Vorlesung Künstliche Intelligenz Wintersemester 2008/09

Teil IV:
Wissensrepräsentation im WWW

Kap.12: Web 2.0

Web 2.0 - Begriffsklärung

Der Begriff „Web 2.0“ bezieht sich primär auf eine veränderte Nutzung und Wahrnehmung des Internets: Die Benutzer erstellen und bearbeiten Inhalte selbst.

Er bezeichnet aus technischer Sicht auch eine Anzahl von Methoden wie

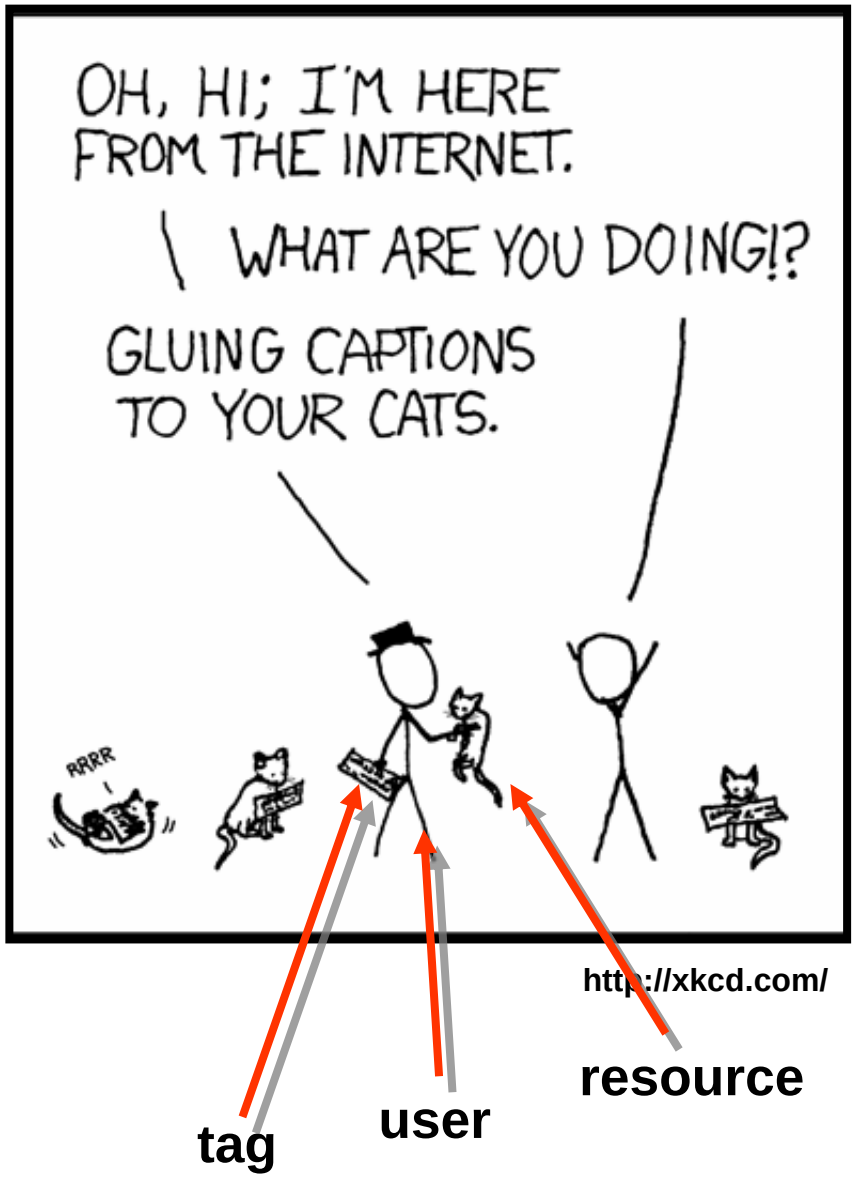
- Web-Service-APIs,
- Ajax (Asynchronous Javascript und XML)
- und Abonnement-Dienste wie RSS.

(Siehe http://de.wikipedia.org/wiki/Web_2.0)

Typen von Web 2.0- Anwendungen

- Wikis (z.B.: Wikipedia)
- Blogs (z.B.: irgendein journalistisches Blog?)
- Photo- und Videoplattformen (z.B.: Youtube, Flickr)
- Social Bookmarking (z.B.: del.icio.us, BibSonomy)
- soziale Online-Netzwerke (z.B.: Xing, Myspace, Facebook, StudiVZ)
- virtuelle Welten (z.B. Second Life, Bailamo)
- Mikroblogs (z.B.: Twitter)

Tagging / Folksonomies



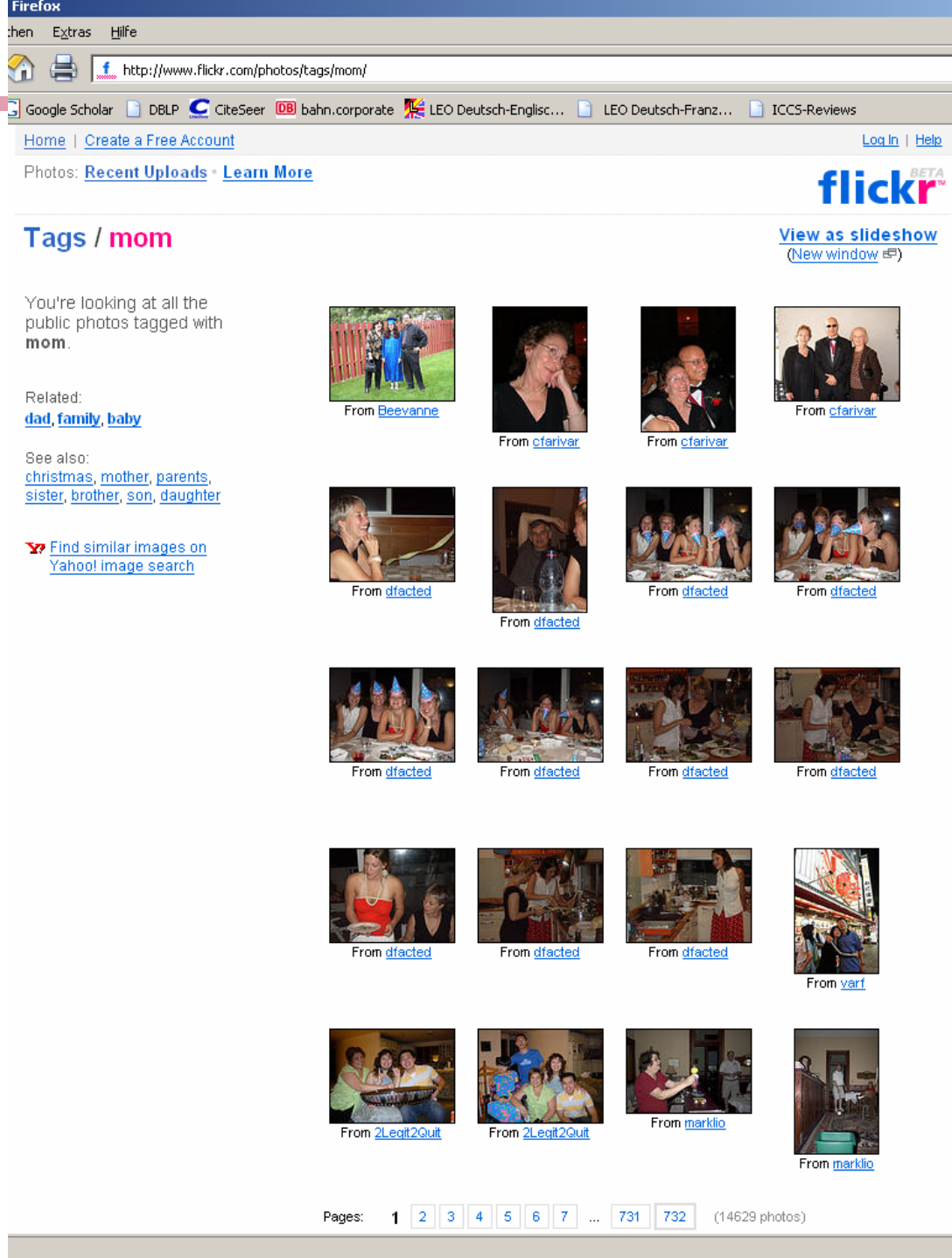
- tagging is a distributed process
- tagging has a small cognitive overhead
- system contents can be browsed by tag
- the system evolves in time: new resources, new users, new tags
- there may be an underlying social network, explicitly exposed or not
- the behavior of users is “selfish”
- users are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

Social Bookmarking Systems

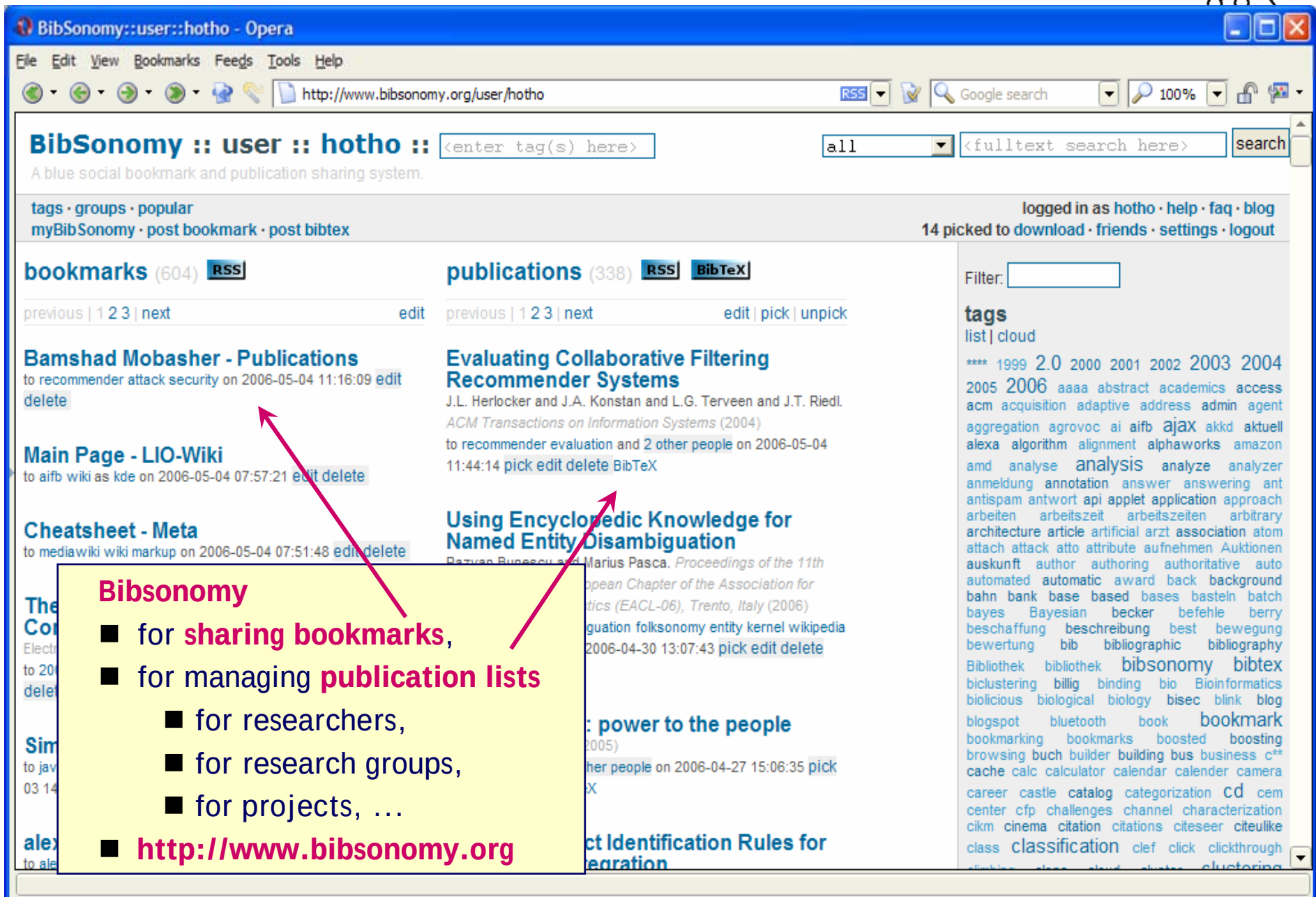
- Collaborative annotation of web resources
- Easy to use, open for everyone
- Joint use leads to converging vocabularies and emergent semantics.

There are many popular folksonomy systems on the web, eg:

- flickr (photos)
- YouTube (videos)
- del.icio.us (bookmarks)



Our system: BibSonomy

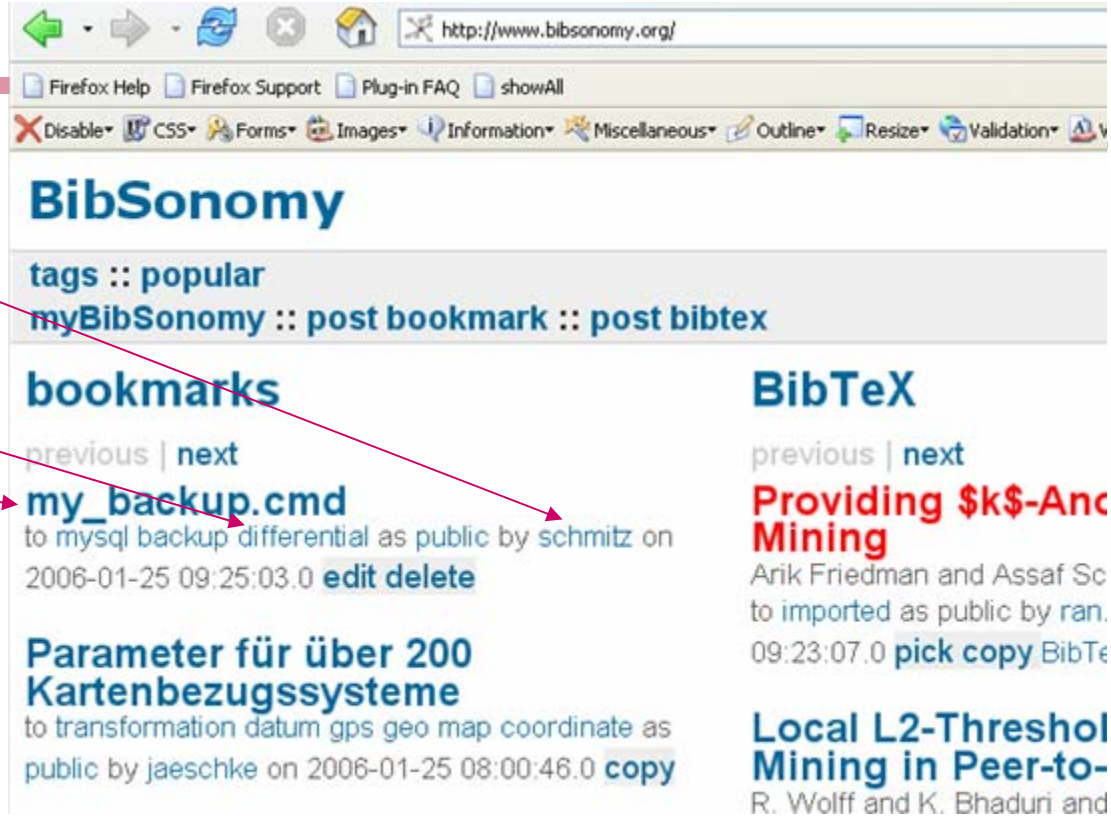


Folksonomies

Folksonomies allow **users**

to assign **tags**

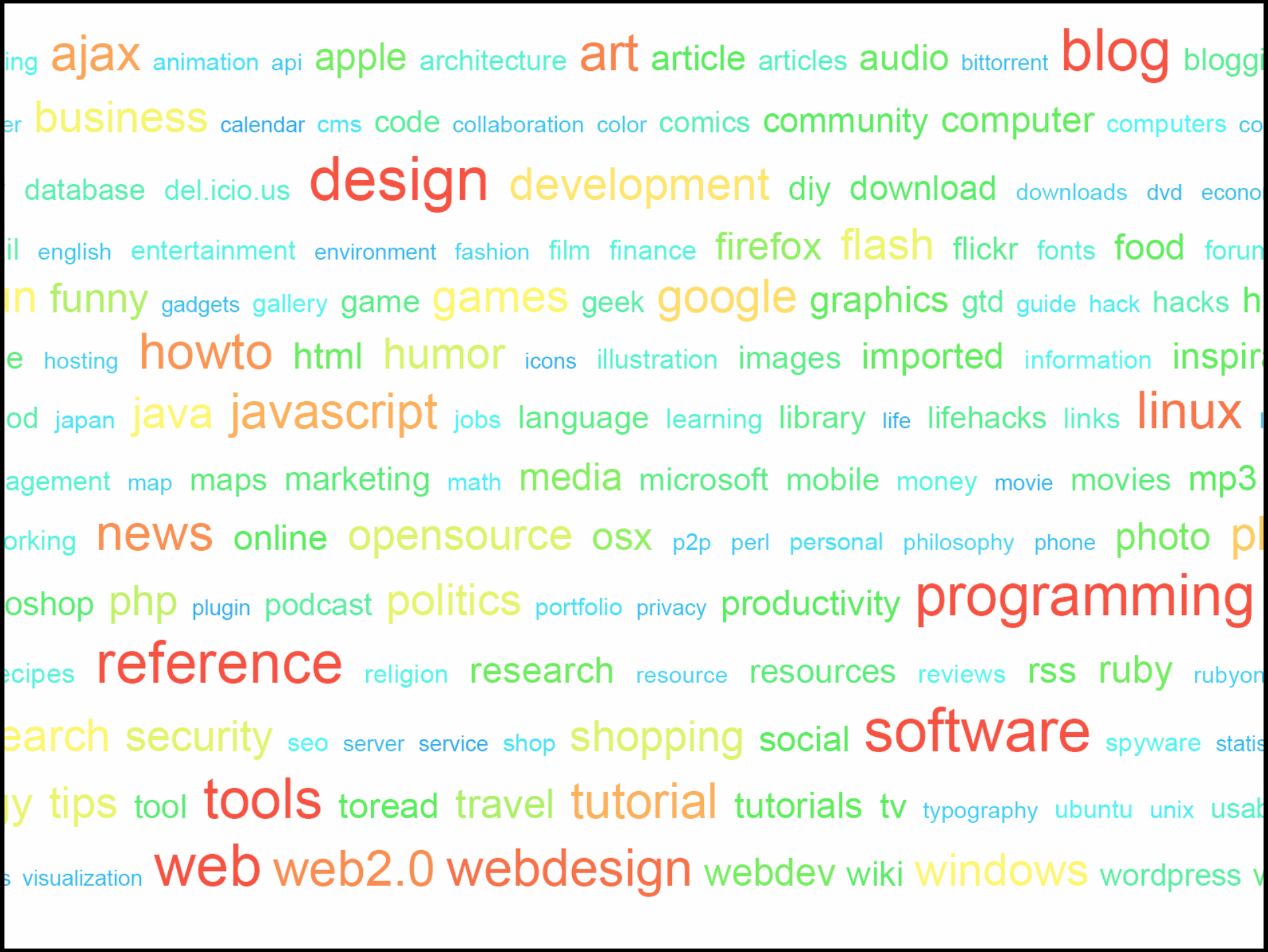
to **resources**.



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.

The *personomy* $\mathbf{P}_u$ of user u is the restriction of $\mathbf{F}$ to u.


ing ajax animation api apple architecture art article articles audio bittorrent blog bloggi
er business calendar cms code collaboration color comics community computer computers co
database del.icio.us design development diy download downloads dvd econo
il english entertainment environment fashion film finance firefox flash flickr fonts food forun
n funny gadgets gallery game games geek google graphics gtd guide hack hacks h
e hosting howto html humor icons illustration images imported information inspir
od japan java javascript jobs language learning library life lifehacks links linux
agement map maps marketing math media microsoft mobile money movie movies mp3
orking news online opensource osx p2p perl personal philosophy phone photo pl
oshop php plugin podcast politics portfolio privacy productivity programming
ecipes reference religion research resource resources reviews rss ruby rubyon
earch security seo server service shop shopping social software spyware statis
y tips tool tools toread travel tutorial tutorials tv typography ubuntu unix usab
s visualization web web2.0 webdesign webdev wiki windows wordpress w

Types of Tags

content/topic of resource (nouns, proper nouns, …)

category of resource

opinion about resource (adjectives)

ownership of resource (user names)

self-reference, relation between resource and user (*mystuff*, *myown*, *citingme*)

task organization (*toread, tobuy*)

social coordination (*for:andrea*)

[see Golder & Huberman ’06]

Probleme und Vorteile des Web2.0 (insbes.Folksonomies)

Probleme:

- keine formale Semantik
- viele Mehrdeutigkeiten, Tippfehler, etc.

Vorteile:

- Viele Beitragende tragen große Mengen an Wissen zusammen
- Hilft gegen den Wissensakquisitions-Flaschenhals

Ziel ist es, die Lücke zwischen dem Semantic Web und dem Web 2.0 zu schliessen. („Bridging the Gap“)

(Dies wird gelegentlich schon als „Web3.0“ bezeichnet.)

Wenn dies (semi-)automatisch gelingt, kann man das Wissen der Vielen („Wisdom of the Crowd“) in eine formale Sprache überführen und somit maschinell verarbeitbar machen.

Tagora

- Semantic Grounding of Measures for Tag Relatedness
- Ontology Learning

EU Project: TAGora – Emergent Semantics in Social Online Communities

C. Cattuto, D. Benz, A. Hotho, G. Stumme: ISWC 2008

Motivation

- 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:

Look up tags in **external thesaurus:**

+ semantically grounded metrics

- "folksonomy jargon" (misspellings, neologisms etc.) not present

Semantic Grounding

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

+ inclusion of complete vocabulary

- semantic interpretation of measures is not clear

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

→ assess their applicability for **tag recommendation, ontology learning**, …

Dataset

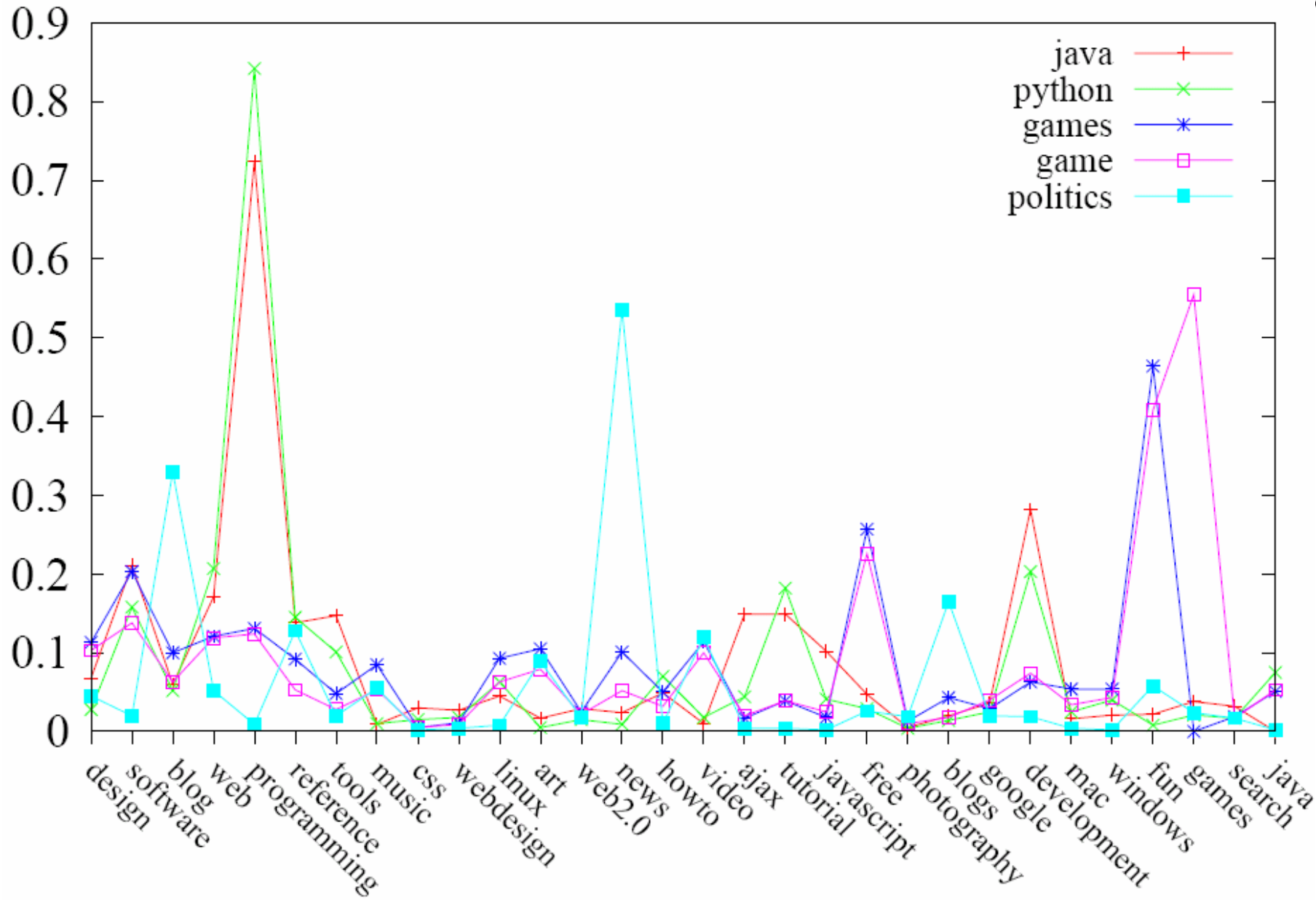
- Del.icio.us crawl 2006
  - $|U| = 667{,}128 \quad |T| = 2{,}454{,}546 \quad |R| = 18{,}782{,}132$
  - $|Y| = 140{,}333{,}714$
- Excerpt: 10,000 most popular tags
  - $|U| = 476{,}378 \quad |T| = 10{,}000 \quad |R| = 12{,}660{,}470$
  - $|Y| = 101{,}491{,}722$
- In the following: **tag rank** = position in most-popular list:
  - 1: design
  - 2: software
  - 3: blog
  - 4: web
  - ...

Relatedness Measures

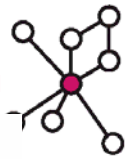
- Take **Co-occurrence frequency** as similarity measure (freq).
- Use **FolkRank** to find related tags (folkrank).
- Describe each tag as a **vector**, whereby each dimension of the vector space corresponds to another tag. Compute similar tags by **cosine similarity** (cosine).

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

Example for cosine measure



Examples of most related tags



Freq

rank	tag	1	2	3	4	5
13	web2.0	ajax	web	tools	blog	webdesign
15	howto	tutorial	reference	tips	linux	programming
28	games	fun	flash	game	free	software
30	java	programming	development	opensource	software	web
39	opensource	software	linux	programming	tools	free
1152	tobuy	shopping	books	book	design	toread

FolkRank

rank	tag	1	2	3	4	5
13	web2.0	web	ajax	tools	design	blog
15	howto	reference	linux	tutorial	programming	software
28	games	game	fun	flash	software	programming
30	java	programming	development	software	ajax	web
39	opensource	software	linux	programming	tools	web
1152	tobuy	toread	shopping	design	books	music

Cosine

rank	tag	1	2	3	4	5
13	web2.0	web2	web-2.0	webapp	“web	web_2.0
15	howto	how-to	guide	tutorials	help	how_to
28	games	game	timewaster	spiel	jeu	bored
30	java	python	perl	code	c++	delphi
39	opensource	open_source	open-source	open.source	oss	foss
1152	tobuy	wishlist	to_buy	buyme	wish-list	iwant

First insights

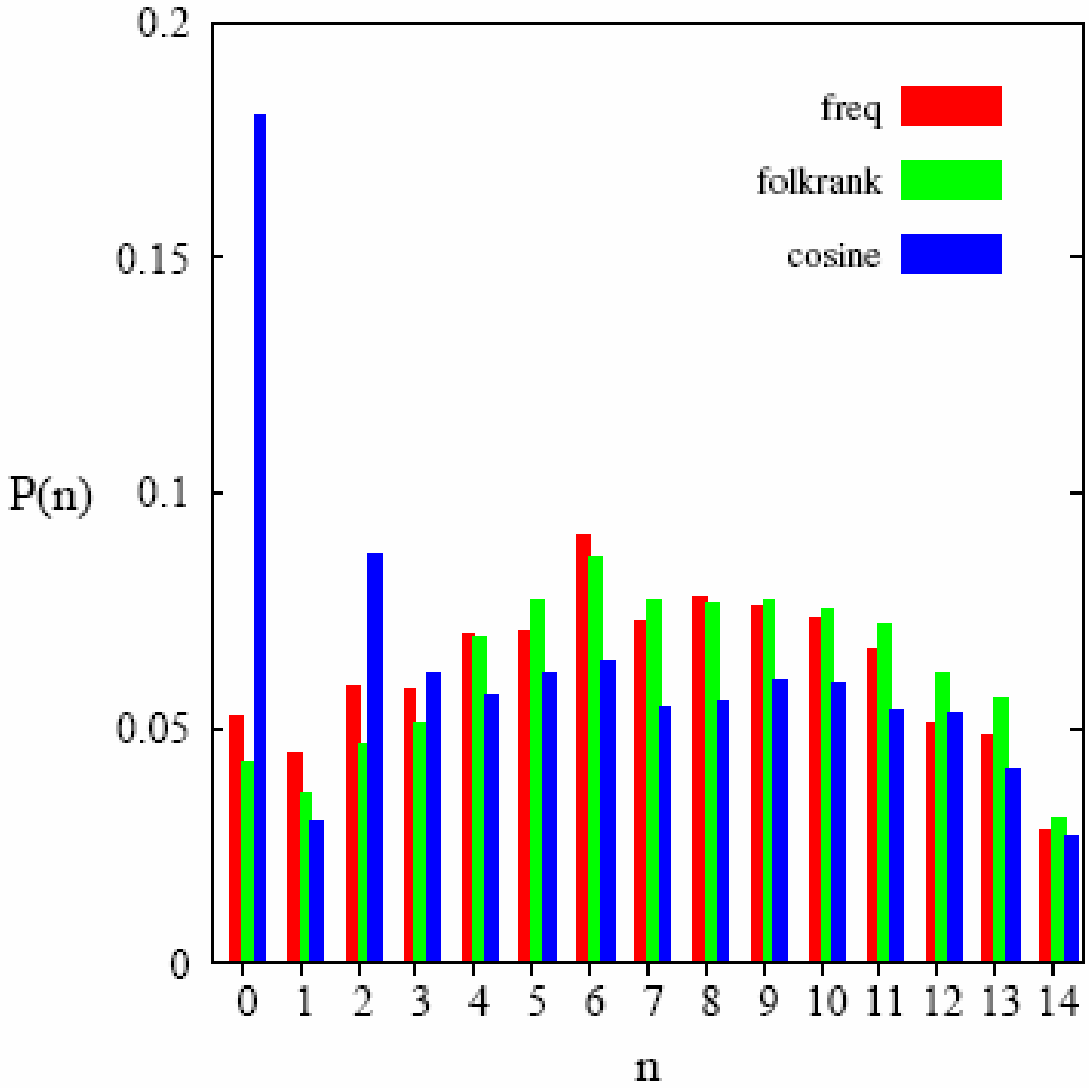
- Freq / FolkRank show bias to high-frequency tags, i.e., to **hyperonyms**.
- Cosine seems to yield more **synomyms** and **“siblings”**.

→ Now: **grounding** of these observations in WordNet.

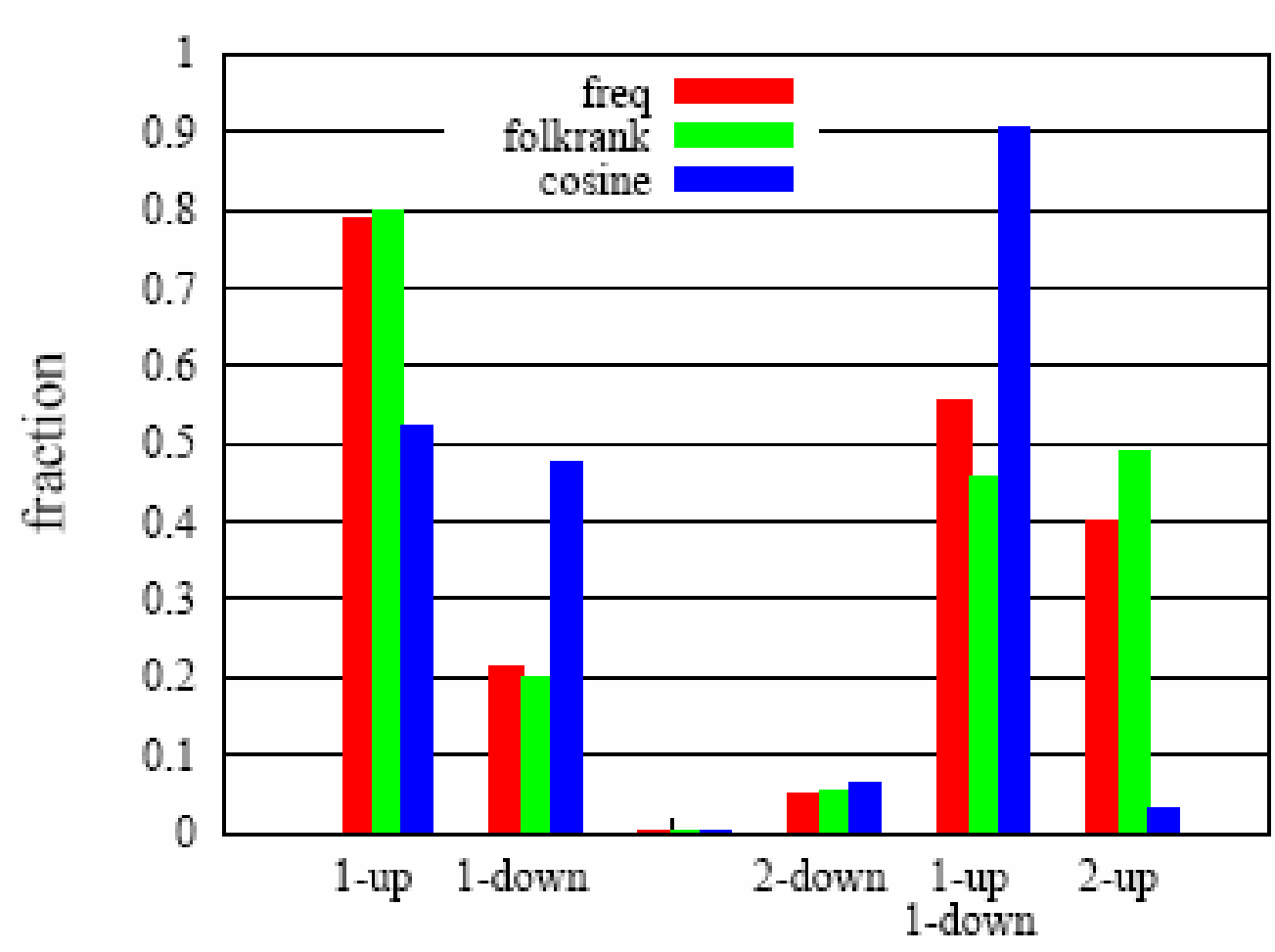
Semantic Grounding in WordNet

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

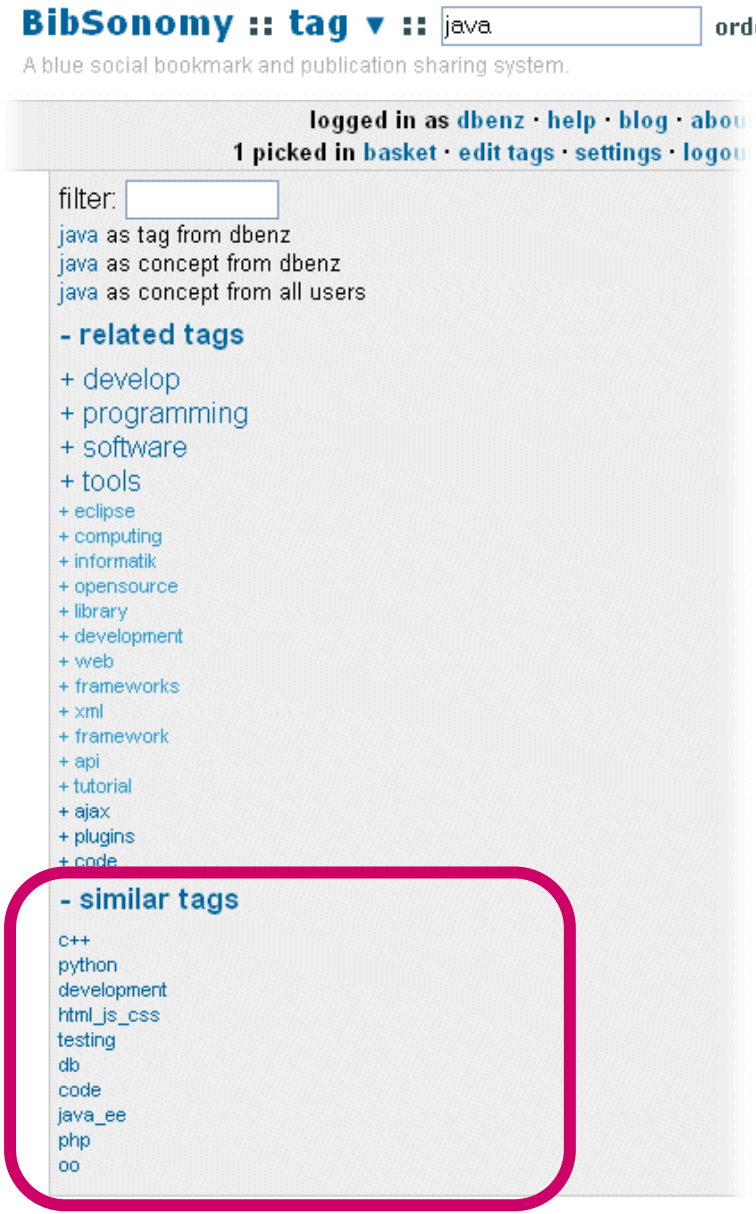
Shortest paths between original tag and most closely related one



Edge composition of shortest paths (for lengths 1 and 2)



Similar tags live on www.bibsonomy.org



Learning Ontologies from Folksonomies

Idea:

- automatically induce a concept hierarchy
- semantics of the relations resembles closely the one of taxonomic relations

Data:

- The tag-tag co-occurrence network of the delicious dataset forms the basis of the experiments (UTC = user-based tag-tag-co-occurrence, RTC = resource based tag-tag-co-occurrence)

Possible approaches:

- Social network analysis
- Set theoretic approaches (association rules, TRIAS)
- Statistical approaches (clustering, similarity measure)

Main steps of an Ontology Learning Algorithm

Filter the tags by an occurrence threshold

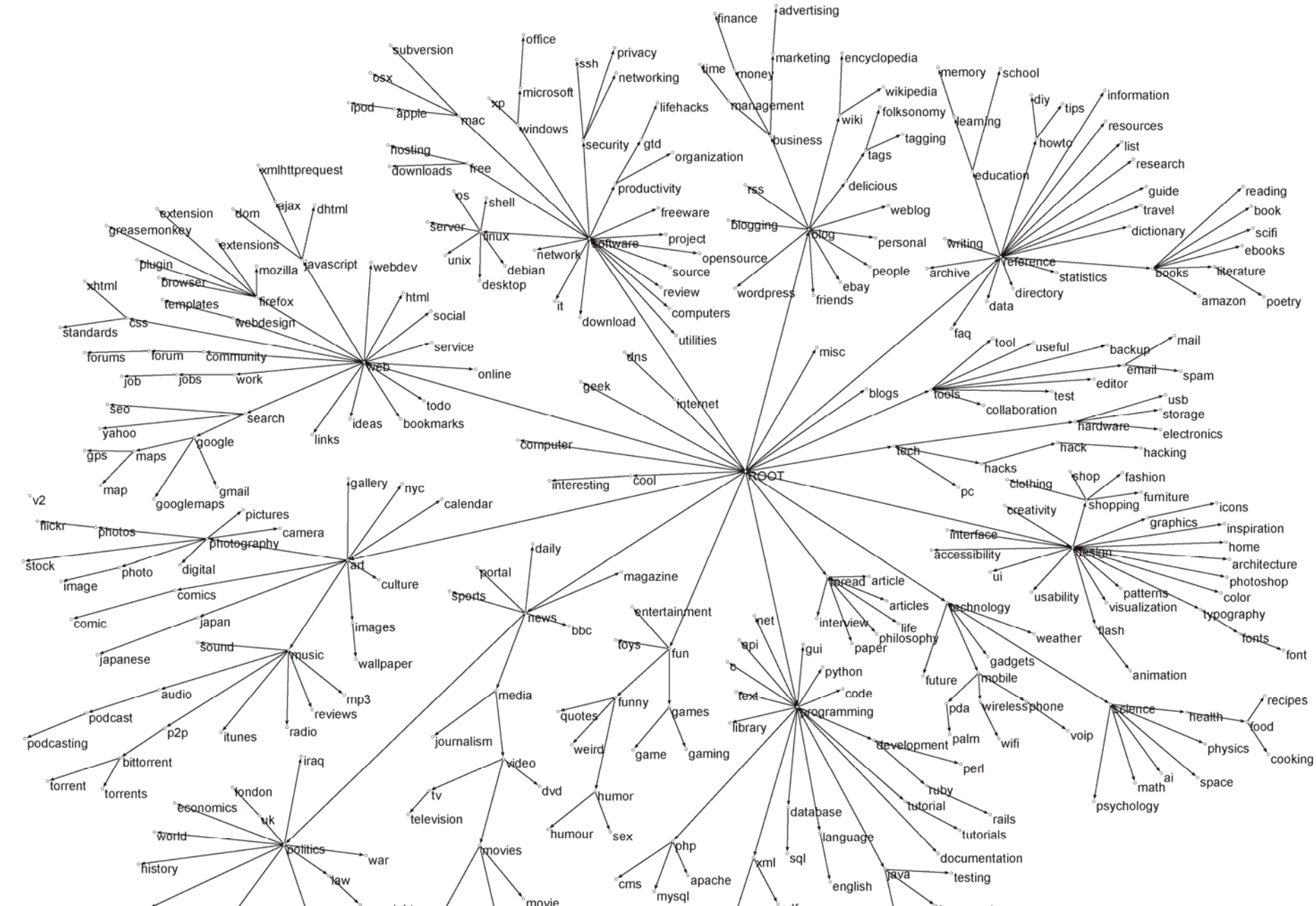
Order the tags in descending order by generality (measured by degree centrality in the UTC network)

Starting from the most general tag, add all tags subsequently to an evolving tree structure:

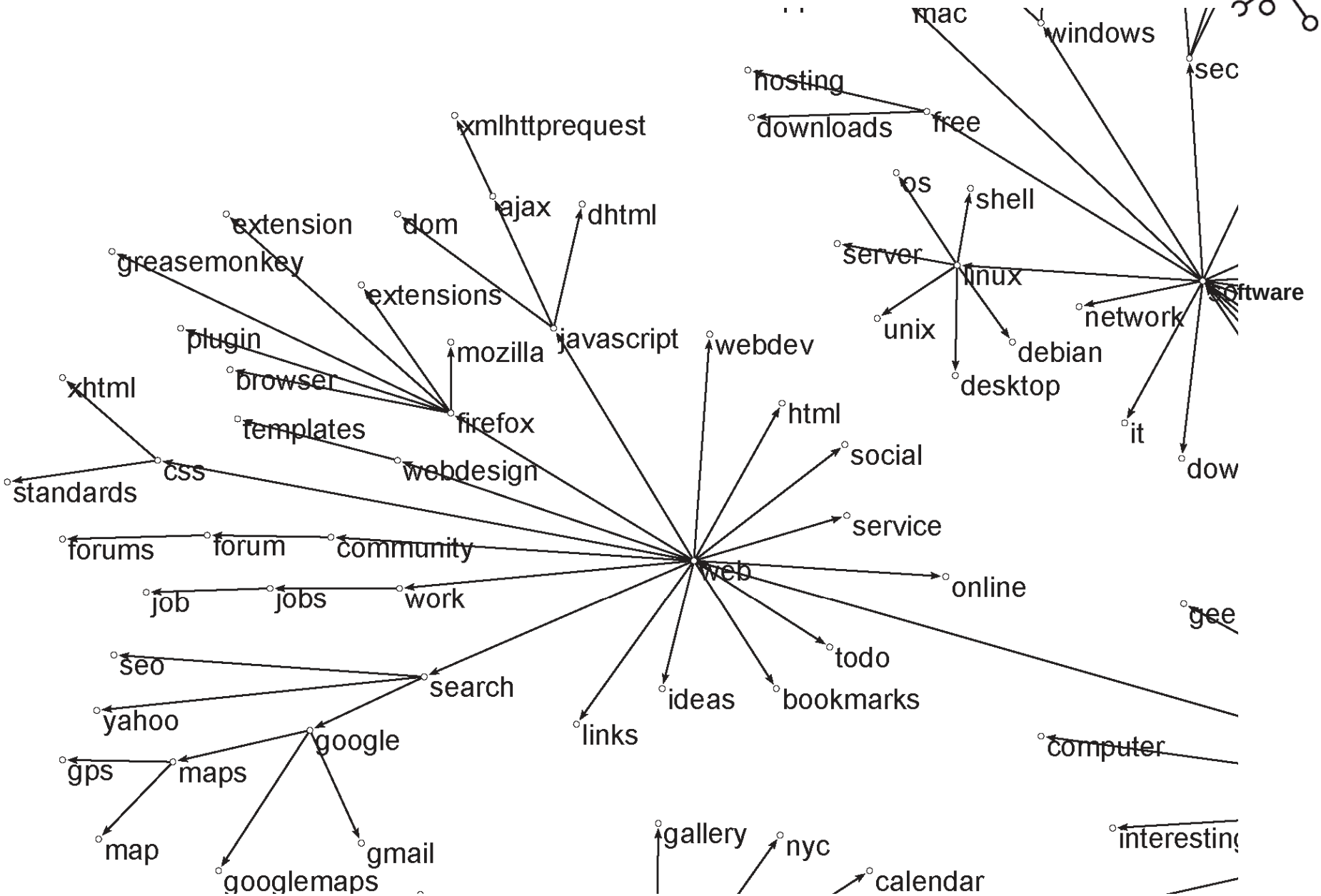
- identify the most similar existing tag
- (decide whether the tags are synonyms or form a compound expression and expand the tree accordingly)

We follow: P. Heymann, H. Garcia-Molina: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. 2006.

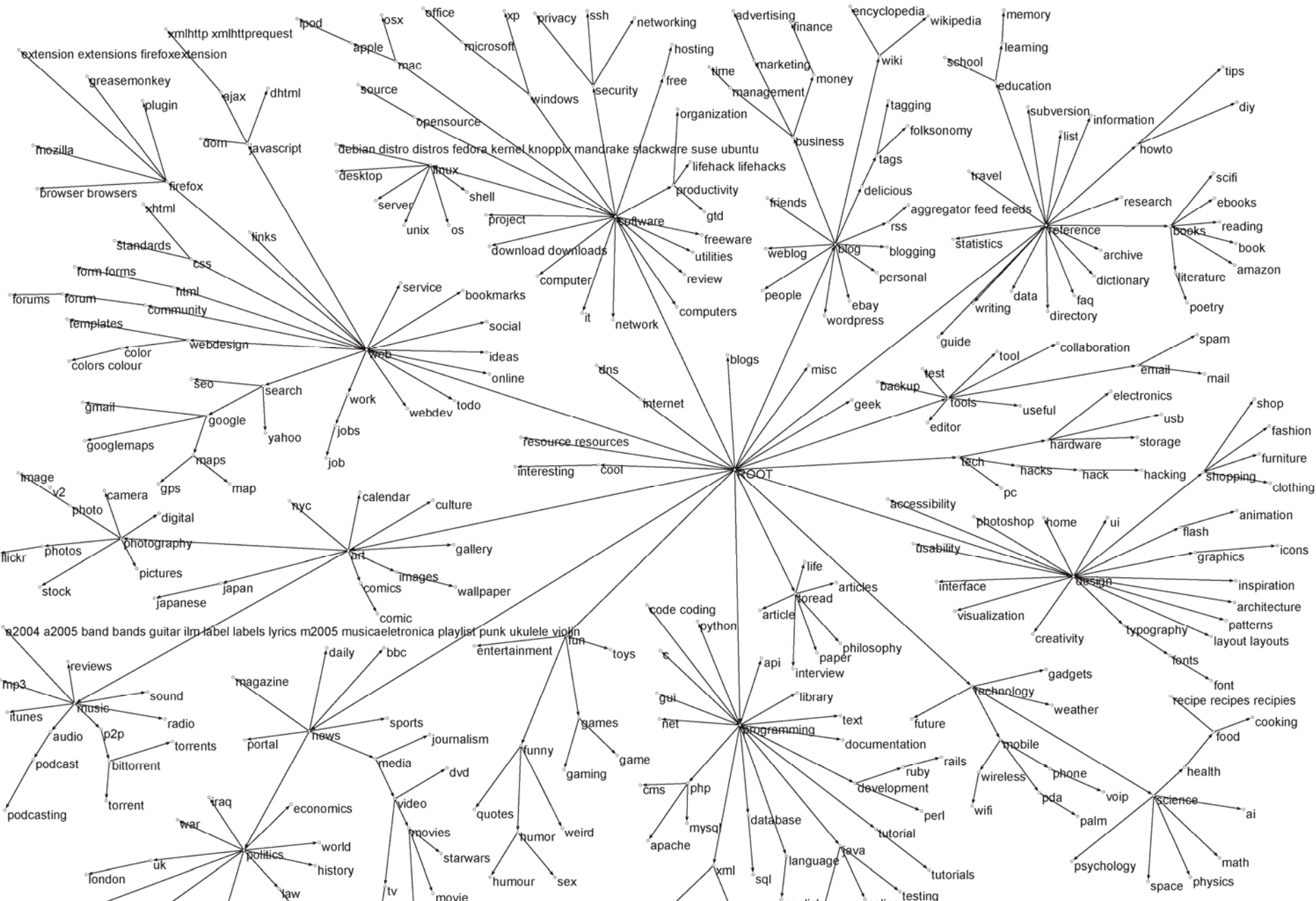
Results for delicious (dataset 2005, 320 tags, used by > 2000 users)



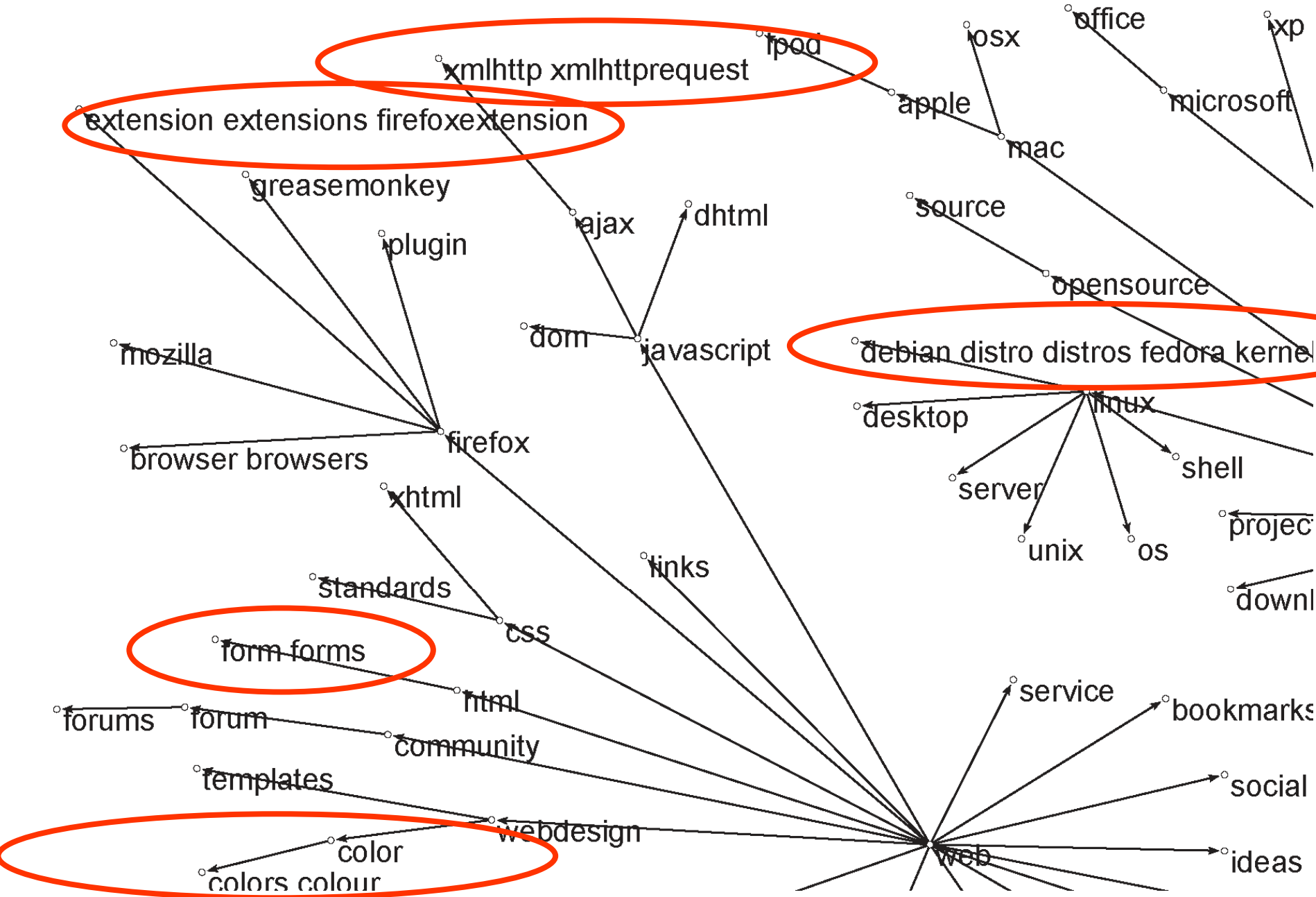
Results for delicious (dataset 2005, 320 tags, used by > 2000 users)



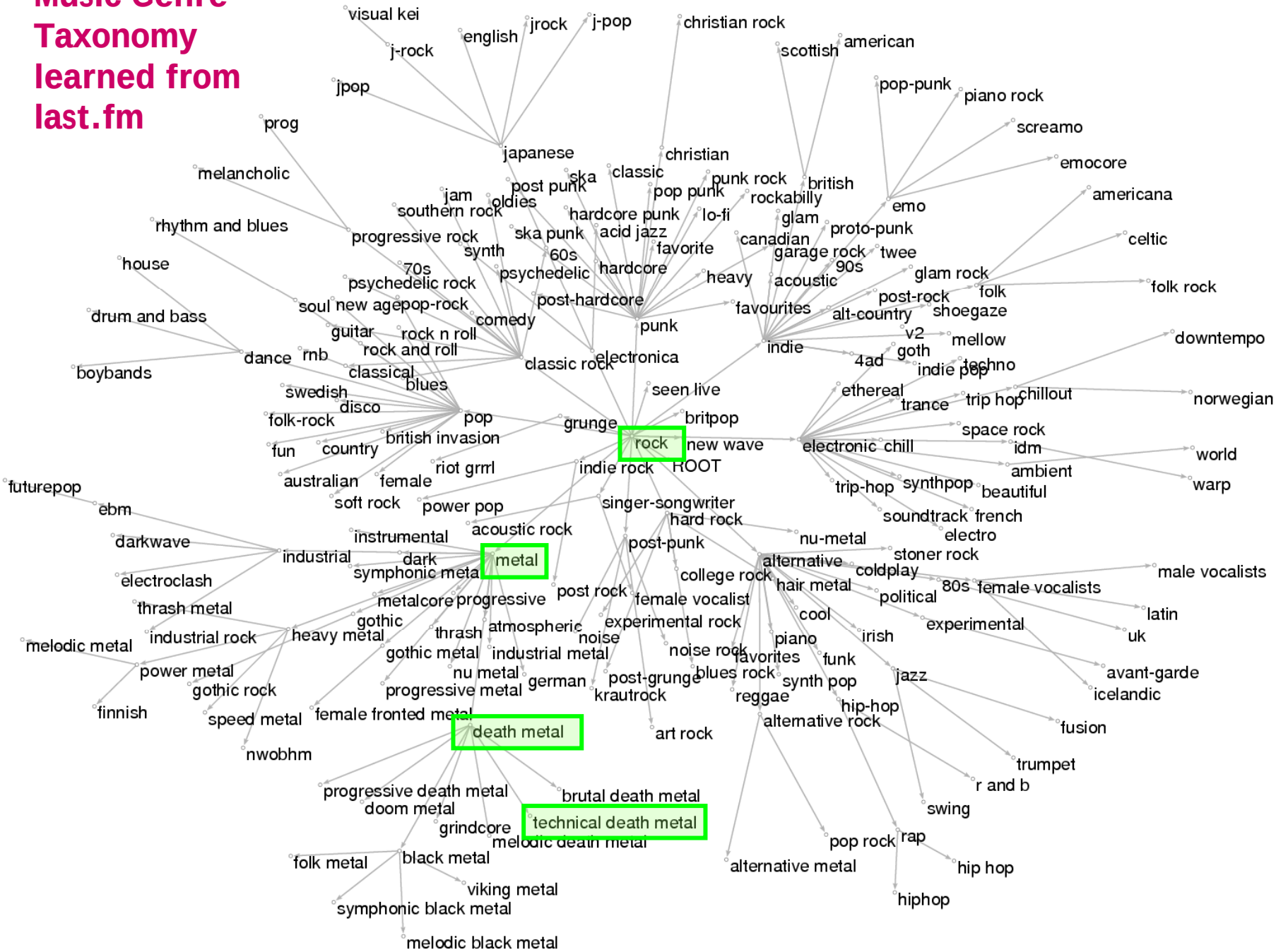
Results for delicious together with similarity pruning



Results for delicious together with similarity pruning



Music Genre Taxonomy learned from last.fm



Conclusion

- Folksonomies overcome the knowledge acquisition bottleneck
  - due to ease of use
  - and therefore of fastly increasing amounts of users.

- Cosine measure seems most suitable to discover synonyms and siblings.

- Similarity measures can be used for Ontology Learning.