Emergent Semantics in BibSonomy

Andreas Hotho[1], Robert Jäschke[1,2], Christoph Schmitz[1], Gerd Stumme[1,2]

[1] Knowledge & Data Engineering Group,
Department of Mathematics and Computer Science,
University of Kassel, Wilhelmshöher Allee 73, D–34121 Kassel, Germany
http://www.kde.cs.uni-kassel.de

[2] Research Center L3S, Expo Plaza 1, D–30539 Hannover, Germany
http://www.l3s.de

Abstract: Social bookmark tools are rapidly emerging on the Web. In such systems users are setting up lightweight conceptual structures called folksonomies. The reason for their immediate success is the fact that no specific skills are needed for participating. In this paper we specify a formal model for folksonomies, briefly describe our own system BibSonomy, which allows for sharing both bookmarks and publication references, and discuss first steps towards emergent semantics.

1 Introduction

Complementing the Semantic Web effort, a new breed of so-called “Web 2.0” applications is currently emerging on the Web. These include user-centric publishing and knowledge management platforms like Wikis, Blogs, and social resource sharing tools.

Social resource sharing systems are web-based systems that allow users to upload their resources, and to label them with arbitrary words, so-called *tags*. The assignment of tags to resources by users is organized in a lightweight knowledge representation, called *folksonomy*. Folksonomies are a bottom-up complement to more formalized Semantic Web technologies, as they rely on *emergent semantics* [SSN$^+$02, Ste98] which result from the converging use of the same vocabulary. The main difference to “classical” ontology engineering approaches is their aim to respect to the largest possible extent the request of non-expert users not to be bothered with any formal modeling overhead.

Folksonomy-based tools, such as the image collection Flickr[1] or the bookmarking system del.icio.us,[2] have acquired large numbers of users within less than two years. Our own system, *BibSonomy*,[3] allows sharing bookmarks and BibTeX entries simultaneously (see Figure 1). The widespread use of these systems shows clearly that they are able to overcome the knowledge acquisition bottleneck, which was a serious handicap for many knowledge-based systems in the past.

[1] `http://www.flickr.com` [2] `http://del.icio.us` [3] `http://www.bibsonomy.org`

This paper summarizes work presented in [HJSS06a], [SHJS06] and [HJSS06b]. It is organized as follows. In Section 2 we introduce a formal model and the BibSonomy system. In Section 3 we present the application of association rule mining in folksonomies as well as the FolkRank – both to support emergent semantics. The paper concludes with a review of related work and an outlook.

2 BibSonomy— A Folksonomy-Based Social Bookmark System

This section briefly describes the BibSonomy system[4] developed by our group. BibSonomy allows a user to share bookmarks (i.e., URLs) as well as publication references. The data model of the publication part is based on BIBTEX, a popular literature management system for LATEX. BibSonomy implements the formal model of a folksonomy which we published in [HJSS06a], and which we briefly recall here.

A folksonomy describes the users, resources, and tags, and the user-based assignment of tags to resources. We formalize this in the following definition:

Definition 1. *A* folksonomy *is a tuple* $\mathbb{F} := (U, T, R, Y, \prec)$ *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.,* $Y \subseteq U \times T \times R$*, whose elements are called tag assignments, and* $\prec$ *is a user-specific subtag/supertag-relation, i. e.,* $\prec \subseteq U \times T \times T$*, called* is-a relation*.*

The personomy $\mathbb{P}_u$ *of a given user* $u \in U$ *is the restriction of* $\mathbb{F}$ *to* u*, i. e.,* $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$ *with* $I_u := \{(t, r) \in T \times R \mid (u, t, r) \in Y\}$*,* $T_u := \pi_1(I_u)$*,* $R_u := \pi_2(I_u)$*, and* $\prec_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \prec\}$*, where* π_i *denotes the projection on the* i*th dimension.*

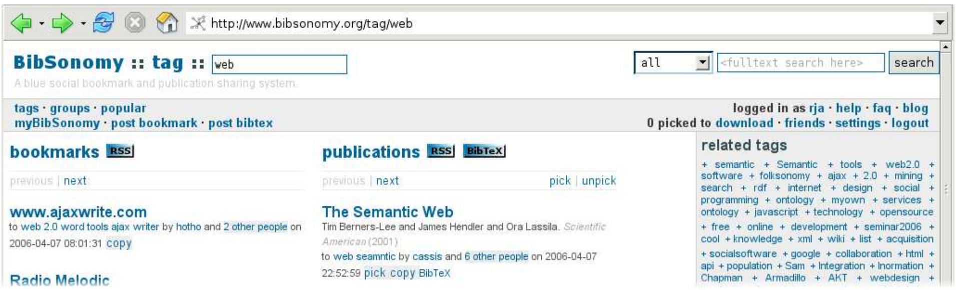
If we disregard the is-a relation, we can simply note a folksonomy as a quadruple $\mathbb{F} := (U, T, R, Y)$. This structure is known in Formal Concept Analysis [Wil82, GW99] as a *triadic context* [LW95]. An equivalent view on this structure is that of a tripartite (undirected) hypergraph $G = (V, E)$, where $V = U \dot{\cup} T \dot{\cup} R$ is the set of nodes, and $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$ is the set of hyperedges.

2.1 User Interface

Figure 1 shows a typical list of bookmark and publication posts containing the tag *web*. The page is divided into four parts: the header (showing navigation links and search boxes), two lists of posts – one for bookmarks and one for publications – sorted by date in descending order, and a list of tags related to the posts. This scheme holds for all pages showing posts and allows for navigation in all dimensions of the folksonomy.

Figures 2 and 3 present a detailed view of bookmark and publication posts from Figure 1. The first line of the bookmark post shows in bold the title of the bookmark which has

[4] http://www.bibsonomy.org



BibSonomy :: tag :: web

A blue social bookmark and publication sharing system.

all <fulltext search here> search

tags · groups · popular
myBibSonomy · post bookmark · post bibtex

logged in as rja · help · faq · blog
0 picked to download · friends · settings · logout

bookmarks RSS

previous | next

www.ajaxwrite.com
to web 2.0 word tools ajax writer by hotho and 2 other people on 2006-04-07 08:01:31 copy

Radio Melodic

publications RSS BibTeX

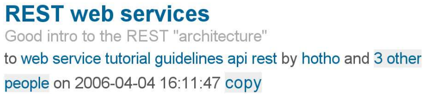
previous | next pick | unpick

The Semantic Web
Tim Berners-Lee and James Hendler and Ora Lassila. *Scientific American* (2001)
to web seamntic by cassis and 6 other people on 2006-04-07 22:52:59 pick copy BibTeX

related tags

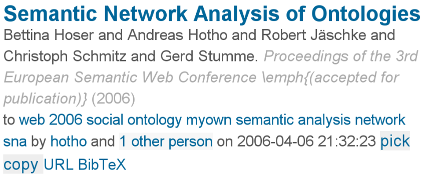
+ semantic + Semantic + tools + web2.0 + software + folksonomy + ajax + 2.0 + mining + search + rdf + internet + design + social + programming + ontology + myown + services + ontology + javascript + technology + opensource + free + online + development + seminar2006 + cool + knowledge + xml + wiki + list + acquisition + socialsoftware + google + collaboration + html + api + population + Sam + Integration + Inormation + Chapman + Armadillo + AKT + webdesign +

Figure 1: BibSonomy displays bookmarks and BIBTEX based bibliographic references simultaneously.



REST web services
Good intro to the REST "architecture"
to web service tutorial guidelines api rest by hotho and 3 other people on 2006-04-04 16:11:47 copy

Figure 2: a single bookmark post



Semantic Network Analysis of Ontologies
Bettina Hoser and Andreas Hotho and Robert Jäschke and Christoph Schmitz and Gerd Stumme. *Proceedings of the 3rd European Semantic Web Conference \emph{(accepted for publication)}* (2006)
to web 2006 social ontology myown semantic analysis network sna by hotho and 1 other person on 2006-04-06 21:32:23 pick copy URL BibTeX

Figure 3: a single publication post

hyperlinks to its URL. The second line is an optional description assigned by the user. The last two lines first show the tags the user has assigned to this post (*web*, *service*, *tutorial*, *guidelines*, *api*, and *rest*), second, the user name (*hotho*), followed by how many users tagged that resource. These parts hyperlink to the corresponding tag pages of the user, the user's overview page, and a page showing all four posts (i. e., the one of user *hotho* and those of the 3 other people) of this resource. The last part shows the posting date and time followed by actions the user can perform on this post – (*edit*, *delete*) for his own posts, or (*copy*) for others. The structure of a publication post displayed in BibSonomy is very similar, showing the bibliographic details instead of the description. The title links to a page providing detailed information on that post. The actions for BIBTEX items include picking the entry for later download, copying it, accessing the URL of the entry or viewing the BIBTEX source code.

2.2 Relations Between Tags

While tagging is popular because it is simple and no specific skills are needed, people using systems like BibSonomy are nevertheless asking for options to structure their tags. A user specific binary relation $\prec$ between tags as described in our folksonomy model is an easy way to arrange tags. Therefore we included this possibility in BibSonomy.

To enable the specification of the $\prec$ relation while tagging, we use the character sequences `<-` and `->`. I. e., if the user u enters t_1`->`t_2, we attach the tags t_1 and t_2 to the respective resource and add the triple (u, t_1, t_2) to the relation $\prec$. The tag t_2`<-`t_1 is interpreted as t_1`->`t_2. This can be read as "t_1 *is a* t_2" or "t_1 is a *subtag* of the *supertag* t_2". There are also other ways to add elements to $\prec$, in particular a relation editor.

The $\prec$ relation is used in several situations. First, the user can structure his tag cloud by showing all subtags of a certain supertag, thus seeing the tags in a hierarchy. Second, BibSonomy offers the option to show on a users tag page not only posts which contain a certain tag, but also posts which contain one of its subtags.

Including this relation raises several questions which are still under discussion:

- How to handle cycles, i.e. $u \in U$ and $t_1, \ldots, t_m \in T$ with $(u, t_i, t_{i+1}) \in \prec$ (for $i = 1, \ldots, m-1$) and $(u, t_m, t_1) \in \prec$?
- How to model equivalence or non-equivalence of tags?
- Should we make use of the transitive closure of the relation? If so: where and how to do it efficiently?
- How to express such queries like "all posts which have the tag *howto* and also one of the subtags of *programming*"? One idea would be "`->programming howto`".

3 Emergent Semantics in Folksonomies

We discuss now the adaptation of a data mining and an information retrieval approach to detect emergent semantics within folksonomy systems. In particular, we present an adaptation of association rule mining [SHJS06] and an adaptation of PageRank [HJSS06b] to the triadic nature of folksonomies. While we intend to implement these techniques within BibSonomy in the near future, we demonstrate our findings on data from the del.icio.us system, as it had a much larger data basis at the time of performing our experiments.

3.1 Association rule mining in Folksonomies

Discussions on the respective mailing lists such as *delicious-discuss*[5] show that there is a demand for more structure on folksonomies beyond flat tags, e. g., in the shape of the abovementioned $\prec$ relation. One possibility of offering such structure, without users having to maintain it themselves, is the application of ontology learning techniques.

Users express the meaning of resources through their tagging behavior. In tagging each resource, often general and more special tags are mixed; i. e., many web pages about XSLT stylesheets are tagged with *xslt* and additionally with *xml*. By computing association rules [AIS93] between tags, these relationships between tags can be extracted.

Regarding Definition 1 we observe that association rules cannot be mined directly on folksonomies, because of their triadic nature. One either has to define some kind of triadic association rules, or to transform the triadic folksonomy into a dyadic relation. In this paper, we follow the latter approach. For a definition of the projections refer to [SHJS06].

We consider one specific projection for demonstration purposes, namely $\mathbb{K}_1 := (U \times$

[5] `http://lists.del.icio.us/pipermail/discuss/`

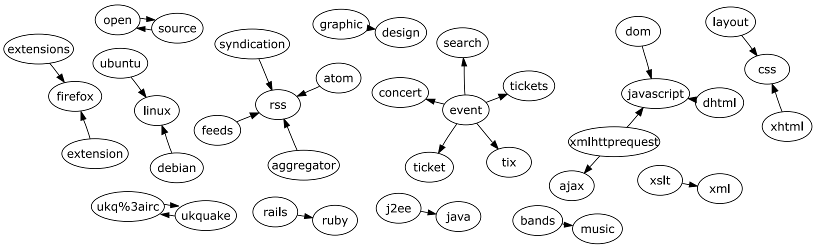


Figure 4: Two element rules between del.icio.us tags with 0,05 % support und 50 % confidence

$R, T, I_1)$ with $I_1 := \{((u,r),t)|(u,t,r) \in Y\}$. An association rule $A \rightarrow B$ in $\mathbb{K}_1$ is read as *users assigning the tags from A to some resources often also assign the tags from B to them.* This type of rules may be used in a recommender system, e. g., for recommending a tag hierarchy. If a user assigns all tags from A then the system suggests him to add also those from B.

We have evaluated this approach on del.icio.us data. To that end, we have extracted $|U| = 75.242$ users, $|T| = 533.191$ tags, and $|R| = 3.158.297$ resources, which are connected in del.icio.us through $|Y| = 17.362.212$ tag assignments.

Figure 4 shows an example from [SHJS06], which was computed on the del.icio.us data. Here we examine which tags occur more often with regard to other tags (e. g., if a user tags a particular resource with *xslt*, he will often tag it with *xml* also). In the nomenclature of association rule mining, the tags are the items, while user-resource combinations are transactions.

In the case at hand, an association rule points towards a subtag-supertag relation, which can be recommended to the user for inclusion in his $\prec$ relation. This approach can be combined with fulltext-based methods (e. g., [CPSTS05]) if the resources are web pages or other fulltext documents.

Figure 4 shows all rules with one element in the premise and one element in the conclusion that we derived from $\mathbb{K}_1$ with a minimum support of 0.05 % and a minimum confidence of 50 %. In the diagram one can see that our interpretation of rules in $\mathbb{K}_1$ holds for these examples: users tagging some webpage with *debian* are likely to tag it with *linux* also, and pages about *bands* are probably also concerned with *music*. These results can be used in a recommender system, aiding the user in choosing the tags which are most helpful in retrieving the resource later.

Another view on these rules is to see them as subsumption relations, so that the rule mining can be used to learn a taxonomic structure. If many resources tagged with *xslt* are also tagged with *xml*, this indicates, for example, that *xml* can be considered a supertopic of *xslt* if one wants to automatically populate the $\prec$ relation. Figure 4 also shows two pairs of tags which occur together very frequently without any distinct direction in the rule: *open source* occurs as a phrase most of the time, while the other pair consists of two tags (*ukquake* and *ukq:irc*), which seem to be added automatically to any resource that is mentioned in a particular chat channel.

3.2 FolkRank

Algorithms spreading weights in graphs like the popular PageRank [BP98] compute node rankings by incorporating the idea that a node is important if there are many edges from other nodes pointing to it and if those nodes are important themselves. We employ the same underlying principle to the tripartite graph of the folksonomy to rank users, tags or resources, e. g., a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. We have developed the *FolkRank* algorithm, a version of PageRank adapted to the structure of a folksonomy.

First we convert the tripartite hypergraph of the folksonomy into an undirected tripartite graph by transforming all co-occurences of tags and users, users and resources, tags and resources into undirected, weighted edges between the respective nodes. Applying plain PageRank with a preference vector (which models the random surfer) to express preference for tags, users or resources gives results dominated by nodes which are important in a ranking without preference vector. This is due to the dominance of these nodes in the folksonomy and the undirected structure where weight swashes back immediately. To compensate this, we compute the winners and losers of the mutual reinforcement of resources when a preference is given by considering the difference to the baseline without a preference vector. We call the resulting weight of an element of the folksonomy the *FolkRank* of that node. More details can be found at [HJSS06b].

3.2.1 Generating Recommendations

The original PageRank paper [BP98] already pointed out the possibility of using the random surfer vector as a personalization mechanism for PageRank computations. FolkRank yields a set of related users and resources for a given tag. Following these observations, FolkRank can be used to generate recommendations within a folksonomy system. These can be presented to the user at different points in the usage of a folksonomy system:

- Documents that are of potential interest to a user can be suggested to him. This kind of recommendation pushes potentially useful content to the user and increases the chance that he finds useful resources that he did not even know by “serendipitous” browsing.
- When using a certain tag, other related tags can be suggested. This can be used, for instance, to speed up the consolidation of different terminologies and thus facilitate the emergence of a common vocabulary.
- Other users that work on related topics can be made explicit, improving thus the knowledge transfer within organizations and fostering the formation of communities.

4 Related Work

A good overview of social bookmarking tools is provided by [HHLS05, LHFH05] whereas [GH05] discusses the structure of folksonomies (especially del.icio.us) and identifies seven types of tags. Visalization of tags over time is the topic of [DKM$^+$06], which also describes algorithms for this task. A related work regarding Blogs is [PMW05], which uses prinicipal component analysis and clustering techniques on a FOAF-network[6] to extract temporal changes of the user structure.

In [Mik05], Mika defines a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Besides calculating measures like the clustering coefficient, (local) betweenness centrality or the network constraint on the extracted one-mode network, Mika uses co-occurence techniques for clustering the folksonomy.

For further related work to our approaches presented in Section 3, we refer to the more comprehensive publications [SHJS06] and [HJSS06b].

5 Conclusion and Outlook

In this paper we described a formal model for folksonomies upon that our BibSonomy system is built. We discussed algorithms for exploring the structure of a folksonomy and described our BibSonomy system.

Since a folksonomy is a rich conceptual structure there are several ways to examine it. Up to now we focused mainly on the graph structure of a folksonomy and exploited and enhanced existing algorithms. With the growing amount of users and the availability of relations between tags, more sophisticated algorithms are needed. With the help of BibSonomy we are able to develop and test them, and let the users profit from our results.

When folksonomy-based systems grow larger, user support has to go beyond enhanced retrieval facilities. Therefore, the internal structure has to become better organized. An obvious approach for this are semantic web technologies. The key question remains though how to exploit their benefits without bothering untrained users with their rigidity. We believe that this will become a fruitful research area for the Semantic Web community for the next years.

Acknowledgement. Part of this research was funded by the EU in the Nepomuk project (FP6-027705).

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of SIGMOD 1993*, pages 207–216. ACM Press, May 1993.

[BP98] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[CPSTS05] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In *Ontology Learning from Text: Meth-*

[6] `http://www.foaf-project.org`

ods, Evaluation and Applications, Frontiers in Artificial Intelligence, pages 59–73. IOS Press, 2005.

[DKM$^+$06] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing Tags over Time. In *Proc. of the 15th International WWW Conference*, 2006.

[GH05] Scott Golder and Bernardo A. Huberman. The Structure of Collaborative Tagging Systems. Technical report, Information Dynamics Lab, HP Labs, Aug 2005.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[HHLS05] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[HJSS06a] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A Social Bookmark and Publication Sharing System. In A. de Moor, S. Polovina, and H. Delugach, editors, *Proc. of the Conceptual Structures Tool Interoperability Workshop at the 14th Int. Conf. on Conceptual Structures*, Aalborg, Denmark, July 2006. Aalborg University Press.

[HJSS06b] 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 *Lecture Notes in Computer Science*, pages 411–426, Heidelberg, June 2006. Springer.

[LHFH05] Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.

[LW95] F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual structures: applications, implementation and theory*, volume 954 of *Lecture Notes in Artificial Intelligence*, pages 32–43. Springer Verlag, 1995.

[Mik05] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536. Springer-Verlag, November 2005.

[PMW05] John C. Paolillo, Sarah Mercure, and Elijah Wright. The Social Semantics of LiveJournal FOAF: Structure and Change from 2004 to 2005. In G. Stumme, B. Hoser, C. Schmitz, and H. Alani, editors, *Proceedings of the 1st Workshop on Semantic Network Analysis at the ISWC 2005 Conference*, pages 69 – 80, November 2005.

[SHJS06] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining Association Rules in Folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, editors, *Data Science and Classification: Proc. of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270. Springer, 2006.

[SSN$^+$02] S. Staab, S. Santini, F. Nack, L. Steels, and A. Maedche. Emergent semantics. *Intelligent Systems, IEEE [see also IEEE Expert]*, 17(1):78–86, 2002.

[Ste98] L. Steels. The Origins of Ontologies and Communication Conventions in Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, 1(2):169–194, October 1998.

[Wil82] Rudolf Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In Ivan Rival, editor, *Ordered sets*, pages 445–470. Reidel, 1982.