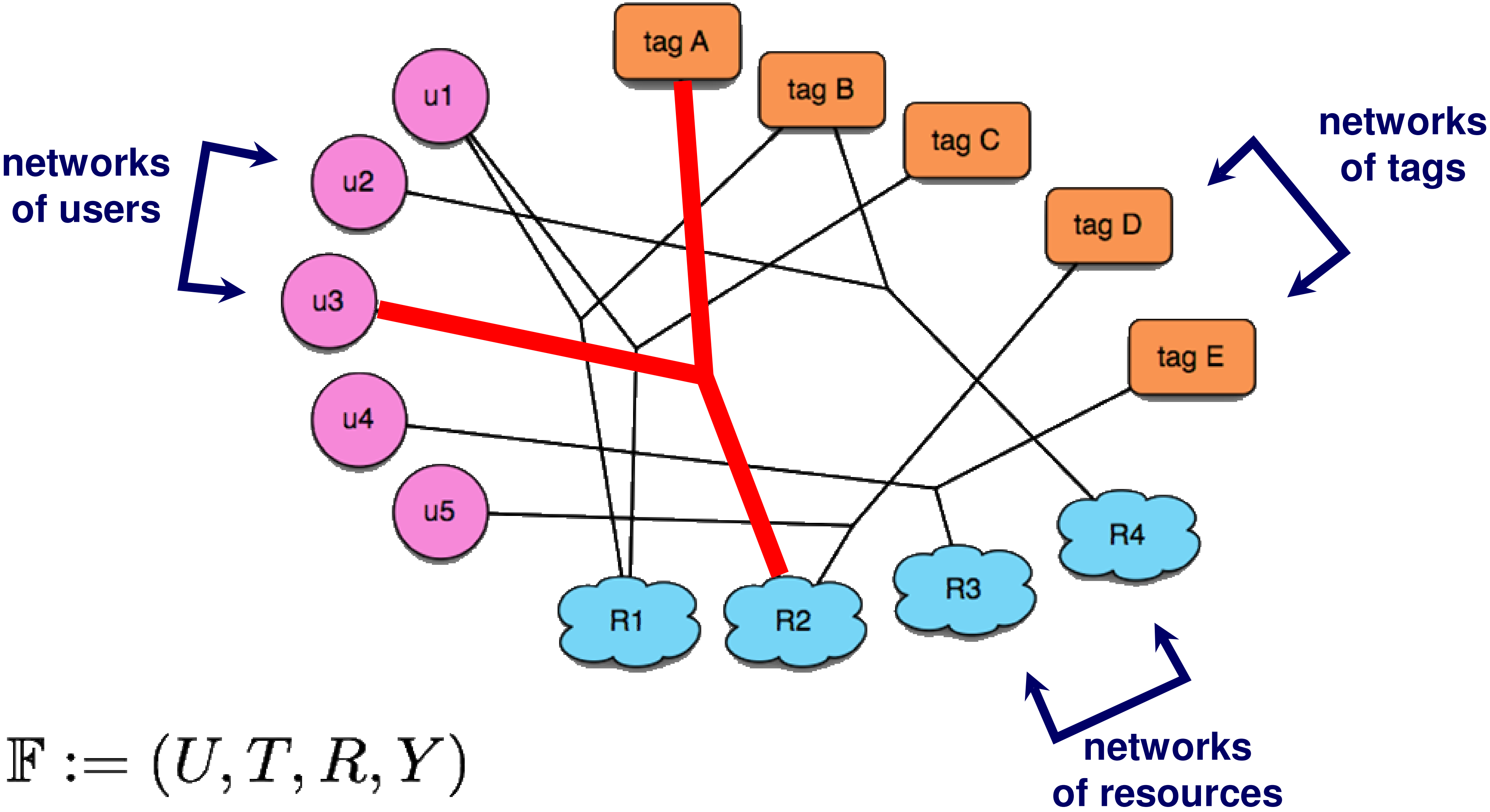
user level (micro)

- Understand the network
- Harvest semantics
- Extract concepts

social similarity



structural unit: (user, resource, tag)



$$\mathbb{F} := (U, T, R, Y)$$

$$Y \subseteq U \times T \times R$$

Topic Definition: Semantic Grounding of Tag similarity

- Final Goal: Understand “**tag semantics**” in a folksonomy, i.e.,
  - Which tags describe the same / a more specific / a more general concept?
- Two basic approaches:

Apply measures **directly to folksonomy structure** (e.g. cooccurrence statistics, …)

+ inclusion of complete vocabulary

- semantic interpretation of measures is not clear

Semantic Grounding →

Look up tags in **external thesaurus:**

+ semantically grounded metrics

- “folksonomy jargon” (misspel-lings, neologisms etc.) not present

→ **Understand characteristics** of (distributional) measures

→ assess their applicability for **concept extraction, ontology learning**, …

Folksonomy Dataset

- Delicious crawl 2006
  - $|U| = 667{,}128$ $|T| = 2{,}454{,}546$ $|R| = 18{,}782{,}132$
  - $|Y| = 140{,}333{,}714$

- Excerpt: 10,000 most popular tags
  - $|U| = 476{,}378$ $|T| = 10{,}000$ $|R| = 12{,}660{,}470$
  - $|Y| = 101{,}491{,}722$

- In the following: **tag rank** = position in most-popular list:
  - 1: design
  - 2: software
  - 3: blog
  - 4: web
  - …

Similarity Measures: Co-occurrence + Cosine

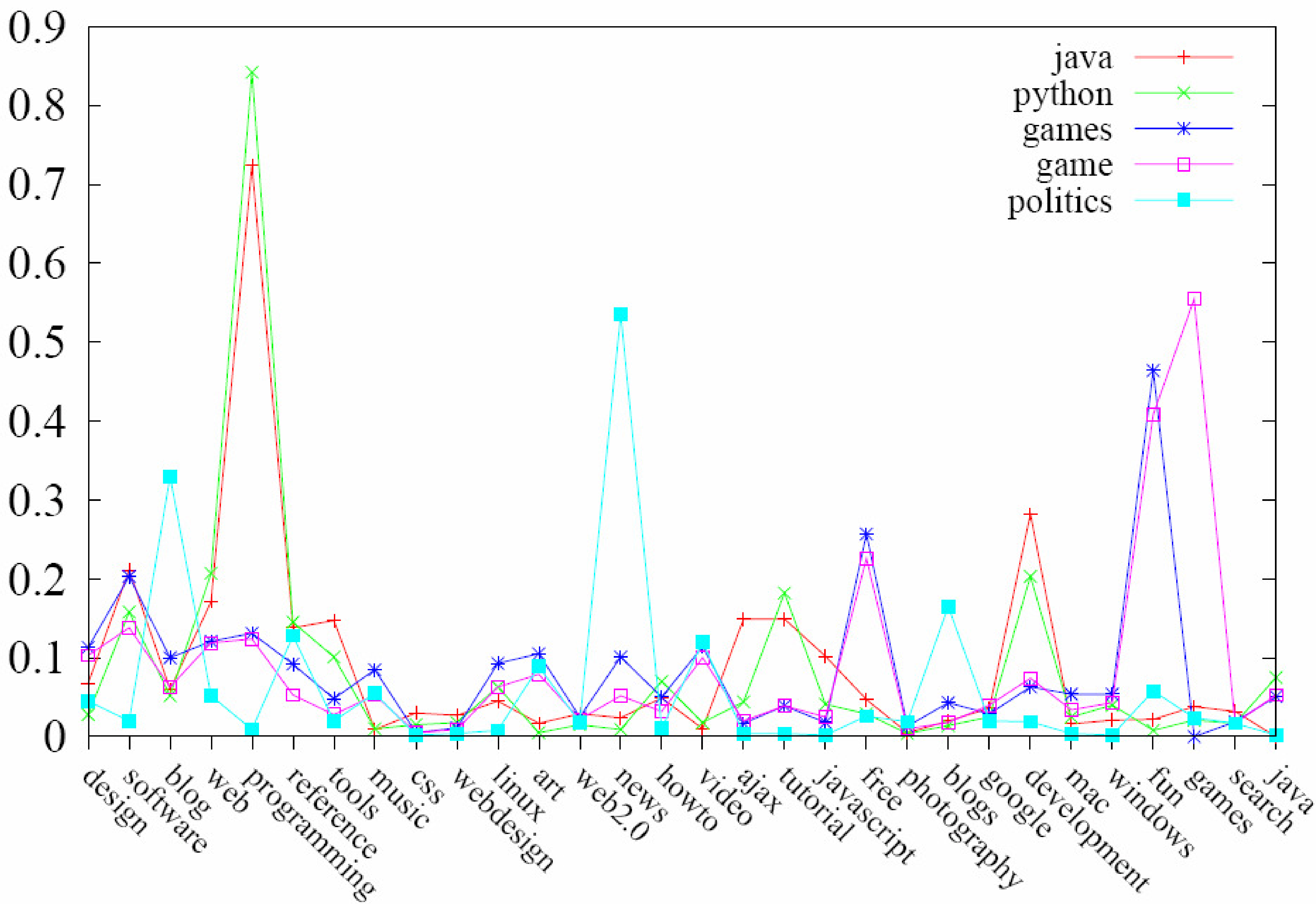
- Take **Co-occurrence frequency** as similarity measure (freq):
  $$freq(t_1, t_2) = |\{(u,r) \in U \times R : (u, t_1, r) \in Y \wedge (u, t_2, r) \in Y\}|$$

- Describe each tag as a **vector**, whereby each dimension of the vector space corresponds to another tag. Compute similar tags by **cosine similarity** (cosine).

  (The same can be done in the user space or the resource space and with TF-IDF.)

	design	software	blog	web	programming	...
JAVA	5	30	1	10	50	...

Example for cosine measure



Most related tags by cooccurrence / cosine simlarity

freq

art	design photography illustration blog graphics
web2.0	ajax web tools blog webdesign
news	blog technology politics media daily
howto	tutorial reference tips linux programming
video	music funny tv software media
ajax	javascript web2.0 web programming webdesign
tutorial	howto programming reference design css
javascript	ajax programming css web webdesign

cosine

art	graphic creative print portfolios nice
web2.0	web2 web-2.0 webapp “web web_2.0
news	blogs people weblog culture future
howto	how-to guide tutorials help how_to
video	entertainment awesome fun cool random
ajax	dhtml dom js ecmascript webdev
tutorial	tutorials tips coding code examples
javascript	webdevelopment webdev example examples webprogramming

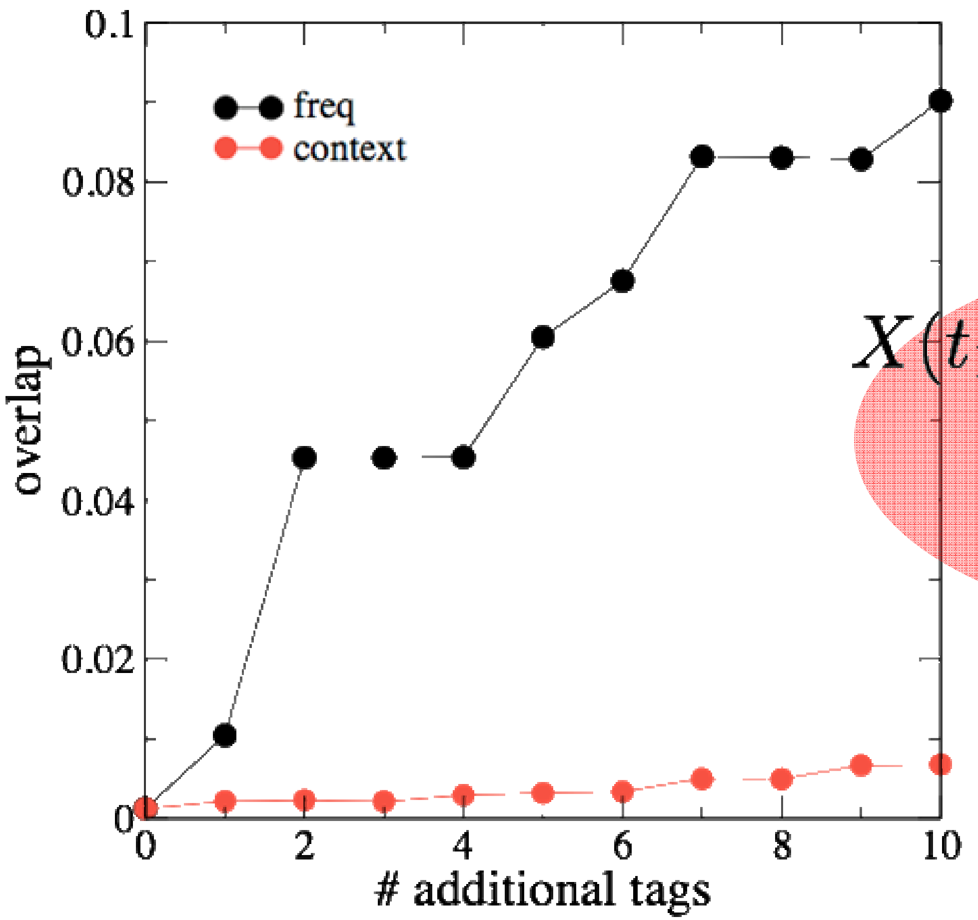
Resource experiment

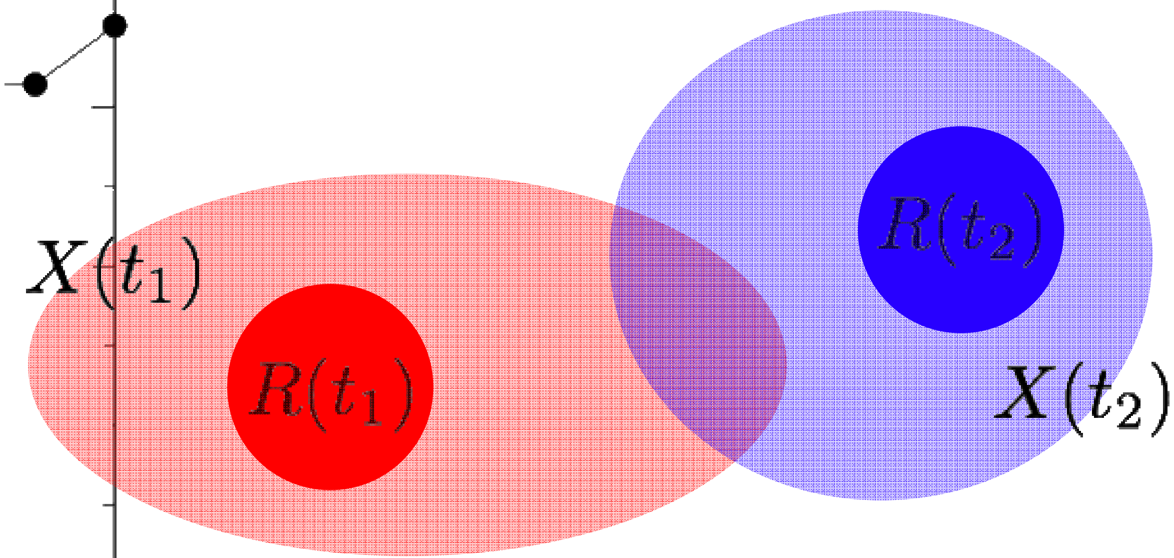
development: dev code coding developement developer

t_1 $t_{1,1}$ $t_{1,2}$

$R(t_1)$

$$\frac{|X(t_1) \cap X(t_2)|}{|X(t_1) \cup X(t_2)|}$$





t_1 : desktop

t_2 : physics

More Similarity Measures: User + Resource Context + Folkrank

- Two further possible context dimensions:
  - Users (***UserContext***)

JAVA

	8	2	0	3	10	
	John	Mary	Joe	Karl	Lucy	…

  - Resources (***ResourceContext***)

JAVA

	20	18	1	3	0	
	java.sun.com	javadev.de	google.com	hacking.com	lwa.de	…

  - (TF-IDF weighting showed no great effect)

- Use **FolkRank** to find related tags (folkrank).
  - Basic Idea: PageRank-like spreading of weights through folksonomy structure + high weights for a particular tag in the random surfer vector

Example: Most related tags for „web2.0“ and „howto“

WEB2.0

Sim. Measure	*1*	*2*	*3*	*4*	*5*
Coocc	ajax	web	tools	blog	webdesign
FolkRank	web	ajax	tools	design	blog
TagContext	web2	web-2.0	webapp	„web	web_2.0
ResourceCont.	web2	web20	2.0	web_2.0	web-2.0
UserContext	ajax	aggregator	rss	google	collaborate

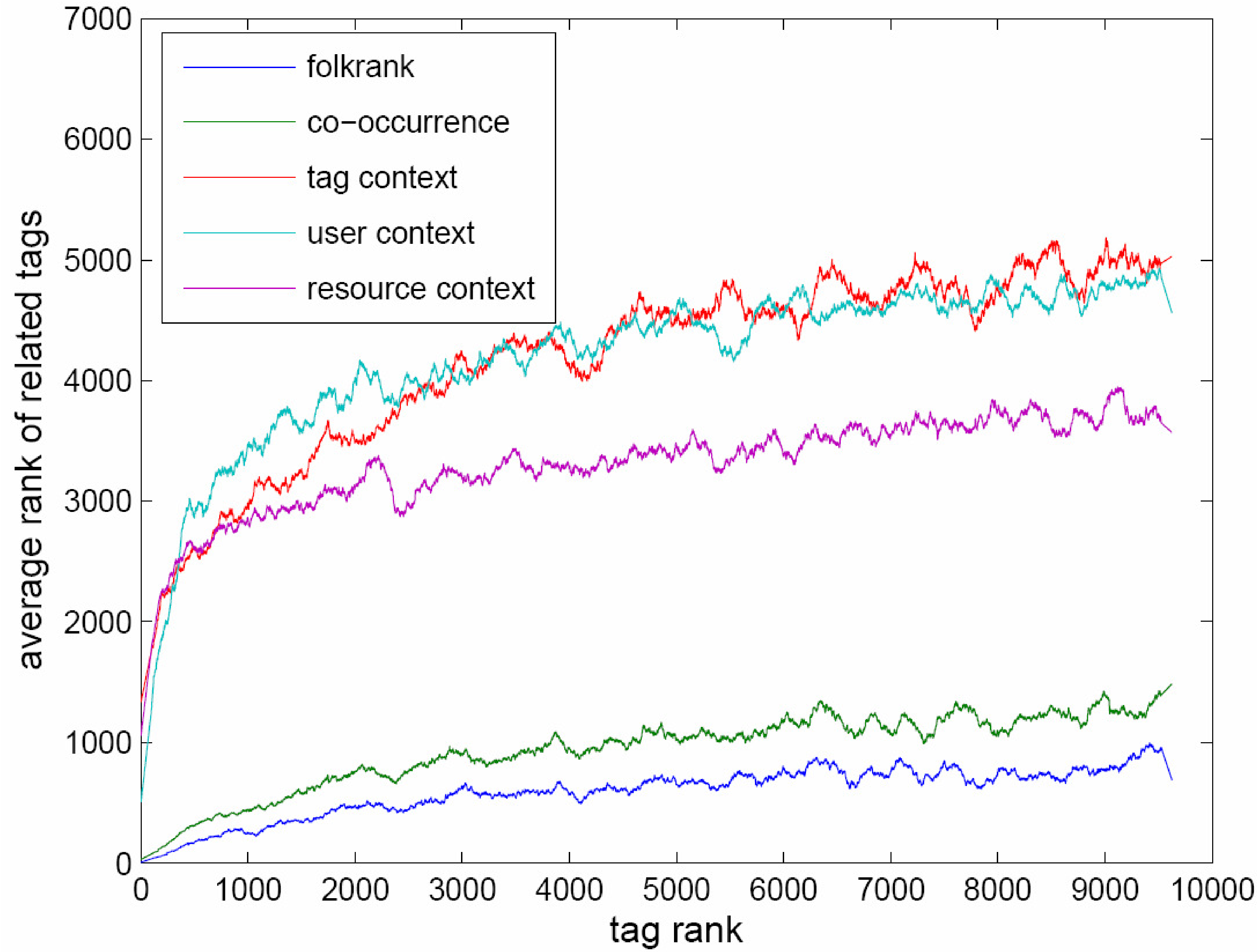
HOWTO

Coocc	tutorial	reference	tips	linux	programming
FolkRank	reference	linux	tutorial	programming	software
TagContext	how-to	guide	tutorials	help	how_to
ResourceCont.	how-to	tutorial	tutorials	tips	diy
UserContext	reference	tutorial	tips	hacks	tools

Qualitative insights: Overlap of 10 most related tags

	coocc	*FolkRank*	*Tag Context*	*Resource Context*
User Context	1.77	1.81	1.35	1.55
Resource Context	3.35	2.65	2.66	
Tag Context	1.69	1.28		
FolkRank	6.81			

Qualitative insights 2: Average rank of related tags



Semantic Grounding in WordNet

- WordNet is a large lexical database for English.
- Words with same meaning are grouped in *synsets*, which are ordered by an *is-a* hierarchy.
- Introduction of single **artificial root node** enables application of graph-based similarity metrics between pairs of nouns / pairs of verbs.
- Inclusion of top *n* Delicious tags in WordNet:
  - 100: 82%
  - 1,000: 79%
  - 5,000: 69%
  - 10,000: 61%

Example of Semantic Grounding

Original tag:

- „java“

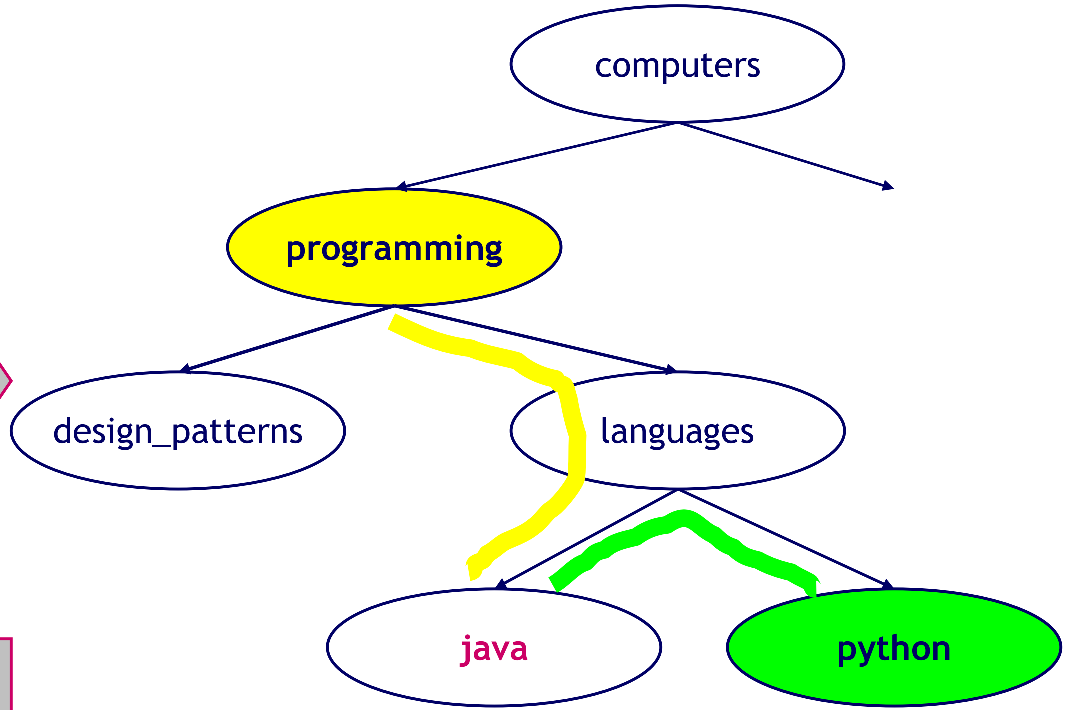
Most similar tag:

- Freq, folkrank: **„programming“**
- Cosine: **„python“**

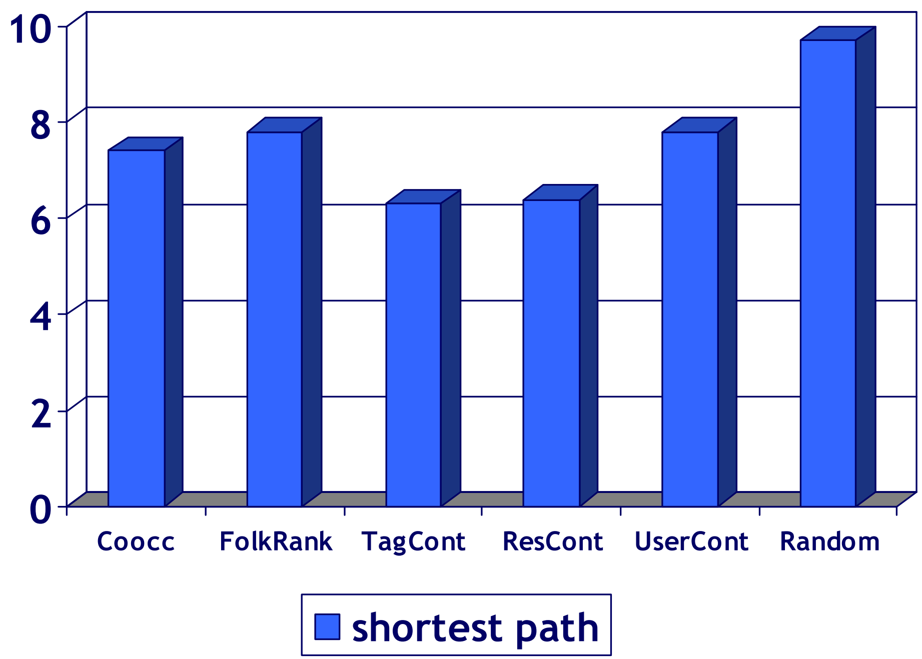
map

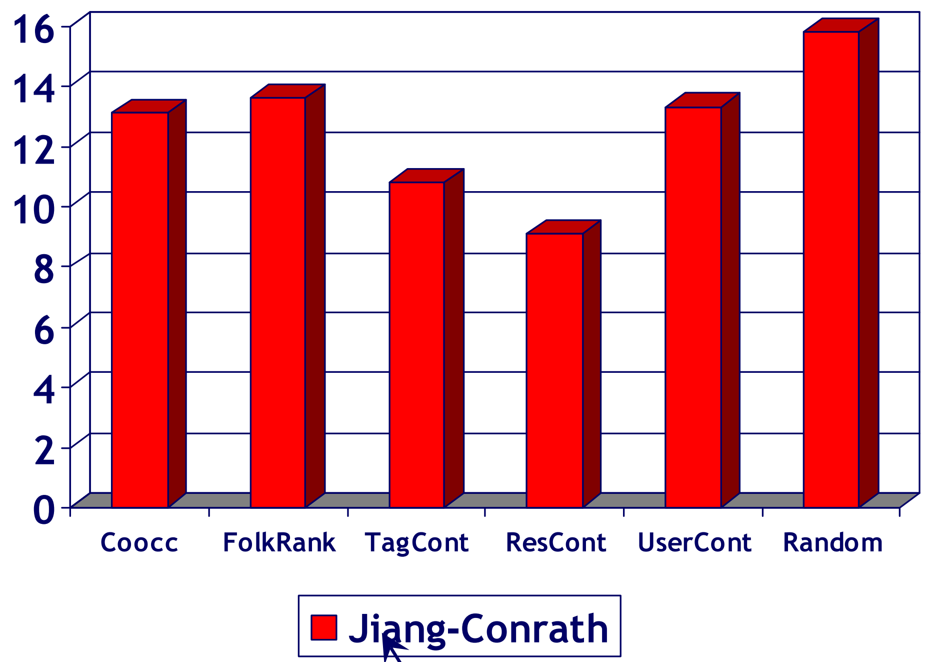
Grounded similarity

Wordnet Synset Hierarchy:



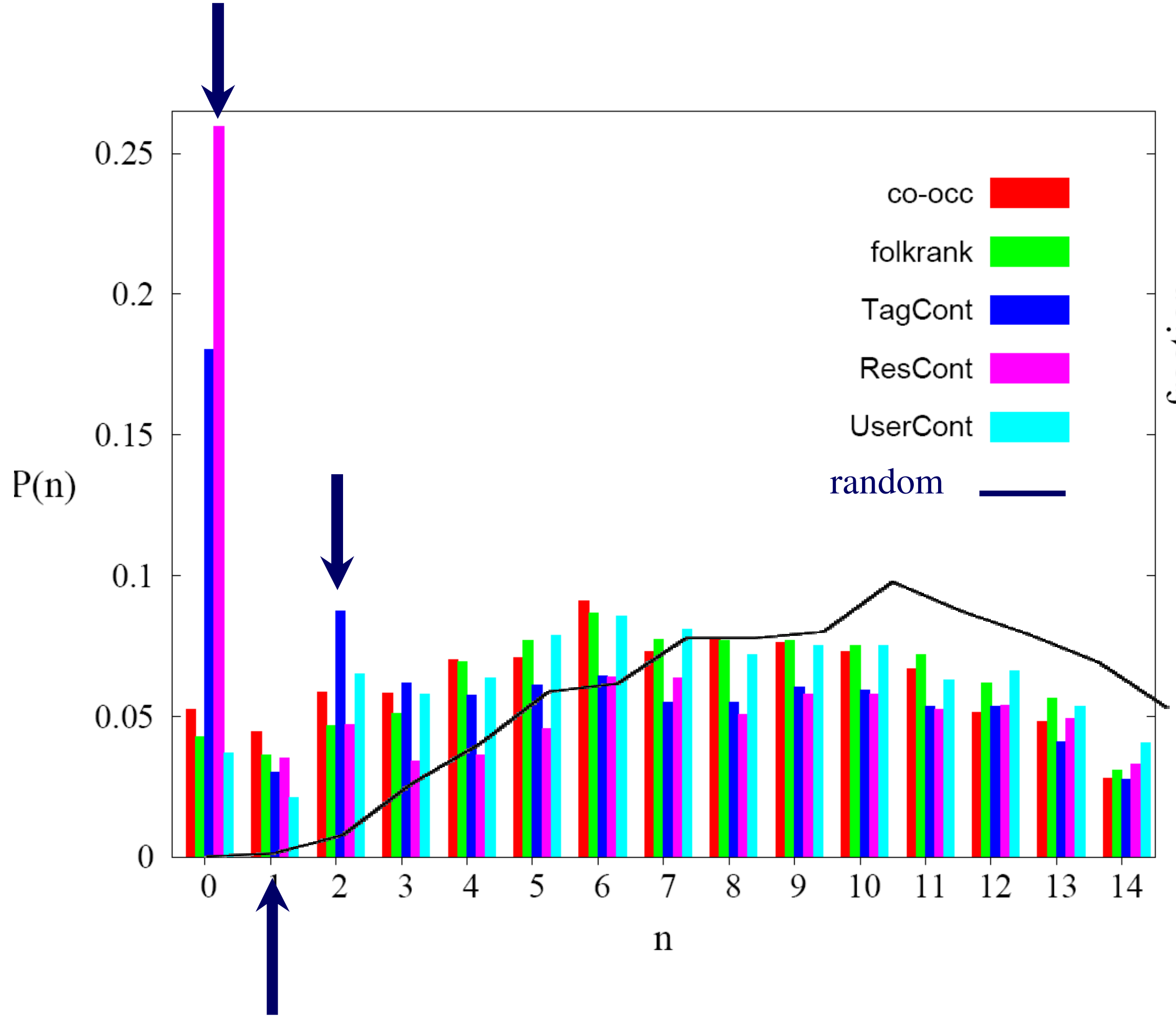
Shortest path between original tag and most closely related one



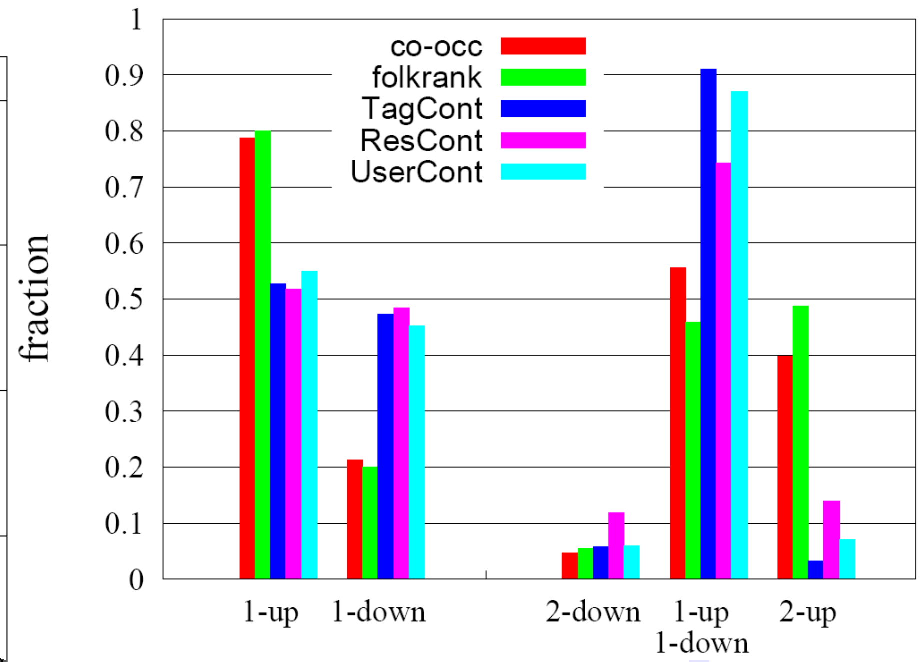


Shown to be the semantically most adequate measure for similarity within WordNet [Budanitsky, Hirst, 2006].

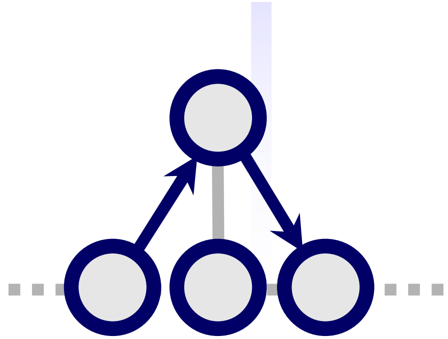
shortest paths in WordNet



length of shortest path to most related tag



siblings



Summary

Analysis of tag similarity measures by mapping to WordNet

Exposed clearly different characteristics:

- freq measure and Folkrank tend to more general tags
- Synonyms and siblings are the result of the cosine measure

Implications for ontology learning:

- Insights can inform the choice of an appropriate measure to extract semantic tag relations
- e.g, FolkRank to find Hyperonyms, Cosine measure for Synonyms

Next Step: Embed these measures in an ontology learning procedure

Ontology Learning

- Dominik Benz and Andreas Hotho. Position Paper: Ontology Learning from Folksonomies.. In Alexander Hinneburg, editor(s), LWA 2007: Lernen - Wissen - Adaption, Halle, September 2007, Workshop Proceedings (LWA), 109-112, Martin-Luther-University Halle-Wittenberg,2007.
- Francis Heylighen. Bootstrapping knowledge representations: from entailment meshes via semantic nets to learning webs. Kybernetes, (30)5/6:691--722, 2001.
- Paul Heymann and Hector Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. 2006-10 2006.
- P. Mika, Ontologies Are Us: A Unified Model of Social Networks and Semantics, Springer, 2005, 522-536.
- P. Schmitz, Inducing Ontology from Flickr Tags. 2006.

Analysis of tagging behaviour

- C. Cattuto, Semiotic dynamics in online social communities. The European Physical Journal C - Particles and Fields, 2006, 46, 33-37
- Shilad Sen and Shyong K. Lam and Al Mamunur Rashid and Dan Cosley and Dan Frankowski and Jeremy Osterhouse and F. Maxwell Harper and John Riedl. tagging, communities, vocabulary, evolution. CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, 181--190,ACM,New York, NY, USA,2006.

More under: http://www.bibsonomy.org/tag/ontology+folksonomy

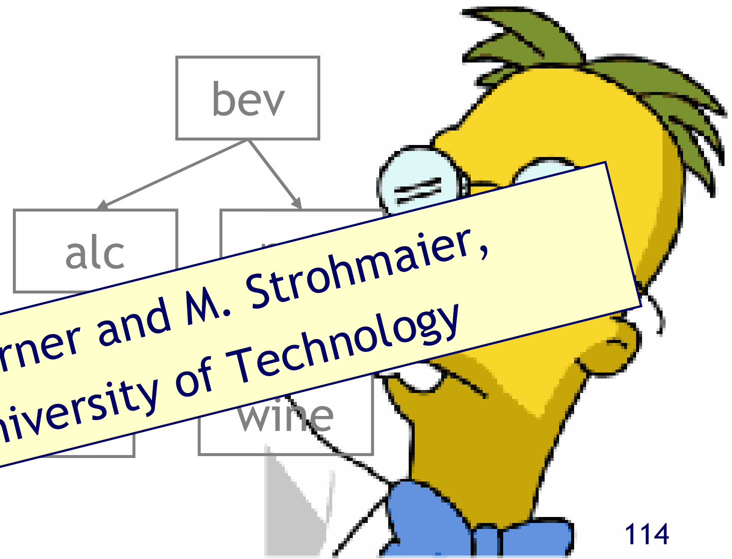
Agenda

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

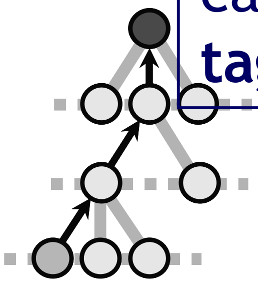
Summary and Outlook



with C. Körner and M. Strohmaier,
Graz University of Technology

The Story

Tag Similarity measures can capture **emergent tag semantics**



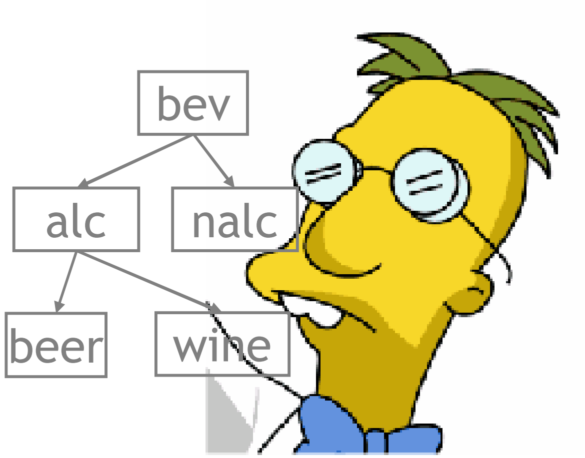
Pragmatics of tagging

Semantic **Implications** of Tagging Pragmatics

Conclusions

Evidence of different ways **HOW** users tag (Tagging **Pragmatics**)

Broad distinction by tagging **motivation** [Strohmaier2009]:



„Describers"...

- tag „verbously" with freely chosen words
- vocabulary not necessarily consistent (synomyms, spelling variants, ...)
- goal: describe content, ease retrieval

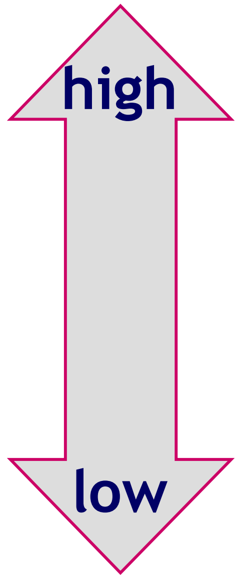
„Categorizers"...

- use a small controlled tag vocabulary
- goal: „ontology-like" categorization by tags, for later browsing
- tags a replacement for folders

Tagging Pragmatics: Measures

How to disinguish between two types of taggers?

Intuition: Describers use open set of many tags,
Categorizers use small set of controlled tags:



Vocabulary size:

$$vocab(u) = |T_u|$$

Tag / Resource ratio:

$$trr(u) = \frac{|T_u|}{|R_u|}$$

Average # tags per post:

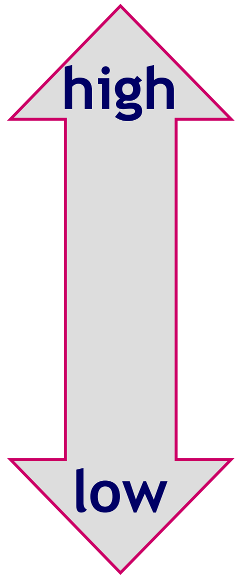
$$tpp(u) = \frac{\sum |T_{ur}|}{|R_u|}$$

- Next Intuition: Describers don‘t care about „abandoned“ tags, Categorizers do

- **Orphan ratio:**

$$orphan(u) = \frac{|T_u^o|}{|T_u|}$$

$$T_u^o = \{t \,|\, |R(t)| \leq n\}, n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil$$

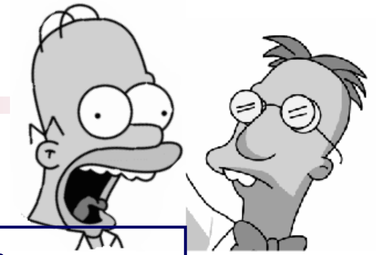


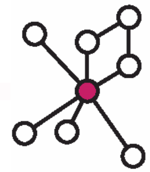
- $R(t)$: set of resources tagged by user u with tag t

Tagging pragmatics: Limitations of measures

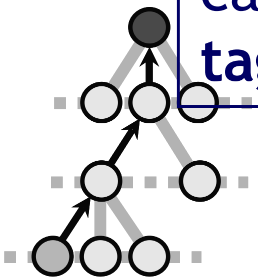
- Real users: no „perfect“ Categorizers / Describers, but **„mixed“ behaviour**
- Possibly influenced by **user interfaces** / recommenders
- Measures are correlated
- But: independent of **semantics**; measures capture **usage patterns**

The Story





Tag Similarity measures can capture **emergent tag semantics**

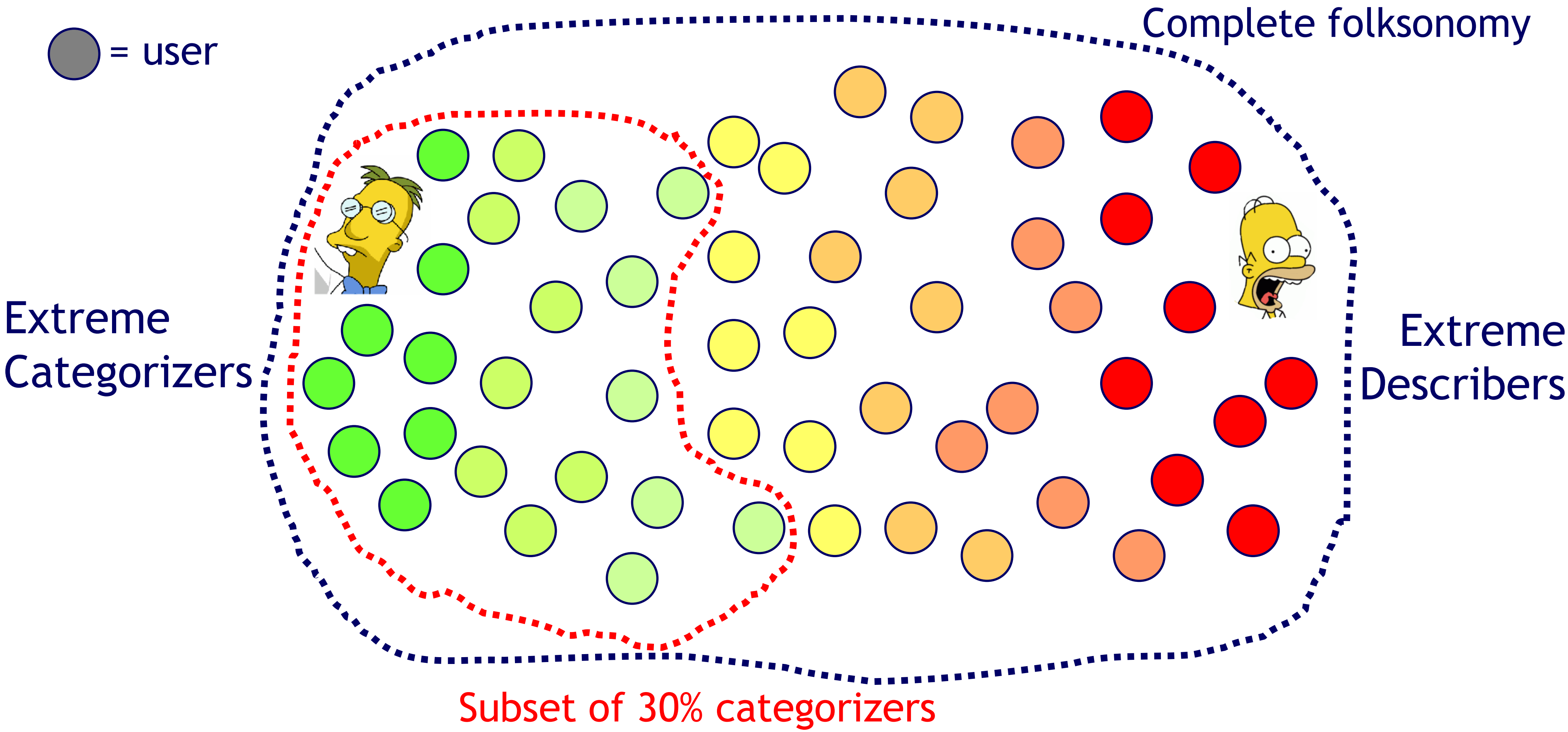


Measures of **tagging pragmatics** differentiate users by tagging motivation

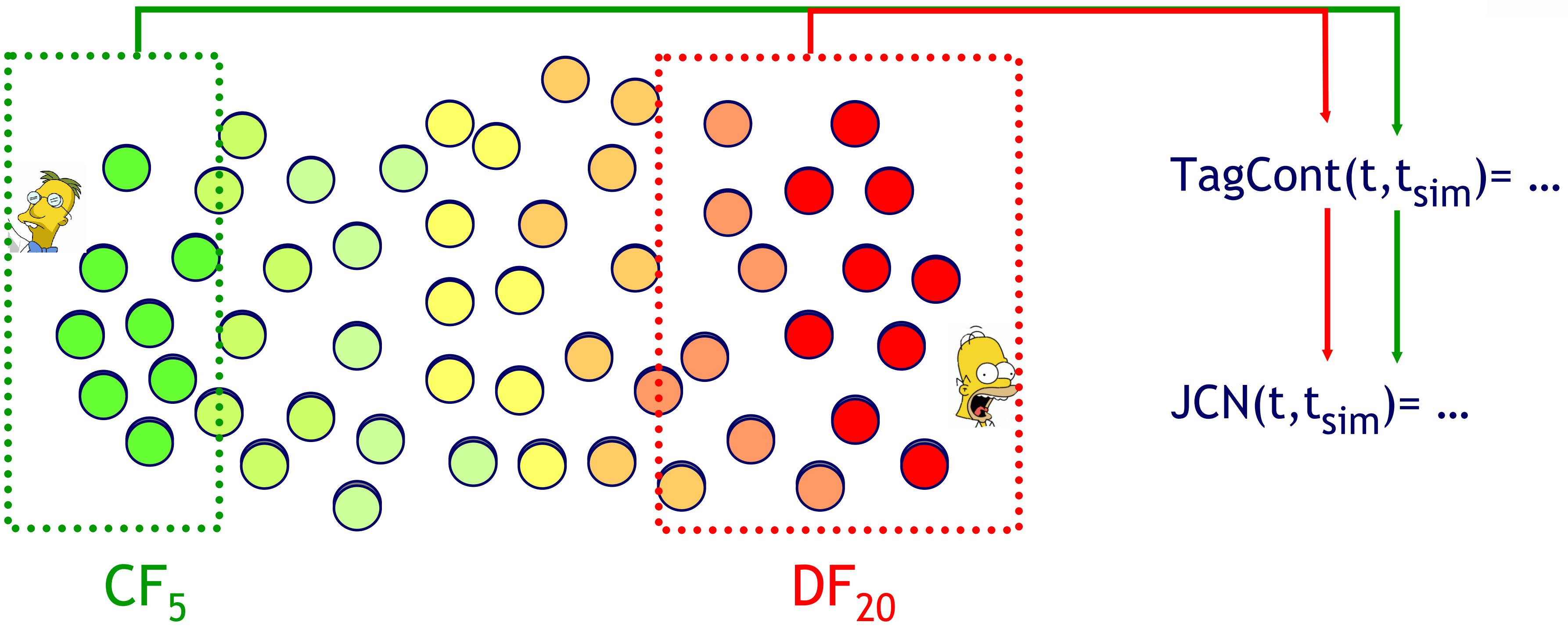
Semantic **Implications** of Tagging Pragmatics

Conclusions

Idea: Can we learn the same (or even better) semantics from the folksonomy induced by a **subset** of describers / categorizers?



Experimental setup



- Apply pragmatic measures *vocab*, *trr*, *tpp*, *orphan* to each user
- Systematically create **„sub-folksonomies“** CF_i / DF_i by subsequently adding i % of Categorizers / Describers (i = 1,2,…,25,30,…,100)
- Compute **similar tags** based on each subset (TagContext Sim.)
- Assess (semantic) **quality** of similar tags by **avg. JCN** distance

Dataset

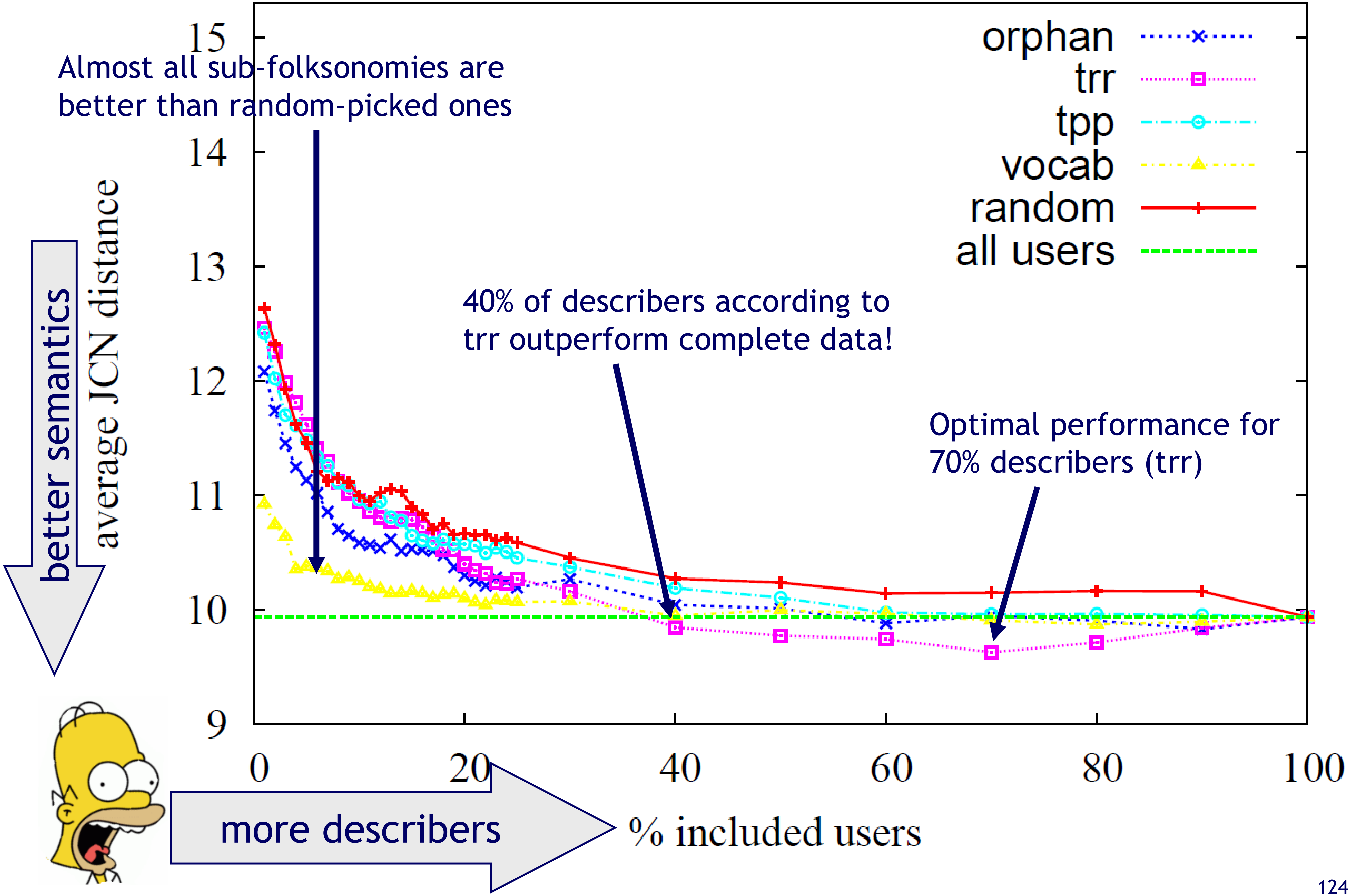
From Social Bookmarking Site **Delicious** in 2006 → ORIGINAL

Two filtering steps (to make measures more meaningful):

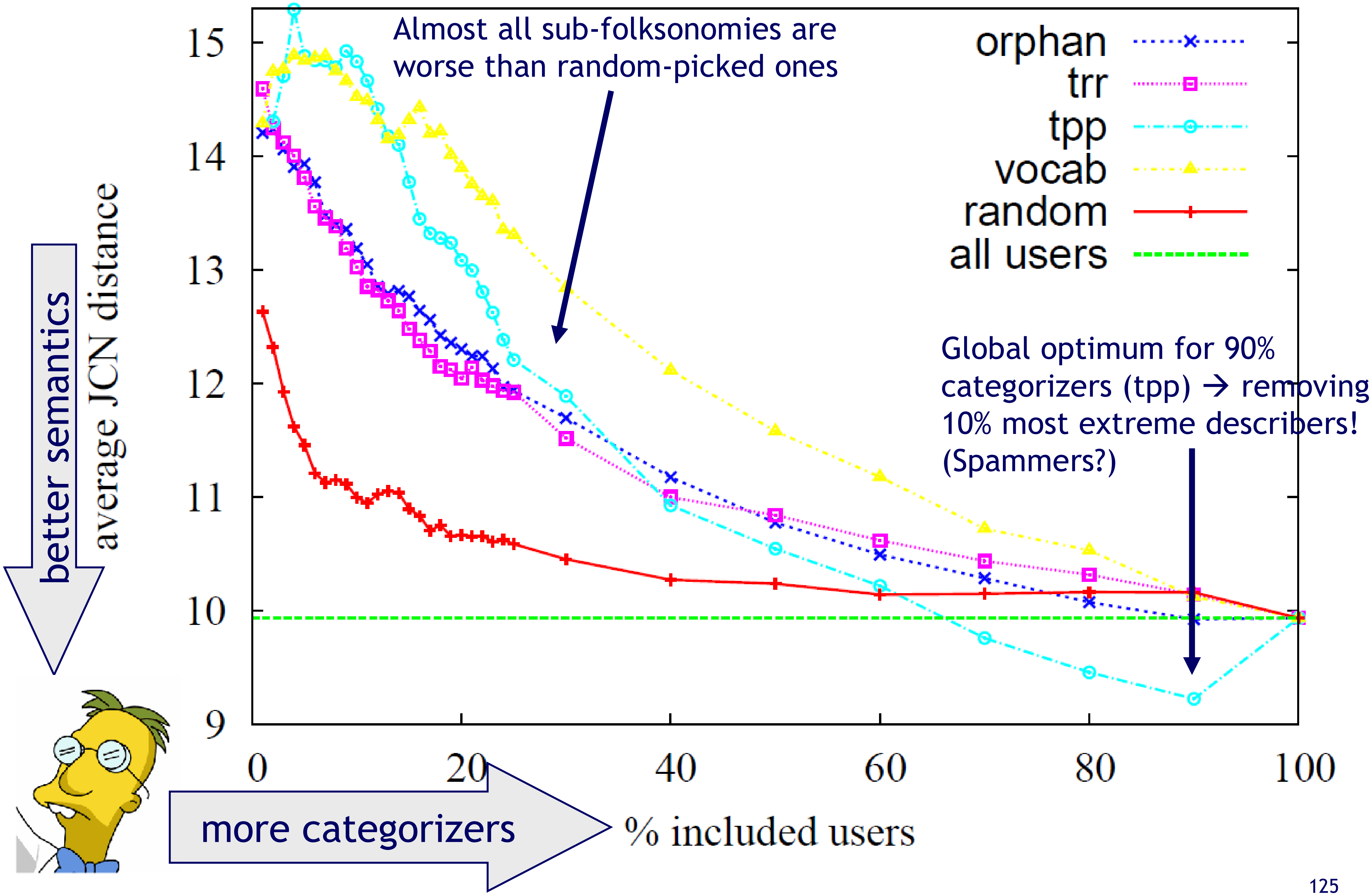
- Restrict to **top 10.000 tags** → FULL
- Keep only users with **> 100 resources** → MIN100RES

dataset	|T|	|U|	|R|	|Y|
ORIGINAL	2,454,546	667,128	18,782,132	140,333,714
FULL	10,000	511,348	14,567,465	117,319,016
MIN100RES	9,944	100,363	12,125,176	96,298,409

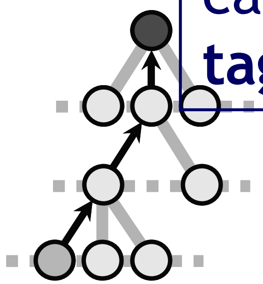
Results – adding Describers (DF_i)



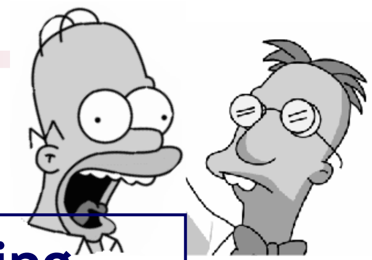
Results – adding Categorizers (CF_i)



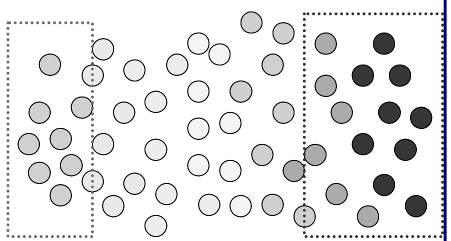
The Story



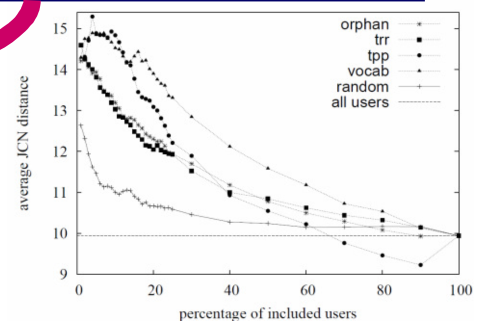
Tag Similarity measures can capture **emergent tag semantics**



Measures of **tagging pragmatics** differentiate users by tagging motivation



Sub-folksonomies introduced by measures of pragmatics show different semantic qualities



Conclusions

- Introduction of **measures** of users‘ **tagging motivation** (Categorizers vs. Describers)
- Evidence for **causal link** between tagging **pragmatics** (HOW people use tags) and tag **semantics** (WHAT tags mean)
- „Mass matters“ for „wisdom of the crowd“, but **composition of crowd** makes a difference („**Verbosity**“ of describers in general better, but with a limitation)
- Relevant for **tag recommendation** and **ontology learning** algorithms

References

[Cattuto2008] Ciro Cattuto, Dominik Benz, Andreas Hotho, Gerd Stumme: *Semantic Grounding of Tag Relatedness in Social Bookmarking Systems*. In: Proc. 7th Intl. Semantic Web Conference (2008), p. 615-631

[Markines2009] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, Gerd Stumme: *Evaluating Similarity Measures for Emergent Semantics of Social Tagging*. In: Proc. 18th Intl. World Wide Web Conference (2009), p.641-641

[Strohmaier2009] Markus Strohmaier, Christian Körner, Roman Kern: *Why do users tag? Detecting users' motivation for tagging in social tagging systems*. Technical Report, Knowledge Management Institute – Graz University of Technology (2009)

Introduction

- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

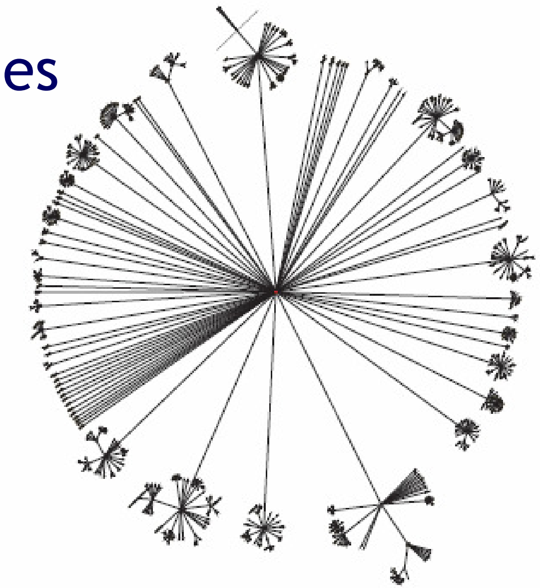
Understanding Folksonomy Data

- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning

- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

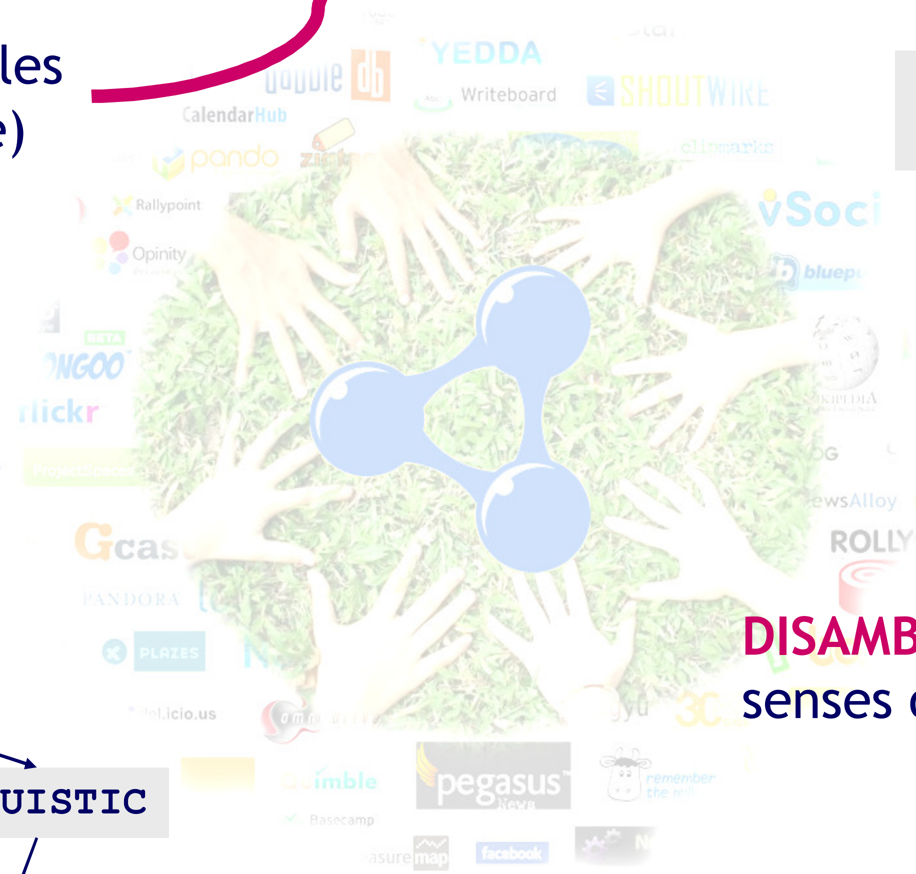


Steps of learning a concept hierarchy from tags

INPUT: tagging triples (tag, user, resource)

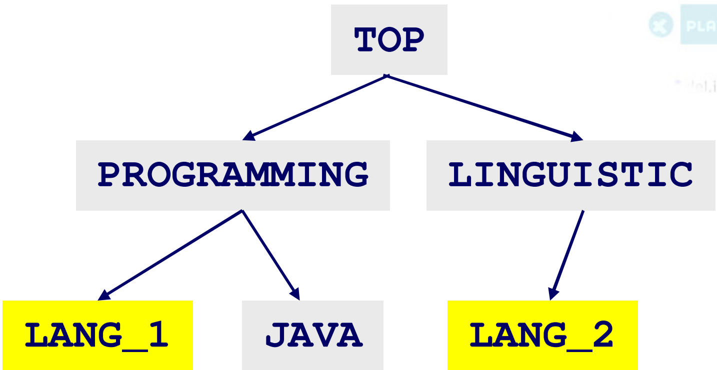
SYNSETIZE: unite tags with same / similar meaning

```
color, colour  -> COLOR
lang, language -> LANG
```



DISAMBIGUATE: differentiate senses of synsetized tags

```
LANG  -> LANG_1 LANG_2
APPLE -> APPLE_1, APPLE_2
```



LEARN HIERARCHY: assemble relations among tags

Learning Approaches

learning of tag relations

Social Network Analysis
- centrality
- clustering coefficient

[Mika, 2005]
[Heymann, 2006]

Statistical approaches
- model of sub-sumption
- association rules

[Schmitz, 2006]
[Schmitz et al., 2006]

Clustering

- e.g. HAC

[Begelmann, 2006]

Tag (co-)occurrence:
most general / resource independent
user-based / resource-based co-occurrence

- Algorithm:
  1. Initial: setup root node
  2. Extract Tags (do filtering, sort by generality)
  3. Iteratively add tags to the ontology, by
     1. Connect the most general one with root
     2. Connect the other with the most similar one of the ontology (only if the maximal degree is small engough)
     3. Connect with root of no similar tag exists (sim > min_sim)
  4. End if all tags are added

Example

	min_occ	min_sim	min_gen	max_children
Input:	2000	0,2	0,55	15

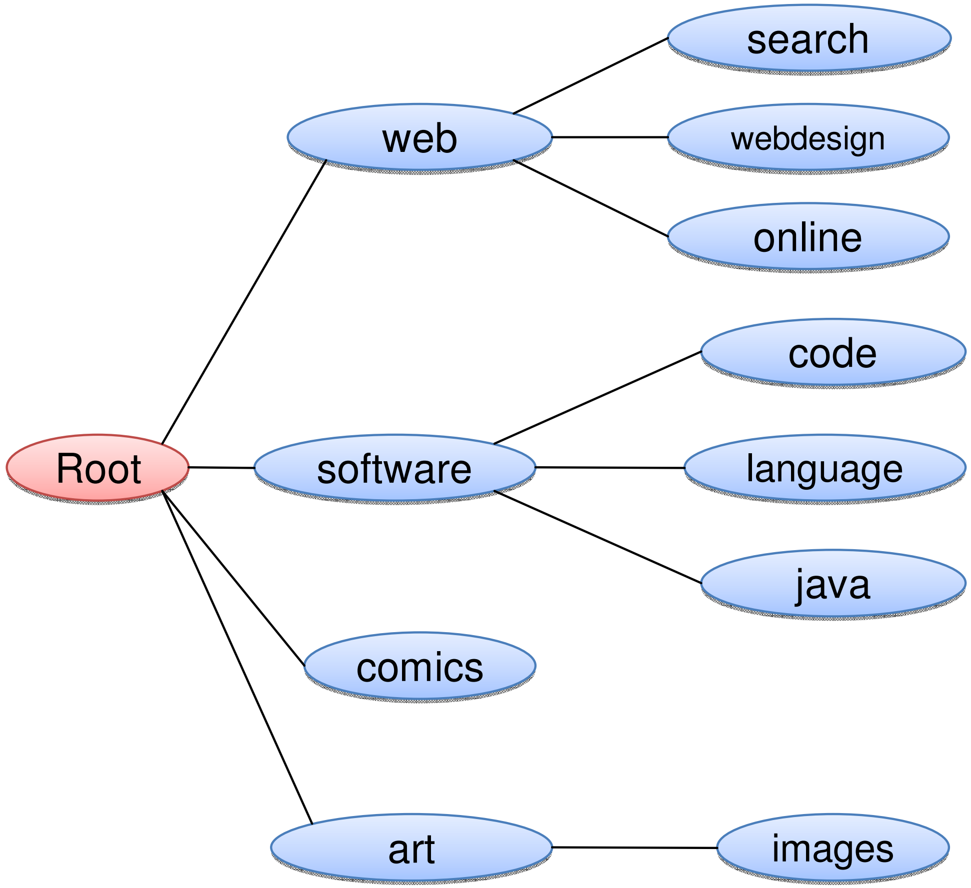
tag ordered by generality

web	0,88
software	0,75
art	0,56
online	0,45
search	0,41
java	0,41
code	0,40
comics	0,37
language	0,36
images	0,36
webdesign	0,35

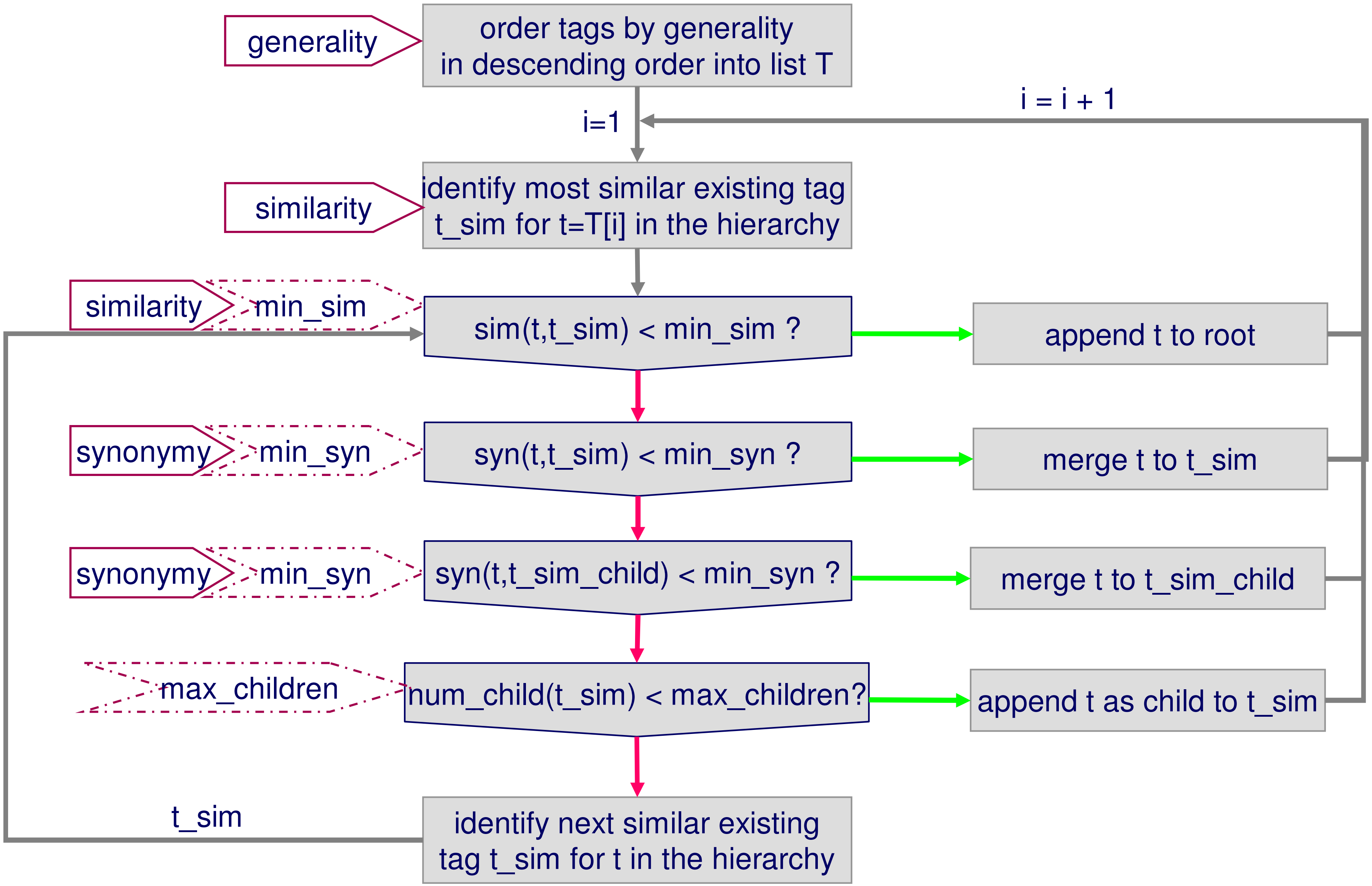
most similar tags

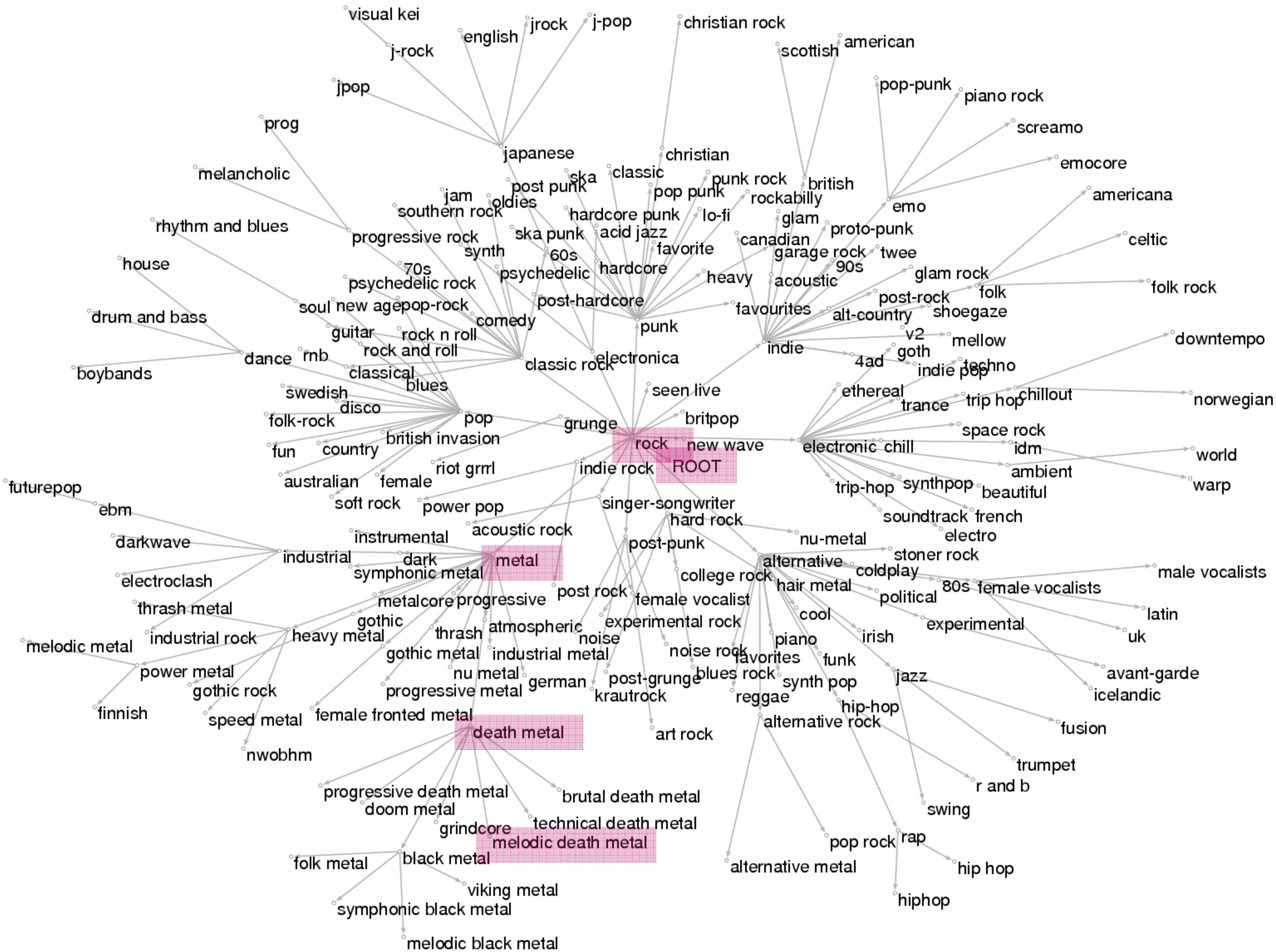
0,65	web
0,48	software
0,30	code

Ontology



Adopted Algorithm by Benz et.al.



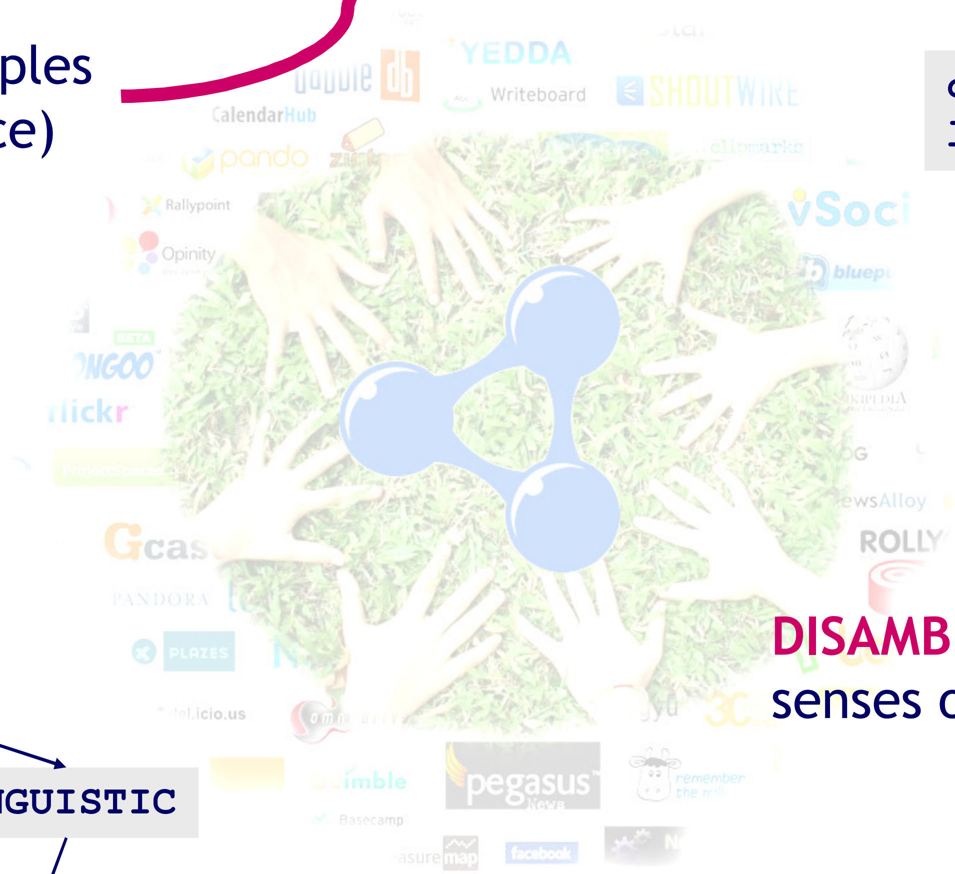

visual kei
jrock
j-pop
christian rock
english
j-rock
scottish
american
jpop
pop-punk
piano rock
prog
screamo
japanese
christian
melancholic
ska
classic
punk rock
emocore
post punk
jam
pop punk
british
oldies
rockabilly
americana
southern rock
hardcore punk
lo-fi
emo
rhythm and blues
glam
proto-punk
ska punk
acid jazz
progressive rock
canadian
celtic
synth
60s
favorite
twee
garage rock
house
70s
hardcore
90s
psychedelic
heavy
glam rock
psychedelic rock
acoustic
folk rock
folk
post-rock
new age
pop-rock
post-hardcore
soul
favourites
shoegaze
drum and bass
comedy
alt-country
punk
guitar
rock n roll
v2
mellow
downtempo
indie
rock and roll
dance
rnb
goth
4ad
electronica
boybands
techno
classic rock
indie pop
classical
blues
seen live
ethereal
swedish
chillout
trip hop
norwegian
trance
disco
pop
britpop
grunge
folk-rock
space rock
british invasion
rock
new wave
fun
country
electronic
chill
idm
world
indie rock
ROOT
riot grrl
ambient
australian
female
trip-hop
synthpop
warp
beautiful
futurepop
soft rock
power pop
singer-songwriter
ebm
soundtrack
french
hard rock
acoustic rock
darkwave
instrumental
nu-metal
electro
post-punk
stoner rock
industrial
dark
metal
alternative
coldplay
male vocalists
symphonic metal
electroclash
college rock
hair metal
80s
female vocalists
post rock
political
metalcore
progressive
female vocalist
latin
thrash metal
cool
gothic
experimental rock
experimental
heavy metal
thrash
atmospheric
irish
industrial rock
noise
piano
uk
melodic metal
gothic metal
industrial metal
noise rock
funk
favorites
avant-garde
power metal
nu metal
blues rock
jazz
german
post-grunge
synth pop
icelandic
gothic rock
progressive metal
krautrock
reggae
finnish
hip-hop
speed metal
female fronted metal
alternative rock
death metal
art rock
fusion
nwobhm
trumpet
progressive death metal
brutal death metal
r and b
doom metal
swing
grindcore
technical death metal
melodic death metal
rap
pop rock
black metal
folk metal
hip hop
alternative metal
viking metal
symphonic black metal
hiphop
melodic black metal

Steps of learning a concept hierarchy from tags

INPUT: tagging triples (tag, user, resource)

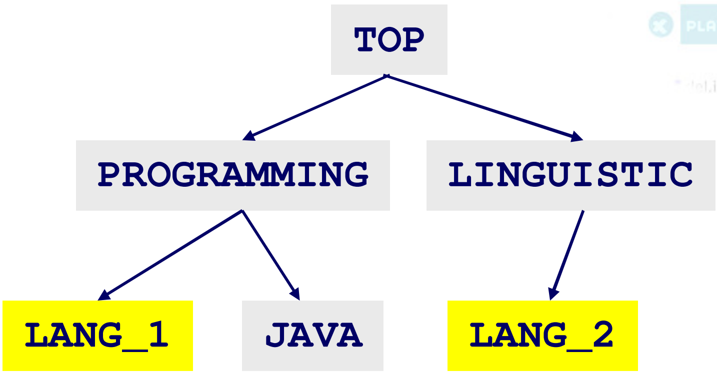
SYNSETIZE: unite tags with same / similar meaning

```
color, colour   -> COLOR
lang, language  -> LANG
```



DISAMBIGUATE: differentiate senses of synsetized tags

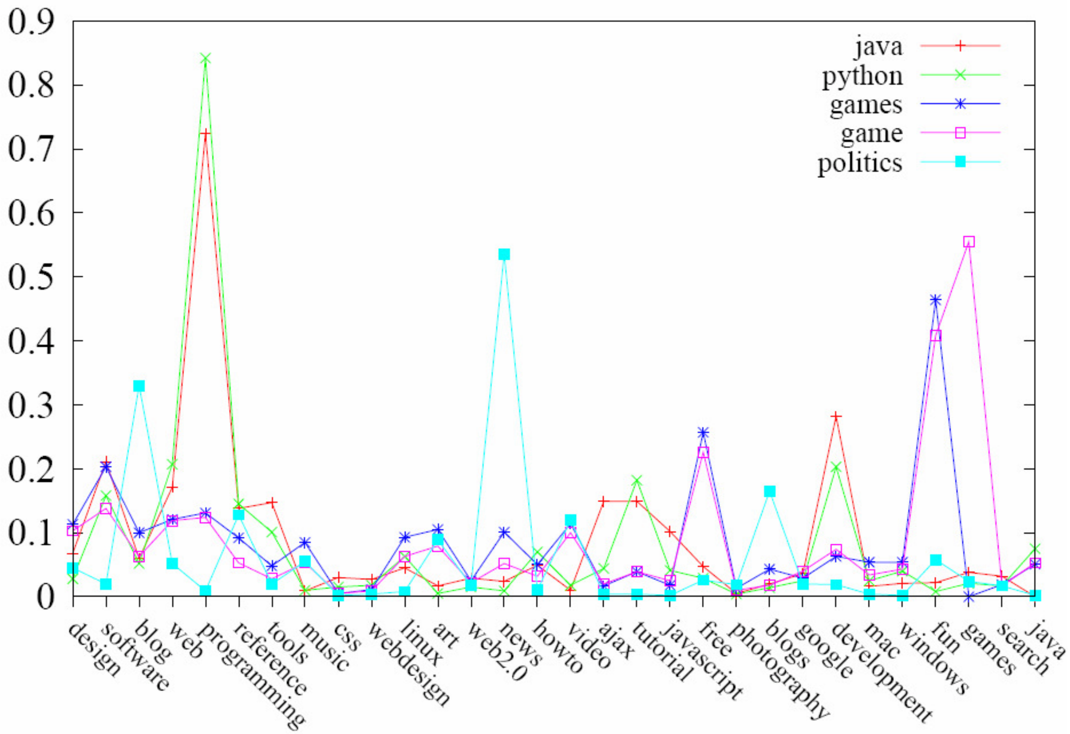
```
LANG   -> LANG_1 LANG_2
APPLE  -> APPLE_1, APPLE_2
```



LEARN HIERARCHY: assemble relations among tags

SYNSETIZE: Unite tags with same / similar meanings

Represent tags by their „**co-occurrence fingerprint**“:

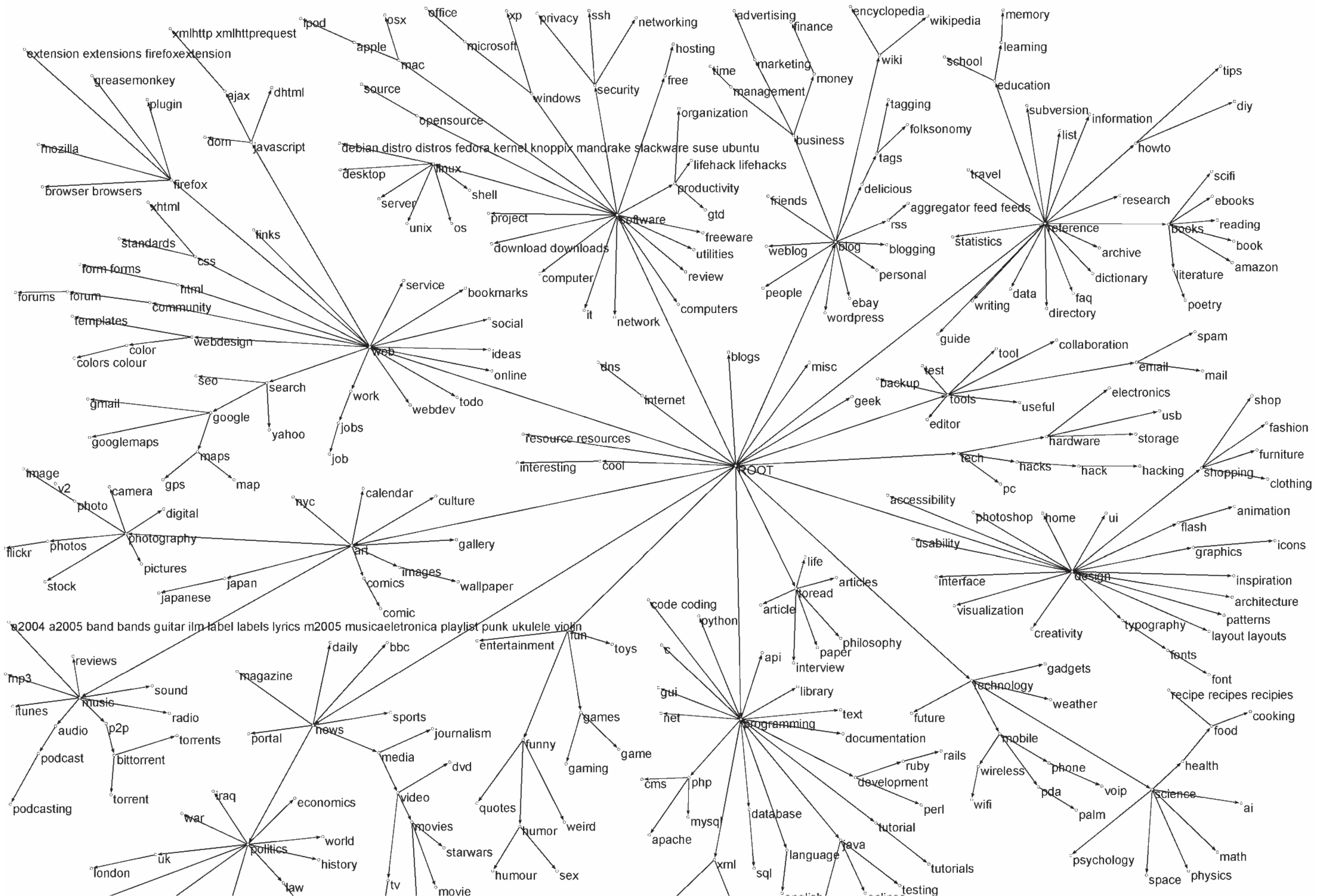


Compute pairwise **cosine similarity** among fingerprint vectors

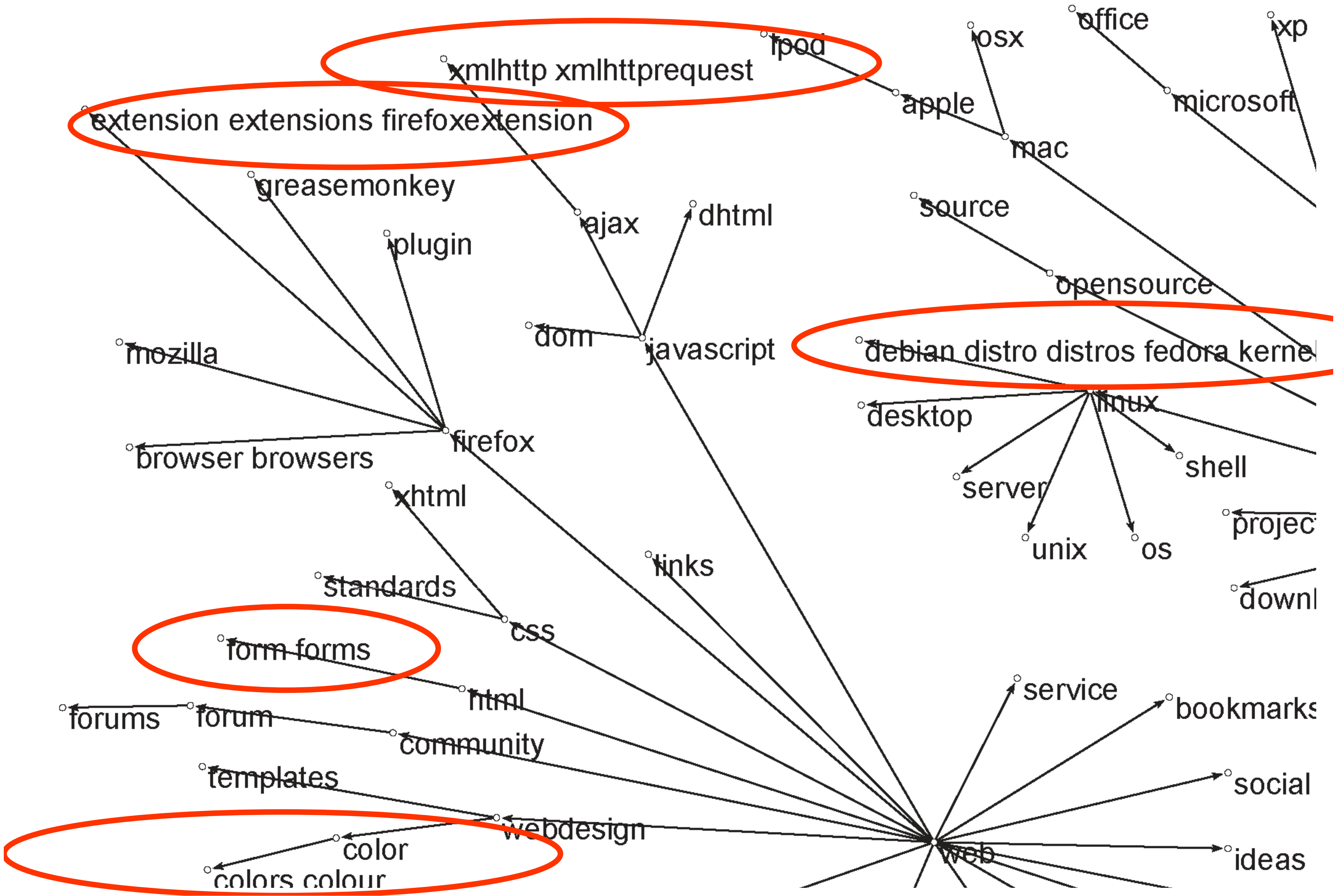
Apply **threshold** → „Synsets“

```
game, games -> GAME
...
```

Results for delicious with SYNSETIZE step

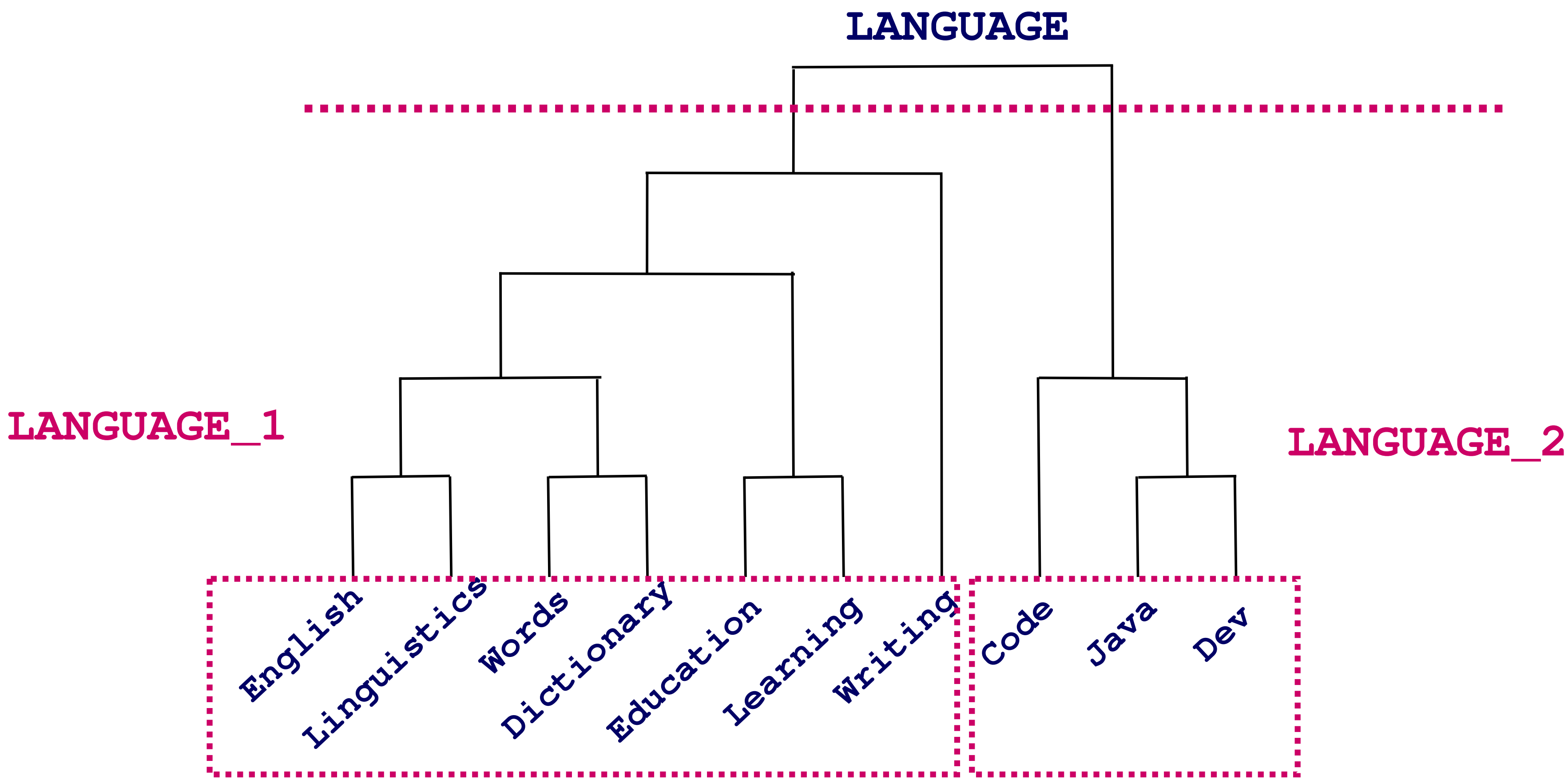


Results for delicious with SYNSETIZE step



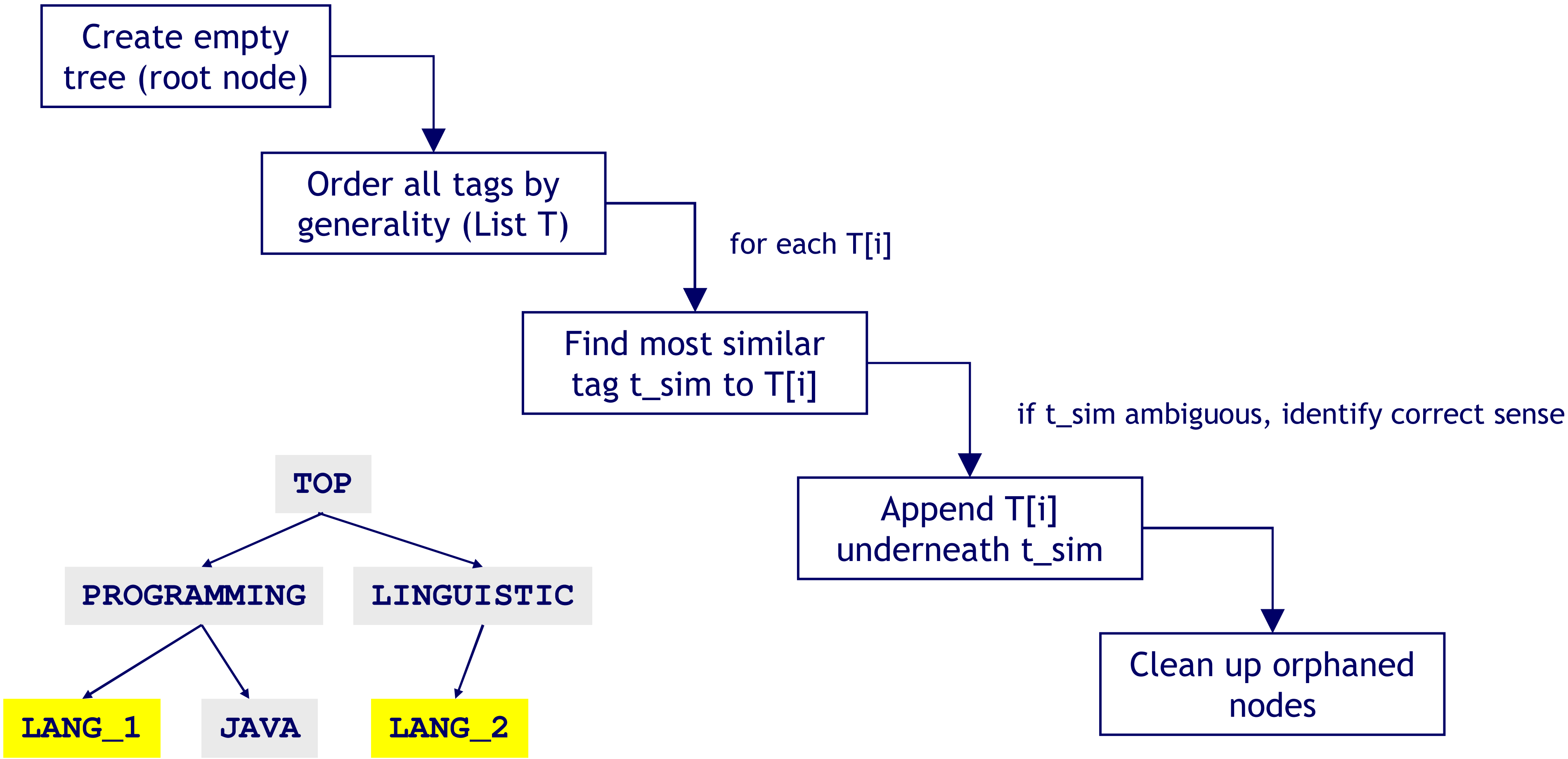
DISAMBIGUATE: differentiate senses of synsetized tags

- Idea: cluster co-occuring tags (hierarchical agglomerative method)
- Represent senses by „preference tags“



LEARN HIERARCHY: assemble tag relations

- Subsequently add tags to evolving tree structure (simplified scheme):



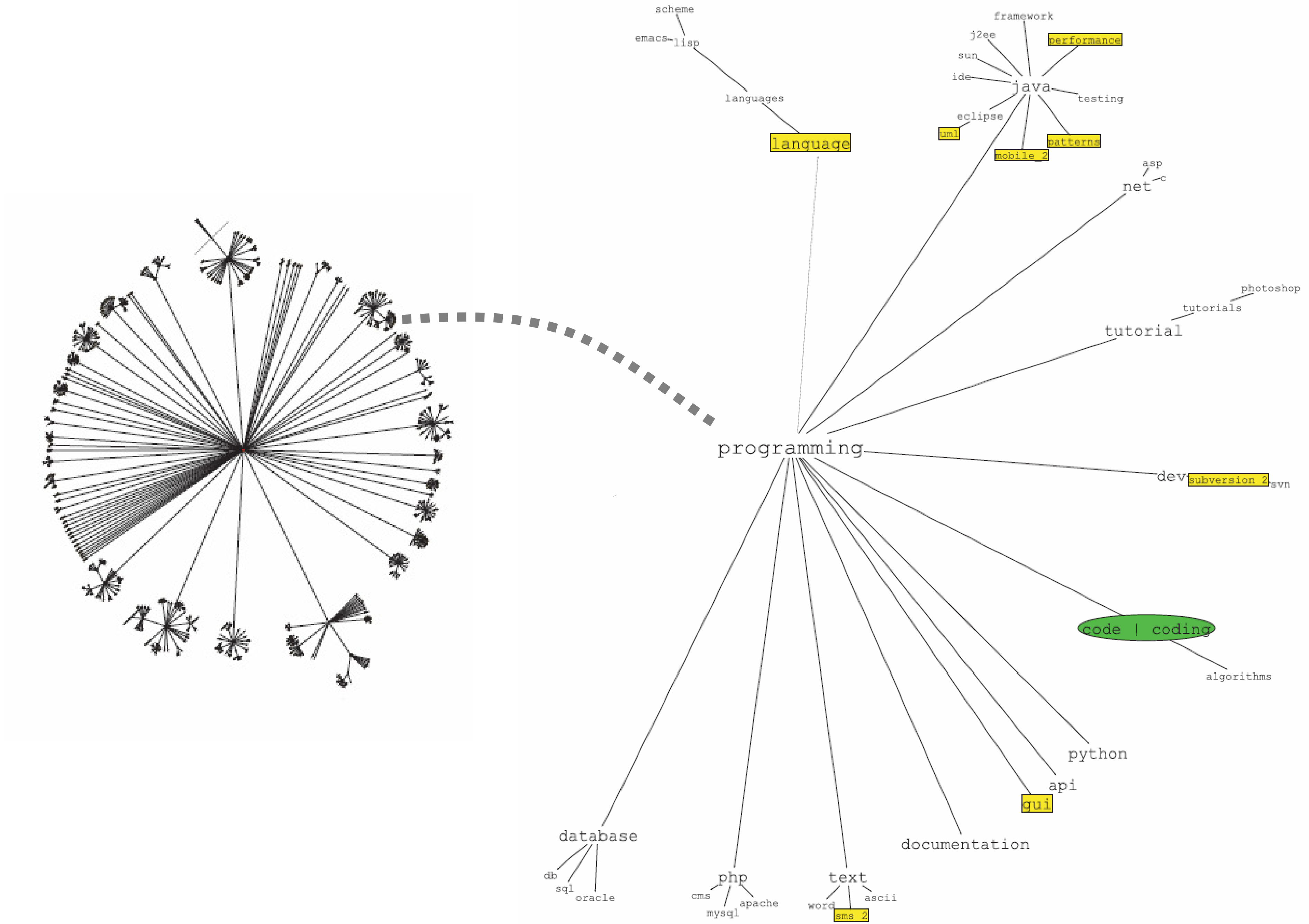
Postprocessing for tags directly connected with root

Replacing, delete or merge misplaced tags directly connected with root

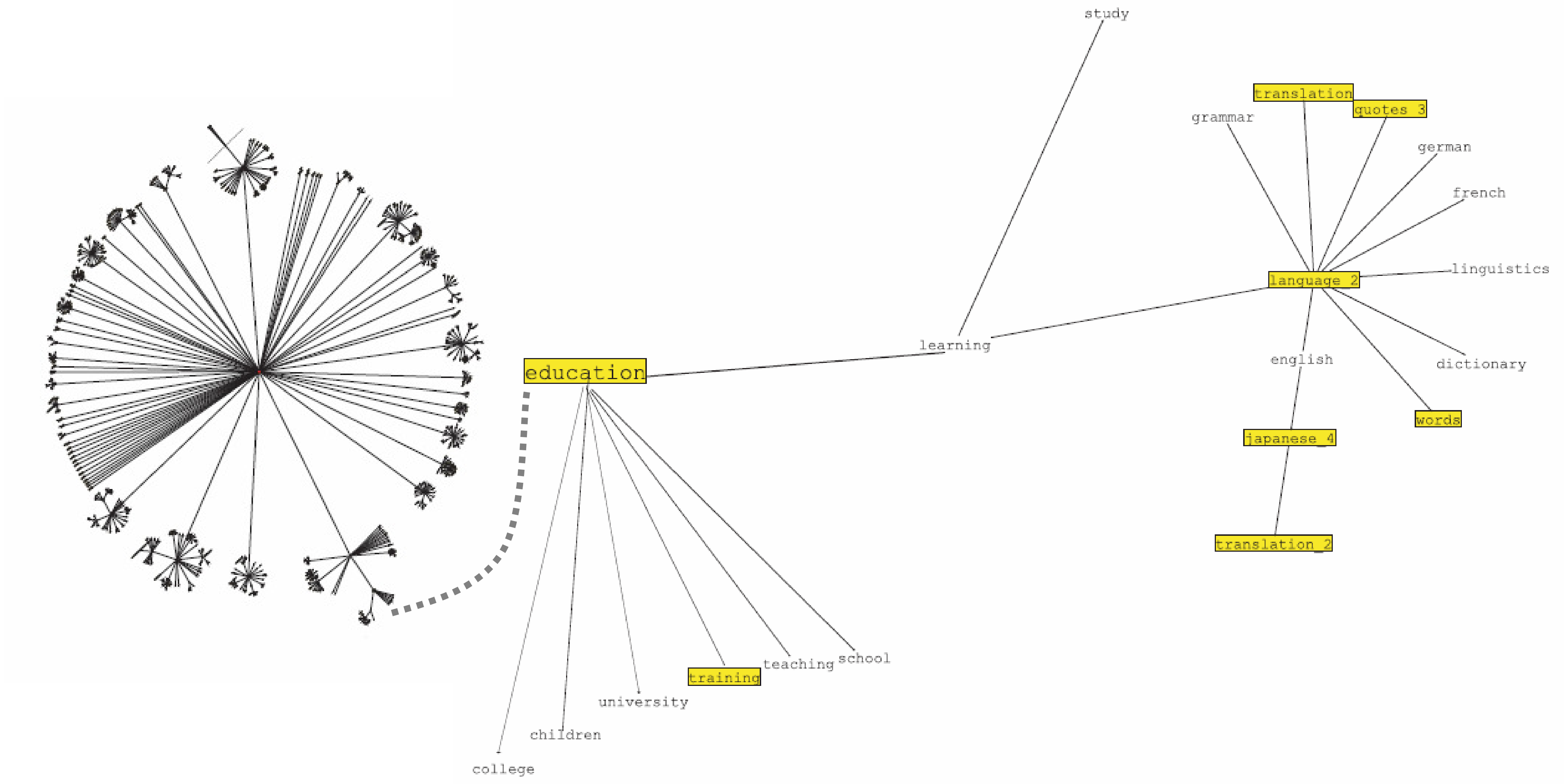
- Observations:
    1. These tags have no child nodes
    2. Concepts with multiple meanings occur several times as such tags
    3. Such tags have often a very low degree

- operations:
    1. delete tags without a child node
    2. merge tags occurring multiple times
    3. tag with low degree are rearranged

Example results: Concept hierarchy from Delicious



Example results: Concept hierarchy from Delicious



Result compared with WordNet

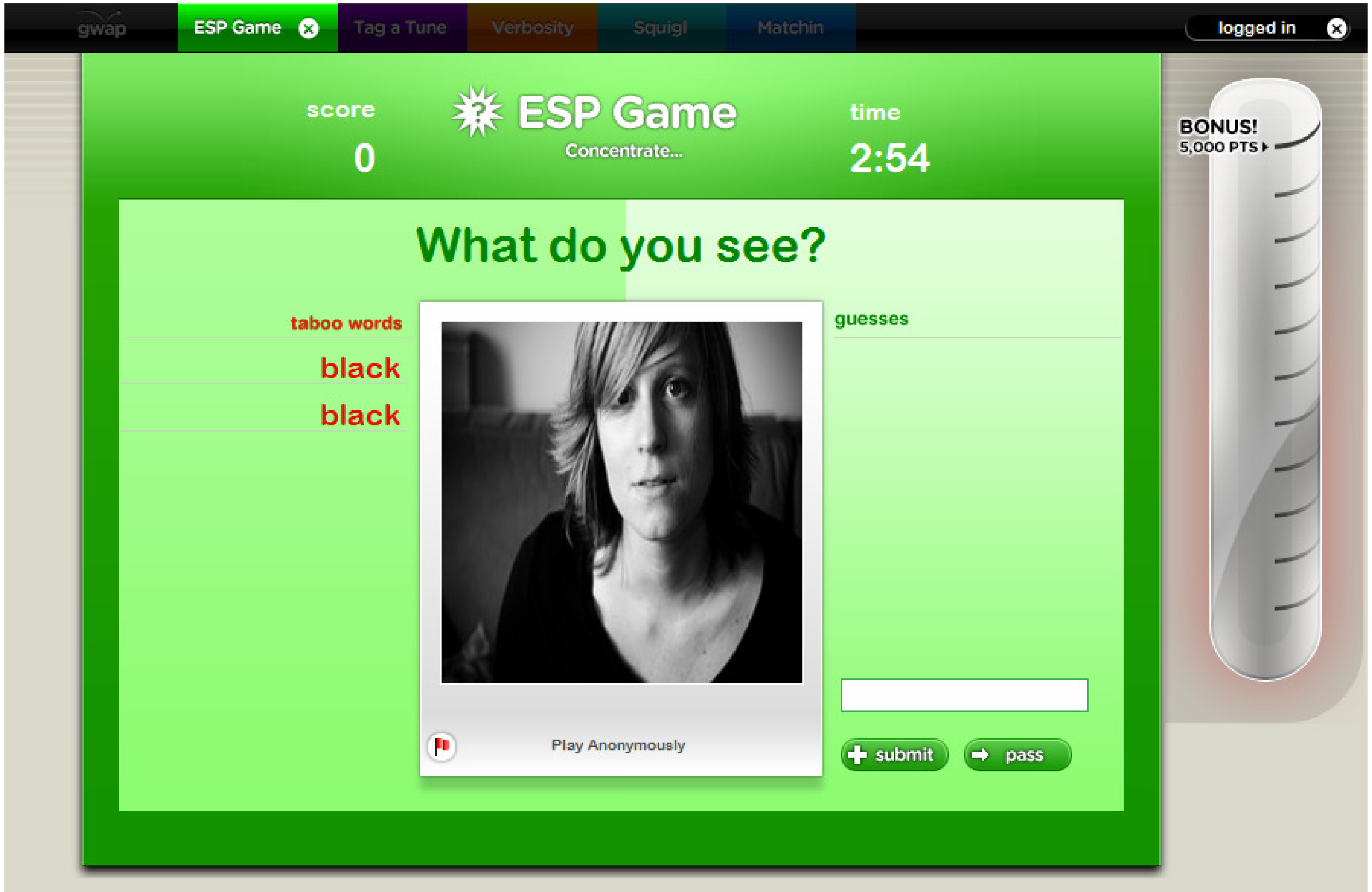
WordNet			
Measure	Benz Algorithm Value	Adopted Algorithm min_syn	Adopted Algorithm value
Taxonomic Precision	0,21867	0,90	0,35153
		0,92	0,35807
		0,94	0,35313
		0,96	0,37295
		0,98	0,35983
		1,00	0,37154
Taxonomic Recall	0,19064	0,90	0,19113
		0,92	0,18900
		0,94	0,19080
		0,96	0,19022
		0,98	0,19003
		1,00	0,18823
Taxonomisches F-Maß	0,20369	0,90	0,24762
		0,92	0,24741
		0,94	0,24774
		0,96	0,25194
		0,98	0,24871
		1,00	0,24987
Taxonomic Overlap	0,10427	0,90	0,12980
		0,92	0,13046
		0,94	0,13027
		0,96	0,13404
		0,98	0,13156
		1,00	0,13283

Result compared with Wikipedia

Wikipedia			
Measure	Benz Algorithm Value	Adopted Algorithm min_syn	Adopted Algorithm value
Taxonomic Precision	0,23634	0,90	0,36091
		0,92	0,37219
		0,94	0,36452
		0,96	0,37503
		0,98	0,36089
		1,00	0,37191
Taxonomic Recall	0,54345	0,90	0,50835
		0,92	0,50138
		0,94	0,50158
		0,96	0,49919
		0,98	0,49797
		1,00	0,49952
Taxonomisches F-Maß	0,32942	0,90	0,42213
		0,92	0,42723
		0,94	0,42220
		0,96	0,42829
		0,98	0,41849
		1,00	0,42637
Taxonomic Overlap	0,19644	0,90	0,26224
		0,92	0,26725
		0,94	0,26449
		0,96	0,27158
		0,98	0,26127
		1,00	0,26903

What kind of other relations can we learn? And how?

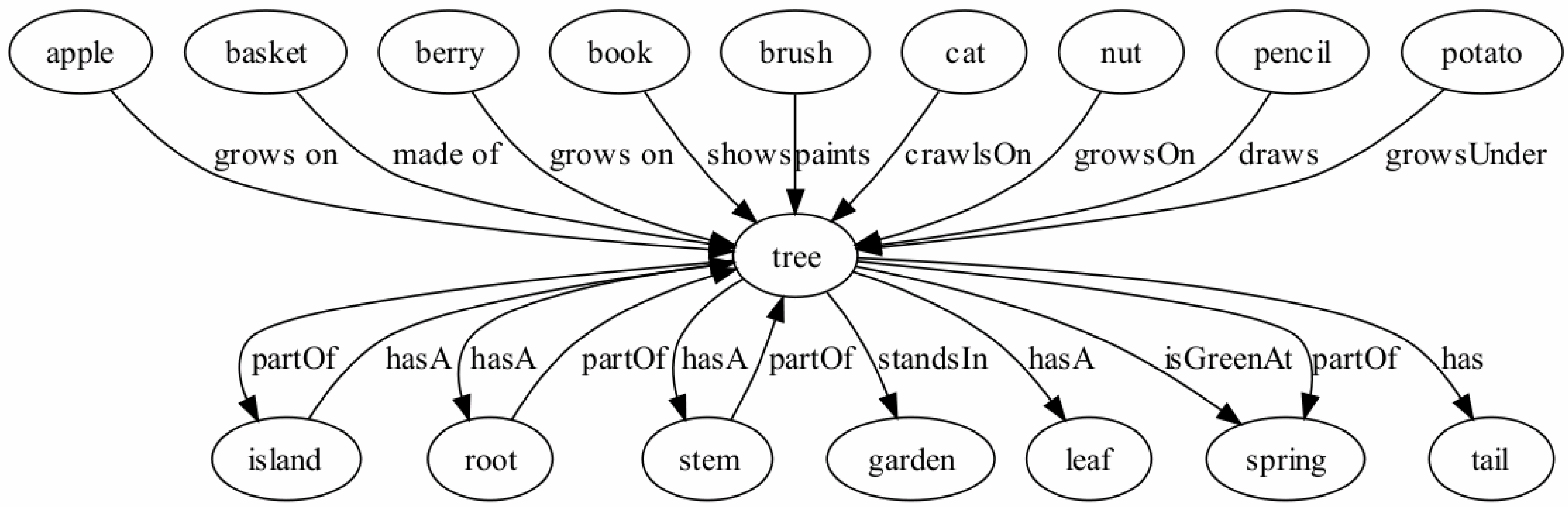
Can we use a game to learn relations?



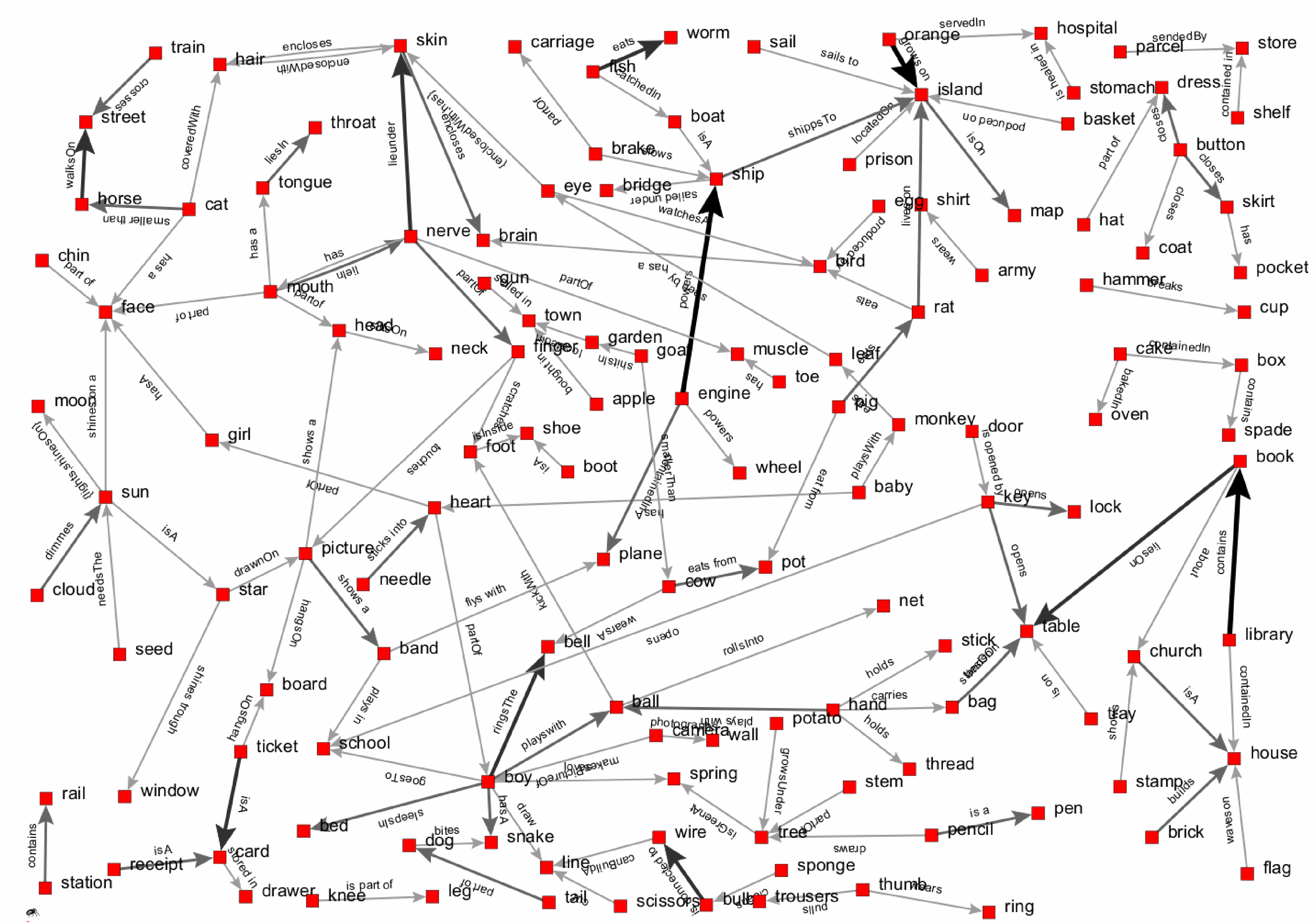
- von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Vienna, Austria, April 24 - 29, 2004). CHI '04. ACM, New York, NY, 319-326.
- Siorpaes, K. and Hepp, M. 2008. Games with a Purpose for the Semantic Web. IEEE Intelligent Systems 23, 3 (May. 2008), 50-60.

Learning Relations with the Image Relation Annotation Game

Learned relations with the concept *tree*



Learning Relations with the Image Relation Annotation Game



Learning Approaches

learning of tag relations

Social Network Analysis
- centrality
- clustering coefficient

[Mika, 2005]
[Heymann, 2006]

Statistical approaches
- model of sub-sumption
- association rules

[Schmitz, 2006]
[Schmitz et al., 2006]

Clustering
- e.g. HAC

[Begelmann, 2006]

Tag (co-)occurrence:
most general / resource independent
user-based / resource-based co-occurrence

Further Learning Approaches

- P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems, 2007.
- T. Eda, M. Yoshikawa, T. Uchiyama, and T. Uchiyama. The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. World Wide Web, 12(4):421–440, December 2009.
- Plangprasopchok and K. Lerman. Constructing folksonomies from user-specified relations on flickr. In WWW '09: Proceedings of the 18th international conference on World wide web, pages 781–790, New York, NY, USA, 2009. ACM.
- P. Schmitz. Inducing ontology from Flickr tags. In Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May 2006.
- L. Specia and E. Motta. Integrating folksonomies with the semantic web. In Proc. of the European Semantic Web Conference (ESWC2007), volume 4519 of LNCS, pages 624–639, Berlin, 2007. Springer.
- J. Tang, H. fung Leung, Q. Luo, D. Chen, and J. Gong. Towards ontology learning from folksonomies. In IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence, pages 2089–2094, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.
- L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies: A quantitative study. pages 168–186. 2006.
- M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, G. Schreiber, and P. Cudr ́-Mauroux, editors, Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, volume 4825 of LNCS, pages 673–686, Berlin, Heidelberg, November 2007. Springer Verlag.

Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Development of Folksonomies

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook

Lessons Learned

- Network measures provide interesting insights into the user behavior of folksonomies
- All types of nodes provide valuable information
- A bunch of factors have influence on emergent folksonomy structure, like recommenders or spam
- First relationships can be extracted by simple data mining approaches
- Relatedness measures on tags in folksonomies are a good basis to extract semantic relations
- The role of users has influence on the emergent semantics
- Several learning approach are able to extract ontologies

Future Work

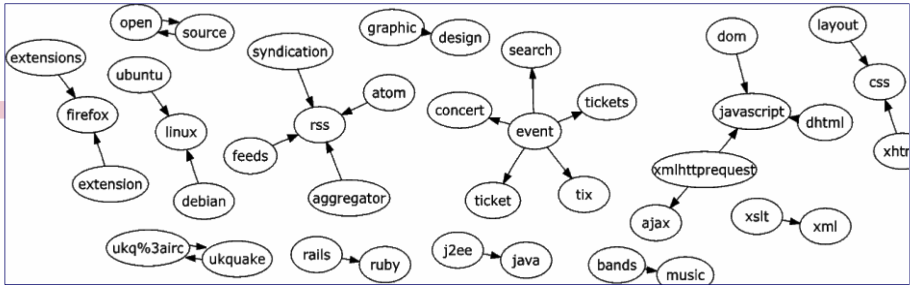
$\forall x, y\,(sufferFrom\,(x, y) \rightarrow ill(x))$	Rules & Axioms	
$cure(dom{:}DOCTOR, range{:}DISEASE)$	Relations	(✓)
is_a(teaching, education)	Taxonomy	✓
TEACHING := <Int, Ext, Lex>	Concepts	✓
{howto, how-to, guide, tutorials, how_to}	(Multilingual) Synonyms	✓
howto guide programming	**Tags**	✓

Future Work

- Learning new relations by using link mining methods
- Extracting rules & axioms e.g. by applying statistical relational learning methods
- Improving synset detection and tag sense discovery component
- Utilizing the information of the annotated resource
- Trying to get feedback from user by allowing semantics within tagging systems
- Combining ontology learning from text with ontology learning from tags
- Using tags to extent existing ontologies

Agenda

The End

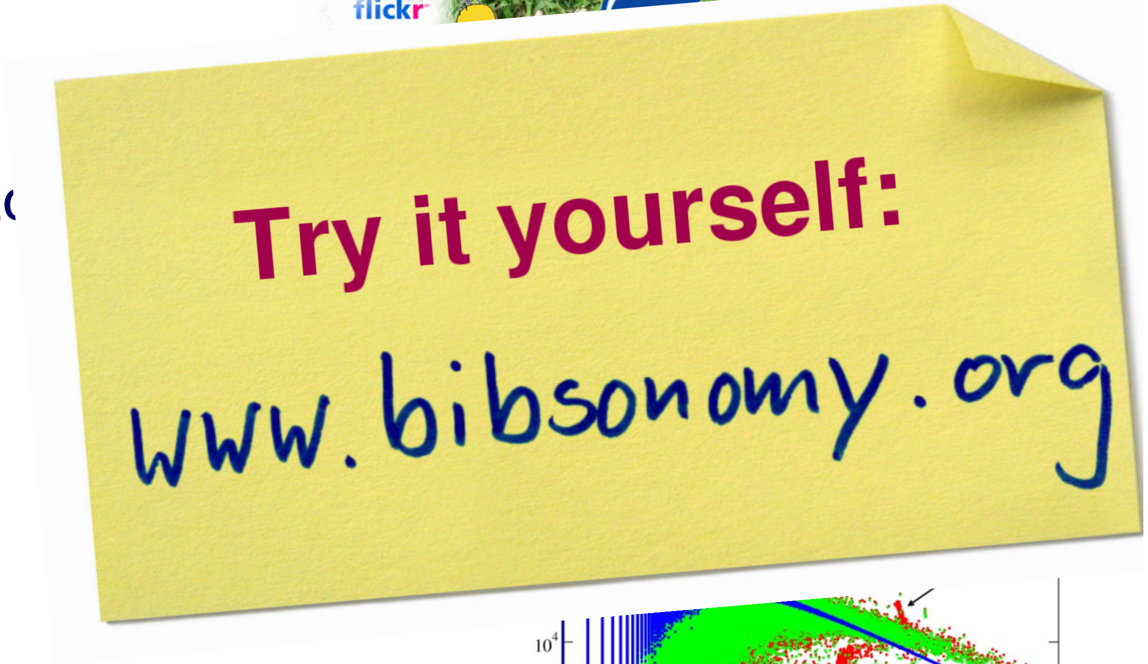


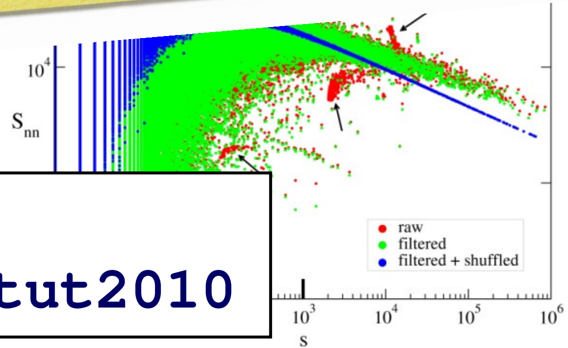
Introduction
- Web 2.0
- Collaborative Tagging Systems and Folksonomies
- Folksonomies and Ontologies

Understanding Folksonomy Data
- Network Properties of Folksonomies
- Types of Tags
- Types of Users
- Types of Resources
- Factors influencing the Devel

Ontology Learning
- Association Rules
- Measures of Tag Relatedness
- Categorizers/Describers
- Learning Approaches

Summary and Outlook





References:
http://www.bibsonomy.org/group/kde/ol_tut2010

Backup

Search engines need

1. to compute the hits for a query
2. and rank them. PageRank algorithm is very successful in the web (see Google):

- Authority values are propagated along the hyperlink according to

$$\boldsymbol{x} \leftarrow d\, A\boldsymbol{x} + (1\text{-}d)\, \boldsymbol{p}$$

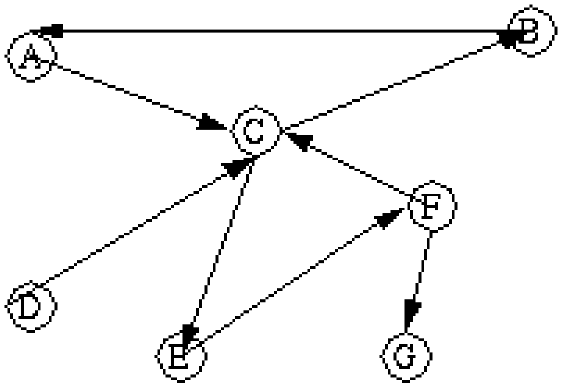
where A is the row-stochastic adjacency matrix of the web graph,
$\boldsymbol{x}$ is the rank vector,
$\boldsymbol{p}$ is the random surfer component
(may be used as preference vector),
$d \in [0,1]$ is a weighting factor.

each row of A is normalized to 1

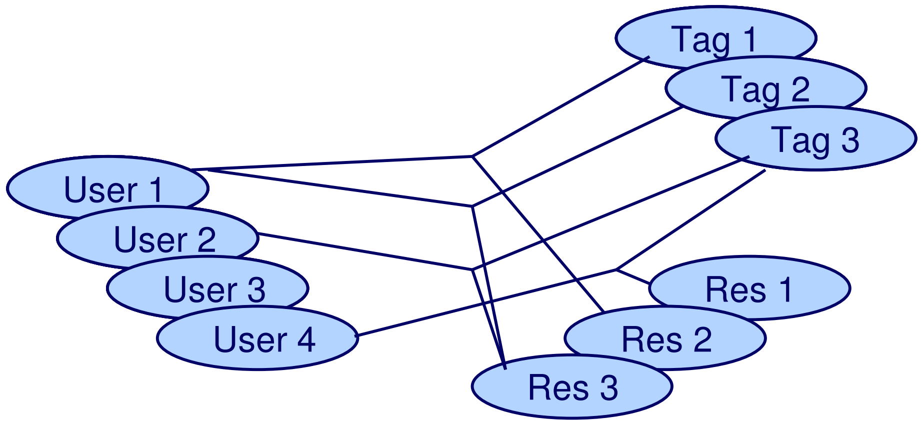
- If $|A|_1 := |\boldsymbol{p}|_1 := 1$ and there are no rank sinks, then the computation of a fixed point equals the computation of the first eigenvector of the matrix $dA + (1\text{-}d)\, \boldsymbol{p}\mathbf{1}^{\mathsf{T}}$.

Search in Folksonomies

- Folksonomies have a different structure as the web graph:



Web graph

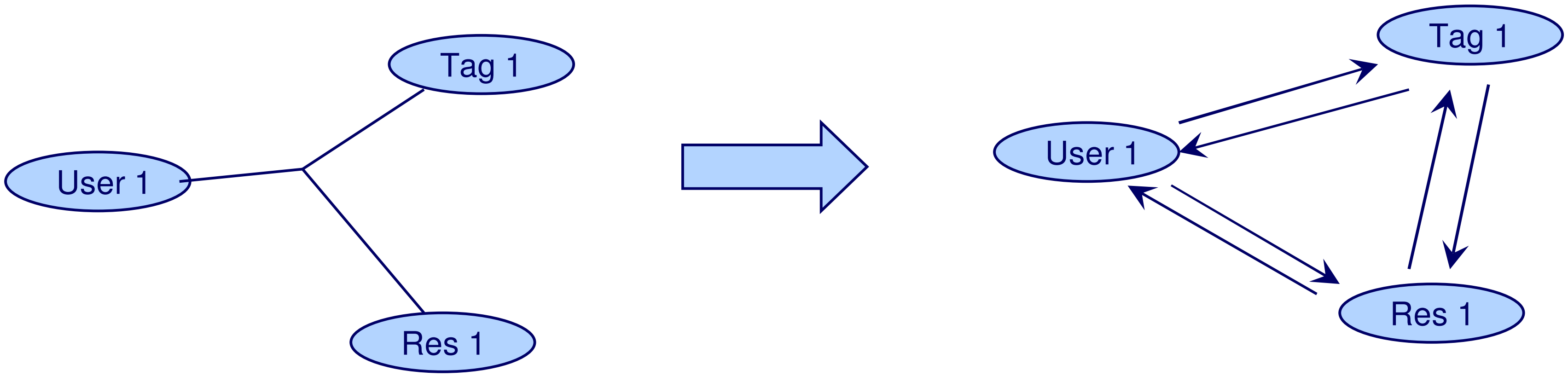


Folksonomies

- **How can a ranking algorithm for this structure look like?**

First Aproach: Adapted PageRank

1. Split each hyperedge into six directed edges.



1. Iterative weight propagation according to PageRank:

$$\boldsymbol{x} \leftarrow d\,A\boldsymbol{x} + (1\text{-}d)\,\boldsymbol{p} \quad .$$

Converting a Folksonomy into an Undirected Graph

Set V of nodes consists of the disjoint union of the sets of tags, users and resources:

$$V = U \cup T \cup R$$

All co-occurrences of users and tags, tags and resources, users and resources become edges between the respective nodes:

- $E = \{\{u,t\} \mid \exists\, r \in R : (u,t,r) \in Y\} \cup$
  $\{\{t,r\} \mid \exists\, u \in U : (u,t,r) \in Y\} \cup$
  $\{\{u,r\} \mid \exists\, t \in T : (u,t,r) \in Y\}$

Problems of folksonomy-adapted PageRank

- dominated by graph structure
- undirected: weight flows back (PageRank $\approx$ edge degree)

Differential approach

- compute rank with and without preferences
- FolkRank = difference between those rankings normalized to [0,1]
  - Let R_{AP} be the fixed point with $p = \mathbf{1}$
  - Let R_{pref} be the fixed point with p representing the high weights for the preferred items
  - $R := R_{pref} - R_{AP}$ is the final weight vector

Results for: "Semantic Web"

PageRank without preference	
Tag	ad. PageRank
system:unfiled	0,0078404
web	0,0044031
blog	0,0042003
design	0,0041828
software	0,0038904
music	0,0037273
programming	0,0037100
css	0,0030766
reference	0,0026019
linux	0,0024779
tools	0,0024147
news	0,0023611
art	0,0023358
blogs	0,0021035
politics	0,0019371
java	0,0018757
javascript	0,0017610
mac	0,0017252
games	0,0015801
photography	0,0015469
fun	0,0015296

PageRank with preference	
Tag	ad. PRank
semanticweb	0,0208605
web	0,0162033
semantic	0,0122028
system:unfiled	0,0088625
semantic_web	0,0072150
rdf	0,0046348
semweb	0,0039897
resources	0,0037884
community	0,0037256
xml	0,0031494
research	0,0026720
programming	0,0025717
css	0,0025290
portal	0,0024118
.imported	0,0020495
imported-bo...	0,0019610
en	0,0018900
science	0,0018166
.idate2005-04-11	0,0017779
newfurl	0,0017578
internet	0,0016122

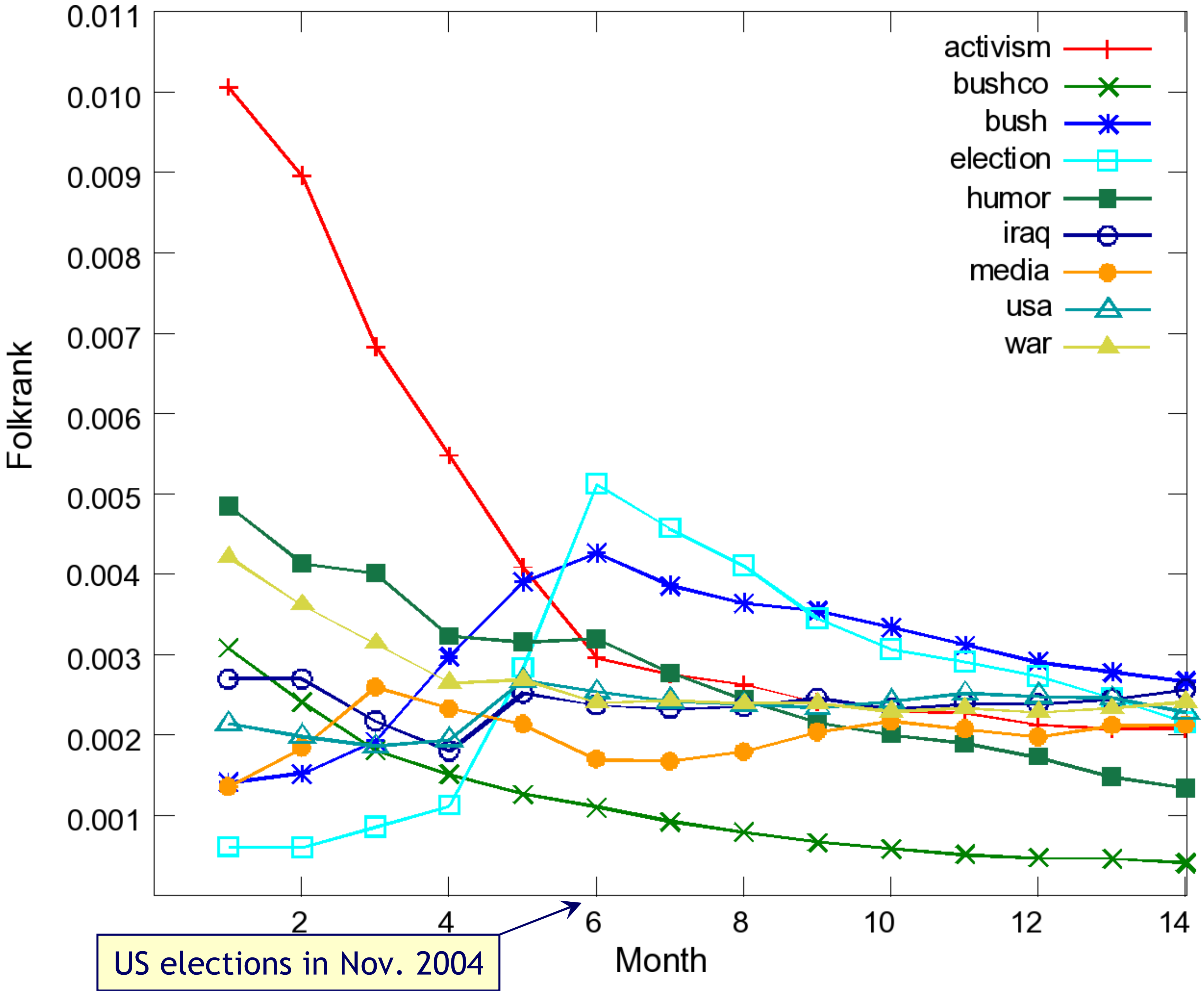
FolkRank with preference	
Tag	FolkRank
semanticweb	0,0207820
semantic	0,0121305
web	0,0118002
semantic_web	0,0071933
rdf	0,0044461
semweb	0,0039308
resources	0,0034209
community	0,0033208
portal	0,0022745
xml	0,0022074
research	0,0020378
imported-bo...	0,0018920
en	0,0018536
.idate2005-04-11	0,0017555
newfurl	0,0017153
tosort	0,0014486
cs	0,0014002
academe	0,0013822
rfid	0,0013456
sem-web	0,0013316
w3c	0,0012994

Rankings for „semanticweb“

for discovering semantic relationships, user comunities, and web pages

Tags		Users		Resources	
semanticweb	0.0208	up4	0.0092	http://www.semanticweb.org/	0.3762
semantic	0.0121	awenger	0.0085	http://flink.semanticweb.org/	0.0006
web	0.0118	j.deville	0.0074	http://simile.mit.edu/piggy-bank/	0.0004
semantic_web	0.0072	chaizzilla	0.0062	http://www.w3.org/2001/sw/	0.0003
rdf	0.0044	elektron	0.0059	http://infomesh.net/2001/swintro/	0.0002
semweb	0.0039	captsolo	0.0055	http://www.ontoweb.org/	0.0002
resources	0.0034	dissipative	0.0050	http://www.aaai.org/AITopics/html/ontol.html	0.0002
community	0.0033	stevag	0.0050	http://del.icio.us/register	0.0002
portal	0.0023	krudd	0.0047	http://mspace.ecs.soton.ac.uk/	0.0002
xml	0.0022	williamteo	0.0037	http://simile.mit.edu/	0.0001
research	0.0020	stevecassidy	0.0036	http://itip.evcc.jp/itipwiki/	0.0001
imported-bookmarks	0.0019	pmika	0.0035	http://www.google.be/	0.0001
en	0.0019	millette	0.0032	http://www.letterjames.de/index.html	0.0001
.idate2005-04-11	0.0018	myren	0.0028	http://www.daml.org/	0.0001
newfurl	0.0017	morningboat	0.0026	http://jena.sourceforge.net/	0.0001
tosort	0.0014	philip.fennell	0.0025	http://www.federalconcierge.com/WritingBusinessCases.html	0.0001
cs	0.0014	webb.	0.0025	http://www.mpuf.org/	0.0001
academe	0.0014	dnaboy76	0.0025	http://www.shirky.com/writings/semantic_syllogism.html	0.0001
rfid	0.0013	mote	0.0024	http://semarts.com.decisivenet.com/	0.0001
sem-web	0.0013	alphajuliet	0.0024	http://www.e-gov.com/	0.0001
w3c	0.0013	nymetbarton	0.0024	http://rdfweb.org/	0.0001

Trends with respect to tag "politics"



Ranking in Folksonomies

- Michail, A. CollaborativeRank: Motivating People to Give Helpful and Timely Ranking Suggestions, School of Computer Science and Engineering, 2005.
- Szekely, B. & Torres, E. Ranking Bookmarks and Bistros: Intelligent Community and Folksonomy Development, 2005.
- Bao, S.; Xue, G.; Wu, X.; Yu, Y.; Fei, B. & Su, Z. Optimizing web search using social annotations, ACM Press, 2007, 501-510.

Ranking in Web 2.0

- Mohammad Nauman and Shahbaz Khan. Using Personalized Web Search for Enhancing Common Sense and Folksonomy Based Intelligent Search Systems. wi, (0):423-426,IEEE Computer Society,Los Alamitos, CA, USA,2007.

Usefulness of Tag Clouds

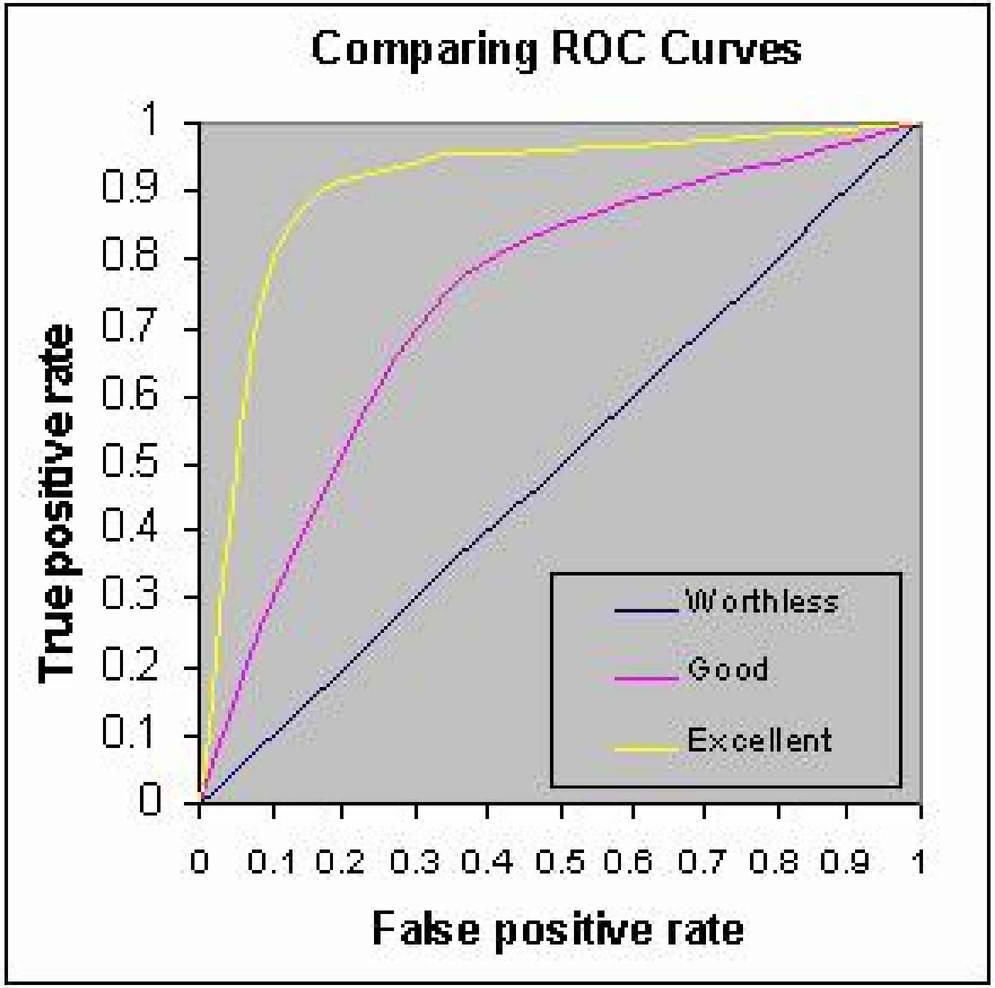
- J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: When is it useful? Journal of Information Science, 016555150607808,CILIP,2007.

Setting

Labelled / Actual	Spam	Non-Spam
Spam	TP	FN
Non-Spam	FP	TN

Precision, Recall, F1

ROC Curve, Area Under Curve (AUC)



[http://gim.unmc.edu/dxtests/ROC3.htm]