Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D3.2 Methods for identifying communities in Social Tagging Systems

Period covered: from 01/06/2006 to 31/05/2008
Start date of project: June 1st, 2006
Due date of deliverable: July 15th, 2008
Distribution: Public

Date of preparation: 31/05/2008
Duration: 36 months
Actual submission date: June 20th, 2008
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: Kassel
Lead contractor for these deliverables: Kassel

Executive Summary

This deliverable summarizes our work on community detection in folksonomies. Our work can be grouped into three parts.

We present an approach that starts with the generation of a hierarchical clustering of the tag space by iteratively applying the k-Means clustering algorithm. The tag clusters on the bottom level are then considered as intensional descriptions of our FolkRank algorithm. For the choice of k-Means as initial clustering algorithm, we provide the semantic grounding, which shows that the average semantic distance of pairs of tags within clusters generated by k-Means is significantly smaller than within randomly generated clusters.

A similar approach has been implemented in BibSonomy. Here, the focus lay on efficiency, since we display the results online on all tag pages. We provide, for each tag, the community of users that are mostly related to this tag.

Contents

List of Figures

Chapter 1

Introduction

This deliverable summarizes our work on community detection in folksonomies. A key question for this topic is what one should consider as a community. The literature provides many alternative definitions, which are used in parallel. Even though all types of community that we considered are defined semantically, i. e., by a tag or a set of tags, this variety is also reflected in this deliverable.

In Chapter 2, we present an approach that was presented at the International Conference on Knowledge Management 2007 (Grahl et al., 2007). There, we start with the generation of a hierarchical clustering of the tag space by iteratively applying the k-Means clustering algorithm. The tag clusters on the bottom level are then considered as intensional descriptions of our FolkRank algorithm. Users can belong to different communities (i. e., , the communities are overlapping) to a different degree.

Chapter 3 provides the semantic grounding for the applicability of k-Means. This (up to now unpublished) work shows that the average semantic distance of pairs of tags within clusters generated by k-Means is significantly smaller than within randomly generated clusters. The semantic distance is measured with the Jiang-Conrath distance in WordNet.

In Chapter 4, we describe an application of the community detection in BibSonomy. Similar to the approach discussed in Chapter 2, we provide, for each tag, the community of users that are mostly related to this tag. The computation is done again by the FolkRank algorithm. One of its benefits is that a user may belong to a community around a tag even if he did not use the tag himself, but used closely related tags or synonyms instead. The resulting communities are shown online in BibSonomy on all tag pages.

Chapter 2

Conceptual Clustering of Social Bookmarking Sites

Currently, social bookmarking systems provide intuitive support for browsing locally their content. A global view is usually presented by a tag cloud of the system, but it does not allow a drill-down along a conceptual hierarchy. After having selected one tag, it leads to a disarranged list of resources, and no further drill-down is possible.

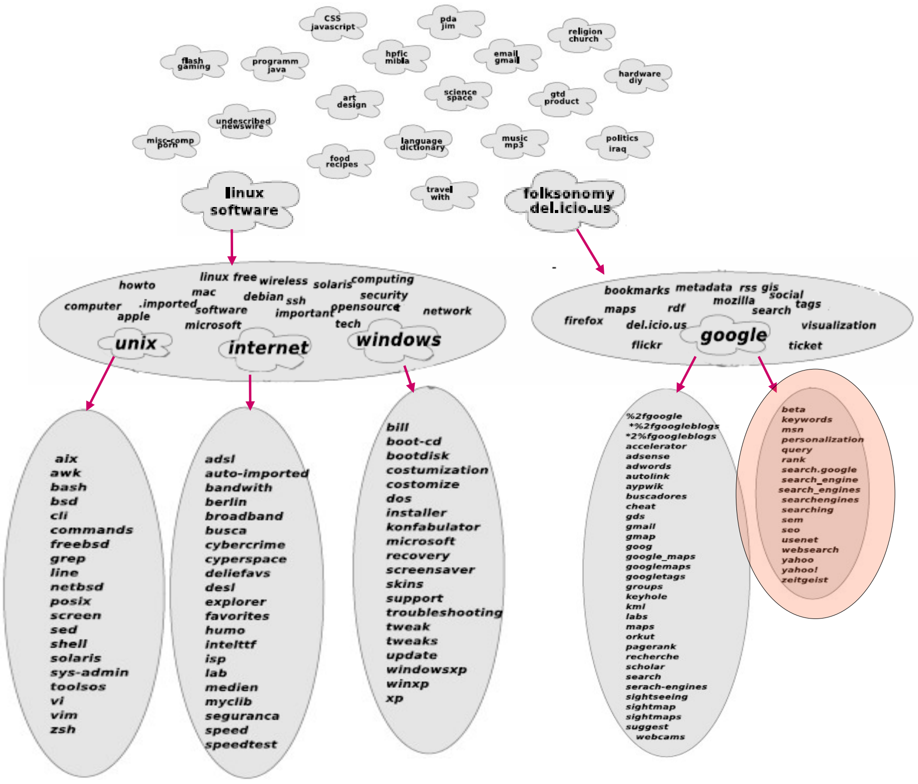


Figure 2.1: A user view of the hierarchy from top to bottom.

The first approach to community detection that we studied is based on an iterative application of k-Means (Forgy, 1965) to a given folksonomy to generate a conceptual hierarchy (see fig. 2.1) in order to provide users in searching. The hierarchy is complemented with a ranked list of users and resources which are most related to each cluster. The rankings are computed using our Folkrank algorithm (Hotho et al., 2006).

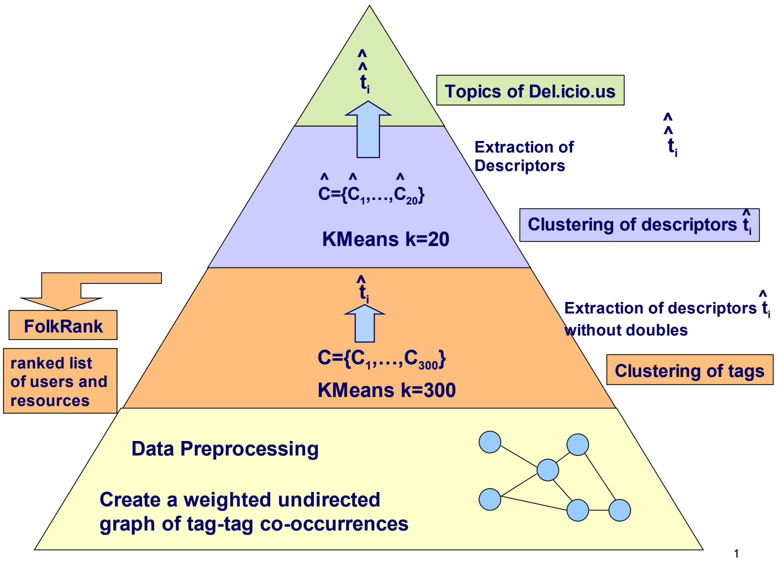
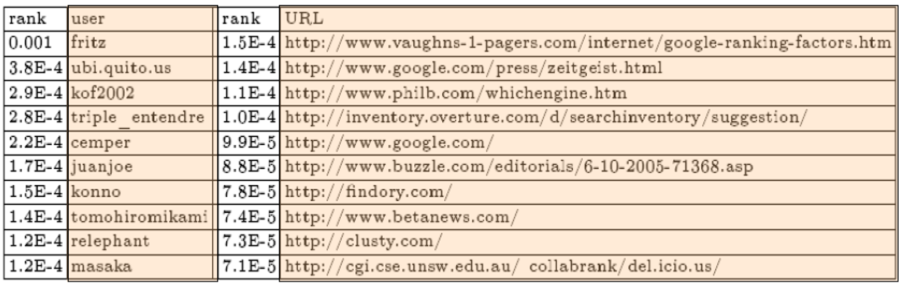

$\hat{\hat{t}}_i$
Topics of Del.icio.us
Extraction of Descriptors
$\hat{\hat{t}}_i$
$C=\{\hat{C}_1,\ldots,\hat{C}_{20}\}$
KMeans k=20
Clustering of descriptors $\hat{t}_i$
$\hat{t}_i$
Extraction of descriptors $\hat{t}_i$ without doubles
FolkRank
ranked list of users and resources
$C=\{C_1,\ldots,C_{300}\}$
KMeans k=300
Clustering of tags
Data Preprocessing
Create a weighted undirected graph of tag-tag co-occurrences
1

Figure 2.2: Generating a conceptual hierarchy from bottom to top.

Our approach (see detailed description in (Grahl et al., 2007)) is visualised in Figure 2.2 for the bookmarking system *del.icio.us*. The algorithmic application is conducted from bottom to top on a weighted tag-tag-concurrence matrix, where the lowest level presents the most fine grained level of the hierarchy – in contrast to the top level, which is a summary of the entire topics of del.icio.us.com of July 2005. To facilitate that users can find others users with similar interests and resources which are related to the personal browsing history, we applied the FolkRank ranking, using the resulting set of tags of the fine grained level as arguments. One example run is shown in Figure 2.3for the red marked cluster on the right side of the hierarchy.



rank	user	rank	URL
0.001	fritz	1.5E-4	http://www.vaughns-1-pagers.com/internet/google-ranking-factors.htm
3.8E-4	ubi.quito.us	1.4E-4	http://www.google.com/press/zeitgeist.html
2.9E-4	kof2002	1.1E-4	http://www.philb.com/whichengine.htm
2.8E-4	triple_entendre	1.0E-4	http://inventory.overture.com/d/searchinventory/suggestion/
2.2E-4	cemper	9.9E-5	http://www.google.com/
1.7E-4	juanjoe	8.8E-5	http://www.buzzle.com/editorials/6-10-2005-71368.asp
1.5E-4	konno	7.8E-5	http://findory.com/
1.4E-4	tomohiromikami	7.4E-5	http://www.betanews.com/
1.2E-4	relephant	7.3E-5	http://clusty.com/
1.2E-4	masaka	7.1E-5	http://cgi.cse.unsw.edu.au/ collabrank/del.icio.us/

Figure 2.3: Ranked list of users and resources.

Chapter 3

A semantically grounded evaluation of the result of the k-Means algorithm on a folksonomy

A key question for the approach presented in the previous chapter is if k-Means is the appropriate clustering algorithm for the task at hand. In this chapter, we present first steps towards the evaluation of the clusterings generated by the k-Means algorithm. The work is unpublished up to now; therefore we provide a longer presentation.

A popular and not already solved problem in unsupervised data mining methods is the evaluation of clustering results. A common way of evaluation is the measurement with so called benchmark data sets like Zachary's karate club (Zachary, 1977) or the Enron Corpus (Klimt and Yang, 2004). The problem of evaluating clusterings of folksonomies is the lack of benchmark data sets. The standard benchmark data are not applicable because of the specific tri-mode structure of folksonomies. Therefore, we focus here on the first steps of evaluating clustering results by considering the semantic distance of pairs of tags within and across clusters.

We will compare the clustering generated by k-Means with two clusterings: The first one serves as null hypothesis; it consists of randomly assigned clusterings, which still have the same size distribution as the k-Means clustering. This clustering is obtained by randomly shuffling the k-Means clustering. The second clustering is one that we consider as optimal: it is using the semantic distance between all pairs of tags as input, and optimises inter- and intra-clustering distances. The optimisation is done with the PAM (partitioning around medoid) clustering algorithm (Kaufman and Rousseeuw, 1990). For each of the three versions, we will compute the macroaverage μ and microaverage ν of the semantic distances between all pairs of tags. As distance measure, we use the Jiang-Conrath distance[1] in WordNet(Fellbaum, 1998), which has been empirically validated by Budanitsky and Hirst (Budanitsky, 2001). We will then assume that k-Means performs good if it is close to the (lower) value of the PAM clustering on the Jiang-Conrath measure and farther away from the (higher) value of the randomly generated clustering. To this end, we performed the following steps.

Dataset and Basic Notations: We have apply our method on two data sets of the social bookmarking system del.icio.us.

Dataset from June 2004 In June 2004, we crawled del.icio.us and obtained a set U of 3,301 users, a set T of 30,416 tags, and a set R of 220,366 resources. There were in total 359,025 posts, i. e., triples of the form (u, S, r), indicating that user $u \in U$ has assigned all tags contained in $S \subseteq T$ to resource $r \in R$. The set $Y \subseteq U \times T \times R$ of all tag assignments consist of 616,819 tag assignments.

[1] In order to make Jiang-Conrath applicable, we restricted the evaluation on all tags that were present in the noun part of WordNet.

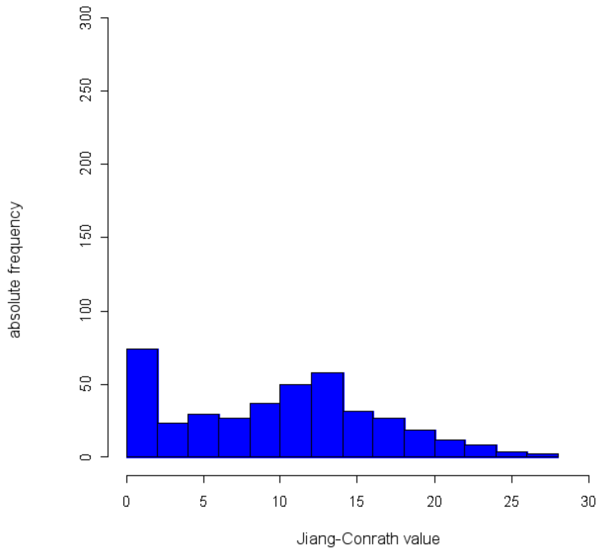


Figure 3.1: Distribution of Jiang-Conrath Distance of tag-tag tuples in del.icio.us 2004.

Dataset from July 2005 Between July 27 and 30, 2005, we also crawled del.icio.us and received totally a set U of 75,085 users, a set T of 456,666 tags, and a set R of 3,006,114 resources. There were in total 7,281,940 posts, i. e., also triples of the form (u, S, r), indicating that user $u \in U$ has assigned all tags contained in $S \subseteq T$ to resource $r \in R$. The set $Y \subseteq U \times T \times R$ of all tag assignments consist overall of 17,362,082 tag assignments.

Data Pre-processing:

Before we started the analysis, we have cut off the long tail of the dataset by using the p-core algorithm (Jäschke et al., 2007) with $p = 10$. This means that all remaining users, tags, and resources, have at least in 10 posts. Regarding the further data pre-processing, we only consider those tags which exist in the noun part of WorldNet. Furthermore we removed those tag assignments, which contain non-alphanumeric characters. For the matrix for the PAM algorithm, we built up an undirected weighted graph, where each pair t_i, t_j of tags is connected by an edge where the weight is the smallest Jiang-Conrath distance of all the occurrences of the two tags in WordNet:

$$W(t_i, t_j) := min(JC(t_i, t_j)) \tag{3.1}$$

k–Means is computed on a two-mode projection of the folksonomy, which is defined as follows. For a given user u, the following tags T_u and resources R_u are connected to u: $T_u = \{t \in T \mid \exists r : (t, u, r) \in Y\}$, $R_u = \{u \in U \mid \exists t : (t, u, r) \in Y\}$. Furthermore, let $tr_u := \{(t, r) \in T \times R \mid (t, u, r) \in Y\}$ the (tag, resource) pairs occurring with u. Each tag $t_i \in T$ is then represented in the $|R|$-dimensional vector space by the vector $\vec{t^i} := (\vec{t^r_j})_{j=1,\ldots,|R|}$ with $\vec{t^i_j} := W(t_i, r_j)$.

The pre-processing reduces the data set of July 2004 to 47,887 tag assignments which contain 1036 users, 452 tags and 2101 resources (30,215 posts); and the data set of July 2005 to 4,993,414 tag assignments which contain 6,373 tags, 35,308 users and 70,567 resources (2,340,286 post).

First Insights. To get a first insight into the behavior of the Jiang-Conrath measure, we evaluated its distribution independent from any clustering structure over all the data set (see fig. 3.1). This evaluation excludes all tags which contain any additional characters, and the synsets assignments are made such that they result in a minimum Jiang-Conrath-distance.

Comparison: Our evaluation is based on the macroaverage μ and microaberage ν of the Jiang-Conrath distances within the clusters:

$$\mu := \frac{1}{|C|} \sum_{c \in C} \left(\frac{1}{|c|} \sum_{t_i, t_j \in c, i<j} JC(t_i, t_j) \right) \tag{3.2}$$

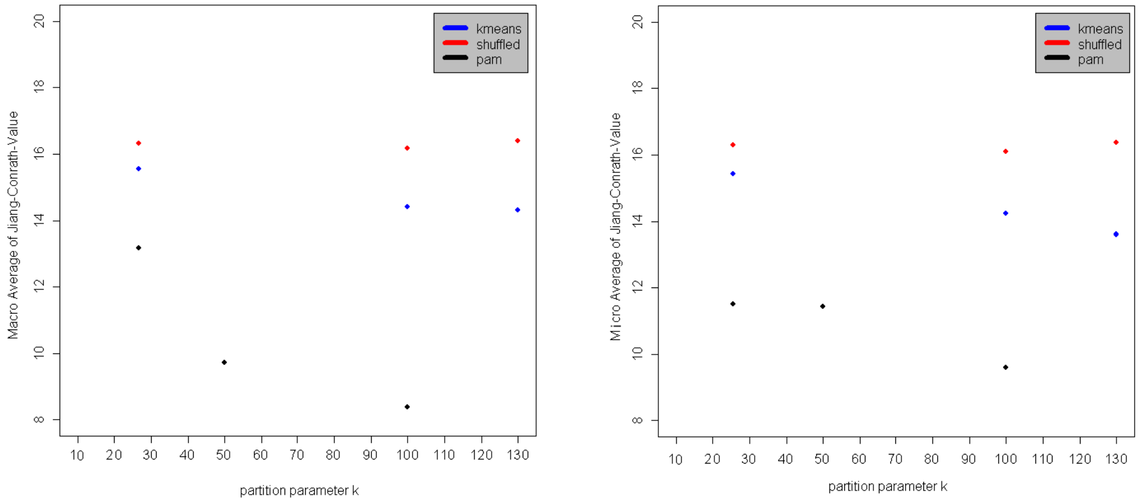
$$\nu := \frac{1}{|T|} \sum_{c \in C} \sum_{t_i, t_j \in c, i<j} JC(t_i, t_j) \tag{3.3}$$

where C is a clustering on the set T of tags and c iterates over the individual clusters of C.

The shuffled clustering is created based on the origin k-Means clustering C by applying a Knuth Shuffle (Knuth, 1981).

Both values, the macro average μ (see eq. 3.2) and micro average ν (see eg. 3.3) gives evidence about the semantic quality of clusters $c_i \in C$. The computation of the Jiang-Conrath distance is based on the tuplewise comparison $t_i \in T_c$ of tags within one cluster. The macro average weighted the size of each cluster and considers the numbers of tupelwise comparison T_c in each cluster. In comparison to the macro average of C, the micro average weights over all tuples T_C concerning a clustering C.

The choice of the partitioning parameter of k-Means was made manually, such that the resulting clusters contain 3-18 elements (for the data set July 2005 and June 2004), which seems to be suitable to human perception.

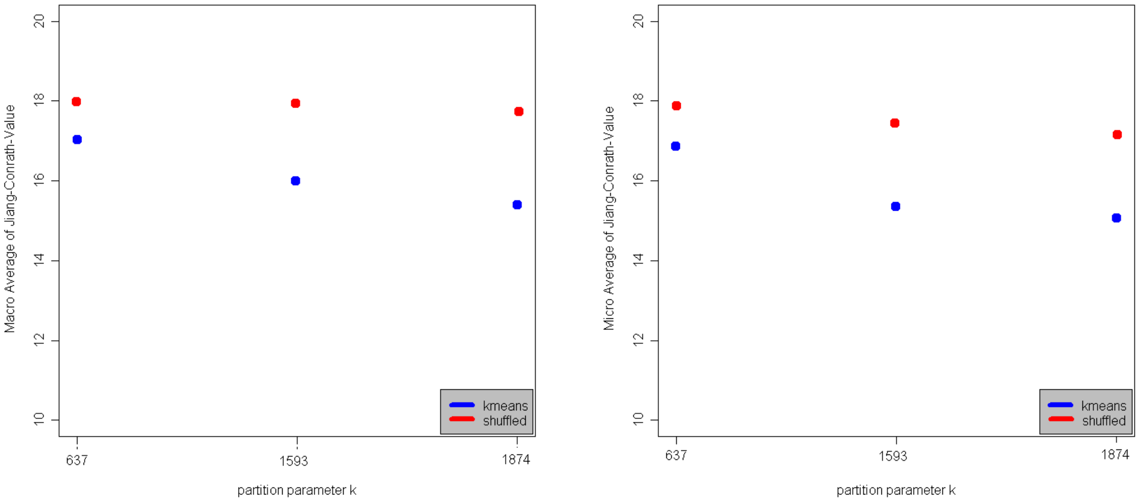


macro average

micro average

Figure 3.2: Comparison of macro and micro average with a random generated clustering of del.icio.us 2004.

Results of data for June 2004 and July 2005. We computed and compare the values in Eq. 3.2 and in Eq. 3.3 for k-Means with different partition parameters, for the null hypothesis provided by the shuffled clustering, and for the PAM application which provides a lower bound. As seen in Figure 3.2, the macro and micro averages of the regular clusterings are different from those of the randomly generated clusterings. In both cases, the values of the regular clustering are lower. The computation of PAM is not conducted for the delicious dataset of July 2005 due to performance reasons.



macro average

micro average

Figure 3.3: Comparison of macro and micro average with a random generated clustering of del.icio.us 2005.

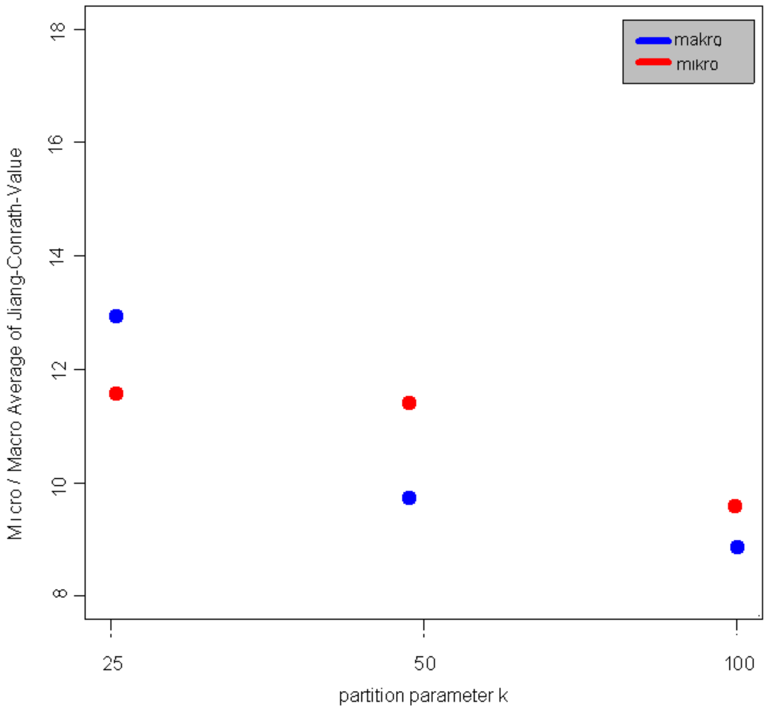


Figure 3.4: Macro -and microaverage of the PAM Clustering.

Chapter 4

Implementation of Community Detection in BibSonomy

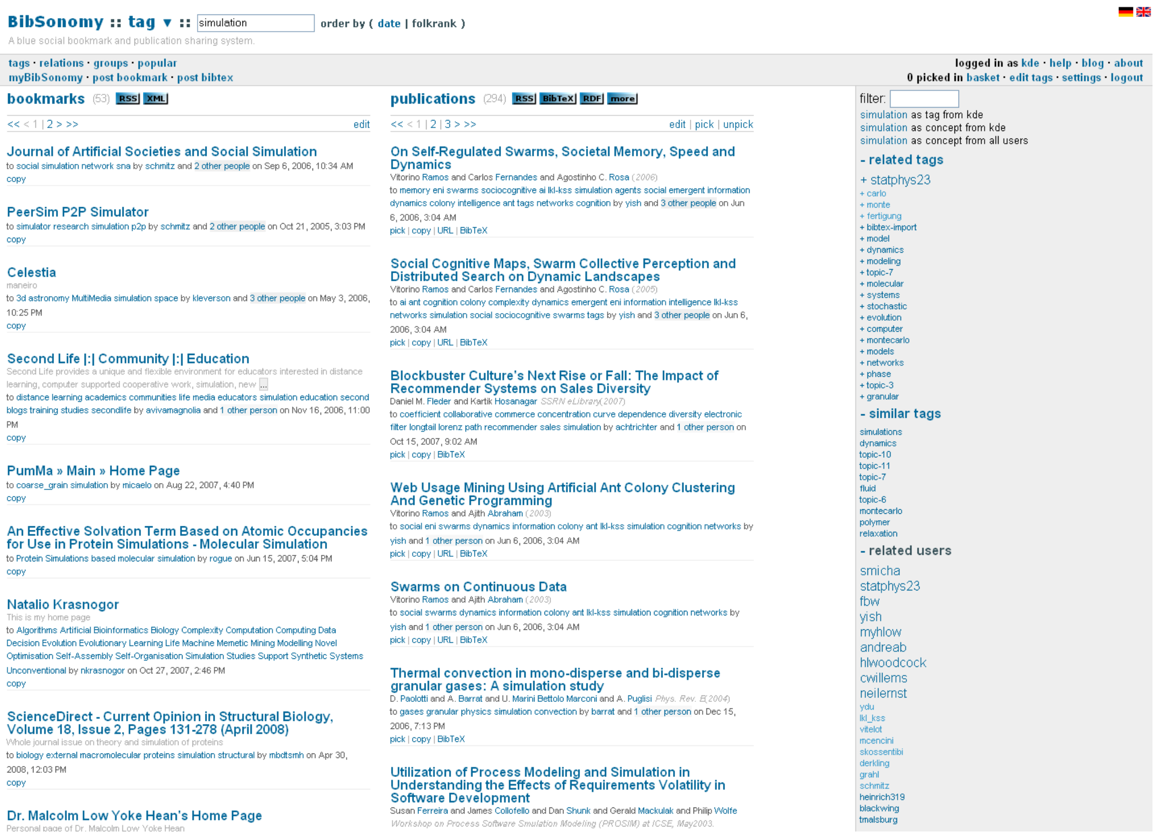
We have implemented a community detection functionality in BibSonomy. It follows an approach similar to the one discussed in Chapter 2, but focusses on efficiency, as this feature should not slow down the generation of the web pages.

We compute communities for each tag. Taking the tag *simulation* as an example, its community is shown in the lower right corner of the web page `http://www.bibsonomy.org/tag/simulation/?order=folkrank` (see Figure 4.1). The given related tags are based on tag-tag co-occurrence, while the similar tags are computed by a more sophisticated context measure. The actual community around a tag (titled *related users* in our system) is computed by the FolkRank algorithm.

As its computation comprises recursively spreading weights through the complete underlying folksonomy, online computation bears the risk of slowing page generation time significantly down. In order to keep our system highly responsive, we decided to source out the FolkRank computation to a background process. This process is started periodically and computes (among others) all possible user communities around all tags currently present in the system. The results are then made available to the online system. This strategy reduces the effort of community detection for a single tag to a simple lookup at page generation time, at the cost of a slight delay with regard to new tags entering the system. However, as the FolkRank communities are recomputed frequently (currently approximately once per week), we believe this tradeoff is justified. Furthermore, for communities around existing tags, our experience suggests that weekly re-computation seems to be sufficiently able to capture changes.

A major benefit of using FolkRank for community detection is that that a user may belong to a community around a tag even if he did not use the tag itself, but closely related tags or synonyms instead. This feature enables users to discover other users interested in the same topic as themselves, but using a differing vocabulary to describe it.

Taking the community list as a starting point, an interested user can click on any member user name to get an overview of this user's resources and tags. Choosing then another tag together with the FolkRank-option, the user can browse through different communities in BibSonomy and discover interesting content and like-minded people.


BibSonomy :: tag ▾ :: simulation order by (date | folkrank)
A blue social bookmark and publication sharing system.
tags · relations · groups · popular
myBibSonomy · post bookmark · post bibtex
logged in as kde · help · blog · about
0 picked in basket · edit tags · settings · logout
bookmarks (53) RSS XML
<< < 1 | 2 > >> edit
publications (294) RSS BibTeX RDF more
<< < 1 | 2 | 3 > >> edit | pick | unpick
filter:

Journal of Artificial Societies and Social Simulation
to social simulation network sna by schmitz and 2 other people on Sep 6, 2006, 10:34 AM
copy

PeerSim P2P Simulator
to simulator research simulation p2p by schmitz and 2 other people on Oct 21, 2005, 3:03 PM
copy

Celestia
maneiro
to 3d astronomy MultiMedia simulation space by kleverson and 3 other people on May 3, 2006, 10:25 PM
copy

Second Life |:| Community |:| Education
Second Life provides a unique and flexible environment for educators interested in distance learning, computer supported cooperative work, simulation, new ...
to distance learning academics communities life media educators simulation education second blogs training studies secondlife by avivamagnolia and 1 other person on Nov 16, 2006, 11:00 PM
copy

PumMa » Main » Home Page
to coarse_grain simulation by micaelo on Aug 22, 2007, 4:40 PM
copy

An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations - Molecular Simulation
to Protein Simulations based molecular simulation by rogue on Jun 15, 2007, 5:04 PM
copy

Natalio Krasnogor
This is my home page
to Algorithms Artificial Bioinformatics Biology Complexity Computation Computing Data Decision Evolution Evolutionary Learning Life Machine Memetic Mining Modelling Novel Optimisation Self-Assembly Self-Organisation Simulation Studies Support Synthetic Systems Unconventional by nkrasnogor on Oct 27, 2007, 2:46 PM
copy

ScienceDirect - Current Opinion in Structural Biology, Volume 18, Issue 2, Pages 131-278 (April 2008)
Whole journal issue on theory and simulation of proteins
to biology external macromolecular proteins simulation structural by mbdtsmh on Apr 30, 2008, 12:03 PM
copy

Dr. Malcolm Low Yoke Hean's Home Page
Personal page of Dr. Malcolm Low Yoke Hean

On Self-Regulated Swarms, Societal Memory, Speed and Dynamics
Vitorino Ramos and Carlos Fernandes and Agostinho C. Rosa (2006)
to memory eni swarms sociocognitive ai lkl-kss simulation agents social emergent information dynamics colony intelligence ant tags networks cognition by yish and 3 other people on Jun 6, 2006, 3:04 AM
pick | copy | URL | BibTeX

Social Cognitive Maps, Swarm Collective Perception and Distributed Search on Dynamic Landscapes
Vitorino Ramos and Carlos Fernandes and Agostinho C. Rosa (2005)
to ai ant cognition colony complexity dynamics emergent eni information intelligence lkl-kss networks simulation social sociocognitive swarms tags by yish and 3 other people on Jun 6, 2006, 3:04 AM
pick | copy | URL | BibTeX

Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity
Daniel M. Fleder and Kartik Hosanagar *SSRN eLibrary*(2007)
to coefficient collaborative commerce concentration curve dependence diversity electronic filter longtail lorenz path recommender sales simulation by achtrichter and 1 other person on Oct 15, 2007, 9:02 AM
pick | copy | BibTeX

Web Usage Mining Using Artificial Ant Colony Clustering And Genetic Programming
Vitorino Ramos and Ajith Abraham (2003)
to social eni swarms dynamics information colony ant lkl-kss simulation cognition networks by yish and 1 other person on Jun 6, 2006, 3:04 AM
pick | copy | URL | BibTeX

Swarms on Continuous Data
Vitorino Ramos and Ajith Abraham (2003)
to social swarms dynamics information colony ant lkl-kss simulation cognition networks by yish and 1 other person on Jun 6, 2006, 3:04 AM
pick | copy | URL | BibTeX

Thermal convection in mono-disperse and bi-disperse granular gases: A simulation study
D. Paolotti and A. Barrat and U. Marini Bettolo Marconi and A. Puglisi *Phys. Rev. E*(2004)
to gases granular physics simulation convection by barrat and 1 other person on Dec 15, 2006, 7:13 PM
pick | copy | BibTeX

Utilization of Process Modeling and Simulation in Understanding the Effects of Requirements Volatility in Software Development
Susan Ferreira and James Collofello and Dan Shunk and Gerald Mackulak and Philip Wolfe
Workshop on Process Software Smulation Modeling (PROSIM) at ICSE, May2003.

simulation as tag from kde
simulation as concept from kde
simulation as concept from all users

- related tags

+ statphys23
+ carlo
+ monte
+ fertigung
+ bibtex-import
+ model
+ dynamics
+ modeling
+ topic-7
+ molecular
+ systems
+ stochastic
+ evolution
+ computer
+ montecarlo
+ models
+ networks
+ phase
+ topic-3
+ granular

- similar tags

simulations
dynamics
topic-10
topic-11
topic-7
fluid
topic-6
montecarlo
polymer
relaxation

- related users

smicha
statphys23
fbw
yish
myhlow
andreab
hlwoodcock
cwillems
neilernst
ydu
lkl_kss
vitelot
mcencini
skossentibi
derkling
grahl
schmitz
heinrich319
blackwing
tmalsburg

Figure 4.1: The BibSonomy community around the tag 'simulation' is shown in the lower right corner. Related tags are based on tag-tag co–occurrence, similar tags are computed by a context measure. The related users are computed by FolkRank.

Bibliography

A. Budanitsky. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, 2001. URL citeseer.ist.psu.edu/budanitsky01semantic.html.

C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.

Miranda Grahl, Andreas Hotho, and Gerd Stumme. Conceptual clustering of social bookmarking sites. In *7th International Conference on Knowledge Management (I-KNOW '07)*, pages 356–364, Graz, Austria, SEP 2007. Know-Center.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006. Springer. URL http://.kde.cs.uni-kassel.de/hotho.

Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514. Springer, 2007. ISBN 978-3-540-74975-2.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an introduction to cluster analysis*. Wiley, 1990.

Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2004. URL http://www.ceas.cc/papers-2004/168.pdf.

Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981.

W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.