UNIKASSEL VERSITÄT

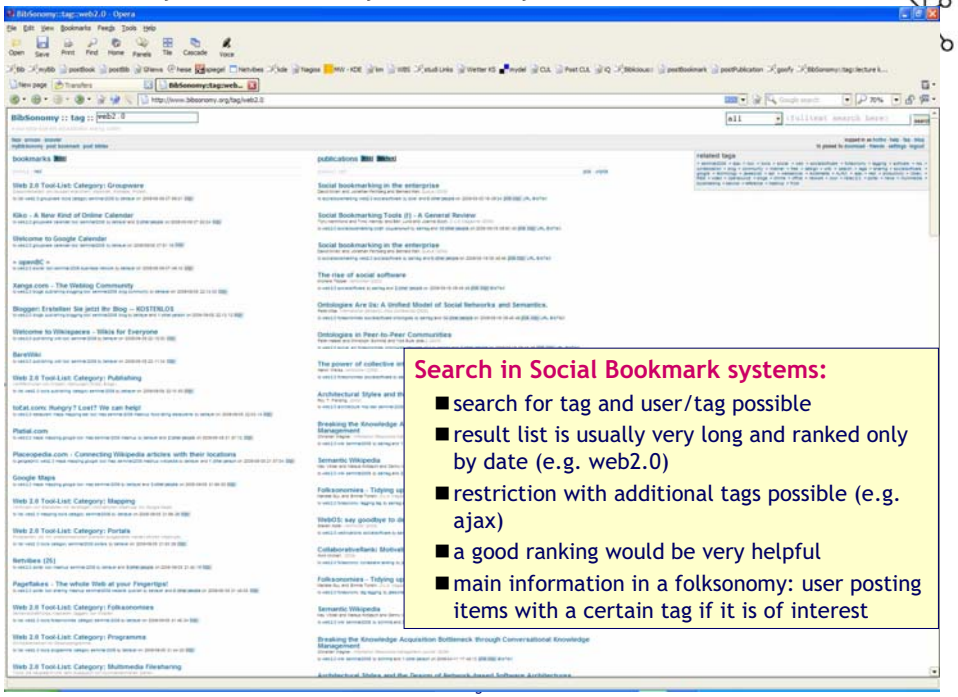
ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Information Retrieval in Folksonomies: Search and Ranking

Andreas Hotho, Robert Jäschke,
Christoph Schmitz, Gerd Stumme

Published in York Sure and John Domingue, editor(s), The Semantic Web: Research and Applications, LNAI 4011, pages 411-426, Springer, Heidelberg, 2006.

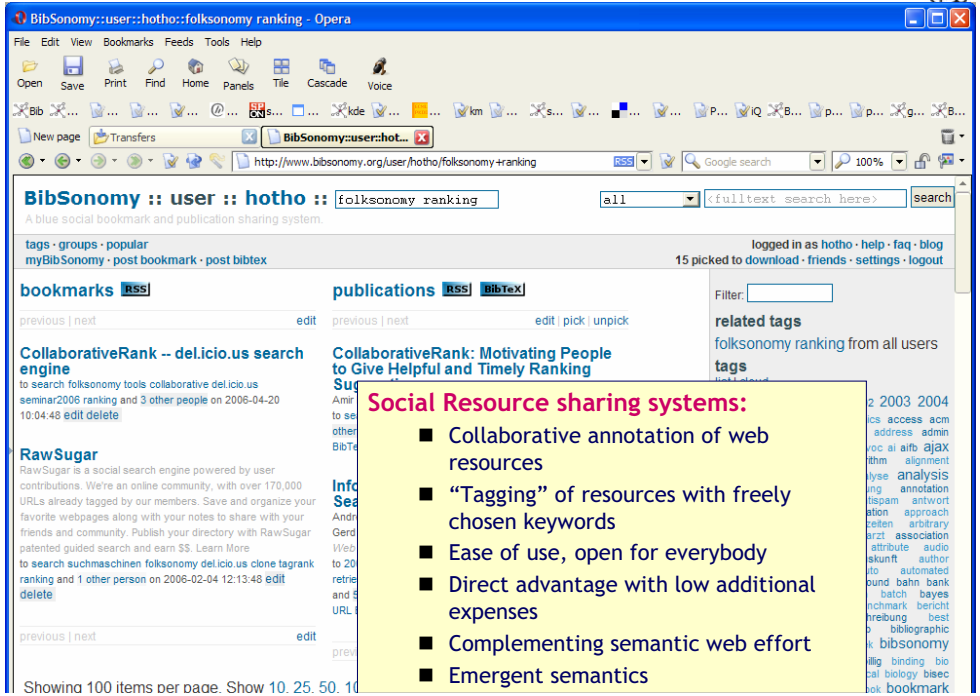
BibSonomy - a Folksonomy/Web 2.0 System



Search in Social Bookmark systems:

- search for tag and user/tag possible
- result list is usually very long and ranked only by date (e.g. web2.0)
- restriction with additional tags possible (e.g. ajax)
- a good ranking would be very helpful
- main information in a folksonomy: user posting items with a certain tag if it is of interest

BibSonomy - a Folksonomy/Web 2.0 System

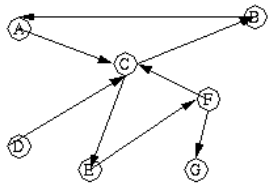


Social Resource sharing systems:

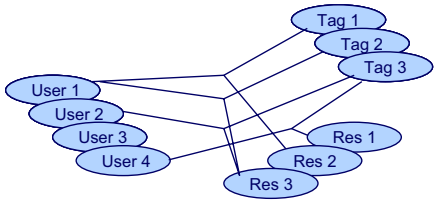
- Collaborative annotation of web resources
- "Tagging" of resources with freely chosen keywords
- Ease of use, open for everybody
- Direct advantage with low additional expenses
- Complementing semantic web effort
- Emergent semantics

Search in Folksonomies

- PageRank in the web: pages are important if a lot of important pages are linking to them
- authority values in a folksonomy are propagated along the hyperlink structure of the folksonomy



Web-Graph



Folksonomies

Formal Model

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

- U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp.
- Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$, called tag assignments (TAS for short)
- and $\prec$ is a user specific subtag/supertag relation, i.e. $\prec \subseteq U \times T \times T$, called subtag/supertag relation.

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:

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

Recall: PageRank

Original PageRank:

Computation of fixed point $\boldsymbol{r}$ of the weight spreading function

$$\boldsymbol{r} := \alpha \boldsymbol{A}\boldsymbol{r} + (1-\alpha)\boldsymbol{e}$$

- $\boldsymbol{A}$ is the row-normalized adjacency matrix reflecting the graph
- $\boldsymbol{e}$: random surfer vector
- α : weighting factor, eg $\alpha = 0.85$

Adaptation to folksonomy:

- each undirected edge → two directed edges

FolkRank: Thematic Ranking in Folksonomies

Problem with the adapted PageRank version:

- Graph is undirected → weight flows in one direction and directly "swashes back"

Idea to solve this is to apply a differential approach:

- Let R_{AP} be the fixed point with $\alpha = 1$
- Let R_{pref} be the fixed point with $\alpha < 1$
- $R := R_{pref} - R_{AP}$ is the final weight vector

Additionally: different weights in random surfer vector allow for topic-specific ranking.

Evaluation on del.icio.us dataset

Crawl of del.icio.us from July 27 to 30, 2005 resulted in a folksonomy with

- $|U|$ = 75,242 users,
- $|T|$ = 533,191 tags and
- $|R|$ = 3,158,297 resources, related by in total
- $|Y|$ = 17,362,212 tag assignments (TAS).

Results: adapted PageRank

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

User	ad. PageRank
shankar	0,0007389
notmuch	0,0007379
fritz	0,0006796
ubi.quito.us	0,0006171
weev	0,0005044
kof2002	0,0004885
ukquake	0,0004844
gearhead	0,0004820
angusf	0,0004797
johncollins	0,0004668
mshook	0,0004556
frizzlebiscuit	0,0004543
rafaspol	0,0004535
xiombarg	0,0004520
tidesonar02	0,0004355
cyrusnews	0,0003829
bldurling	0,0003727
onpause_tv_anytime	0,0003600
cataracte	0,0003462
triple_entendre	0,0003419
kayodeok	0,0003407

Results: adapted PageRank

http://slashdot.org/	0,0002613
http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html	0,0002320
http://script.aculo.us/	0,0001770
http://www.adaptivepath.com/publications/essays/archives/000385.php	0,0001654
http://johnvey.com/features/deliciousdirector/	0,0001593
http://en.wikipedia.org/wiki/Main_Page	0,0001407
http://www.flickr.com/	0,0001376
http://www.goodfonts.org/	0,0001349
http://www.43folders.com/	0,0001160
http://www.csszengarden.com/	0,0001149
http://wellstyled.com/tools/colorscheme2/index-en.html	0,0001108
http://pro.html.it/esempio/nifty/	0,0001070
http://www.alistapart.com/	0,0001059
http://postsecret.blogspot.com/	0,0001058
http://www.beelerspace.com/index.php?p=890	0,0001035
http://www.techsupportalert.com/best_46_free_utilities.htm	0,0001034
http://www.alvit.de/web-dev/	0,0001020
http://www.technorati.com/	0,0001015
http://www.lifehacker.com/	0,0001009
http://www.lucazappa.com/brilliantMaker/buttonImage.php	0,0000992
http://www.engadget.com/	0,0000984

Results: boomerang

Preference for tag: boomerang

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
boomerang	0,4036883
shop	0,0069058
lang:de	0,0050943
software	0,0016797
java	0,0016389
programming	0,0016296
web	0,0016043
reference	0,0014713
system:unfiled	0,0014199
wood	0,0012378
kassel	0,0011969
linux	0,0011442
construction	0,0011023
plans	0,0010226
network	0,0009460
rdf	0,0008506
css	0,0008266
design	0,0008248
dclicious	0,0008097
injuries	0,0008087
pitching	0,0007999

FolkRank with preference

Tag	FolkRank
boomerang	0,4036867
shop	0,0066477
lang:de	0,0050860
wood	0,0012236
kassel	0,0011964
construction	0,0010828
plans	0,0010085
injuries	0,0008078
pitching	0,0007982
rdf	0,0006619
semantic	0,0006533
material	0,0006279
trifly	0,0005691
network	0,0005568
webring	0,0005552
sna	0,0005073
socialnetworkanalysis	0,0004822
cinema	0,0004726
erie	0,0004525
riparian	0,0004467
erosion	0,0004425

Results: Semantic Web

Preference for ressource: http://www.semanticweb.org

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

Results: Semantic Web

http://www.semanticweb.org/	0,3761957
http://flink.semanticweb.org/	0,0005566
http://simile.mit.edu/piggy-bank/	0,0003828
http://www.w3.org/2001/sw/	0,0003216
http://infomesh.net/2001/swintro/	0,0002162
http://del.icio.us/register	0,0001745
http://mspace.ecs.soton.ac.uk/	0,0001712
http://www.adaptivepath.com/publications/essays/archives/000385.php	0,0001637
http://www.ontoweb.org/	0,0001617
http://www.aaai.org/AITopics/html/ontol.html	0,0001613
http://simile.mit.edu/	0,0001395
http://itip.evcc.jp/itipwiki/	0,0001256
http://www.google.be/	0,0001224
http://www.letterjames.de/index.html	0,0001224
http://www.daml.org/	0,0001216
http://shirky.com/writings/ontology_overrated.html	0,0001195
http://jena.sourceforge.net/	0,0001167
http://www.alistapart.com/	0,0001102
http://www.federalconcierge.com/WritingBusinessCases.html	0,0001060
http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html	0,0001059
http://www.shirky.com/writings/semantic_syllogism.html	0,0001052

Conclusion

- Folksonomies might overcome the knowledge aquisition bottleneck through ease of use and growing amount of users
- Our ranking is just based on the structure of the folksonomy – contents of resources not used
- Suitable for intranets
  - resources are typically not hyperlinked
  - community building is important
- Recommend probably relevant documents to user