Mining Association Rules in Folksonomies

Christoph Schmitz, Andreas Hotho, Robert Jäschke, Gerd Stumme

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

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. These systems provide currently relatively few structure. We discuss in this paper, how association rule mining can be adopted to analyze and structure folksonomies, and how the results can be used for ontology learning and supporting emergent semantics. We demonstrate our approach on a large scale dataset stemming from an online system.

1 Introduction

A new family 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 systems. In this paper, we focus on resource sharing systems, which all use the same kind of lightweight knowledge representation, called *folksonomy*. The word 'folksonomy' is a blend of the words 'taxonomy' and 'folk', and stands for conceptual structures created by the people.

Resource sharing systems, such as Flickr[1] or del.icio.us,[2] have acquired large numbers of users (from discussions on the del.icio.us mailing list, one can approximate the number of users on del.icio.us to be more than one hundred thousand) within less than two years. The reason for their immediate success is the fact that no specific skills are needed for participating, and that these tools yield immediate benefit for each individual user (e.g. organizing ones bookmarks in a browser-independent, persistent fashion) without too much overhead. Large numbers of users have created huge amounts of information within a very short period of time. As these systems grow larger, however, the users feel the need for more structure for better organizing their resources. For instance, approaches for tagging tags, or for bundling them, are currently discussed on the corresponding news groups. Currently, however, there is a lack of theoretical foundations adapted to the new opportunities which has to be overcome.

A first step towards more structure within such systems is to discover knowledge that is already implicitly present by the way different users assign tags to resources. This knowledge may be used for recommending both a hierarchy on the already existing tags, and additional tags, ultimately leading towards *emergent semantics* (Staab et al. (2002); Steels (1998)) by converging use of the same vocabulary. In this sense, knowledge discovery (KDD) techniques are a promising tool for bottom-up building of conceptual structures.

[1] http://www.flickr.com/ [2] http://del.icio.us

Fig. 1. Bibsonomy displays bookmarks and (BIBTEX based) bibliographic references simultaneously.

In this paper, we will focus on a selected KDD technique, namely association rules. Since folksonomies provide a three-dimensional dataset (users, tags, and resources) instead of a usual two-dimensional one (items and transactions), we present first a systematic overview of projecting a folksonomy onto a two-dimensional structure. Then we will show the results of mining rules from two selected projections on the del.icio.us system.

This paper is organized as follows. Section 2 reviews recent developments in the area of social bookmark systems, and presents a formal model. In Section 3, we briefly recall the notions of association rules, before providing a systematic overview over the projections of a folksonomy onto a two-dimensional dataset in Section 4. In Section 5, we present the results of mining association rules on data of the del.icio.us system. Section 6 concludes the paper with a discussion of further research topics on knowledge discovery within folksonomies.

2 Social Resource Sharing and Folksonomies

Social resource sharing systems are web-based systems that allow users to upload their resources, and to label them with names. The systems can be distinguished according to what kind of resources are supported. Flickr,[3] for instance, allows the sharing of photos, del.icio.us[4] the sharing of bookmarks, CiteULike[5] and Connotea[6] the sharing of bibliographic references, and 43Things[7] even the sharing of goals in private life. Our own upcoming system, called *BibSonomy*,[8] will allow to share simultaneously bookmarks and BIBTEX entries (see Fig. 1).

In their core, these systems are all very similar. Once a user is logged in, he can add a resource to the system, and assign arbitrary labels, so-called *tags*, to it. We call the collection of all his assignments his *personomy*, and the collection of all personomies is called *folksonomy*. The user can also explore the personomies of other users in all dimensions: for a given user he can see the resources that user has uploaded, together with the tags he has assigned to them (see Fig. 1); when clicking on a resource he

[3] http://www.flickr.com/ [4] http://del.icio.us/ [5] http://www.citeulike.org/
[6] http://www.connotea.org/ [7] http://www.43things.com/ [8] http://www.bibsonomy.org

sees which other users have uploaded this resource and how they tagged it; and when clicking on a tag he sees who assigned it to which resources.

The systems allow for additional functionality. For instance, one can copy a resource from another user, and label it with one owns tags. Overall, these systems provide a very intuitive navigation through the data.

2.1 State of the Art

There are currently virtually no scientific publications about folksonomy-based web collaboration systems. Among the rare exceptions are Hammond et al. (2005) and Lund et al. (2005) who provide good overviews of social bookmarking tools with special emphasis on folksonomies, and Mathes (2004) who discusses strengths and limitations of folksonomies. The main discussion on folksonomies and related topics is currently only going on mailing lists, e.g. Connotea (2005). To the best of our knowledge, the ideas presented in this paper have not been explored before, but there is a lot of recent work dealing with folksonomies.

Mika (2005) 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 concept network.

There are several systems working on top of del.icio.us to explore the underlying folksonomy. CollaborativeRank[9] provides ranked search results on top of del.icio.us bookmarks. The ranking takes into account, how early someone bookmarked an URL and how many people followed him or her. Other systems show popular sites (Populicious[10]) or focus on graphical representations (Cloudalicious[11], Grafolicious[12]) of statistics about del.icio.us.

2.2 A Formal Model for Folksonomies

A folksonomy basically describes users, resources, tags, and allows users to assign (arbitrary) tags to resources. We present here a formal definition of folksonomies, which is also underlying our BibSonomy system.

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., $Y \subseteq U \times T \times R$, called assignments, and
- $\prec$ is a user-specific *subtag/supertag-relation*, i. e., $\prec \subseteq U \times ((T \times T) \setminus \{(t,t) \mid t \in T\})$.

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 $\prec_u := \{(t_1,t_2) \in T \times T \mid (u,t_1,t_2) \in \prec\}$.

[9] http://collabrank.org/ [10] http://populicio.us/ [11] http://cloudalicio.us/
[12] http://www.neuroticweb.com/recursos/del.icio.us-graphs/

Users are typically described by their user ID, and tags may be arbitrary strings. What is considered as a resource depends on the type of system. In del.icio.us, for instance, the resources are URLs, and in Flickr, the resources are pictures. In our BibSonomy system, we have two types of resources, bookmarks and BIBTEX entries. From an implementation point of view, resources are internally represented by some ID.

In this paper, we do not make use of the subtag/supertag relation for sake of simplicity. I. e., $\prec = \emptyset$, and we will simply note a folksonomy as a quadruple $\mathbb{F} := (U, T, R, Y)$. This structure is known in Formal Concept Analysis (Wille (1982); Ganter and Wille (1999)) as a *triadic context* (Lehmann and Wille (1995); Stumme (2005)). An equivalent view on folksonomy data 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.

2.3 Del.ico.us — A Folksonomy-Based Social Bookmark System

In order to evaluate our folksonomy mining approach, we have analyzed the popular social bookmarking sytem del.icio.us. Del.icio.us is a server-based system with a simple-to-use interface that allows users to organize and share bookmarks on the internet. It is able to store in addition to the URL a description, a note, and tags (i. e., arbitrary labels). We chose del.icio.us rather than our own system, Bibsonomy, as the latter is going online only after the time of writing of this article. For our experiments, we collected from the del.ico.us system $|U| = 75,242$ users, $|T| = 533,191$ tags and $|R| = 3,158,297$ resources, related by in total $|Y| = 17,362,212$ triples.

3 Association Rule Mining

We assume here, that the reader is familiar with the basics of association rule mining introduced by Agrawal et al. (1993). As the work presented in this paper is on the conceptual rather than on the computational level, we refrain in particular from describing the vast area of developing efficient algorithms. Many of the existing algorithms can be found at the Frequent Itemset Mining Implementations Repository.[13] Instead, we just recall the definition of the association rule mining problem, which was initially stated by Agrawal et al. (1993), in order to clarify the notations used in the following. We will not use the original terminology of Srikant et al, but rather exploit the vocabulary of Formal Concept Analysis (FCA) (Wille (1982)), as it better fits with the formal folksonomy model introduced in Definition 1.[14]

Definition 2. A *formal context* is a dataset $\mathbb{K} := (G, M, I)$ consisting of a set G of *objects*, a set M of *attributes*, and a binary relation $I \subseteq G \times M$, where $(g, m) \in I$ is read as "object g has attribute m".

In the usual basket analysis scenario, M is the set of items sold by a supermarket, G is the set of all transactions, and, for a given transaction $g \in G$, the set $g^I := \{m \in M | (g, m) \in I\}$ contains all items bought in that transaction.

[13] http://fimi.cs.helsinki.fi/ [14] For a detailed discussion about the role of FCA for association rule mining see (Stumme (2002)).

Definition 3. For a set X of attributes, we define $A' := \{g \in G \mid \forall m \in X : (g, m) \in I\}$. The *support* of A is calculated by $\text{supp}(A) := \frac{|A'|}{|G|}$.

Definition 4 (Association Rule Mining Problem (Agrawal et al. (1993))). Let $\mathbb{K}$ be a formal context, and $\text{minsupp}, \text{minconf} \in [0, 1]$, called *minimum support* and *minimum confidence thresholds*, resp. The *association rule mining problem* consists now of determining all pairs $A \rightarrow B$ of subsets of M whose *support* $\text{supp}(A \rightarrow B) := \text{supp}(A \cup B)$ is above the threshold minsupp, and whose *confidence* $\text{conf}(A \rightarrow B) := \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$ is above the threshold minconf.

As the rules $A \rightarrow B$ and $A \rightarrow B \setminus A$ carry the same information, and in particular have same support and same confidence, we will consider in this paper the additional constraint prevalent in the data mining community, that premise A and conclusion B are to be disjoint.[15]

When comparing Definitions 1 and 2, 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 formal context. In this paper, we follow the latter approach.

4 Projecting the Folksonomy onto two Dimensions

As discussed in the previous section, we have to reduce the three-dimensional folksonomy to a two-dimensional formal context before we can apply any association rule mining technique. Several such projections have already been introduced in Lehmann and Wille (1995). In Stumme (2005), we provide a more complete approach, which we here adapt slightly to the association rule mining scenario.

As we want to analyze all facets of the folksonomy, we want to allow to use any (combination) of the three sets U, T, and R as the set of objects – on which the support is computed – at some point in time, depending on the task on hand. Therefore, we will not fix the roles of the three sets in advance. Instead, we consider a triadic context as symmetric structure, where all three sets are of equal importance. For easier handling, we therefore denote the folksonomy $\mathbb{F} := (U, T, R, Y)$ alternatively by $\mathbb{F} := (X_1, X_2, X_3, Y)$ in the following.

We determine the set of objects – i. e., the set on which the support will be counted – by a permutation σ on the set $\{1, 2, 3\}$. The choice of a permutation indicates, together with one of the aggregation modes '$\mathfrak{D}$', 'W', '$\exists n$' with $n \in \mathbb{N}$, and '$\forall$', on which formal context $\mathbb{K} := (G, M, I)$ the association rules are computed.

- $\mathbb{K}^{\sigma,\mathfrak{D}} := (X_{\sigma(1)} \times X_{\sigma(3)}, X_{\sigma(2)}, I)$ with $((x_{\sigma(1)}, x_{\sigma(3)}), x_{\sigma(2)}) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.
- $\mathbb{K}^{\sigma,\mathsf{W}} := (X_{\sigma(1)}, X_{\sigma(2)} \times X_{\sigma(3)}, I)$ with $(x_{\sigma(1)}, (x_{\sigma(2)}, x_{\sigma(3)})) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.

[15] In FCA, in contrast, one often requires A to be a subset of B, as this better fits with the notion of *closed itemsets* which arose of applying FCA to the association mining problem(Pasquier et al. (1999); Zaki and Hsiao (1999); Stumme (1999)).

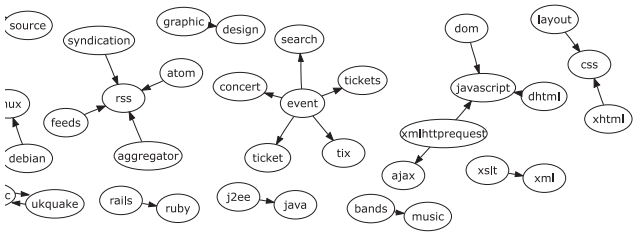

Fig. 2. All rules with two elements of $\mathbb{K}_1$ with .05 % support, 50 % confidence

- $\mathbb{K}^{\sigma,\exists n} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if there exist n different $x_{\sigma(3)} \in X_{\sigma(3)}$ with $(x_1, x_2, x_3) \in Y$.
- $\mathbb{K}^{\sigma,\forall} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if for all $x_{\sigma(3)} \in X_{\sigma(3)}$ holds $(x_1, x_2, x_3) \in Y$. This mode is equivalent to '$\exists n$' with $n = |X_{\sigma(3)}|$.

These projections are complemented by the following way to 'cut slices' out of the folksonomy. A slice is obtained by selecting one dimension (out of user/tag/resource), and then fixing in this dimension one particular instance.

- Let $x := x_{\sigma(3)} \in X_{\sigma(3)}$. $\mathbb{K}^{\sigma,x} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.

In the next section, we discuss for two of these projections the kind of rules one obtains from mining the formal context that is resulting from the projection.

5 Mining Association Rules on the Projected Folksonomy

After having performed one of the projections described in the previous section, one can apply the standard association rule mining techniques as described in Section 3. Due to space restrictions, we have to focus on a subset of projections. In particular, we address the two projections $\mathbb{K}^{\sigma_i,\mathfrak{D}}$ with $\sigma_1 := \mathrm{id}$ and $\sigma_2 := (1 \mapsto 1, 2 \mapsto 3, 3 \mapsto 2)$. We obtain the two dyadic contexts $\mathbb{K}_1 := (U \times R, T, I_1)$ with $I_1 := \{((u, r), t) | (u, t, r) \in Y\}$ and $\mathbb{K}_2 := (T \times U, R, I_2)$ with $I_2 := \{(t, u), r) | (u, t, r) \in Y\}$.

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. If a user assigns all tags from A then the system suggests him to add also those from B.

Figure 2 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

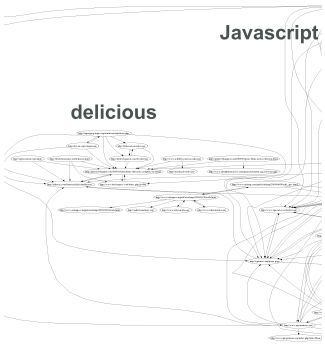

Fig. 3. Rules with two elements of $\mathbb{K}_2$ with 0.05 % support, and 10 % confidence. (The online version of this article allows to zoom into the diagram.)

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 2 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.

The second example are association rules $A \rightarrow B$ in $\mathbb{K}_2$ which are read as *Users labelling the resources in A with some tags often also assign these tags to the resources in B*. In essence both resources have to have something in common. Figure 3 shows parts of the resulting graph for applying association rules with 0.05 % support, and 10 % confidence on $\mathbb{K}_2$. Only associations rules with one element in premise and one element in conclusion are considered in the graph. In Figure 3 we identified four major areas in the graph which we labeled with the topics *delicious*, *Javascript*, *Ajax*, and *CSS*. The topics can be derived by applying the FolkRank (Hotho et al. (2006)) on some of the resources of interest, which also yields relevant users and other resources for the respective area, such that communities of interest can be identified.

6 Conclusion

In this paper, we have presented a formal model of folksonomies as a triadic context – or, equivalently, a tripartite hypergraph. In order to apply association rule mining to folksonomies, we have systematically explored possible projections of the folksonomy structure into the standard notion of "shopping baskets" used in rule mining. For two selected projections, we demonstrated the outcome of rule mining on a large-scale folksonomy dataset. The rules can be applied for different purposes, such as recommending tags, users, or resources, populating the supertag relation of the folksonomy, and community detection.

Future work includes the tighter integration of the various techniques we used here, namely, association rule mining, FolkRank ranking, and graph clustering, to further contribute to the abovementioned applications.

Bibliography

AGRAWAL, R., IMIELINSKI, T. and SWAMI, A. (1993): Mining association rules between sets of items in large databases. In: *Proc. of SIGMOD 1993*, pp. 207–216. ACM Press.

CONNOTEA (2005): Connotea Mailing List. https://lists.sourceforge.net/lists/listinfo/connotea-discuss.

GANTER, B. and WILLE, R. (1999): *Formal Concept Analysis : Mathematical foundations*. Springer.

HAMMOND, T., HANNAY, T., LUND, B. and SCOTT, J. (2005): Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4).

HOTHO, A., JÄSCHKE, R., SCHMITZ, C. and STUMME, G. (2006): Information Retrieval in Folksonomies: Search and Ranking. In: *Proc. ESWC 2006 (submitted)*.

LEHMANN, F. and WILLE, R. (1995): A triadic approach to Formal Concept Analysis. In: G. Ellis, R. Levinson, W. Rich and J. F. Sowa (Eds.), *Conceptual Structures: Applications, Implementation and Theory*, vol. 954 of *Lecture Notes in Computer Science*. Springer. ISBN 3-540-60161-9.

LUND, B., HAMMOND, T., FLACK, M. and HANNAY, T. (2005): Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4).

MATHES, A. (2004): Folksonomies – Cooperative Classification and Communication Through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

MIKA, P. (2005): Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Y. Gil, E. Motta, V. R. Benjamins and M. A. Musen (Eds.), *ISWC 2005*, vol. 3729 of *LNCS*, pp. 522–536. Springer-Verlag, Berlin Heidelberg.

PASQUIER, N., BASTIDE, Y., TAOUIL, R. and LAKHAL, L. (1999): Closed set based discovery of small covers for association rules. In: *Actes des 15èmes journées Bases de Données Avancées (BDA'99)*, pp. 361–381.

STAAB, S., SANTINI, S., NACK, F., STEELS, L. and MAEDCHE, A. (2002): Emergent semantics. *Intelligent Systems, IEEE*, 17(1):78.

STEELS, L. (1998): The Origins of Ontologies and Communication Conventions in Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, 1(2):169.

STUMME, G. (1999): Conceptual Knowledge Discovery with Frequent Concept Lattices. FB4-Preprint 2043, TU Darmstadt.

STUMME, G. (2002): Efficient Data Mining Based on Formal Concept Analysis. In: A. Hameurlain, R. Cicchetti and R. Traunmüller (Eds.), *Proc. DEXA 2002*, vol. 2453 of *LNCS*, pp. 534–546. Springer, Heidelberg.

STUMME, G. (2005): A Finite State Model for On-Line Analytical Processing in Triadic Contexts. In: B. Ganter and R. Godin (Eds.), *ICFCA*, vol. 3403 of *Lecture Notes in Computer Science*, pp. 315–328. Springer. ISBN 3-540-24525-1.

WILLE, R. (1982): Restructuring lattices theory : An approach based on hierarchies of concepts. In: I. Rival (Ed.), *Ordered Sets*, pp. 445–470. Reidel, Dordrecht-Boston.

ZAKI, M. J. and HSIAO, C.-J. (1999): ChARM: An efficient algorithm for closed association rule mining. Technical Report 99–10. Tech. rep., Computer Science Dept., Rensselaer Polytechnic.