Leveraging Publication Metadata and Social Data into FolkRank for Scientific Publication Recommendation

Stephan Doerfel,[1] Robert Jäschke,[2] Andreas Hotho,[3] Gerd Stumme[1]

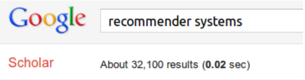
[1]Knowledge and Data Engineering Group, University of Kassel

[2]L3S Research Center, Hannover

[3]Data Mining and Information Retrieval Group, University of Würzburg

RSWeb: 9th September 2012

Publication Overload



- # papers doubles every 10 years
- # journals doubles every 15 years
- → information overload

typical approach: scientific article recommender
e.g., in collaborative tagging systems

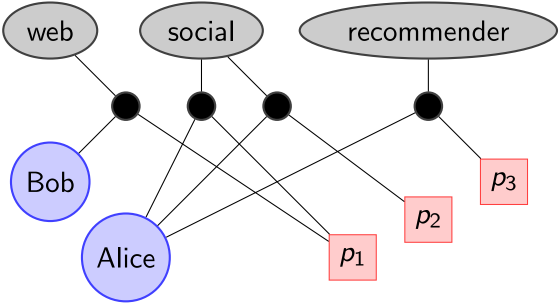
Agenda

Agenda

Item Recommendations in Folksonomies

Folksonomy

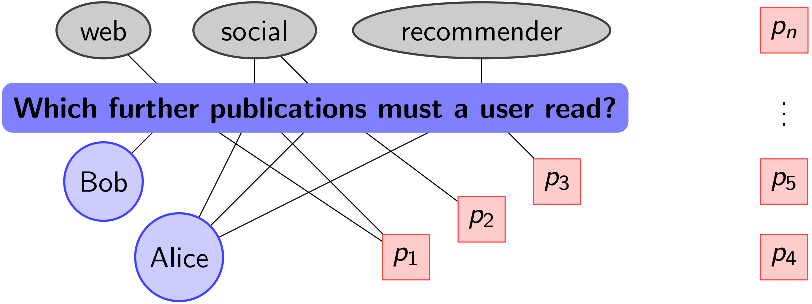
A *folksonomy* is a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., and Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$, whose elements are called *tag assignments*.



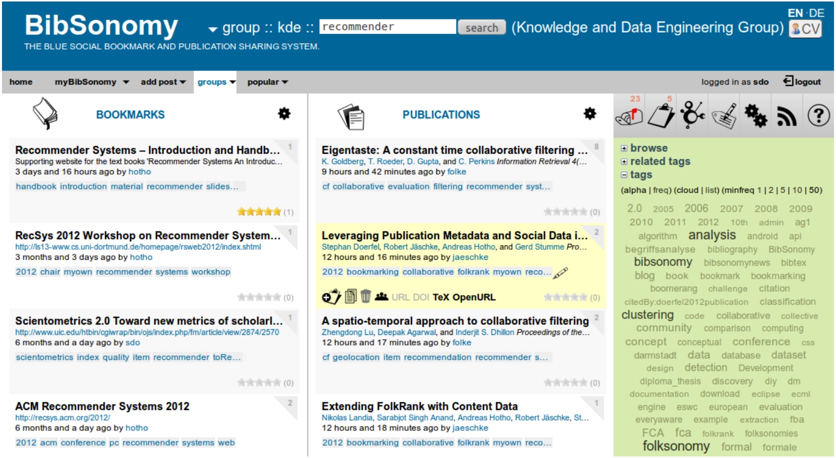
Item Recommendations in Folksonomies

Folksonomy

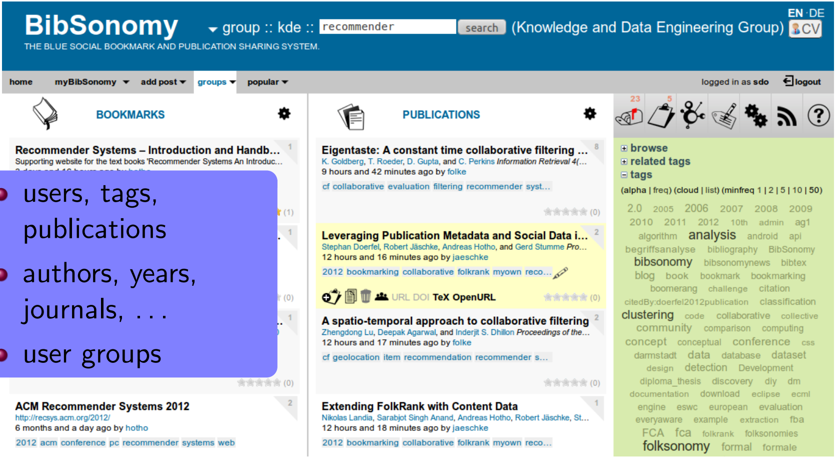
A *folksonomy* is a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., and Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$, whose elements are called *tag assignments*.



BibSonomy


BibSonomy
THE BLUE SOCIAL BOOKMARK AND PUBLICATION SHARING SYSTEM.
group :: kde ::
recommender
search
(Knowledge and Data Engineering Group)
EN · DE
CV
home
myBibSonomy
add post
groups
popular
logged in as sdo
logout
BOOKMARKS
Recommender Systems – Introduction and Handb…
Supporting website for the text books 'Recommender Systems An Introduc…
3 days and 16 hours ago by hotho
handbook introduction material recommender slides…
(1)
RecSys 2012 Workshop on Recommender System…
http://ls13-www.cs.uni-dortmund.de/homepage/rsweb2012/index.shtml
3 months and 3 days ago by hotho
2012 chair myown recommender systems workshop
(0)
Scientometrics 2.0 Toward new metrics of scholarl…
http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570
6 months and a day ago by sdo
scientometrics index quality item recommender toRe…
(0)
ACM Recommender Systems 2012
http://recsys.acm.org/2012/
6 months and a day ago by hotho
2012 acm conference pc recommender systems web
PUBLICATIONS
Eigentaste: A constant time collaborative filtering …
K. Goldberg, T. Roeder, D. Gupta, and C. Perkins Information Retrieval 4(…
9 hours and 42 minutes ago by folke
cf collaborative evaluation filtering recommender syst…
(0)
Leveraging Publication Metadata and Social Data i…
Stephan Doerfel, Robert Jäschke, Andreas Hotho, and Gerd Stumme Pro…
12 hours and 16 minutes ago by jaeschke
2012 bookmarking collaborative folkrank myown reco…
URL DOI TeX OpenURL
(0)
A spatio-temporal approach to collaborative filtering
Zhengdong Lu, Deepak Agarwal, and Inderjit S. Dhillon Proceedings of the…
12 hours and 17 minutes ago by folke
cf geolocation item recommendation recommender s…
(0)
Extending FolkRank with Content Data
Nikolas Landia, Sarabjot Singh Anand, Andreas Hotho, Robert Jäschke, St…
12 hours and 18 minutes ago by jaeschke
2012 bookmarking collaborative folkrank myown reco…
browse
related tags
tags
(alpha | freq) (cloud | list) (minfreq 1 | 2 | 5 | 10 | 50)
2.0 2005 2006 2007 2008 2009
2010 2011 2012 10th admin ag1
algorithm analysis android api
begriffsanalyse bibliography BibSonomy
bibsonomy bibsonomynews bibtex
blog book bookmark bookmarking
boomerang challenge citation
citedBy:doerfel2012publication classification
clustering code collaborative collective
community comparison computing
concept conceptual conference css
darmstadt data database dataset
design detection Development
diploma_thesis discovery diy dm
documentation download eclipse ecml
engine eswc european evaluation
everyaware example extraction fba
FCA fca folkrank folksonomies
folksonomy formal formale

BibSonomy



- users, tags, publications
- authors, years, journals, . . .
- user groups

FolkRank

$$\mathbb{F} = (U, T, R, Y) \;\rightarrow\; \text{undirected graph } G_{\mathbb{F}} = (U \,\dot\cup\, T \,\dot\cup\, R, E)$$
$$(u, t, r) \in Y \;\rightarrow\; \{u, t\}, \{u, r\}, \{t, r\} \in E$$

$$\textit{Adapted PageRank (APR)}\ \vec{w}^d \;:\; \vec{w}^d_{i+1} \leftarrow dA^{\mathrm{T}}\vec{w}^d_i + (1-d)\vec{p}$$
$$\textit{FolkRank} \;:\; \vec{w}^d - \vec{w}^1$$

$$A = \text{the row-stochastic adjacency matrix of } G_{\mathbb{F}}$$
$$\vec{p} = \text{preference vector}$$
$$d \in [0, 1] = \text{parameter for the influence of } \vec{p}$$

Modifications to Folkrank

Adding new Dimensions to the Graph

$\mathbb{F} + M := (U, T, R, M, Y')$ extends the folksonomy $\mathbb{F}$ where

- Y' is a relation $Y' \subseteq U \times T \times R \times M$ and
- each triple of Y is extended with those elements of M that one of the elements of the triple is associated with.

Modifications to the Preference Vector

Select certain users, resources or tags and assign them some weight in $\vec{p}$.

Agenda

Agenda

Four Datasets

Publication posts of BibSonomy[1]

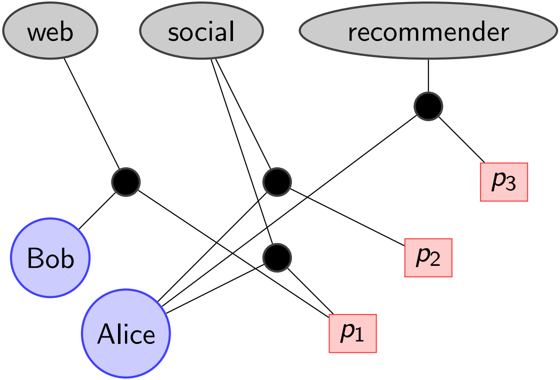
- D_{08} challenge dataset from 2008
- D_{12} recent dataset from 2012
- D_R each publication in ≥ 2 posts
- D_{UR} each publications in ≥ 2 posts, each user ≥ 20 posts

dataset	users	publications
$D_{12,R}$	2,886	29,921
$D_{12,UR}$	541	25,072
$D_{08,R}$	729	13,001
$D_{08,UR}$	150	11,689

[1] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/

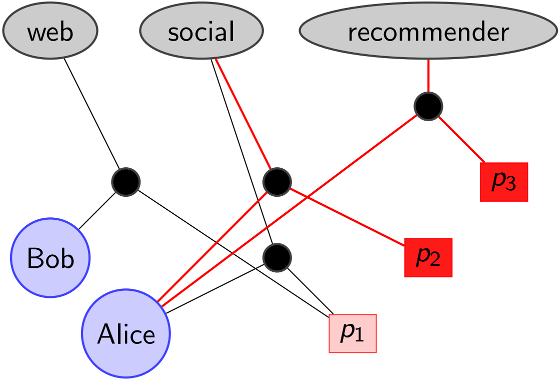
Agenda

Experimental Methodology



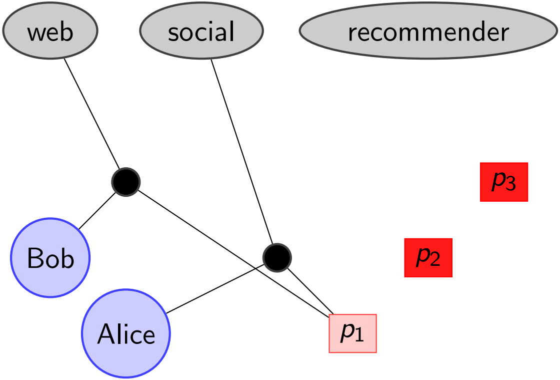
Leave*N*PostsOut

Experimental Methodology



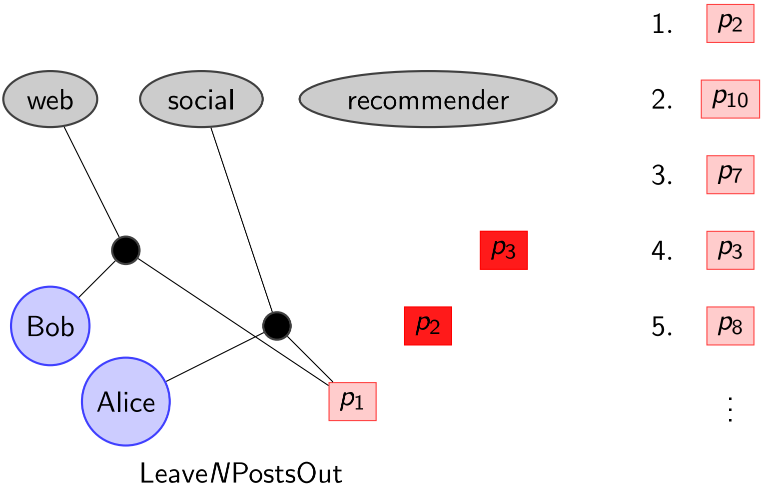
Leave$N$PostsOut

Experimental Methodology

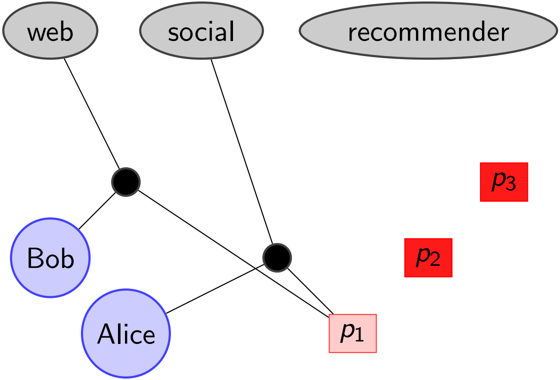


Leave$N$PostsOut

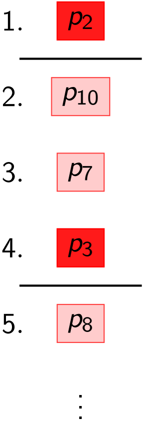
Experimental Methodology



Experimental Methodology



LeaveNPostsOut



MAP = Mean Average Precision

Baselines

Most Popular

- global ranking
- suggesting the most often bookmarked publications to a user

User-Based Collaborative Filtering

- users are represented as vectors in
  - the tag vector space → CF_T
  - or the resource vector space → CF_R
- using a similarity function, one determines a set of similar users and recommends publications that are popular among them

Experiment: Who has the publications to recommend?

How many similar users does it take, to find the 10 left-out resources? – The average coverage of the withheld resources in differently sized neighborhoods of similar users:

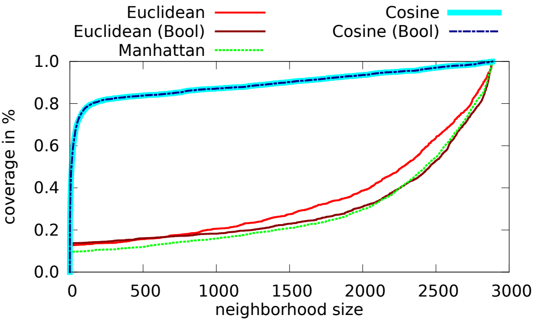


Figure: $D_{12,R}$ as resource vector space.

→ Winner: Cosine similarity in the resource vector space

Agenda

Baselines and plain FolkRank and APR

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
CF_R $k = 10$	**0.109**	**0.141**	**0.120**	**0.152**
FolkRank	0.090	0.118	0.099	0.129
adapted PageRank (APR)	0.066	0.058	0.070	0.062
CF_T $k = 4$	0.062	0.081	0.060	0.088
most popular	0.006	0.013	0.007	0.013

$$\mathrm{CF}_R > \mathit{FolkRank} > \begin{cases} \mathit{APR} \\ \mathrm{CF}_T \end{cases} > \text{most popular}$$

Agenda

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

first authors > last authors > all authors

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

comparable to plain FolkRank

Including New Dimensions

MAP scores	$D_{12,R}$	$D_{08,R}$	$D_{12,UR}$	$D_{08,UR}$
(plain) FolkRank	**0.090**	**0.118**	0.099	**0.129**
$\mathbb{F}$ + first authors	0.089	0.113	**0.102**	0.126
$\mathbb{F}$ + last authors	0.086	0.108	0.097	0.120
$\mathbb{F}$ + all authors	0.085	0.103	0.096	0.115
$\mathbb{F}$ + group	0.085	0.117	0.093	0.128

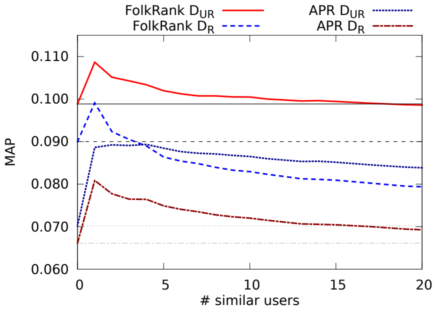
$\mathbb{F}$ + posting year, publication year, tag clusters, venues, ...

Agenda

Exploiting Similar Users

Motivation: Good results of CF_R.

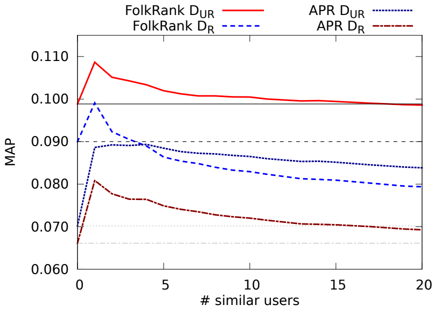
For user u select the k most similar users, insert their similarity value to u into $\vec{p}$.



All scenarios profit from the *inclusion of* at least very small *user neighborhoods*.

Exploiting Similar Users

Motivation: Good results of CF_R.
For user u select the k most similar users, insert their similarity value to u into $\vec{p}$.



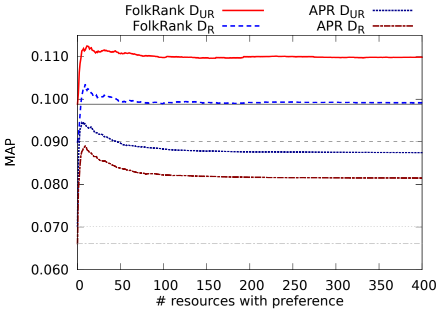
- *Best results*: only the single most similar user gets additional preference
- APR profits even more from the inclusion of similar users, also for larger neighborhoods
- Using the *Euclidean distance decreases* MAP.

All scenarios profit from the *inclusion of* at least very small *user neighborhoods*.

Exploiting Recent Resources

Motivation: User's interests vary during the use of the system.

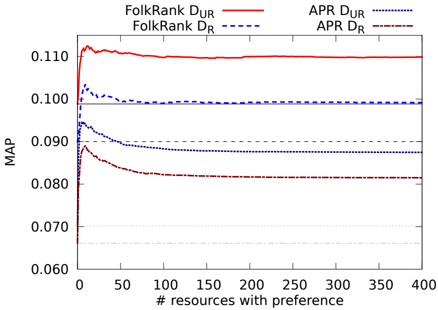
For user u select the k most recently posted resources, assign the same weight to them in $\vec{p}$.



Comparable to the scores of *CF_R*.

Exploiting Recent Resources

Motivation: User's interests vary during the use of the system.
For user u select the k most recently posted resources, assign the same weight to them in $\vec{p}$.



- *Significantly exceeds FolkRank* (for ≥ 3 resources on all datasets).
- Highest MAP scores for different number of recent resources.
- Constant MAP for larger numbers of recent resources.

Comparable to the scores of CF_R.

Agenda

Conclusions

- Cosine similarity is the measure of choice (CF_R or $\vec{p}$ in FolkRank).
- Generally FolkRank below CF_R but better than CF_T.
- Using resources seems to be more beneficial than using tags.
- Authors or groups as additional dimension → scores comparable to plain FolkRank.
- Small user neighborhoods can improve FolkRank recommendations.
- Recency of a post is a valuable indicator for the current interests of a user (best FolkRank results).

Future Work

- Repeat the more successful experiments on further datasets.
- Investigate whether certain types of users can benefit more from the inclusion of certain data than others.
- Truly capture a recommender's performance: Online-Evaluation in BibSonomy
- Use different versions of FolkRank as candidates for hybrid-recommenders.

http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/

Future Work

- Repeat the more successful experiments on further datasets.
- Investigate whether certain types of users can benefit more from the inclusion of certain data than others.
- Truly capture a recommender's performance: Online-Evaluation in BibSonomy
- Use different versions of FolkRank as candidates for hybrid-recommenders.

Thank you for your attention!

http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/